Linear Algebra, Theory And Applications

Kenneth Kuttler

August 7, 2023

Contents

Preface

This is a book on linear algebra and matrix theory. While it is self contained, it will work best for those who have already had some exposure to linear algebra. It is also assumed that the reader has had calculus. Some optional topics require more analysis than this, however.

I think that the subject of linear algebra is likely the most significant topic discussed in undergraduate mathematics courses. Part of the reason for this is its usefulness in unifying so many different topics. Linear algebra is essential in analysis, applied math, and even in theoretical mathematics. This is the point of view of this book, more than a presentation of linear algebra for its own sake. This is why there are numerous applications, some fairly unusual.

This book features an ugly, elementary, and complete treatment of determinants early in the book. Thus it might be considered as Linear algebra done wrong. I have done this because of the usefulness of determinants. However, all major topics are also presented in an alternative manner which is independent of determinants.

The book has an introduction to various numerical methods used in linear algebra. This is done because of the interesting nature of these methods. The presentation here emphasizes the reasons why they work. It does not discuss many important numerical considerations necessary to use the methods effectively. These considerations are found in numerical analysis texts.

In the exercises, you may occasionally see $\uparrow$ at the beginning. This means you ought to have a look at the exercise above it. Some exercises develop a topic sequentially. There are also a few exercises which appear more than once in the book. I have done this deliberately because I think that these illustrate exceptionally important topics and because some people don't read the whole book from start to finish but instead jump in to the middle somewhere. There is one on a theorem of Sylvester which appears no fewer than 3 times. Then it is also proved in the text. There are multiple proofs of the Cayley Hamilton theorem, some in the exercises. Some exercises also are included for the sake of emphasizing something which has been done in the preceding chapter.

Chapter 1

Preliminaries

1.1 Sets and Set Notation

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

If A and B are sets with the property that every element of A is an element of B, then A is a subset of B. For example, $\{1, 2, 3, 8\}$ is a subset of $\{1, 2, 3, 4, 5, 8\}$, in symbols, $\{1, 2, 3, 8\} \subseteq \{1, 2, 3, 4, 5, 8\}$. 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.

1.2 Functions

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

Definition 1.2.1 *Consider two sets, D and R along with a rule which assigns a unique element of R to every element of D. This rule is called a* ***function*** *and it is denoted by a letter such as f. Given $x \in D$, $f(x)$ is the name of the thing in R which results from doing f to x. Then D is called the* ***domain*** *of f. In order to specify that D pertains to f, the notation $D(f)$ may be used. The set R is sometimes called the* ***range*** *of f. These days it is referred to as the* ***codomain****. The set of all elements of R which are of the form $f(x)$ for some $x \in D$ is therefore, a subset of R. This is sometimes referred to as the image of f. When this set equals R, the function f is said to be* ***onto****, also* ***surjective****. If whenever $x \neq y$ it follows $f(x) \neq f(y)$, the function is called* ***one to one.*** *, also* ***injective*** *It is common notation to write $f : D \mapsto R$ to denote the situation just described in this definition where f is a function defined on a domain D which has values in a codomain R. Sometimes you may also see something like $D \overset{f}{\mapsto} R$ to denote the same thing.*

1.3 The Number Line and Algebra of the Real Numbers

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

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

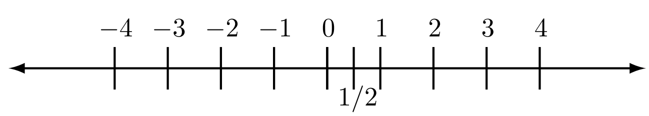
the natural numbers,

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

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

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

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



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

Axiom 1.3.1 $x + y = y + x$, *(commutative law for addition)*

Axiom 1.3.2 $x + 0 = x$, *(additive identity).*

Axiom 1.3.3 *For each $x \in \mathbb{R}$, there exists $-x \in \mathbb{R}$ such that $x + (-x) = 0$, (existence of additive inverse).*

Axiom 1.3.4 $(x + y) + z = x + (y + z)$, *(associative law for addition).*

Axiom 1.3.5 $xy = yx,$ (*commutative law for multiplication).*

Axiom 1.3.6 $(xy)\,z = x\,(yz)\,,$ (*associative law for multiplication).*

Axiom 1.3.7 $1x = x,$ (*multiplicative identity).*

Axiom 1.3.8 *For each $x \neq 0$, there exists x^{-1} such that $xx^{-1} = 1$.(existence of multiplicative inverse).*

Axiom 1.3.9 $x\,(y+z) = xy + xz$.(*distributive law).*

These axioms are known as the field axioms and any set (there are many others besides $\mathbb{R}$) which has two such operations satisfying the above axioms is called a field. Division and subtraction are defined in the usual way by $x - y \equiv x + (-y)$ and $x/y \equiv x\left(y^{-1}\right)$. We assume $0 \neq 1$ so that the axioms will describe someting useful.

Here is a little proposition which derives some familiar facts.

Proposition 1.3.10 *0 and 1 are unique. Also $-x$ is unique and x^{-1} is unique. Furthermore, $0x = x0 = 0$ and $-x = (-1)\,x$.*

Proof: Suppose $0'$ is another additive identity. Then

$$0' = 0' + 0 = 0.$$

Thus 0 is unique. Say $1'$ is another multiplicative identity. Then

$$1 = 1'1 = 1'.$$

Now suppose y acts like the additive inverse of x. Then

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

Finally,

$$0x = (0 + 0)\,x = 0x + 0x$$

and so

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

Finally

$$x + (-1)\,x = (1 + (-1))\,x = 0x = 0$$

and so by uniqueness of the additive inverse, $(-1)\,x = -x$. ■

1.4 Ordered Fields

The real numbers $\mathbb{R}$ are an example of an ordered field. More generally, here is a definition.

Definition 1.4.1 *Let F be a field. It is an ordered field if there exists an order, $<$ which satisfies*

1. *For any x, y, exactly one of the following holds: $x = y$, $x < y$, or $y < x$.*
2. *If $x < y$ and either $z < w$ or $z = w$, then, $x + z < y + w$.*
3. *If $0 < x, 0 < y$, then $xy > 0$.*

With this definition, the familiar properties of order can be proved. The following proposition lists many of these familiar properties. The relation '$a > b$' has the same meaning as '$b < a$'.

Proposition 1.4.2 *The following are obtained. Recall $-x$ is the symbol for the additive inverse of x.*

1. *If $x < y$ and $y < z$, then $x < z$.*
2. *If $x > 0$ and $y > 0$, then $x + y > 0$.*
3. *If $x > 0$, then $-x < 0$.*
4. *If $x \neq 0$, either x or $-x$ is > 0.*
5. *If $x < y$, then $-x > -y$.*
6. *If $x \neq 0$, then $x^2 > 0$.*
7. *If $0 < x < y$ then $x^{-1} > y^{-1}$.*

Proof: First consider 1, called the transitive law. Suppose that $x < y$ and $y < z$. Then from the axioms, $x + y < y + z$ and so, adding $-y$ to both sides, it follows

$$x < z$$

Next consider 2. Suppose $x > 0$ and $y > 0$. Then from 2,

$$0 = 0 + 0 < x + y.$$

Next consider 3. It is assumed $x > 0$ so

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

Now consider 4. If $x < 0$, then

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

Consider the 5. Since $x < y$, it follows from 2

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

and so by 4 and Proposition 1.3.10,

$$(-1)\,(y + (-x)) < 0$$

Also from Proposition 1.3.10 $(-1)\,(-x) = -\,(-x) = x$ and so

$$-y + x < 0.$$

Hence

$$-y < -x.$$

Consider 6. If $x > 0$, there is nothing to show. It follows from the definition. If $x < 0$, then by 4, $-x > 0$ and so by Proposition 1.3.10 and the definition of the order,

$$(-x)^2 = (-1)\,(-1)\,x^2 > 0$$

By this proposition again, $(-1)(-1) = -(-1) = 1$ and so $x^2 > 0$ as claimed.

Note this shows that $1 > 0$ because 1 equals 1^2.

Finally, consider 7. First, if $x > 0$ then if $x^{-1} < 0$, it would follow $(-1)x^{-1} > 0$ and so $x(-1)x^{-1} = (-1)1 = -1 > 0$. However, this would require

$$0 > 1 = 1^2 > 0$$

from what was just shown. Therefore, $x^{-1} > 0$. Now the assumption implies $y + (-1)x > 0$ and so multiplying by x^{-1},

$$yx^{-1} + (-1)xx^{-1} = yx^{-1} + (-1) > 0$$

Now multiply by y^{-1}, which by the above satisfies $y^{-1} > 0$, to obtain

$$x^{-1} + (-1)y^{-1} > 0$$

and so

$$x^{-1} > y^{-1}. \blacksquare$$

In an ordered field the symbols $\leq$ and $\geq$ have the usual meanings. Thus $a \leq b$ means $a < b$ or else $a = b$, etc.

1.5 The Complex Numbers

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

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

and

$$(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad).$$

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

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

You should prove the following theorem.

Theorem 1.5.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms listed on Page 10.*

Note that if $x + iy$ is a complex number, it can be written as

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

Now $\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right)$ is a point on the unit circle and so there exists a unique $\theta \in [0, 2\pi)$ such that this ordered pair equals $(\cos\theta, \sin\theta)$. Letting $r = \sqrt{x^2 + y^2}$, it follows that the complex number can be written in the form

$$x + iy = r(\cos\theta + i\sin\theta)$$

This is called the polar form of the complex number.

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

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

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

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

Definition 1.5.2 *Define the absolute value of a complex number as follows.*

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

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

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

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

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

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

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

Now note that

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

and so

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

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

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

The polar form of the complex number is then

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

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

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

Theorem 1.5.4 *Let $r \geq 0$ 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 1.5.5 *Let z be a non zero complex number. Then there are always exactly k k^{th} roots of z in $\mathbb{C}$.*

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

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

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

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

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

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

for l an integer. Thus

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

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

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

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

Example 1.5.6 *Find the three cube roots of i.*

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

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

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

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

and

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

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

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

Example 1.5.7 *Factor the polynomial* $x^3 - 27$.

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

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

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

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

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

The real and complex numbers both are fields satisfying the axioms on Page 10 and it is usually one of these two fields which is used in linear algebra. The numbers are often called scalars. However, it turns out that all algebraic notions work for any field and there are many others. For this reason, I will often refer to the field of scalars as $\mathbb{F}$ although $\mathbb{F}$ will usually be either the real or complex numbers. If there is any doubt, assume it is the field of complex numbers which is meant.

1.6 The Fundamental Theorem of Algebra

The reason the complex numbers are so significant in linear algebra is that they are algebraically complete. This means that every polynomial $\sum_{k=0}^{n} a_k z^k,\ n \geq 1, a_n \neq 0$, having coefficients a_k in $\mathbb{C}$ has a root in in $\mathbb{C}$. I will give next a simple explanation of why it is reasonable to believe in this theorem followed by a legitimate proof. The first completely correct proof of this theorem was given in 1806 by Argand although Gauss is often credited with proving it earlier and many others worked on it in the 1700's.

Theorem 1.6.1 *Let* $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ *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

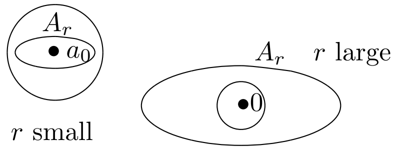
$$p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

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,

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

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

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

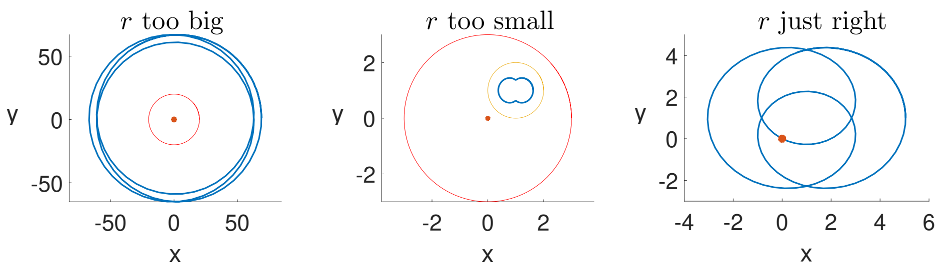
Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

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

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

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

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 is in the exterior of a 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, $1.386(\cos t + i \sin t)$ is a solution of the equation $p(z) = 0$ or very close to one.

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

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

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 $|q(z)| \geq 1$. ■

1.7 Exercises

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

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

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

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

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

6. De Moivre's theorem says $[r(\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt)$ for n a positive integer. Does this formula continue to hold for all integers, n, even negative integers? Explain.

7. You already know formulas for $\cos(x + y)$ and $\sin(x + y)$ and these were used to prove De Moivre's theorem. Now using De Moivre's theorem, derive a formula for $\sin(5x)$ and one for $\cos(5x)$. **Hint:** Use the binomial theorem.

8. If z and w are two complex numbers and the polar form of z involves the angle θ while the polar form of w involves the angle ϕ, show that in the polar form for zw the angle involved is $\theta + \phi$. Also, show that in the polar form of a complex number, z, $r = |z|$.

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

10. Write $x^3 + 27$ in the form $(x + 3)\left(x^2 + ax + b\right)$ where $x^2 + ax + b$ cannot be factored any more using only real numbers.

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

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

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

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

15. I claim that $1 = -1$. Here is why: $-1 = i^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)^2} = \sqrt{1} = 1$. This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?

16. De Moivre's theorem is really a grand thing. I plan to use it now for rational exponents, not just integers.
$$1 = 1^{(1/4)} = (\cos 2\pi + i \sin 2\pi)^{1/4} = \cos(\pi/2) + i \sin(\pi/2) = i.$$
Therefore, squaring both sides it follows $1 = -1$ as in the previous problem. What does this tell you about De Moivre's theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?

17. Show that $\mathbb{C}$ cannot be considered an ordered field. **Hint:** Consider $i^2 = -1$. Recall that $1 > 0$ by Proposition 1.4.2.

18. Say $a + ib < x + iy$ if $a < x$ or if $a = x$, then $b < y$. This is called the lexicographic order. Show that any two different complex numbers can be compared with this order. What goes wrong in terms of the other requirements for an ordered field.

19. With the order of Problem 18, consider for $n \in \mathbb{N}$ the complex number $1 - \frac{1}{n}$. Show that with the lexicographic order just described, each of $1 - in$ is an upper bound to all these numbers. Therefore, this is a set which is "bounded above" but has no least upper bound with respect to the lexicographic order on $\mathbb{C}$.

1.8 Completeness of $\mathbb{R}$

Recall the following important definition from calculus, completeness of $\mathbb{R}$.

Definition 1.8.1 *A non empty set, $S \subseteq \mathbb{R}$ is bounded above (below) if there exists $x \in \mathbb{R}$ such that $x \geq (\leq)\, s$ for all $s \in S$. If S is a nonempty set in $\mathbb{R}$ which is bounded above, then a number, l which has the property that l is an upper bound and that every other upper bound is no smaller than l is called a least upper bound, l.u.b. (S) or often $\sup(S)$. If S is a nonempty set bounded below, define the greatest lower bound, g.l.b. (S) or $\inf(S)$ similarly. Thus g is the g.l.b. (S) means g is a lower bound for S and it is the largest of all lower bounds. If S is a nonempty subset of $\mathbb{R}$ which is not bounded above, this information is expressed by saying $\sup(S) = +\infty$ and if S is not bounded below, $\inf(S) = -\infty$.*

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

Axiom 1.8.2 *(completeness) Every nonempty set of real numbers which is bounded above has a least upper bound and every nonempty set of real numbers which is bounded below has a greatest lower bound.*

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

Proposition 1.8.3 *Let S be a nonempty set and suppose $\sup(S)$ exists. Then for every $\delta > 0$,*
$$S \cap (\sup(S) - \delta, \sup(S)] \neq \emptyset.$$
If $\inf(S)$ exists, then for every $\delta > 0$,
$$S \cap [\inf(S), \inf(S) + \delta) \neq \emptyset.$$

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

1.9 Well Ordering and Archimedean Property

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

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

In particular, the natural numbers defined as

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

is well ordered.

The above axiom implies the principle of mathematical induction.

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

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

Example 1.9.4 *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.

Definition 1.9.5 *The Archimedean property states that whenever $x \in \mathbb{R}$, and $a > 0$, there exists $n \in \mathbb{N}$ such that $na > x$.*

Proposition 1.9.6 *$\mathbb{R}$ has the Archimedean property.*

Proof: Suppose it is not true. Then there exists $x \in \mathbb{R}$ and $a > 0$ such that $na \leq x$ for all $n \in \mathbb{N}$. Let $S = \{na : n \in \mathbb{N}\}$. By assumption, this is bounded above by x. By completeness, it has a least upper bound y. By Proposition 1.8.3 there exists $n \in \mathbb{N}$ such that

$$y - a < na \leq y.$$

Then $y = y - a + a < na + a = (n+1)a \leq y$, a contradiction. ■

Theorem 1.9.7 *Suppose $x < y$ and $y - x > 1$. Then there exists an integer $l \in \mathbb{Z}$, such that $x < l < y$. If x is an integer, there is no integer y satisfying $x < y < x + 1$.*

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

Now suppose $y - x > 1$ and let

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

The set S is nonempty by the Archimedean property. Let k be the smallest element of S. Therefore, $k - 1 < y$. Either $k - 1 \leq x$ or $k - 1 > x$. If $k - 1 \leq x$, then

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

contrary to the assumption that $y - x > 1$. Therefore, $x < k - 1 < y$. Let $l = k - 1$. ■

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

Theorem 1.9.8 *If $x < y$ then there exists a rational number r such that $x < r < y$.*

Proof: Let $n \in \mathbb{N}$ be large enough that

$$n(y - x) > 1.$$

Thus $(y - x)$ added to itself n times is larger than 1. Therefore,

$$n(y - x) = ny + n(-x) = ny - nx > 1.$$

It follows from Theorem 1.9.7 there exists $m \in \mathbb{Z}$ such that

$$nx < m < ny$$

and so take $r = m/n$. ■

Definition 1.9.9 *A set $S \subseteq \mathbb{R}$ is dense in $\mathbb{R}$ if whenever $a < b$, $S \cap (a, b) \neq \emptyset$.*

Thus the above theorem says $\mathbb{Q}$ is "dense" in $\mathbb{R}$.

Theorem 1.9.10 *Suppose $0 < a$ and let $b \geq 0$. Then there exists a unique integer p and real number r such that $0 \leq r < a$ and $b = pa + r$.*

Proof: Let $S \equiv \{n \in \mathbb{N} : an > b\}$. By the Archimedean property this set is nonempty. Let $p + 1$ be the smallest element of S. Then $pa \leq b$ because $p + 1$ is the smallest in S. Therefore,

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

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

To verify uniqueness of p and r, suppose p_i and r_i, $i = 1, 2$, both work and $r_2 > r_1$. Then a little algebra shows

$$p_1 - p_2 = \frac{r_2 - r_1}{a} \in (0, 1).$$

Thus $p_1 - p_2$ is an integer between 0 and 1, contradicting Theorem 1.9.7. The case that $r_1 > r_2$ cannot occur either by similar reasoning. Thus $r_1 = r_2$ and it follows that $p_1 = p_2$. ■

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

1.10 Division

First recall Theorem 1.9.10, the Euclidean algorithm.

Theorem 1.10.1 *Suppose* $0 < a$ *and let* $b \geq 0$. *Then there exists a unique integer* p *and unique real number* r *such that* $0 \leq r < a$ *and* $b = pa + r$.

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

Definition 1.10.2 *The number,* a *divides the number,* b *if in Theorem 1.9.10,* $r = 0$. *That is there is zero remainder. The notation for this is* $a|b$, *read* a *divides* b *and* a *is called a factor of* b. *A prime number is one which has the property that the only numbers which divide it are itself and 1. The greatest common divisor of two positive integers,* m, n *is that number,* p *which has the property that* p *divides both* m *and* n *and also if* q *divides both* m *and* n, *then* q *divides* p. *Two integers are relatively prime if their greatest common divisor is one. The greatest common divisor of* m *and* n *is denoted as* (m, n).

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

Theorem 1.10.3 *Let* m, n *be two positive integers and define*

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

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

Proof: First note that both m and n are in S so it is a nonempty set of positive integers. By well ordering, there is a smallest element of S, called $p = x_0 m + y_0 n$. Either p divides m or it does not. If p does not divide m, then by Theorem 1.9.10,

$$m = pq + r$$

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

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

However, this is a contradiction because p was the smallest element of S. Thus $p|m$. Similarly $p|n$.

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

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

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

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

$$n = qm + r, \ r < m \tag{1.1}$$

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

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

Thus d divides both m and r. If k divides both m and r, then from the equation of 1.1 it follows k also divides n. Therefore, k divides d by the definition of the greatest common divisor. Thus d is the greatest common divisor of m and r but $m + r < m + n$. This yields another pair of positive integers for which d is still the greatest common divisor but the sum of these integers is strictly smaller than the sum of the first two. Now you can do the same thing to these integers. Eventually the process must end because the sum gets strictly smaller each time it is done. It ends when there are not two positive integers produced. That is, one is a multiple of the other. At this point, the greatest common divisor is the smaller of the two numbers.

Procedure 1.10.4 *To find the greatest common divisor of m, n where $0 < m < n$, replace the pair $\{m, n\}$ with $\{m, r\}$ where $n = qm + r$ for $r < m$. This new pair of numbers has the same greatest common divisor. Do the process to this pair and continue doing this till you obtain a pair of numbers where one is a multiple of the other. Then the smaller is the sought for greatest common divisor.*

Example 1.10.5 *Find the greatest common divisor of* 165 *and* 385.

Use the Euclidean algorithm to write

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

Thus the next two numbers are 55 and 165. Then

$$165 = 3 \times 55$$

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

Example 1.10.6 *Find the greatest common divisor of 1237 and 4322.*

Use the Euclidean algorithm

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

Now the two new numbers are 1237,611. Then

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

The two new numbers are 15,611. Then

$$611 = 40(15) + 11$$

The two new numbers are 15,11. Then

$$15 = 1(11) + 4$$

The two new numbers are 11,4

$$2(4) + 3$$

The two new numbers are $4, 3$. Then

$$4 = 1(3) + 1$$

The two new numbers are $3, 1$. Then

$$3 = 3 \times 1$$

and so 1 is the greatest common divisor. Of course you could see this right away when the two new numbers were 15 and 11. Recall the process delivers numbers which have the same greatest common divisor.

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

Theorem 1.10.7 *If p is a prime and $p|ab$ then either $p|a$ or $p|b$.*

Proof: Suppose p does not divide a. Then since p is prime, the only factors of p are 1 and p so follows $(p, a) = 1$ and therefore, there exists integers, x and y such that

$$1 = ax + yp.$$

Multiplying this equation by b yields

$$b = abx + ybp.$$

Since $p|ab$, $ab = pz$ for some integer z. Therefore,

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

and this shows p divides b. ■

Theorem 1.10.8 *(Fundamental theorem of arithmetic) Let $a \in \mathbb{N}\backslash\{1\}$. Then $a = \prod_{i=1}^{n} p_i$ where p_i are all prime numbers. Furthermore, this prime factorization is unique except for the order of the factors.*

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

Suppose

$$\prod_{i=1}^{n} p_i = \prod_{j=1}^{m} q_j$$

where the p_i and q_j are all prime, there is no way to reorder the q_k such that $m = n$ and $p_i = q_i$ for all i, and $n + m$ is the smallest positive integer such that this happens. Then by Theorem 1.10.7, $p_1|q_j$ for some j. Since these are prime numbers this requires $p_1 = q_j$. Reordering if necessary it can be assumed that $q_j = q_1$. Then dividing both sides by $p_1 = q_1$,

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

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

There is a similar division result for polynomials. This will be discussed more intensively later. For now, here is a definition and the division theorem.

Definition 1.10.9 *A polynomial is an expression of the form* $a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$, $a_n \neq 0$ *where the* a_i *come from a field of scalars. Two polynomials are equal means that **the coefficients match for each power of*** λ***.*** *The degree of a polynomial is the largest power of* λ*. Thus the degree of the above polynomial is* n*. Addition of polynomials is defined in the usual way as is multiplication of two polynomials. The leading term in the above polynomial is* $a_n\lambda^n$*. The coefficient of the leading term is called the leading coefficient. It is called a monic polynomial if* $a_n = 1$*.*

Lemma 1.10.10 *Let* $f(\lambda)$ *and* $g(\lambda) \neq 0$ *be polynomials. Then there exist polynomials,* $q(\lambda)$ *and* $r(\lambda)$ *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 $\hat{q}(\lambda) = q(\lambda)$. ■

1.11 Systems of Equations

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

$$x + y = 7 \text{ and } 2x - y = 8. \tag{1.2}$$

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

$$x + y = 7, \overbrace{2x - y + (-2)(x + y) = 8 + (-2)(7)}^{-3y=-6}. \tag{1.3}$$

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

Why exactly does the replacement of one equation with a multiple of another added to it not change the solution set? The two equations of 1.2 are of the form

$$E_1 = f_1, E_2 = f_2 \tag{1.4}$$

where E_1 and E_2 are expressions involving the variables. The claim is that if a is a number, then 1.4 has the same solution set as

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{1.5}$$

Why is this?

If (x, y) solves 1.4 then it solves the first equation in 1.5. Also, it satisfies $aE_1 = af_1$ and so, since it also solves $E_2 = f_2$ it must solve the second equation in 1.5. If (x, y) solves 1.5 then it solves the first equation of 1.4. Also $aE_1 = af_1$ and it is given that the second equation of 1.5 is verified. Therefore, $E_2 = f_2$ and it follows (x, y) is a solution of the second equation in 1.4. This shows the solutions to 1.4 and 1.5 are exactly the same which means they have the same solution set. Of course the same reasoning applies with no change if there are many more variables than two and many more equations than two. It is still the case that when one equation is replaced with a multiple of another one added to itself, the solution set of the whole system does not change.

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

Example 1.11.1 *Find the solutions to the system,*

$$\begin{aligned} x + 3y + 6z &= 25 \\ 2x + 7y + 14z &= 58 \\ 2y + 5z &= 19 \end{aligned} \tag{1.6}$$

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

$$\begin{aligned} x + 3y + 6z &= 25 \\ y + 2z &= 8 \\ 2y + 5z &= 19 \end{aligned} \tag{1.7}$$

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

$$\begin{aligned} x + 3y + 6z &= 25 \\ y + 2z &= 8 \\ z &= 3 \end{aligned} \tag{1.8}$$

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

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

In 1.8 you could have continued as follows. Add (-2) times the bottom equation to the middle and then add (-6) times the bottom to the top. This yields

$$\begin{aligned} x + 3y &= 19 \\ y &= 6 \\ z &= 3 \end{aligned}$$

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

$$x = 1,\ y = 6,\ z = 3,$$

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

It is foolish to write the variables every time you do these operations. It is easier to write the system 1.6 as the following "augmented matrix"

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

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

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

Now when you replace an equation with a multiple of another equation added to itself, you are just taking a row of this augmented matrix and replacing it with a multiple of another row added to it. Thus the first step in solving 1.6 would be to take (-2) times the first row of the augmented matrix above and add it to the second row,

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

Note how this corresponds to 1.7. Next take (-2) times the second row and add to the third,

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

which is the same as 1.8. You get the idea I hope. Write the system as an augmented matrix and follow the procedure of either switching rows, multiplying a row by a non zero number, or replacing a row by a multiple of another row added to it. Each of these operations leaves the solution set unchanged. These operations are called row operations.

Definition 1.11.2 *The row operations consist of the following*

1. *Switch two rows.*

2. *Multiply a row by a nonzero number.*

3. *Replace a row by a multiple of another row added to it.*

It is important to observe that any row operation can be "undone" by another inverse row operation. For example, if $\mathbf{r}_1, \mathbf{r}_2$ are two rows, and $\mathbf{r}_2$ is replaced with $\mathbf{r}_2' = \alpha\mathbf{r}_1 + \mathbf{r}_2$ using row operation 3, then you could get back to where you started by replacing the row $\mathbf{r}_2'$ with $-\alpha$ times $\mathbf{r}_1$ and adding to $\mathbf{r}_2'$. In the case of operation 2, you would simply multiply the row that was changed by the inverse of the scalar which multiplied it in the first place, and in the case of row operation 1, you would just make the same switch again and you would be back to where you started. In each case, the row operation which undoes what was done is called the **inverse row operation.**

Example 1.11.3 *Give the complete solution to the system of equations,* $5x+10y-7z = -2$, $2x + 4y - 3z = -1$, *and* $3x + 6y + 5z = 9$.

The augmented matrix for this system is

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

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

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

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

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

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

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

Example 1.11.4 *Give the complete solution to the system of equations,* $3x - y - 5z = 9$, $y - 10z = 0$, *and* $-2x + y = -6$.

The augmented matrix of this system is

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

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

$$\left(\begin{array}{cccc} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 1 & -10 & 0 \end{array}\right)$$

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

$$\left(\begin{array}{cccc} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

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

$$\left(\begin{array}{cccc} 1 & 0 & -5 & 3 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

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

Definition 1.11.5 *Since $z = t$ where t is arbitrary, the variable z is called a* ***free variable.***

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

Definition 1.11.6 *A system of linear equations is a list of equations,*

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

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

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

Example 1.11.7 *Give the complete solution to the system of equations, $-41x + 15y = 168$, $109x - 40y = -447$, $-3x + y = 12$, and $2x + z = -1$.*

The augmented matrix is

$$\left(\begin{array}{cccc} -41 & 15 & 0 & 168 \\ 109 & -40 & 0 & -447 \\ -3 & 1 & 0 & 12 \\ 2 & 0 & 1 & -1 \end{array}\right).$$

To solve this multiply the top row by 109, the second row by 41, add the top row to the second row, and multiply the top row by 1/109. Note how this process combined several row operations. This yields

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

Next take 2 times the third row and replace the fourth row by this added to 3 times the fourth row. Then take (-41) times the third row and replace the first row by this added to 3 times the first row. Then switch the third and the first rows. This yields

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

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

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

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

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

1.12 Exercises

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

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

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

4. Give the complete solution to the system of equations, $x+2y+6z = 5, 3x+2y+6z = 7$,$-4x + 5y + 15z = -7$.

5. Give the complete solution to the system of equations

$$\begin{aligned} x + 2y + 3z &= 5, 3x + 2y + z = 7, \\ -4x + 5y + z &= -7, x + 3z = 5. \end{aligned}$$

6. Give the complete solution of the system of equations,

$$\begin{aligned} x + 2y + 3z &= 5,\ 3x + 2y + 2z = 7 \\ -4x + 5y + 5z &= -7,\ x = 5 \end{aligned}$$

7. Give the complete solution of the system of equations

$$\begin{aligned} x+y+3z &= 2,\ 3x-y+5z=6 \\ -4x+9y+z &= -8,\ x+5y+7z=2 \end{aligned}$$

8. Determine a such that there are infinitely many solutions and then find them. Next determine a such that there are no solutions. Finally determine which values of a correspond to a unique solution. The system of equations for the unknown variables x, y, z is

$$\begin{gathered} 3za^2-3a+x+y+1=0 \\ 3x-a-y+z\left(a^2+4\right)-5=0 \\ za^2-a-4x+9y+9=0 \end{gathered}$$

9. Find the solutions to the following system of equations for x, y, z, w.

$$y+z=2, z+w=0, y-4z-5w=2, 2y+z-w=4$$

10. Find all solutions to the following equations.

$$\begin{aligned} x+y+z &= 2,\ z+w=0, \\ 2x+2y+z-w &= 4,\ x+y-4z-5z=2 \end{aligned}$$

1.13 $\mathbb{F}^n$

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

Definition 1.13.1 *Define* $\mathbb{F}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{F} \text{ for } j = 1, \cdots, n\}$. $(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ *if and only if for all* $j = 1, \cdots, n$, $x_j = y_j$. *When* $(x_1, \cdots, x_n) \in \mathbb{F}^n$, *it is conventional to denote* $(x_1, \cdots, x_n)$ *by the single bold face letter* $\mathbf{x}$. *The numbers* x_j *are called the coordinates. The set*

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

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

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

1.14 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 1.14.1 *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*

$$a\mathbf{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{1.9}$$

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

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{1.10}$$

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

Theorem 1.14.2 *For $\mathbf{v}, \mathbf{w} \in \mathbb{F}^n$ and α, β scalars, (real numbers), the following hold.*

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{1.11}$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{1.12}$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{1.13}$$

the existence of an additive identity,

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{1.14}$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \tag{1.15}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{1.16}$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \tag{1.17}$$

$$1\mathbf{v} = \mathbf{v}. \tag{1.18}$$

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

You should verify that these properties all hold. As usual subtraction is defined as $\mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y})$. The conclusions of the above theorem are called the vector space axioms.

1.15 Exercises

1. Verify all the properties 1.11-1.18.
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. If you have trouble drawing this, describe it in words.
   (a) $(1, 2, 0)$
   (b) $(-2, -2, 1)$
   (c) $(-2, 3, -2)$

1.16 The Inner Product in $\mathbb{F}^n$

When $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, there is something called an inner product. In case of $\mathbb{R}$ it is also called the dot product. This is also often referred to as the scalar product.

Definition 1.16.1 *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{F}^n$ *define* $\mathbf{a} \cdot \mathbf{b}$ *as*

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

This will also be denoted as $(\mathbf{a}, \mathbf{b})$*. Often it is also denoted as* $\langle \mathbf{a}, \mathbf{b} \rangle$*. The notation with the dot is more usually used when the field is* $\mathbb{R}$*.*

With this definition, there are several important properties satisfied by the inner product. In the statement of these properties, α and β will denote scalars and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ will denote vectors or in other words, points in $\mathbb{F}^n$.

Proposition 1.16.2 *The inner product satisfies the following properties.*

$$\mathbf{a} \cdot \mathbf{b} = \overline{\mathbf{b} \cdot \mathbf{a}} \tag{1.19}$$

$$\mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \tag{1.20}$$

$$(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \tag{1.21}$$

$$\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) = \overline{\alpha} (\mathbf{c} \cdot \mathbf{a}) + \overline{\beta} (\mathbf{c} \cdot \mathbf{b}) \tag{1.22}$$

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \tag{1.23}$$

You should verify these properties. Also be sure you understand that 1.22 follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 1.16.3 *Find* $(1, 2, 0, -1) \cdot (0, i, 2, 3)$.

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

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

Theorem 1.16.4 *The inner product satisfies the inequality*

$$|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|. \tag{1.24}$$

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

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

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

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

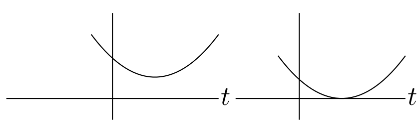
$$f\left(t\right) = \left(\mathbf{a}+t\theta\mathbf{b}\right)\cdot\left(\mathbf{a}+t\theta\mathbf{b}\right).$$

Then by 1.20, $f\left(t\right) \geq 0$ for all $t \in \mathbb{R}$. Also from 1.21,1.22,1.19, and 1.23

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

$$= \left|\mathbf{a}\right|^2 + 2t\operatorname{Re}\overline{\theta}\left(\mathbf{a}\cdot\mathbf{b}\right) + \left|\mathbf{b}\right|^2 t^2 = \left|\mathbf{a}\right|^2 + 2t\left|\mathbf{a}\cdot\mathbf{b}\right| + \left|\mathbf{b}\right|^2 t^2$$

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



Then the quadratic formula requires that

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

since otherwise the function, $f\left(t\right)$ would have two real zeros and would necessarily have a graph which dips below the t axis. This proves 1.24.

It is clear from the axioms of the inner product that equality holds in 1.24 whenever one of the vectors is a scalar multiple of the other. It only remains to verify this is the only way equality can occur. If either vector equals zero, then equality is obtained in 1.24 so it can be assumed both vectors are non zero. Then if equality is achieved, it follows $f\left(t\right)$ has exactly one real zero because the discriminant vanishes. Therefore, for some value of $t, \mathbf{a}+t\theta\mathbf{b} = \mathbf{0}$ showing that $\mathbf{a}$ is a multiple of $\mathbf{b}$. ■

You should note that the entire argument was based only on the properties of the inner product listed in 1.19 - 1.23. This means that whenever something satisfies these properties, the Cauchy Schwarz inequality holds. There are many other instances of these properties besides vectors in $\mathbb{F}^n$. Also note that 1.24 holds if 1.20 is simplified to $\mathbf{a}\cdot\mathbf{a} \geq 0$.

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

Theorem 1.16.5 *(Triangle inequality) For* $\mathbf{a}, \mathbf{b} \in \mathbb{F}^n$

$$\left|\mathbf{a}+\mathbf{b}\right| \leq \left|\mathbf{a}\right| + \left|\mathbf{b}\right| \tag{1.25}$$

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

$$\left|\left|\mathbf{a}\right| - \left|\mathbf{b}\right|\right| \leq \left|\mathbf{a}-\mathbf{b}\right| \tag{1.26}$$

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

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

$$= \left|\mathbf{a}\right|^2 + 2\operatorname{Re}\left(\mathbf{a}\cdot\mathbf{b}\right) + \left|\mathbf{b}\right|^2 \leq \left|\mathbf{a}\right|^2 + 2\left|\mathbf{a}\cdot\mathbf{b}\right| + \left|\mathbf{b}\right|^2$$

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

Taking square roots of both sides you obtain 1.25.

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

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

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

To get the other form of the triangle inequality,

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

so

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

Therefore,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| \tag{1.27}$$

Similarly,

$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}|. \tag{1.28}$$

It follows from 1.27 and 1.28 that 1.26 holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either 1.27 or 1.28 and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|$. ■

1.17 What is Linear Algebra?

The above preliminary considerations form the necessary scaffolding upon which linear algebra is built. Linear algebra is the study of a certain algebraic structure called a vector space described in a special case in Theorem 1.14.2 and in more generality below along with special functions known as linear transformations. These linear transformations preserve certain algebraic properties.

A good argument could be made that linear algebra is the most useful subject in all of mathematics and that it exceeds even courses like calculus in its significance. It is used extensively in applied mathematics and engineering. Continuum mechanics, for example, makes use of topics from linear algebra in defining things like the strain and in determining appropriate constitutive laws. It is fundamental in the study of statistics. For example, principal component analysis is really based on the singular value decomposition discussed in this book. It is also fundamental in pure mathematics areas like number theory, functional analysis, geometric measure theory, and differential geometry. Even calculus cannot be correctly understood without it. For example, the derivative of a function of many variables is an example of a linear transformation, and this is the way it must be understood as soon as you consider functions of more than one variable.

1.18 Exercises

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

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

3. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, define $\mathbf{a} \cdot \mathbf{b} \equiv \sum_{k=1}^{n} \beta_k a_k b_k$ where $\beta_k > 0$ for each k. Show this satisfies the axioms of the inner product. What does the Cauchy Schwarz inequality say in this case.

4. In Problem 3 above, suppose you only know $\beta_k \geq 0$. Does the Cauchy Schwarz inequality still hold? If so, prove it.

5. Let f, g be continuous functions and define $f \cdot g \equiv \int_0^1 f(t) \overline{g(t)} dt$. Show this satisfies the axioms of a inner product if you think of continuous functions in the place of a vector in $\mathbb{F}^n$. What does the Cauchy Schwarz inequality say in this case?

6. Show that if f is a real valued continuous function, $\left(\int_a^b f(t)\, dt \right)^2 \leq (b-a) \int_a^b f(t)^2\, dt$.

Chapter 2

Linear Transformations

2.1 Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out that such rectangular arrays of numbers are important from many other different points of view. Numbers are also called scalars. In general, scalars are just elements of some field. However, in the first part of this book, the field will typically be either the real numbers or the complex numbers.

A matrix is a rectangular array of numbers. Several of them are referred to as matrices. For example, here is a matrix.

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

This matrix is a 3×4 matrix because there are three rows and four columns. The first row is $(1\ 2\ 3\ 4)$, the second row is $(5\ 2\ 8\ 7)$ and so forth. The first column is $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. The convention in dealing with matrices is to always list the rows first and then the columns. Also, you can remember the columns are like columns in a Greek temple. They stand up right while the rows just 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 in which the i denotes the row and the j denotes the column. Using this notation on the above matrix, $a_{23} = 8, a_{32} = -9, a_{12} = 2$, etc.

There are various operations which are done on matrices. They can sometimes be added, multiplied by a scalar and sometimes multiplied. To illustrate scalar multiplication, consider the following example.

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

The new matrix is obtained by multiplying every entry of the original matrix by the given scalar. If A is an $m \times n$ matrix $-A$ is defined to equal $(-1) A$.

Two matrices which are the same size can be added. When this is done, the result is the matrix which is obtained by adding corresponding entries. Thus

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 4 \\ 2 & 8 \\ 6 & -4 \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ 5 & 12 \\ 11 & -2 \end{pmatrix}.$$

Two matrices are equal exactly when they are the same size and the corresponding entries are identical. Thus

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

because they are different sizes. As noted above, you write (c_{ij}) for the matrix C whose ij^{th} entry is c_{ij}. In doing arithmetic with matrices you must define what happens in terms of the c_{ij} sometimes called the entries of the matrix or the components of the matrix.

The above discussion stated for general matrices is given in the following definition.

Definition 2.1.1 *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. Then $A + B = C$ where*

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

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

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

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

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

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

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

$$A + B = B + A, \tag{2.1}$$

the commutative law of addition,

$$(A + B) + C = A + (B + C), \tag{2.2}$$

the associative law for addition,

$$A + 0 = A, \tag{2.3}$$

the existence of an additive identity,

$$A + (-A) = 0, \tag{2.4}$$

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

$$\alpha (A + B) = \alpha A + \alpha B, \tag{2.5}$$

$$(\alpha + \beta) A = \alpha A + \beta A, \tag{2.6}$$

$$\alpha (\beta A) = \alpha\beta (A), \tag{2.7}$$

$$1A = A. \tag{2.8}$$

The above properties, 2.1 - 2.8 are known as the vector space axioms and the fact that the $m \times n$ matrices satisfy these axioms is what is meant by saying this set of matrices with addition and scalar multiplication as defined above forms a vector space.

Definition 2.1.2 *Matrices which are $n \times 1$ or $1 \times n$ are especially called vectors and are often denoted by a bold letter. Thus*

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

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

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

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

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

It equals

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

Thus it is what is called a **linear combination** of the columns. These will be discussed more later. Motivated by this example, here is the definition of how to multiply an $m \times n$ matrix by an $n \times 1$ matrix (vector).

Definition 2.1.3 *Let $A = A_{ij}$ be an $m \times n$ matrix and let $\mathbf{v}$ be an $n \times 1$ matrix,*

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}, \ A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$$

where $\mathbf{a}_i$ is an $m \times 1$ 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.

$$v_1 \mathbf{a}_1 + v_2 \mathbf{a}_2 + \cdots + v_n \mathbf{a}_n \equiv \sum_{j=1}^{n} v_j \mathbf{a}_j \tag{2.9}$$

If the j^{th} column of A is

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

then 2.9 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 2.1.4 *Compute*

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

First of all, this is of the form $(3 \times 4)(4 \times 1)$ and so the result should be a (3×1). Note how the inside numbers cancel. To get the entry in the second row and first and only column, compute

$$\begin{aligned} \sum_{k=1}^{4} a_{2k} v_k &= a_{21} v_1 + a_{22} v_2 + a_{23} v_3 + a_{24} v_4 \\ &= 0 \times 1 + 2 \times 2 + 1 \times 0 + (-2) \times 1 = 2. \end{aligned}$$

You should do the rest of the problem and verify

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

With this done, the next task is to multiply an $m \times n$ matrix times an $n \times p$ matrix. Before doing so, the following may be helpful.

$$(m \times n)\ (n \times p) = m \times p$$

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

The number of columns on the left equals the number of rows on the right.

Definition 2.1.5 *Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. Then B is of the form*

$$B = (\mathbf{b}_1, \cdots, \mathbf{b}_p)$$

where $\mathbf{b}_k$ is an $n \times 1$ matrix. Then an $m \times p$ matrix AB is defined as follows:

$$AB \equiv (A\mathbf{b}_1, \cdots, A\mathbf{b}_p) \tag{2.10}$$

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

Example 2.1.6 *Multiply the following.*

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

The first thing you need to check before doing anything else is whether it is possible to do the multiplication. The first matrix is a 2×3 and the second matrix is a 3×3. Therefore, is it possible to multiply these matrices. According to the above discussion it should be a 2×3 matrix of the form

$$\left(\overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}}^{\text{First column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}}^{\text{Second column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}}^{\text{Third column}} \right)$$

You know how to multiply a matrix times a vector and so you do so to obtain each of the three columns. Thus

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

Here is another example.

Example 2.1.7 *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.

$$\boxed{\textbf{Order Matters!}}$$

Matrix multiplication is not commutative. This is very different than multiplication of numbers!

2.1.1 The ij^{th} Entry of a Product

It is important to describe matrix multiplication in terms of entries of the matrices. What is the ij^{th} entry of AB? It would be the i^{th} entry of the j^{th} column of AB. Thus it would be the i^{th} entry of $A\mathbf{b}_j$. Now

$$\mathbf{b}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}$$

and from the above definition, the i^{th} entry is

$$\sum_{k=1}^{n} A_{ik} B_{kj}. \tag{2.11}$$

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 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 harmonizes with Definition 2.1.3.

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

Definition 2.1.8 *Let $A = (A_{ij})$ be an $m \times n$ matrix and let $B = (B_{ij})$ be an $n \times p$ matrix. Then AB is an $m \times p$ matrix and*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}. \tag{2.12}$$

Two matrices, A and B are said to be conformable in a particular order if they can be multiplied in that order. Thus if A is an $r \times s$ matrix and B is a $s \times p$ then A and B are conformable in the order AB. The above formula for $(AB)_{ij}$ says that it equals the i^{th} row of A times the j^{th} column of B.

Example 2.1.9 *Multiply if possible* $\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix}$.

First check to see if this is possible. It is of the form $(3 \times 2)(2 \times 3)$ and since the inside numbers match, it must be possible to do this and the result should be a 3×3 matrix. The answer is of the form

$$\left(\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)$$

where the commas separate the columns in the resulting product. Thus the above product equals

$$\begin{pmatrix} 16 & 15 & 5 \\ 13 & 15 & 5 \\ 46 & 42 & 14 \end{pmatrix},$$

a 3×3 matrix as desired. In terms of the ij^{th} entries and the above definition, the entry in the third row and second column of the product should equal

$$\sum_j a_{3k} b_{k2} = a_{31} b_{12} + a_{32} b_{22} = 2 \times 3 + 6 \times 6 = 42.$$

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 2.1.10 *Multiply if possible* $\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix}$.

This is not possible because it is of the form $(3 \times 2)(3 \times 3)$ and the middle numbers don't match.

Example 2.1.11 *Multiply if possible* $\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}$.

This is possible because in this case it is of the form $(3 \times 3)(3 \times 2)$ and the middle numbers do match. When the multiplication is done it equals

$$\begin{pmatrix} 13 & 13 \\ 29 & 32 \\ 0 & 0 \end{pmatrix}.$$

Check this and be sure you come up with the same answer.

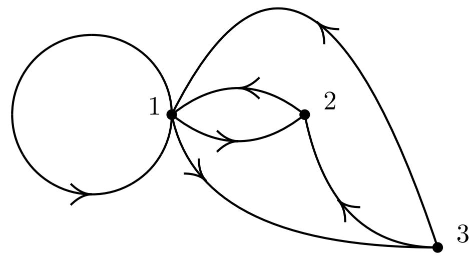
Example 2.1.12 *Multiply if possible* $\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix}$.

In this case you are trying to do $(3 \times 1)(1 \times 4)$. The inside numbers match so you can do it. Verify

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

2.1.2 Digraphs

Consider the following graph illustrated in the picture.



There are three locations in this graph, labelled 1,2, and 3. The directed lines represent a way of going from one location to another. Thus there is one way to go from location 1 to location 1. There is one way to go from location 1 to location 3. It is not possible to go from location 2 to location 3 although it is possible to go from location 3 to location 2. Lets refer to moving along one of these directed lines as a step. The following 3×3 matrix is a numerical way of writing the above graph. This is sometimes called a digraph, short for directed graph.

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

Thus a_{ij}, the entry in the i^{th} row and j^{th} column represents the number of ways to go from location i to location j in one step.

Problem: Find the number of ways to go from i to j using exactly k steps.

Denote the answer to the above problem by a_{ij}^k. We don't know what it is right now unless $k = 1$ when it equals a_{ij} described above. However, if we did know what it was, we could find a_{ij}^{k+1} as follows.

$$a_{ij}^{k+1} = \sum_r a_{ir}^k a_{rj}$$

This is because if you go from i to j in $k+1$ steps, you first go from i to r in k steps and then for each of these ways there are a_{rj} ways to go from there to j. Thus $a_{ir}^k a_{rj}$ gives the number of ways to go from i to j in $k+1$ steps such that the k^{th} step leaves you at location r. Adding these gives the above sum. Now you recognize this as the ij^{th} entry of the product of two matrices. Thus

$$a_{ij}^2 = \sum_r a_{ir} a_{rj}, \quad a_{ij}^3 = \sum_r a_{ir}^2 a_{rj}$$

and so forth. From the above definition of matrix multiplication, this shows that if A is the matrix associated with the directed graph as above, then a_{ij}^k is just the ij^{th} entry of A^k where A^k is just what you would think it should be, A multiplied by itself k times.

Thus in the above example, to find the number of ways of going from 1 to 3 in two steps you would take that matrix and multiply it by itself and then take the entry in the first row and third column. Thus

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

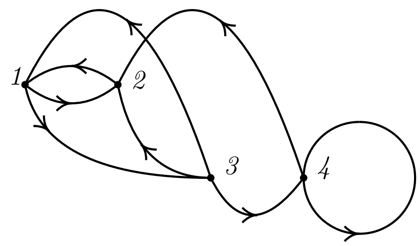
and you see there is exactly one way to go from 1 to 3 in two steps. You can easily see this is true from looking at the graph also. Note there are three ways to go from 1 to 1 in 2 steps. Can you find them from the graph? What would you do if you wanted to consider 5 steps?

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}^5 = \begin{pmatrix} 28 & 19 & 13 \\ 13 & 9 & 6 \\ 19 & 13 & 9 \end{pmatrix}$$

There are 19 ways to go from 1 to 2 in five steps. Do you think you could list them all by looking at the graph? I don't think you could do it without wasting a lot of time.

Of course there is nothing sacred about having only three locations. Everything works just as well with any number of locations. In general if you have n locations, you would need to use a $n \times n$ matrix.

Example 2.1.13 *Consider the following directed graph.*



Write the matrix which is associated with this directed graph and find the number of ways to go from 2 to 4 in three steps.

Here you need to use a 4×4 matrix. The one you need is

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

Then to find the answer, you just need to multiply this matrix by itself three times and look at the entry in the second row and fourth column.

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

There is exactly one way to go from 2 to 4 in three steps.

How many ways would there be of going from 2 to 4 in five steps?

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}^5 = \begin{pmatrix} 5 & 9 & 5 & 4 \\ 5 & 4 & 1 & 3 \\ 9 & 10 & 4 & 6 \\ 4 & 6 & 3 & 3 \end{pmatrix}$$

There are three ways. Note there are 10 ways to go from 3 to 2 in five steps.

This is an interesting application of the concept of the ij^{th} entry of the product matrices.

2.1.3 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 they be equal then?

Example 2.1.14 *Compare* $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ *and* $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

The first product is

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

the second product is

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

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

Proposition 2.1.15 *If all multiplications and additions make sense, the following hold for matrices,* A, B, C *and* a, b *scalars.*

$$A(aB + bC) = a(AB) + b(AC) \tag{2.13}$$

$$(B + C)A = BA + CA \tag{2.14}$$

$$A(BC) = (AB)C \tag{2.15}$$

Proof: Using the above definition of matrix multiplication,

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

showing that $A(B + C) = AB + AC$ as claimed. Formula 2.14 is entirely similar.

Consider 2.15, the associative law of multiplication. Before reading this, review the definition of matrix multiplication in terms of entries of the matrices.

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

Another important operation on matrices is that of taking the transpose. The following example shows what is meant by this operation, denoted by placing a T as an exponent on the matrix.

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

What happened? The first column became the first row and the second column became the second row. Thus the 3×2 matrix became a 2×3 matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. This motivates the following definition of the transpose of a matrix.

Definition 2.1.16 *Let A be an $m \times n$ matrix. Then A^T denotes the $n \times m$ matrix which is defined as follows.*

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

The transpose of a matrix has the following important property.

Lemma 2.1.17 *Let A be an $m \times n$ matrix and let B be a $n \times p$ matrix. Then*

$$(AB)^T = B^T A^T \tag{2.16}$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \tag{2.17}$$

Proof: From the definition,

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

2.17 is left as an exercise. ■

Definition 2.1.18 *An $n \times n$ matrix A is said to be symmetric if $A = A^T$. It is said to be skew symmetric if $A^T = -A$.*

Example 2.1.19 *Let*

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

Then A is symmetric.

Example 2.1.20 *Let*

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

Then A is skew symmetric.

There is a special matrix called I and defined by

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

where δ_{ij} is the 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 2.1.21 *Suppose A is an $m \times n$ matrix and I_n is the $n \times n$ identity matrix. Then $AI_n = A$. If I_m is the $m \times m$ identity matrix, it also follows that $I_m A = A$.*

Proof:

$$\begin{aligned} (AI_n)_{ij} &= \sum_k A_{ik}\delta_{kj} \\ &= A_{ij} \end{aligned}$$

and so $AI_n = A$. The other case is left as an exercise for you.

Definition 2.1.22 *An $n \times n$ matrix A has an inverse A^{-1} if and only if there exists a matrix, denoted as A^{-1} such that $AA^{-1} = A^{-1}A = I$ where $I = (\delta_{ij})$ for*

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

Such a matrix is called invertible.

If it acts like an inverse, then it is the inverse. This is the message of the following proposition.

Proposition 2.1.23 *Suppose $AB = BA = I$. Then $B = A^{-1}$.*

Proof: From the definition B is an inverse for A. Could there be another one B'?

$$B' = B'I = B'(AB) = (B'A)B = IB = B.$$

Thus, the inverse, if it exists, is unique. ■

2.1.4 Finding The Inverse of a Matrix

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

Example 2.1.24 *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 multiply on the left by the inverse A and conclude the vector $(-1,1)^T = (0,0)^T$. Thus the answer is that A does not have an inverse.

Suppose you want to find B such that $AB = I$. Let

$$B = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{pmatrix}$$

Also the i^{th} column of I is

$$\mathbf{e}_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}^T$$

Thus, if $AB = I$, $\mathbf{b}_i$, the i^{th} column of B must satisfy the equation $A\mathbf{b}_i = \mathbf{e}_i$. The augmented matrix for finding $\mathbf{b}_i$ is $(A|\mathbf{e}_i)$. Thus, by doing row operations till A becomes I, you end up with $(I|\mathbf{b}_i)$ where $\mathbf{b}_i$ is the solution to $A\mathbf{b}_i = \mathbf{e}_i$. Now the same sequence of row operations works regardless of the right side of the agumented matrix $(A|\mathbf{e}_i)$ and so you can save trouble by simply doing the following.

$$(A|I) \overset{\text{row operations}}{\rightarrow} (I|B)$$

and the i^{th} column of B is $\mathbf{b}_i$, the solution to $A\mathbf{b}_i = \mathbf{e}_i$. Thus $AB = I$.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the Gauss Jordan procedure. It produces the inverse if the matrix has one. Actually, it produces the right inverse.

Procedure 2.1.25 *Suppose A is an $n \times n$ matrix. To find A^{-1} if it exists, form the augmented $n \times 2n$ matrix,*

$$(A|I)$$

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

$$(I|B) \tag{2.18}$$

if possible. When this has been done, $B = A^{-1}$. The matrix A has an inverse exactly when it is possible to do row operations and end up with one like 2.18.

As described above, the following is a description of what you have just done.

$$A \overset{R_q R_{q-1} \cdots R_1}{\rightarrow} I$$
$$I \overset{R_q R_{q-1} \cdots R_1}{\rightarrow} B$$

where those R_i sympolize row operations. It follows that you could undo what you did by doing the inverse of these row operations in the opposite order. Thus

$$I \overset{R_1^{-1} \cdots R_{q-1}^{-1} R_q^{-1}}{\rightarrow} A$$
$$B \overset{R_1^{-1} \cdots R_{q-1}^{-1} R_q^{-1}}{\rightarrow} I$$

Here R^{-1} is the row operation which undoes the row operation R. Therefore, if you form $(B|I)$ and do the inverse of the row operations which produced I from A in the reverse order, you would obtain $(I|A)$. By the same reasoning above, it follows that A is a right inverse of B and so $BA = I$ also. It follows from Proposition 2.1.23 that $B = A^{-1}$. Thus the procedure produces **the** inverse whenever it works.

If it is possible to do row operations and end up with $A \overset{\text{row operations}}{\to} I$, then the above argument shows that A has an inverse. Conversely, if A has an inverse, can it be found by the above procedure? In this case there exists a unique solution $\mathbf{x}$ to the equation $A\mathbf{x} = \mathbf{y}$. In fact it is just $\mathbf{x} = I\mathbf{x} = A^{-1}\mathbf{y}$. Thus in terms of augmented matrices, you would expect to obtain

$$(A|\mathbf{y}) \to \left(I|A^{-1}\mathbf{y}\right)$$

That is, you would expect to be able to do row operations to A and end up with I.

The details will be explained fully when a more careful discussion is given which is based on more fundamental considerations. For now, it suffices to observe that whenever the above procedure works, it finds the inverse.

Example 2.1.26 *Let* $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. *Find* A^{-1}.

Form the augmented matrix

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

Now do row operations 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 2.1.27 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. *Find* A^{-1}.

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

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

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

$$\left(\begin{array}{cccccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -5 & -7 & -3 & 0 & 1 \end{array}\right).$$

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

$$\left(\begin{array}{cccccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right)$$

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

$$\left(\begin{array}{cccccc} -7 & -14 & 0 & -6 & 5 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

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

$$\left(\begin{array}{cccccc} -7 & 0 & 0 & 1 & -2 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

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

$$\left(\begin{array}{cccccc} 1 & 0 & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right).$$

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 2.1.28 *Let* $A = \left(\begin{array}{ccc} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{array}\right)$. *Find* A^{-1}.

Write the augmented matrix $(A|I)$

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

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.

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

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

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

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

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

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

2.2 Exercises

1. In 2.1 - 2.8 describe $-A$ and 0.

2. Let A be an $n \times n$ matrix. Show A equals the sum of a symmetric and a skew symmetric matrix.

3. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form a_{ii}. It runs from the upper left down to the lower right.

4. Using only the properties 2.1 - 2.8 show $-A$ is unique.

5. Using only the properties 2.1 - 2.8 show 0 is unique.

6. Using only the properties 2.1 - 2.8 show $0A = 0$. Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for $m \times n$ matrices.

7. Using only the properties 2.1 - 2.8 and previous problems show $(-1)A = -A$.

8. Prove 2.17.

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

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

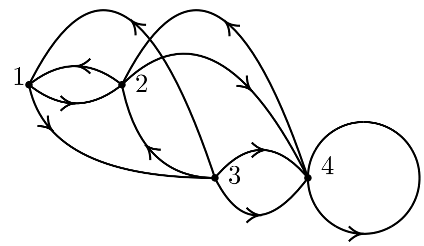
11. Use the result of Problem 10 to verify directly that $(AB)^T = B^T A^T$ without making any reference to subscripts.

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

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

14. Let $A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \end{pmatrix}$, and $C = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0 \end{pmatrix}$. Find if possible the following products. AB, BA, AC, CA, CB, BC.

15. Consider the following digraph.



Write the matrix associated with this digraph and find the number of ways to go from 3 to 4 in three steps.

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

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

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

19. 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 $\mathbf{x} = A^{-1}\mathbf{b}$.

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

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

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

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

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

    (a) $(A - B)^2 = A^2 - 2AB + B^2$
    (b) $(AB)^2 = A^2B^2$
    (c) $(A + B)^2 = A^2 + 2AB + B^2$
    (d) $(A + B)^2 = A^2 + AB + BA + B^2$
    (e) $A^2B^2 = A(AB)B$
    (f) $(A + B)^3 = A^3 + 3A^2B + 3AB^2 + B^3$
    (g) $(A + B)(A - B) = A^2 - B^2$
    (h) None of the above. They are all wrong.
    (i) All of the above. They are all right.

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

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

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

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

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

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

2.3 Linear Transformations

By 2.13, if A is an $m \times n$ matrix, then for $\mathbf{v}, \mathbf{u}$ vectors in $\mathbb{F}^n$ and a, b scalars,

$$A\left(\overbrace{a\mathbf{u} + b\mathbf{v}}^{\in \mathbb{F}^n}\right) = aA\mathbf{u} + bA\mathbf{v} \in \mathbb{F}^m \tag{2.19}$$

Definition 2.3.1 *A function, $A : \mathbb{F}^n \to \mathbb{F}^m$ is called a linear transformation if for all $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$ and a, b scalars, 2.19 holds.*

From 2.19, matrix multiplication defines a linear transformation as just defined. It turns out this is the only type of linear transformation available. Thus if A is a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$, there is always a matrix which produces A. Before showing this, here is a simple definition.

Definition 2.3.2 *A vector, $\mathbf{e}_i \in \mathbb{F}^n$ is defined as follows:*

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

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

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

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

These vectors have a significant property.

Lemma 2.3.3 *Let $\mathbf{v} \in \mathbb{F}^n$. Thus $\mathbf{v}$ is a list of numbers arranged vertically, $v_1, \cdots, v_n$. Then*

$$\mathbf{e}_i^T \mathbf{v} = v_i. \tag{2.20}$$

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

$$\mathbf{e}_i^T A \mathbf{e}_j = A_{ij} \tag{2.21}$$

Proof: First note that $\mathbf{e}_i^T$ is a $1 \times n$ matrix and $\mathbf{v}$ is an $n \times 1$ matrix so the above multiplication in 2.20 makes perfect sense. It equals

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

as claimed.

Consider 2.21. From the definition of matrix multiplication, and noting that $(\mathbf{e}_j)_k = \delta_{kj}$

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

by the first part of the lemma. ■

Theorem 2.3.4 *Let $L : \mathbb{F}^n \to \mathbb{F}^m$ be a linear transformation. Then there exists a unique $m \times n$ matrix A such that*

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

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

$$\mathbf{e}_i^T L \mathbf{e}_k \tag{2.22}$$

Stated in another way, the k^{th} column of A equals $L\mathbf{e}_\mathbf{k}$.

Proof: By the lemma,

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

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

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

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

showing $A = B$. ■

Corollary 2.3.5 *A linear transformation, $L : \mathbb{F}^n \to \mathbb{F}^m$ is completely determined by the vectors $\{L\mathbf{e}_1, \cdots, L\mathbf{e}_n\}$.*

Proof: This follows immediately from the above theorem. The unique matrix determining the linear transformation which is given in 2.22 depends only on these vectors. ■

For a different proof of this theorem and corollary, see the following section.

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

Example 2.3.6 *Find the linear transformation, $L : \mathbb{R}^2 \to \mathbb{R}^2$ which has the property that $L\mathbf{e}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $L\mathbf{e}_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$. From the above theorem and corollary, this linear transformation is that determined by matrix multiplication by the matrix*

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

2.4 Geometrically Defined Linear Transformations

If T is any linear transformation which maps $\mathbb{F}^n$ to $\mathbb{F}^m$, there is always an $m \times n$ matrix $A \equiv [T]$ with the property that

$$A\mathbf{x} = T\mathbf{x} \tag{2.23}$$

for all $\mathbf{x} \in \mathbb{F}^n$. What is the form of A? Suppose $T : \mathbb{F}^n \to \mathbb{F}^m$ is a linear transformation and you want to find the matrix defined by this linear transformation as described in 2.23. 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 $T(\mathbf{e}_i)$. This proves the existence part of the following theorem.

Theorem 2.4.1 *Let T be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix A satisfying 2.23 is given by*

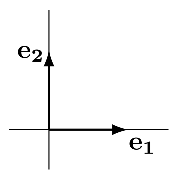
$$\begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix}$$

where $T\mathbf{e}_i$ is the i^{th} column of A.

Proof: It remains to verify uniqueness. However, if A is a matrix which works, $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$, then $T\mathbf{e}_i \equiv A\mathbf{e}_i = \mathbf{a}_i$ and so the matrix is of the form claimed above. ■

Example 2.4.2 *Determine the matrix for the transformation mapping $\mathbb{R}^2$ to $\mathbb{R}^2$ which consists of rotating every vector counter clockwise through an angle of θ.*

Let $\mathbf{e}_1 \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{e}_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. These identify the geometric vectors which point along the positive x axis and positive y axis as shown.



From Theorem 2.4.1, you only need to find $T\mathbf{e}_1$ and $T\mathbf{e}_2$, the first being the first column of the desired matrix A and the second being the second column. From drawing a picture and doing a little geometry, you see that

$$T\mathbf{e}_1 = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, T\mathbf{e}_2 = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

Therefore, from Theorem 2.4.1,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Example 2.4.3 *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

$$\begin{aligned} A_{\theta+\phi} &= \begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = A_\theta A_\phi \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \end{aligned}$$

Therefore,

$$\begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = \begin{pmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\cos\theta\sin\phi - \sin\theta\cos\phi \\ \sin\theta\cos\phi + \cos\theta\sin\phi & \cos\theta\cos\phi - \sin\theta\sin\phi \end{pmatrix}.$$

Don't these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

Example 2.4.4 *Find the matrix of the linear transformation which rotates vectors in $\mathbb{R}^3$ counter-clockwise about the positive z axis.*

Let T be the name of this linear transformation. In this case, $T\mathbf{e}_3 = \mathbf{e}_3, T\mathbf{e}_1 = (\cos\theta, \sin\theta, 0)^T$, and $T\mathbf{e}_2 = (-\sin\theta, \cos\theta, 0)^T$. Therefore, the matrix of this transformation is just

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.24}$$

In Physics it is important to consider the work done by a force field on an object. This involves the concept of projection onto a vector. Suppose you want to find the projection of a vector, $\mathbf{v}$ onto the given vector, $\mathbf{u}$, denoted by $\text{proj}_{\mathbf{u}}(\mathbf{v})$ 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 $\mathbf{v} \to \text{proj}_{\mathbf{u}}(\mathbf{v})$ 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 2.4.5 *Let the projection map be defined above and let $\mathbf{u} = (1,2,3)^T$. Find the matrix of this linear transformation with respect to the usual basis.*

You can find this matrix in the same way as in earlier examples. $\text{proj}_{\mathbf{u}}(\mathbf{e}_i)$ gives the i^{th} column of the desired matrix. Therefore, it is only necessary to find

$$\text{proj}_{\mathbf{u}}(\mathbf{e}_i) \equiv \left(\frac{\mathbf{e}_i\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u}$$

For the given vector in the example, this implies the columns of the desired matrix are

$$\frac{1}{14}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \frac{2}{14}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \frac{3}{14}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Hence the matrix is

$$\frac{1}{14}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

Example 2.4.6 *Find the matrix of the linear transformation which reflects all vectors in $\mathbb{R}^3$ through the xz plane.*

As illustrated above, you just need to find $T\mathbf{e}_i$ where T is the name of the transformation. But $T\mathbf{e}_1 = \mathbf{e}_1, T\mathbf{e}_3 = \mathbf{e}_3$, and $T\mathbf{e}_2 = -\mathbf{e}_2$ so the matrix is

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

Example 2.4.7 *Find the matrix of the linear transformation which first rotates counter clockwise about the positive z axis and then reflects through the xz plane.*

This linear transformation is just the composition of two linear transformations having matrices

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

respectively. Thus the matrix desired is

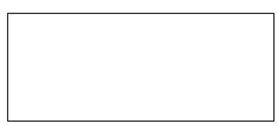
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ -\sin\theta & -\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.5 The Null Space of a Linear Transformation

The null space or kernel of a matrix or linear transformation is given in the following definition. Essentially, it is just the set of all vectors which are sent to the zero vector by the linear transformation.

Definition 2.5.1 *Let $L : \mathbb{F}^n \to \mathbb{F}^m$ be a linear transformation and let its matrix be the $m \times n$ matrix A. Then* $\ker(L) \equiv \{\mathbf{x} \in \mathbb{F}^n : L\mathbf{x} = \mathbf{0}\}$. *Sometimes people also write this as $N(A)$, the null space of A.*

Then there is a fundamental result in the case where $m < n$. In this case, the matrix A of the linear transformation looks like the following.



Theorem 2.5.2 *Let A be an $m \times n$ matrix where $m < n$. Then $N(A)$ contains nonzero vectors.*

Proof: First consider the case where A is a $1 \times n$ matrix for $n > 1$. Say

$$A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$$

If $a_1 = 0$, consider the vector $\mathbf{x} = \mathbf{e}_1$. If $a_1 \neq 0$, let

$$\mathbf{x} = \begin{pmatrix} b \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

where b is chosen to satisfy the equation

$$a_1 b + \sum_{k=2}^{n} a_k = 0$$

Suppose now that the theorem is true for any $m \times n$ matrix with $n > m$ and consider an $(m \times 1) \times n$ matrix A where $n > m + 1$. If the first column of A is $\mathbf{0}$, then you could let $\mathbf{x} = \mathbf{e}_1$ as above. If the first column is not the zero vector, then by doing row operations, the equation $A\mathbf{x} = \mathbf{0}$ can be reduced to the equivalent system

$$A_1 \mathbf{x} = \mathbf{0}$$

where A_1 is of the form

$$A_1 = \begin{pmatrix} 1 & \mathbf{a}^T \\ \mathbf{0} & B \end{pmatrix}$$

where B is an $m \times (n-1)$ matrix. Since $n > m + 1$, it follows that $(n-1) > m$ and so by induction, there exists a nonzero vector $\mathbf{y} \in \mathbb{F}^{n-1}$ such that $B\mathbf{y} = \mathbf{0}$. Then consider the vector

$$\mathbf{x} = \begin{pmatrix} b \\ \mathbf{y} \end{pmatrix}$$

$A_1\mathbf{x}$ has for its top entry the expression $b + \mathbf{a}^T\mathbf{y}$. Letting $B = \begin{pmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_m^T \end{pmatrix}$, the i^{th} entry of $A_1\mathbf{x}$ for $i > 1$ is of the form $\mathbf{b}_i^T\mathbf{y} = 0$. Thus if b is chosen to satisfy the equation $b + \mathbf{a}^T\mathbf{y} = 0$, then $A_1\mathbf{x} = \mathbf{0}$.■

2.6 Subspaces and Spans

Definition 2.6.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ be vectors in $\mathbb{F}^n$. A linear combination is any expression of the form*

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

where the c_i are scalars. The set of all linear combinations of these vectors is called $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. *A nonempty $V \subseteq \mathbb{F}^n$, is is called a subspace if whenever α, β are scalars and $\mathbf{u}$ and $\mathbf{v}$ are vectors of V, it follows $\alpha\mathbf{u} + \beta\mathbf{v} \in V$. That is, it is "closed under the algebraic operations of vector addition and scalar multiplication". The empty set is never a subspace by definition. A linear combination of vectors is said to be trivial if all the scalars in the linear combination equal zero. A set of vectors is said to be linearly independent if the only linear combination of these vectors which equals the zero vector is the trivial linear combination. Thus $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is called linearly independent if whenever*

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

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

Proposition 2.6.2 *Let $V \subseteq \mathbb{F}^n$. Then V is a subspace if and only if it is a vector space itself with respect to the same operations of scalar multiplication and vector addition.*

Proof: Suppose first that V is a subspace. All algebraic properties involving scalar multiplication and vector addition hold for V because these things hold for $\mathbb{F}^n$. Is $\mathbf{0} \in V$? Yes it is. This is because $0\mathbf{v} \in V$ and $0\mathbf{v} = \mathbf{0}$. By assumption, for α a scalar and $\mathbf{v} \in V, \alpha\mathbf{v} \in V$. Therefore, $-\mathbf{v} = (-1)\,\mathbf{v} \in V$. Thus V has the additive identity and additive inverse. By assumption, V is closed with respect to the two operations. Thus V is a vector space. If $V \subseteq \mathbb{F}^n$ is a vector space, then by definition, if α, β are scalars and $\mathbf{u}, \mathbf{v}$ vectors in V, it follows that $\alpha\mathbf{v} + \beta\mathbf{u} \in V$. ■

Thus, from the above, subspaces of $\mathbb{F}^n$ are just subsets of $\mathbb{F}^n$ which are themselves vector spaces.

Lemma 2.6.3 *A set of vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

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

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

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

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

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

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

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

contrary to assumption. ■

The following is called the exchange theorem.

Theorem 2.6.4 *(Exchange Theorem) Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be a linearly independent set of vectors such that each $\mathbf{x}_i$ is in span$(\mathbf{y}_1, \cdots, \mathbf{y}_s)$. Then $r \leq s$.*

Proof 1: Suppose not. Then $r > s$. By assumption, there exist scalars a_{ji} such that

$$\mathbf{x}_i = \sum_{j=1}^{s} a_{ji} \mathbf{y}_j$$

The matrix whose ji^{th} entry is a_{ji} has more columns than rows. Therefore, by Theorem 2.5.2 there exists a **nonzero** vector $\mathbf{b} \in \mathbb{F}^r$ such that $A\mathbf{b} = \mathbf{0}$. Thus

$$0 = \sum_{i=1}^{r} a_{ji} b_i, \text{ each } j.$$

Then

$$\sum_{i=1}^{r} b_i \mathbf{x}_i = \sum_{i=1}^{r} b_i \sum_{j=1}^{s} a_{ji} \mathbf{y}_j = \sum_{j=1}^{s} \left(\sum_{i=1}^{r} a_{ji} b_i \right) \mathbf{y}_j = \mathbf{0}$$

contradicting the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent.

Proof 2: Define $\text{span}\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{2.25}$$

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 2.25 for $\mathbf{y}_k$ and obtain

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

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

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

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

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

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

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

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{2.26}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, 2.26 can be solved for one of the $\mathbf{z}_i$, say $\mathbf{z}_k$, in terms of $\mathbf{x}_{l+1}$ and the other $\mathbf{z}_i$ and just as in the above argument, replace that $\mathbf{z}_i$ with $\mathbf{x}_{l+1}$ to obtain

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

Continue this way, eventually obtaining

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

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

Proof 3: 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 them in terms of $\mathbf{x}_{m+1}, \mathbf{x}_m, \cdots, \mathbf{x}_1$, and the other $\mathbf{z}_k$. Therefore, there exists

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{j-1}\} \subseteq \{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$$

such that

$$\text{span}(\mathbf{y}_1, \cdots, \mathbf{y}_s) = \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_{m+1}, \mathbf{z}_1, \cdots, \mathbf{z}_{j-1})$$

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. ■

Definition 2.6.5 *The set of vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a basis for $\mathbb{F}^n$ if* $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = \mathbb{F}^n$ *and $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent.*

Corollary 2.6.6 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ and $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ be two bases[1] of $\mathbb{F}^n$. Then $r = s = n$.*

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

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

for $i = 1, 2, \cdots, n$ are a basis for $\mathbb{F}^n$. ■

Lemma 2.6.7 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ be a set of vectors. Then $V \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_r)$ is a subspace.*

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

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

Is it also in V?

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

so the answer is yes. ■

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

Definition 2.6.8 *A finite set of vectors,* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is a basis for a subspace* V *of* $\mathbb{F}^n$ *if* $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = V$ *and* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is linearly independent.*

Corollary 2.6.9 *Let* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *and* $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ *be two bases for* V*. Then* $r = s$*.*

Proof: From the exchange theorem, $r \leq s$ and $s \leq r$. ■

Definition 2.6.10 *Let* V *be a subspace of* $\mathbb{F}^n$*. Then* $\dim(V)$ *read as the dimension of* V *is the number of vectors in a basis.*

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

Lemma 2.6.11 *Suppose* $\mathbf{v} \notin \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ *and* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ *is linearly independent. Then* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}\}$ *is also linearly independent.*

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

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

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

Theorem 2.6.12 *Let* V *be a nonzero subspace of* $\mathbb{F}^n$*. Then* V *has a basis.*

Proof: Let $\mathbf{v}_1 \in V$ where $\mathbf{v}_1 \neq 0$. If $\operatorname{span}\{\mathbf{v}_1\} = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\operatorname{span}\{\mathbf{v}_1\}$. By Lemma 2.6.11 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} \neq V$, then there exists $\mathbf{v}_3 \notin \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem. ■

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

Corollary 2.6.13 *Let* V *be a subspace of* $\mathbb{F}^n$ *and let* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ *be a linearly independent set of vectors in* V*. Then either it is a basis for* V *or there exist vectors,* $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_s$ *such that* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_s\}$ *is a basis for* V*.*

Proof: This follows immediately from the proof of Theorem 2.6.12. You do exactly the same argument except you start with $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ rather than $\{\mathbf{v}_1\}$. ■

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

Theorem 2.6.14 *Let* V *be a subspace of* $\mathbb{F}^n$ *and suppose* $\operatorname{span}(\mathbf{u}_1 \cdots, \mathbf{u}_p) = V$ *where the* $\mathbf{u}_i$ *are nonzero vectors. Then there exist vectors* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ *such that* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$ *and* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ *is a basis for* V*.*

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

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

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

2.7 An Application to Matrices

The following is a theorem of major significance.

Theorem 2.7.1 *Suppose A is an $n \times n$ matrix. Then A is one to one (injective) if and only if A is onto (surjective). 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 $c_k = 0$ because the $\mathbf{e}_k$ are clearly linearly independent.

Therefore, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ must be a basis for $\mathbb{F}^n$ because if not there would exist a vector, $\mathbf{y} \notin \operatorname{span}(A\mathbf{e}_1, \cdots, A\mathbf{e}_n)$ and then by Lemma 2.6.11, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n, \mathbf{y}\}$ would be an independent set of vectors having $n+1$ vectors in it, contrary to the exchange theorem. It follows that for $\mathbf{y} \in \mathbb{F}^n$ there exist constants, c_i such that

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

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

Next suppose A is onto. This means the span of the columns of A equals $\mathbb{F}^n$. If these columns are not linearly independent, then by Lemma 2.6.3 on Page 61, one of the columns is a linear combination of the others and so the span of the columns of A equals the span of the $n-1$ other columns. This violates the exchange theorem because $\{\mathbf{e}_1, \cdots \mathbf{e}_n\}$ would be a linearly independent set of vectors contained in the span of only $n-1$ vectors. Therefore, the columns of A must be independent and this is equivalent to saying that $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. This implies A is one to one because if $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} - \mathbf{y} = \mathbf{0}$.

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

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

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

This 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

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix} \tag{2.27}$$

Then

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

but

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

2.8 Matrices and Calculus

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

Definition 2.8.1 *Let $A(t)$ be an $m \times n$ matrix. Say $A(t) = (A_{ij}(t))$. Suppose also that $A_{ij}(t)$ is a differentiable function for all i, j. Then define $A'(t) \equiv \left(A'_{ij}(t)\right)$. That is, $A'(t)$ is the matrix which consists of replacing each entry by its derivative. Such an $m \times n$ matrix in which the entries are differentiable functions is called a differentiable matrix.*

The next lemma is just a version of the product rule.

Lemma 2.8.2 *Let $A(t)$ be an $m \times n$ matrix and let $B(t)$ be an $n \times p$ matrix with the property that all the entries of these matrices are differentiable functions. Then*

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

Proof: This is like the usual proof one sees in a calculus course.

$$\frac{1}{h}(A(t+h) B(t+h) - A(t) B(t)) =$$

$$\frac{1}{h}(A(t+h) B(t+h) - A(t+h) B(t)) + \frac{1}{h}(A(t+h) B(t) - A(t) B(t))$$

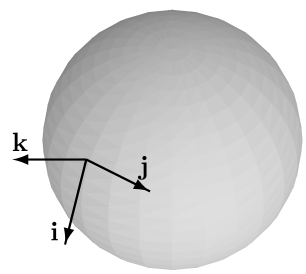
$$= A(t+h) \frac{B(t+h) - B(t)}{h} + \frac{A(t+h) - A(t)}{h} B(t)$$

and now, using the fact that the entries of the matrices are all differentiable, one can pass to a limit in both sides as $h \to 0$ and conclude that

$$(A(t) B(t))' = A'(t) B(t) + A(t) B'(t) \blacksquare$$

2.8.1 The Coriolis Acceleration

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



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

In general, let $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ be the usual fixed vectors in space and let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be an orthonormal basis of vectors for each t, like the vectors described in the first paragraph. It is assumed these vectors are C^1 functions of t. Letting the positive x axis extend in the direction of $\mathbf{i}(t)$, the positive y axis extend in the direction of $\mathbf{j}(t)$, and the positive z axis extend in the direction of $\mathbf{k}(t)$, yields a moving coordinate system. Now let $\mathbf{u}$ be a vector and let t_0 be some reference time. For example you could let $t_0 = 0$. Then define the components of $\mathbf{u}$ with respect to these vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ at time t_0 as

$$\mathbf{u} \equiv u^1 \mathbf{i}(t_0) + u^2 \mathbf{j}(t_0) + u^3 \mathbf{k}(t_0).$$

Let $\mathbf{u}(t)$ be defined as the vector which has the same components with respect to $\mathbf{i}, \mathbf{j}, \mathbf{k}$ but at time t. Thus

$$\mathbf{u}(t) \equiv u^1 \mathbf{i}(t) + u^2 \mathbf{j}(t) + u^3 \mathbf{k}(t).$$

and the vector has changed although the components have not.

This is exactly the situation in the case of the apparently fixed basis vectors on the earth if $\mathbf{u}$ is a position vector from the given spot on the earth's surface to a point regarded as fixed with the earth due to its keeping the same coordinates relative to the coordinate axes which are fixed with the earth. Now define a linear transformation $Q(t)$ mapping $\mathbb{R}^3$ to $\mathbb{R}^3$ by

$$Q(t)\,\mathbf{u} \equiv u^1 \mathbf{i}(t) + u^2 \mathbf{j}(t) + u^3 \mathbf{k}(t)$$

where

$$\mathbf{u} \equiv u^1 \mathbf{i}(t_0) + u^2 \mathbf{j}(t_0) + u^3 \mathbf{k}(t_0)$$

Thus letting $\mathbf{v}$ be a vector defined in the same manner as $\mathbf{u}$ and α, β, scalars,

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

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

showing that $Q(t)$ is a linear transformation. Also, $Q(t)$ preserves all distances because, since the vectors, $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ form an orthonormal set,

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

Lemma 2.8.3 *Suppose $Q(t)$ is a real, differentiable $n\times n$ matrix which preserves distances. Then $Q(t)Q(t)^T = Q(t)^T Q(t) = I$. Also, if $\mathbf{u}(t) \equiv Q(t)\mathbf{u}$, then there exists a vector, $\mathbf{\Omega}(t)$ such that*

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).$$

The symbol $\times$ refers to the cross product.

Proof: Recall that $(\mathbf{z}\cdot\mathbf{w}) = \frac{1}{4}\left(|\mathbf{z}+\mathbf{w}|^2 - |\mathbf{z}-\mathbf{w}|^2\right)$. Therefore,

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

This implies

$$\left(Q(t)^T Q(t)\mathbf{u}\cdot\mathbf{w}\right) = (\mathbf{u}\cdot\mathbf{w})$$

for all $\mathbf{u},\mathbf{w}$. Therefore, $Q(t)^T Q(t)\mathbf{u} = \mathbf{u}$ and so $Q(t)^T Q(t) = Q(t)Q(t)^T = I$. This proves the first part of the lemma.

It follows from the product rule, Lemma 2.8.2 that

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

and so

$$Q'(t)Q(t)^T = -\left(Q'(t)Q(t)^T\right)^T. \tag{2.28}$$

From the definition, $Q(t)\mathbf{u} = \mathbf{u}(t)$,

$$\mathbf{u}'(t) = Q'(t)\mathbf{u} = Q'(t)\overbrace{Q(t)^T\mathbf{u}(t)}^{=\mathbf{u}}.$$

Then writing the matrix of $Q'(t)Q(t)^T$ with respect to fixed in space orthonormal basis vectors, $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$, where these are the usual basis vectors for $\mathbb{R}^3$, it follows from 2.28 that the matrix of $Q'(t)Q(t)^T$ is of the form

$$\begin{pmatrix} 0 & -\omega_3(t) & \omega_2(t) \\ \omega_3(t) & 0 & -\omega_1(t) \\ -\omega_2(t) & \omega_1(t) & 0 \end{pmatrix}$$

for some time dependent scalars ω_i. Therefore,

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

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

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t) = Q'(t)\, Q(t)^T \mathbf{u}(t) \tag{2.29}$$

where

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

because

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

$$\mathbf{i}^*\left(w_2 u^3 - w_3 u^2\right) + \mathbf{j}^*\left(w_3 u^1 - w_1^3\right) + \mathbf{k}^*\left(w_1 u^2 - w_2 u^1\right).$$

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

Theorem 2.8.4 *Let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be as described. Then there exists a unique vector $\mathbf{\Omega}(t)$ such that if $\mathbf{u}(t)$ is a vector whose components are constant with respect to $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$, then*

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).$$

Proof: It only remains to prove uniqueness. Suppose $\mathbf{\Omega}_1$ also works. Then $\mathbf{u}(t) = Q(t)\,\mathbf{u}$ and so $\mathbf{u}'(t) = Q'(t)\,\mathbf{u}$ and

$$Q'(t)\,\mathbf{u} = \mathbf{\Omega} \times Q(t)\,\mathbf{u} = \mathbf{\Omega}_1 \times Q(t)\,\mathbf{u}$$

for all $\mathbf{u}$. Therefore,

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

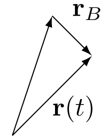
for all $\mathbf{u}$ and since $Q(t)$ is one to one and onto, this implies $(\mathbf{\Omega} - \mathbf{\Omega}_1) \times \mathbf{w} = \mathbf{0}$ for all $\mathbf{w}$ and thus $\mathbf{\Omega} - \mathbf{\Omega}_1 = \mathbf{0}$. ■

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

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

where

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



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

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

By , 2.29, if $\mathbf{e} \in \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$, $\mathbf{e}' = \mathbf{\Omega} \times \mathbf{e}$ because the components of these vectors with respect to $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are constant. Therefore,

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

and consequently,

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

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

$$\mathbf{a} = \mathbf{v}' = \mathbf{R}'' + x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k} + \overbrace{x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'}^{\mathbf{\Omega}\times\mathbf{v}_B} + \mathbf{\Omega}' \times \mathbf{r}_B$$

$$+\mathbf{\Omega} \times \left(\overbrace{x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}}^{\mathbf{v}_B} + \overbrace{x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}'}^{\mathbf{\Omega}\times\mathbf{r}_B(t)} \right)$$

$$= \mathbf{R}'' + \mathbf{a}_B + \mathbf{\Omega}' \times \mathbf{r}_B + 2\mathbf{\Omega} \times \mathbf{v}_B + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B).$$

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

$$\mathbf{a}_B = \mathbf{a} - \mathbf{R}'' - \mathbf{\Omega}' \times \mathbf{r}_B - 2\mathbf{\Omega} \times \mathbf{v}_B - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B). \tag{2.30}$$

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

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

2.8.2 The Coriolis Acceleration on the Rotating Earth

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

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

Therefore, the vector of Theorem 2.8.4 is $\mathbf{\Omega} = \omega \mathbf{k}^*$ and so

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

since $\mathbf{\Omega}$ does not depend on t. Formula 2.30 implies

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R}) - 2\mathbf{\Omega} \times \mathbf{v}_B - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B). \tag{2.31}$$

In this formula, you can totally ignore the term $\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B)$ because it is so small whenever you are considering motion near some point on the earth's surface. To see this, note $\omega \overbrace{(24)(3600)}^{\text{seconds in a day}} = 2\pi$, and so $\omega = 7.2722 \times 10^{-5}$ in radians per second. If you are using seconds to measure time and feet to measure distance, this term is therefore, no larger than

$$\left(7.2722 \times 10^{-5}\right)^2 |\mathbf{r}_B| .$$

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

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

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

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

Note that

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

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

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

While the vector $\mathbf{\Omega}$ is quite small, if the relative velocity, $\mathbf{v}_B$ is large, the Coriolis acceleration could be significant. This is described in terms of the vectors $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ next.

Letting (ρ, θ, ϕ) be the usual spherical coordinates of the point $\mathbf{p}(t)$ on the surface taken with respect to $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ the usual way with ϕ the polar angle, it follows the $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ coordinates of this point are

$$\begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}.$$

It follows,

$$\mathbf{i} = \cos(\phi)\cos(\theta)\,\mathbf{i}^* + \cos(\phi)\sin(\theta)\,\mathbf{j}^* - \sin(\phi)\,\mathbf{k}^*$$
$$\mathbf{j} = -\sin(\theta)\,\mathbf{i}^* + \cos(\theta)\,\mathbf{j}^* + 0\mathbf{k}^*$$

and

$$\mathbf{k} = \sin(\phi)\cos(\theta)\,\mathbf{i}^* + \sin(\phi)\sin(\theta)\,\mathbf{j}^* + \cos(\phi)\,\mathbf{k}^*.$$

It is necessary to obtain $\mathbf{k}^*$ in terms of the vectors, $\mathbf{i},\mathbf{j},\mathbf{k}$. Thus the following equation needs to be solved for a,b,c to find $\mathbf{k}^* = a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$

$$\overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{\mathbf{k}^*} = \begin{pmatrix} \cos(\phi)\cos(\theta) & -\sin(\theta) & \sin(\phi)\cos(\theta) \\ \cos(\phi)\sin(\theta) & \cos(\theta) & \sin(\phi)\sin(\theta) \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{2.32}$$

The first column is $\mathbf{i}$, the second is $\mathbf{j}$ and the third is $\mathbf{k}$ in the above matrix. The solution is $a = -\sin(\phi), b = 0$, and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

$$2(\mathbf{\Omega} \times \mathbf{v}_B) = 2\omega \left(\overbrace{-\sin(\phi)\,\mathbf{i}+0\mathbf{j}+\cos(\phi)\,\mathbf{k}}^{\mathbf{k}^*} \right) \times (x'\mathbf{i}+y'\mathbf{j}+z'\mathbf{k}).$$

This equals

$$2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]. \tag{2.33}$$

Remember ϕ is fixed and pertains to the fixed point, $\mathbf{p}(t)$ on the earth's surface. Therefore, if the acceleration $\mathbf{a}$ is due to gravity,

$$\mathbf{a}_B = \mathbf{g}-2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$$

where $\mathbf{g} = -\frac{GM(\mathbf{R}+\mathbf{r}_B)}{|\mathbf{R}+\mathbf{r}_B|^3} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$ as explained above. The term $\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$ is pretty small and so it will be neglected. However, the Coriolis force will not be neglected.

Example 2.8.5 *Suppose a rock is dropped from a tall building. Where will it strike?*

Assume $\mathbf{a} = -g\mathbf{k}$ and the $\mathbf{j}$ component of $\mathbf{a}_B$ is approximately

$$-2\omega(x'\cos\phi+z'\sin\phi).$$

The dominant term in this expression is clearly the second one because x' will be small. Also, the $\mathbf{i}$ and $\mathbf{k}$ contributions will be very small. Therefore, the following equation is descriptive of the situation.

$$\mathbf{a}_B = -g\mathbf{k}-2z'\omega\sin\phi\mathbf{j}.$$

$z' = -gt$ approximately. Therefore, considering the $\mathbf{j}$ component, this is

$$2gt\omega\sin\phi.$$

Two integrations give $\left(\omega g t^3/3\right)\sin\phi$ for the $\mathbf{j}$ component of the relative displacement at time t.

This shows the rock does not fall directly towards the center of the earth as expected but slightly to the east.

Example 2.8.6 *In 1851 Foucault set a pendulum vibrating and observed the earth rotate out from under it. It was a very long pendulum with a heavy weight at the end so that it would vibrate for a long time without stopping*[2]*. This is what allowed him to observe the earth rotate out from under it. Clearly such a pendulum will take 24 hours for the plane of vibration to appear to make one complete revolution at the north pole. It is also reasonable to expect that no such observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?*

[2]There is such a pendulum in the Eyring building at BYU and to keep people from touching it, there is a little sign which says Warning! 1000 ohms.

Using 2.33, in 2.31,

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$$
$$-2\omega\left[(-y'\cos\phi)\,\mathbf{i} + (x'\cos\phi + z'\sin\phi)\,\mathbf{j} - (y'\sin\phi)\,\mathbf{k}\right].$$

Neglecting the small term, $\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$, this becomes

$$= -g\mathbf{k} + \mathbf{T}/m - 2\omega\left[(-y'\cos\phi)\,\mathbf{i} + (x'\cos\phi + z'\sin\phi)\,\mathbf{j} - (y'\sin\phi)\,\mathbf{k}\right]$$

where $\mathbf{T}$, the tension in the string of the pendulum, is directed towards the point at which the pendulum is supported, and m is the mass of the pendulum bob. The pendulum can be thought of as the position vector from $(0,0,l)$ to the surface of the sphere $x^2+y^2+(z-l)^2 = l^2$. Therefore,

$$\mathbf{T} = -T\frac{x}{l}\mathbf{i} - T\frac{y}{l}\mathbf{j} + T\frac{l-z}{l}\mathbf{k}$$

and consequently, the differential equations of relative motion are

$$x'' = -T\frac{x}{ml} + 2\omega y'\cos\phi$$

$$y'' = -T\frac{y}{ml} - 2\omega(x'\cos\phi + z'\sin\phi)$$

and

$$z'' = T\frac{l-z}{ml} - g + 2\omega y'\sin\phi.$$

If the vibrations of the pendulum are small so that for practical purposes, $z'' = z = 0$, the last equation may be solved for T to get

$$gm - 2\omega y'\sin(\phi)\,m = T.$$

Therefore, the first two equations become

$$x'' = -(gm - 2\omega m y'\sin\phi)\frac{x}{ml} + 2\omega y'\cos\phi$$

and

$$y'' = -(gm - 2\omega m y'\sin\phi)\frac{y}{ml} - 2\omega(x'\cos\phi + z'\sin\phi).$$

All terms of the form xy' or $y'y$ can be neglected because it is assumed x and y remain small. Also, the pendulum is assumed to be long with a heavy weight so that x' and y' are also small. With these simplifying assumptions, the equations of motion become

$$x'' + g\frac{x}{l} = 2\omega y'\cos\phi$$

and

$$y'' + g\frac{y}{l} = -2\omega x'\cos\phi.$$

These equations are of the form

$$x'' + a^2x = by',\ y'' + a^2y = -bx' \tag{2.34}$$

where $a^2 = \frac{g}{l}$ and $b = 2\omega\cos\phi$. Then it is fairly tedious but routine to verify that for each constant, c,

$$x = c\sin\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right),\ y = c\cos\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right) \tag{2.35}$$

yields a solution to 2.34 along with the initial conditions,

$$x(0)=0, y(0)=0, x'(0)=0, y'(0)=\frac{c\sqrt{b^2+4a^2}}{2}. \tag{2.36}$$

It is clear from experiments with the pendulum that the earth does indeed rotate out from under it causing the plane of vibration of the pendulum to appear to rotate. The purpose of this discussion is not to establish these self evident facts but to predict how long it takes for the plane of vibration to make one revolution. Therefore, there will be some instant in time at which the pendulum will be vibrating in a plane determined by **k** and **j**. (Recall **k** points away from the center of the earth and **j** points East.) At this instant in time, defined as $t=0$, the conditions of 2.36 will hold for some value of c and so the solution to 2.34 having these initial conditions will be those of 2.35 by uniqueness of the initial value problem. Writing these solutions differently,

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c\begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix} \sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$$

This is very interesting! The vector, $c\begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix}$ always has magnitude equal to $|c|$ but its direction changes very slowly because b is very small. The plane of vibration is determined by this vector and the vector **k**. The term $\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$ changes relatively fast and takes values between -1 and 1. This is what describes the actual observed vibrations of the pendulum. Thus the plane of vibration will have made one complete revolution when $t=T$ for

$$\frac{bT}{2} \equiv 2\pi.$$

Therefore, the time it takes for the earth to turn out from under the pendulum is

$$T = \frac{4\pi}{2\omega\cos\phi} = \frac{2\pi}{\omega}\sec\phi.$$

Since ω is the angular speed of the rotating earth, it follows $\omega=\frac{2\pi}{24}=\frac{\pi}{12}$ in radians per hour. Therefore, the above formula implies

$$T = 24\sec\phi.$$

I think this is really amazing. You could actually determine latitude, not by taking readings with instruments using the North Star but by doing an experiment with a big pendulum. You would set it vibrating, observe T in hours, and then solve the above equation for ϕ. Also note the pendulum would not appear to change its plane of vibration at the equator because $\lim_{\phi\to\pi/2}\sec\phi=\infty$.

The Coriolis acceleration is also responsible for the phenomenon of the next example.

Example 2.8.7 *It is known that low pressure areas rotate counterclockwise as seen from above in the Northern hemisphere but clockwise in the Southern hemisphere. Why?*

Neglect accelerations other than the Coriolis acceleration and the following acceleration which comes from an assumption that the point $\mathbf{p}(t)$ is the location of the lowest pressure.

$$\mathbf{a} = -a(r_B)\mathbf{r}_B$$

where $r_B=r$ will denote the distance from the fixed point $\mathbf{p}(t)$ on the earth's surface which is also the lowest pressure point. Of course the situation could be more complicated but this will suffice to explain the above question. Then the acceleration observed by a person on the earth relative to the apparently fixed vectors, $\mathbf{i}, \mathbf{k}, \mathbf{j}$, is

$$\mathbf{a}_B = -a(r_B)(x\mathbf{i}+y\mathbf{j}+z\mathbf{k}) - 2\omega\left[-y'\cos(\phi)\,\mathbf{i}+ (x'\cos(\phi)+z'\sin(\phi))\,\mathbf{j}- (y'\sin(\phi)\,\mathbf{k})\right]$$

Therefore, one obtains some differential equations from $\mathbf{a}_B = x''\mathbf{i}+y''\mathbf{j}+z''\mathbf{k}$ by matching the components. These are

$$\begin{aligned} x'' + a(r_B)\,x &= 2\omega y'\cos\phi \\ y'' + a(r_B)\,y &= -2\omega x'\cos\phi - 2\omega z'\sin(\phi) \\ z'' + a(r_B)\,z &= 2\omega y'\sin\phi \end{aligned}$$

Now remember, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are fixed relative to the earth and so are constant vectors. Therefore, from the properties of the determinant and the above differential equations,

$$(\mathbf{r}_B' \times \mathbf{r}_B)' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x' & y' & z' \\ x & y & z \end{vmatrix}' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x'' & y'' & z'' \\ x & y & z \end{vmatrix}$$

$$= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a(r_B)\,x + 2\omega y'\cos\phi & -a(r_B)\,y - 2\omega x'\cos\phi - 2\omega z'\sin(\phi) & -a(r_B)\,z + 2\omega y'\sin\phi \\ x & y & z \end{vmatrix}$$

Then the $\mathbf{k}^{th}$ component of this cross product equals

$$\omega\cos(\phi)\left(y^2+x^2\right)' + 2\omega x z'\sin(\phi)\,.$$

The first term will be negative because it is assumed $\mathbf{p}(t)$ is the location of low pressure causing y^2+x^2 to be a decreasing function. If it is assumed there is not a substantial motion in the $\mathbf{k}$ direction, so that z is fairly constant and the last term can be neglected, then the $\mathbf{k}^{th}$ component of $(\mathbf{r}_B' \times \mathbf{r}_B)'$ is negative provided $\phi \in \left(0, \frac{\pi}{2}\right)$ and positive if $\phi \in \left(\frac{\pi}{2}, \pi\right)$. Beginning with a point at rest, this implies $\mathbf{r}_B' \times \mathbf{r}_B = \mathbf{0}$ initially and then the above implies its $\mathbf{k}^{th}$ component is negative in the upper hemisphere when $\phi < \pi/2$ and positive in the lower hemisphere when $\phi > \pi/2$. Using the right hand and the geometric definition of the cross product, this shows clockwise rotation in the lower hemisphere and counter clockwise rotation in the upper hemisphere.

Note also that as ϕ gets close to $\pi/2$ near the equator, the above reasoning tends to break down because $\cos(\phi)$ becomes close to zero. Therefore, the motion towards the low pressure has to be more pronounced in comparison with the motion in the $\mathbf{k}$ direction in order to draw this conclusion.

2.9 Exercises

1. Show the map $T : \mathbb{R}^n \to \mathbb{R}^m$ 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.

2. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$.

3. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.

4. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $-\pi/3$.

5. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$.

6. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/12$. **Hint:** Note that $\pi/12 = \pi/3 - \pi/4$.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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 $\pi/4$.

12. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$ and next reflects every vector across the x axis. Compare with the above problem.

13. 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 $\pi/6$.

14. 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 $\pi/6$.

15. 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 $5\pi/12 = 2\pi/3 - \pi/4$.

16. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, -2, 3)^T$.

17. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, 5, 3)^T$.

18. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, 0, 3)^T$.

19. Give an example of a 2×2 matrix A which has all its entries nonzero and satisfies $A^2 = A$. A matrix which satisfies $A^2 = A$ is called idempotent.

20. Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix where $n < m$. Show that AB cannot have an inverse.

21. 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 $A\mathbf{x} = \mathbf{0}$.

22. If A is a linear transformation, and $A\mathbf{x}_p = \mathbf{b}$, show that the general solution to the equation $A\mathbf{x} = \mathbf{b}$ is of the form $\mathbf{x}_p + \mathbf{y}$ where $\mathbf{y} \in \ker(A)$. By this I mean to show that whenever $A\mathbf{z} = \mathbf{b}$ there exists $\mathbf{y} \in \ker(A)$ such that $\mathbf{x}_p + \mathbf{y} = \mathbf{z}$. For the definition of $\ker(A)$ see Problem 21.

23. Using Problem 21, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \\ 18 \\ 7 \end{pmatrix}$$

24. Using Problem 21, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 13 \\ 7 \end{pmatrix}$$

25. 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.

26. If $\mathbf{u} = (1, 2, 3)^T$, as in Example 2.4.5 and $T_{\mathbf{u}}$ is given in the above problem, find the matrix $A_{\mathbf{u}}$ which satisfies $A_{\mathbf{u}}\mathbf{x} = T_{\mathbf{u}}(\mathbf{x})$.

27. 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.

28. In spite of Problem 27 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.

29. You want to add $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ to every point in $\mathbb{R}^3$ and then rotate about the x axis clockwise through the angle of $30°$. Find what happens to the point $\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$.

30. You are given a linear transformation $T:\mathbb{F}^n\to\mathbb{F}^m$ 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 ($A\mathbf{x}=T\mathbf{x}$) must be of the form
$$\begin{pmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_m\end{pmatrix}\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}^{-1}$$

31. 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 $T\mathbf{x}=A\mathbf{x}$.

32. 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 $T\mathbf{x}=A\mathbf{x}$.

33. 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 $T\mathbf{x}=A\mathbf{x}$.

34. 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 $T\mathbf{x}=A\mathbf{x}$.

35. 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 $T\mathbf{x} = A\mathbf{x}$.

36. Suppose V is a subspace of $\mathbb{F}^n$ and $T : V \to \mathbb{F}^p$ is a nonzero linear transformation. Show that there exists a basis for $\text{Im}(T) \equiv T(V)$

$$\{T\mathbf{v}_1, \cdots, T\mathbf{v}_m\}$$

and that in this situation,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$$

is linearly independent.

37. ↑In the situation of Problem 36 where V is a subspace of $\mathbb{F}^n$, show that there exists $\{\mathbf{z}_1, \cdots, \mathbf{z}_r\}$ a basis for $\ker(T)$. (Recall Theorem 2.6.12. Since $\ker(T)$ is a subspace, it has a basis.) Now for an arbitrary $T\mathbf{v} \in T(V)$, explain why

$$T\mathbf{v} = a_1 T\mathbf{v}_1 + \cdots + a_m T\mathbf{v}_m$$

and why this implies

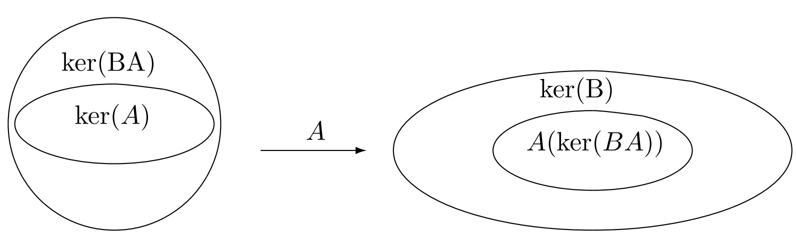
$$\mathbf{v} - (a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m) \in \ker(T).$$

Then explain why $V = \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_m, \mathbf{z}_1, \cdots, \mathbf{z}_r)$.

38. ↑In the situation of the above problem, show $\{\mathbf{v}_1, \cdots, \mathbf{v}_m, \mathbf{z}_1, \cdots, \mathbf{z}_r\}$ is a basis for V and therefore, $\dim(V) = \dim(\ker(T)) + \dim(T(V))$.

39. ↑Let A be a linear transformation from V to W and let B be a linear transformation from W to U where V, W, U are all subspaces of some $\mathbb{F}^p$. Explain why

$$A(\ker(BA)) \subseteq \ker(B), \ \ker(A) \subseteq \ker(BA).$$



40. ↑Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis of $\ker(A)$ and let

$$\{A\mathbf{y}_1, \cdots, A\mathbf{y}_m\}$$

be a basis of $A(\ker(BA))$. Let $\mathbf{z} \in \ker(BA)$. Explain why

$$Az \in \text{span}\{A\mathbf{y}_1, \cdots, A\mathbf{y}_m\}$$

and why there exist scalars a_i such that

$$A\left(z-\left(a_1\mathbf{y}_1+\cdots+a_m\mathbf{y}_m\right)\right)=0$$

and why it follows $z-\left(a_1\mathbf{y}_1+\cdots+a_m\mathbf{y}_m\right)\in\operatorname{span}\left\{\mathbf{x}_1,\cdots,\mathbf{x}_n\right\}$. Now explain why

$$\ker\left(BA\right)\subseteq\operatorname{span}\left\{\mathbf{x}_1,\cdots,\mathbf{x}_n,\mathbf{y}_1,\cdots,\mathbf{y}_m\right\}$$

and so

$$\dim\left(\ker\left(BA\right)\right)\leq\dim\left(\ker\left(B\right)\right)+\dim\left(\ker\left(A\right)\right).$$

This important inequality is due to Sylvester. Show that equality holds if and only if $A(\ker BA)=\ker(B)$.

41. Generalize the result of the previous problem to any finite product of linear mappings.

42. If $W\subseteq V$ for W,V two subspaces of $\mathbb{F}^n$ and if $\dim\left(W\right)=\dim\left(V\right)$, show $W=V$.

43. Let V be a subspace of $\mathbb{F}^n$and let $V_1,\cdots,V_m$ be subspaces, each contained in V. Then

$$V=V_1\oplus\cdots\oplus V_m \tag{2.37}$$

if every $v\in V$ can be written in a unique way in the form

$$v=v_1+\cdots+v_m$$

where each $v_i\in V_i$. This is called a direct sum. If this uniqueness condition does not hold, then one writes

$$V=V_1+\cdots+V_m$$

and this symbol means all vectors of the form

$$v_1+\cdots+v_m,\ v_j\in V_j \text{ for each } j.$$

Show 2.37 is equivalent to saying that if

$$0=v_1+\cdots+v_m,\ v_j\in V_j \text{ for each } j,$$

then each $v_j=0$. Next show that in the situation of 2.37, if $\beta_i=\left\{u_1^i,\cdots,u_{m_i}^i\right\}$ is a basis for V_i, then $\left\{\beta_1,\cdots,\beta_m\right\}$ is a basis for V.

44. ↑Suppose you have finitely many linear mappings $L_1,L_2,\cdots,L_m$ which map V to V where V is a subspace of $\mathbb{F}^n$ and suppose they commute. That is, $L_iL_j=L_jL_i$ for all i,j. Also suppose L_k is one to one on $\ker\left(L_j\right)$ whenever $j\neq k$. Letting P denote the product of these linear transformations, $P=L_1L_2\cdots L_m$, first show

$$\ker\left(L_1\right)+\cdots+\ker\left(L_m\right)\subseteq\ker\left(P\right)$$

Next show $L_j:\ker\left(L_i\right)\to\ker\left(L_i\right)$. Then show

$$\ker\left(L_1\right)+\cdots+\ker\left(L_m\right)=\ker\left(L_1\right)\oplus\cdots\oplus\ker\left(L_m\right).$$

Using Sylvester's theorem, and the result of Problem 42, show

$$\ker\left(P\right)=\ker\left(L_1\right)\oplus\cdots\oplus\ker\left(L_m\right)$$

Hint: By Sylvester's theorem and the above problem,

$$\begin{aligned}\dim\left(\ker\left(P\right)\right) &\leq \sum_i\dim\left(\ker\left(L_i\right)\right)\\ &= \dim\left(\ker\left(L_1\right)\oplus\cdots\oplus\ker\left(L_m\right)\right)\leq\dim\left(\ker\left(P\right)\right)\end{aligned}$$

Now consider Problem 42.

45. Let $\mathcal{M}(\mathbb{F}^n, \mathbb{F}^n)$ denote the set of all $n \times n$ matrices having entries in $\mathbb{F}$. With the usual operations of matrix addition and scalar multiplications, explain why $\mathcal{M}(\mathbb{F}^n, \mathbb{F}^n)$ can be considered as $\mathbb{F}^{n^2}$. Give a basis for $\mathcal{M}(\mathbb{F}^n, \mathbb{F}^n)$. If $A \in \mathcal{M}(\mathbb{F}^n, \mathbb{F}^n)$, explain why there exists a monic (leading coefficient equals 1) polynomial of the form
$$\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0$$
such that
$$A^k + a_{k-1}A^{k-1} + \cdots + a_1 A + a_0 I = 0$$
The minimal polynomial of A is the polynomial like the above, for which $p(A) = 0$ which has smallest degree. I will discuss the uniqueness of this polynomial later. **Hint:** Consider the matrices $I, A, A^2, \cdots, A^{n^2}$. There are $n^2 + 1$ of these matrices. Can they be linearly independent? Now consider all polynomials and pick one of smallest degree and then divide by the leading coefficient.

46. ↑Suppose the field of scalars is $\mathbb{C}$ and A is an $n \times n$ matrix. From the preceding problem, and the fundamental theorem of algebra, this minimal polynomial factors
$$(\lambda - \lambda_1)^{r_1}(\lambda - \lambda_2)^{r_2} \cdots (\lambda - \lambda_k)^{r_k}$$
where r_j is the algebraic multiplicity of λ_j, and the λ_j are distinct. Thus
$$(A - \lambda_1 I)^{r_1}(A - \lambda_2 I)^{r_2} \cdots (A - \lambda_k I)^{r_k} = 0$$
and so, letting $P = (A - \lambda_1 I)^{r_1}(A - \lambda_2 I)^{r_2} \cdots (A - \lambda_k I)^{r_k}$ and $L_j = (A - \lambda_j I)^{r_j}$ apply the result of Problem 44 to verify that
$$\mathbb{C}^n = \ker(L_1) \oplus \cdots \oplus \ker(L_k)$$
and that $A : \ker(L_j) \to \ker(L_j)$. In this context, $\ker(L_j)$ is called the generalized eigenspace for λ_j. You need to verify the conditions of the result of this problem hold.

47. In the context of Problem 46, show there exists a nonzero vector $\mathbf{x}$ such that
$$(A - \lambda_j I)\mathbf{x} = \mathbf{0}.$$
This is called an eigenvector and the λ_j is called an eigenvalue. **Hint:**There must exist a vector $\mathbf{y}$ such that
$$(A - \lambda_1 I)^{r_1}(A - \lambda_2 I)^{r_2} \cdots (A - \lambda_j I)^{r_j - 1} \cdots (A - \lambda_k I)^{r_k}\mathbf{y} = \mathbf{z} \neq \mathbf{0}$$
Why? Now what happens if you do $(A - \lambda_j I)$ to $\mathbf{z}$?

48. Suppose $Q(t)$ is an orthogonal matrix. This means $Q(t)$ is a real $n \times n$ matrix which satisfies
$$Q(t)Q(t)^T = I$$
Suppose also the entries of $Q(t)$ are differentiable. Show $\left(Q^T\right)' = -Q^T Q' Q^T$.

49. Remember the Coriolis force was $2\boldsymbol{\Omega} \times \mathbf{v}_B$ where $\boldsymbol{\Omega}$ was a particular vector which came from the matrix $Q(t)$ as described above. Show that
$$Q(t) = \begin{pmatrix} \mathbf{i}(t) \cdot \mathbf{i}(t_0) & \mathbf{j}(t) \cdot \mathbf{i}(t_0) & \mathbf{k}(t) \cdot \mathbf{i}(t_0) \\ \mathbf{i}(t) \cdot \mathbf{j}(t_0) & \mathbf{j}(t) \cdot \mathbf{j}(t_0) & \mathbf{k}(t) \cdot \mathbf{j}(t_0) \\ \mathbf{i}(t) \cdot \mathbf{k}(t_0) & \mathbf{j}(t) \cdot \mathbf{k}(t_0) & \mathbf{k}(t) \cdot \mathbf{k}(t_0) \end{pmatrix}.$$
There will be no Coriolis force exactly when $\boldsymbol{\Omega} = \mathbf{0}$ which corresponds to $Q'(t) = 0$. When will $Q'(t) = 0$?

50. An illustration used in many beginning physics books is that of firing a rifle horizontally and dropping an identical bullet from the same height above the perfectly flat ground followed by an assertion that the two bullets will hit the ground at exactly the same time. Is this true on the rotating earth assuming the experiment takes place over a large perfectly flat field so the curvature of the earth is not an issue? Explain. What other irregularities will occur? Recall the Coriolis acceleration is $2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$ where $\mathbf{k}$ points away from the center of the earth, $\mathbf{j}$ points East, and $\mathbf{i}$ points South.

Chapter 3

Determinants

3.1 Basic Techniques and Properties

Let A be an $n \times n$ matrix. The determinant of A, denoted as $\det(A)$ is a number. If the matrix is a 2×2 matrix, this number is very easy to find.

Definition 3.1.1 *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 3.1.2 *Find* $\det \begin{pmatrix} 2 & 4 \\ -1 & 6 \end{pmatrix}$.

From the definition this is just $(2)(6) - (-1)(4) = 16$.

Assuming the determinant has been defined for $k \times k$ matrices for $k \leq n-1$, it is now time to define it for $n \times n$ matrices.

Definition 3.1.3 *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.*

Now here is the definition of the determinant given recursively.

Theorem 3.1.4 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{3.1}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Note that for a $n \times n$ matrix, you will need $n!$ terms to evaluate the determinant in this way. If $n = 10$, this is $10! = 3,628,800$ terms. This is a lot of terms.

In addition to the difficulties just discussed, why is the determinant well defined? Why should you get the same thing when you expand along any row or column? I think you should regard this claim that you always get the same answer by picking any row or column with considerable skepticism. It is incredible and not at all obvious. However, it requires a little effort to establish it. This is done in the section on the theory of the determinant which follows.

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.

Definition 3.1.5 *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.*

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

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 3.1.6 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

Proof: The corollary is true if the matrix is one to one. Suppose it is $n \times n$. Then the matrix is of the form

$$\begin{pmatrix} m_{11} & \mathbf{a} \\ \mathbf{0} & M_1 \end{pmatrix}$$

where M_1 is $(n-1) \times (n-1)$. Then expanding along the first row, you get $m_{11} \det(M_1) + 0$. Then use the induction hypothesis to obtain that $\det(M_1) = \prod_{i=2}^{n} m_{ii}$. ■

Example 3.1.7 *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, this is -6.

There are many properties satisfied by determinants. Some of the most important are listed in the following theorem.

Theorem 3.1.8 *If two rows or two columns in an $n \times n$ matrix A are switched, the determinant of the resulting matrix equals (-1) times the determinant of the original matrix. If A is an $n \times n$ matrix in which two rows are equal or two columns are equal then $\det(A) = 0$. Suppose the i^{th} row of A equals $(xa_1 + yb_1, \cdots, xa_n + yb_n)$. Then*

$$\det(A) = x \det(A_1) + y \det(A_2)$$

where the i^{th} row of A_1 is $(a_1, \cdots, a_n)$ and the i^{th} row of A_2 is $(b_1, \cdots, b_n)$, all other rows of A_1 and A_2 coinciding with those of A. In other words, $\det$ is a linear function of each row A. The same is true with the word "row" replaced with the word "column". In addition to this, if A and B are $n \times n$ matrices, then

$$\det(AB) = \det(A)\det(B),$$

and if A is an $n \times n$ matrix, then

$$\det(A) = \det\left(A^T\right).$$

This theorem implies the following corollary which gives a way to find determinants. As I pointed out above, the method of Laplace expansion will not be practical for any matrix of large size.

Corollary 3.1.9 *Let A be an $n \times n$ matrix and let B be the matrix obtained by replacing the i^{th} row (column) of A with the sum of the i^{th} row (column) added to a multiple of another row (column). Then* $\det(A) = \det(B)$*. If B is the matrix obtained from A be replacing the i^{th} row (column) of A by a times the i^{th} row (column) then* $a\det(A) = \det(B)$*.*

Here is an example which shows how to use this corollary to find a determinant.

Example 3.1.10 *Find the determinant of the matrix*

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

First take -1 times the first row and add to the second and the third. The resulting matrix is

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

It has the same determinant as the original matrix. Next switch the bottom two rows to get

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

It has determinant which is -1 times the determinant of the original matrix. Hence the original matrix has determinant equal to 1.

The theorem about expanding a matrix along any row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 2.1.22 on Page 48. The following theorem gives a formula for the inverse of a matrix. It is proved in the next section.

Theorem 3.1.11 *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.

Theorem 3.1.11 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix A. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

Example 3.1.12 *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. This is seen to be 12. The cofactor matrix of A is

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

Each entry of A was replaced by its cofactor. Therefore, from the above theorem, the inverse of A should equal

$$\frac{1}{12}\begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}.$$

This way of finding inverses is especially useful in the case where it is desired to find the inverse of a matrix whose entries are functions.

Example 3.1.13 *Suppose*

$$A(t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix}$$

Find $A(t)^{-1}$.

First note $\det(A(t)) = e^t$. A routine computation using the above theorem shows that this inverse is

$$\frac{1}{e^t}\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t\cos t & e^t\sin t \\ 0 & -e^t\sin t & e^t\cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

This formula for the inverse also implies a famous 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 case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^n a_{ij}^{-1} y_j = \sum_{j=1}^n \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det\begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Procedure 3.1.14 *Suppose A is an $n \times n$ matrix and it is desired to solve the system $A\mathbf{x} = \mathbf{y}, \mathbf{y} = (y_1, \cdots, y_n)^T$ for $\mathbf{x} = (x_1, \cdots, x_n)^T$. Then Cramer's rule says*

$$x_i = \frac{\det A_i}{\det A}$$

where A_i is obtained from A by replacing the i^{th} column of A with the column $(y_1, \cdots, y_n)^T$.

The following theorem is of fundamental importance and ties together many of the ideas presented above. It is proved in the next section.

Theorem 3.1.15 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. *A is one to one.*
2. *A is onto.*
3. $\det(A) \neq 0$.

3.2 Exercises

1. Find the determinants of the following matrices.

   (a) $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 2 \\ 0 & 9 & 8 \end{pmatrix}$ (The answer is 31.)

   (b) $\begin{pmatrix} 4 & 3 & 2 \\ 1 & 7 & 8 \\ 3 & -9 & 3 \end{pmatrix}$ (The answer is 375.)

   (c) $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 3 \\ 4 & 1 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{pmatrix}$, (The answer is -2.)

2. If A^{-1} exists, what is the relationship between $\det(A)$ and $\det(A^{-1})$. Explain your answer.

3. Let A be an $n \times n$ matrix where n is odd. Suppose also that A is skew symmetric. This means $A^T = -A$. Show that $\det(A) = 0$.

4. Is it true that $\det(A + B) = \det(A) + \det(B)$? If this is so, explain why it is so and if it is not so, give a counter example.

5. Let A be an $r \times r$ matrix and suppose there are $r - 1$ rows (columns) such that all rows (columns) are linear combinations of these $r - 1$ rows (columns). Show $\det(A) = 0$.

6. Show $\det(aA) = a^n \det(A)$ where here A is an $n \times n$ matrix and a is a scalar.

7. Suppose A is an upper triangular matrix. Show that A^{-1} exists if and only if all elements of the main diagonal are non zero. Is it true that A^{-1} will also be upper triangular? Explain. Is everything the same for lower triangular matrices?

8. Let A and B be two $n \times n$ matrices. $A \sim B$ (A is similar to B) means there exists an invertible matrix S such that $A = S^{-1}BS$. Show that if $A \sim B$, then $B \sim A$. Show also that $A \sim A$ and that if $A \sim B$ and $B \sim C$, then $A \sim C$.

9. In the context of Problem 8 show that if $A \sim B$, then $\det(A) = \det(B)$.

10. Let A be an $n \times n$ matrix and let $\mathbf{x}$ be a nonzero vector such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar, λ. When this occurs, the vector, $\mathbf{x}$ is called an eigenvector and the scalar, λ is called an eigenvalue. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $A\mathbf{x} = \lambda\mathbf{x}$, then $(\lambda I - A)\mathbf{x} = \mathbf{0}$. Explain why this shows that $(\lambda I - A)$ is not one to one and not onto. Now use Theorem 3.1.15 to argue $\det(\lambda I - A) = 0$. What sort of equation is this? How many solutions does it have?

11. Suppose $\det(\lambda I - A) = 0$. Show using Theorem 3.1.15 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$.

12. Let $F(t) = \det\begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$. Verify
$$F'(t) = \det\begin{pmatrix} a'(t) & b'(t) \\ c(t) & d(t) \end{pmatrix} + \det\begin{pmatrix} a(t) & b(t) \\ c'(t) & d'(t) \end{pmatrix}.$$
Now suppose
$$F(t) = \det\begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix}.$$
Use Laplace expansion and the first part to verify $F'(t) =$
$$\det\begin{pmatrix} a'(t) & b'(t) & c'(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix} + \det\begin{pmatrix} a(t) & b(t) & c(t) \\ d'(t) & e'(t) & f'(t) \\ g(t) & h(t) & i(t) \end{pmatrix}$$
$$+ \det\begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g'(t) & h'(t) & i'(t) \end{pmatrix}.$$
Conjecture a general result valid for $n \times n$ matrices and explain why it will be true. Can a similar thing be done with the columns?

13. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix
$$A = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \end{pmatrix}.$$

14. Let A be an $r \times r$ matrix and let B be an $m \times m$ matrix such that $r + m = n$. Consider the following $n \times n$ block matrix
$$C = \begin{pmatrix} A & 0 \\ D & B \end{pmatrix}.$$

where the D is an $m \times r$ matrix, and the 0 is a $r \times m$ matrix. Letting I_k denote the $k \times k$ identity matrix, tell why

$$C = \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} \begin{pmatrix} I_r & 0 \\ 0 & B \end{pmatrix}.$$

Now explain why $\det(C) = \det(A)\det(B)$. **Hint:** Part of this will require an explanation of why

$$\det \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} = \det(A).$$

See Corollary 3.1.9.

15. Suppose Q is an orthogonal matrix. This means Q is a real $n \times n$ matrix which satisfies $QQ^T = I$. Find the possible values for $\det(Q)$.

16. Suppose $Q(t)$ is an orthogonal matrix. This means $Q(t)$ is a real $n \times n$ matrix which satisfies $Q(t)Q(t)^T = I$ Suppose $Q(t)$ is continuous for $t \in [a, b]$, some interval. Also suppose $\det(Q(t)) = 1$. Show that $\det(Q(t)) = 1$ for all $t \in [a, b]$.

3.3 The Mathematical Theory of Determinants

It is easiest to give a different definition of the determinant which is clearly well defined and then prove the one which involves Laplace expansion. Let $(i_1, \cdots, i_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1, 2, 3)$ and $(2, 1, 3)$ are different. There will be some repetition between this section and the earlier section on determinants. The main purpose is to give all the missing proofs. Two books which give a good introduction to determinants are Apostol [1] and Rudin [23]. A recent book which also has a good introduction is Baker [3]

3.3.1 The Function sgn

The following Lemma will be essential in the definition of the determinant.

Lemma 3.3.1 *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.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{3.2}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{3.3}$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1. *Also, in the case where* $n > 1$ *and* $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ *so that every number from* $\{1, \cdots, n\}$ *appears in the ordered list,* $(i_1, \cdots, i_n)$,

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta} \operatorname{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{3.4}$$

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 1	i_2 2	$\cdots$	i_p p	$\cdots$	i_q q	$\cdots$	i_n n
i_1 1	i_2 2	$\cdots$	i_q p	$\cdots$	i_p q	$\cdots$	i_n n
j_1 1	j_2 2	$\cdots$	j_p p	$\cdots$	j_q q	$\cdots$	j_n n

Then

$$\text{sgn}_n(j_1, \cdots, j_n) \equiv \text{sign}\left(\prod_{r<s}(j_s - j_r)\right)$$

$$= \text{sign}\left(\overbrace{(i_p - i_q)}^{\text{both } p,q} \prod_{p<j<q} \overbrace{(i_j - i_q) \prod_{p<j<q}(i_p - i_j)}^{\text{one of } p,q} \overbrace{\prod_{r<s,r,s\notin\{p,q\}}(i_s - i_r)}^{\text{neither } p \text{ nor } q}\right)$$

The last product consists of the product of terms which were in $\prod_{r<s}(i_s - i_r)$ while the two products in the middle both introduce $q - p - 1$ minus signs. Thus their product is positive. The first factor is of opposite sign to the $i_q - i_p$ which occured in $\text{sgn}_n(i_1, \cdots, i_n)$. Therefore, this switch introduced a minus sign and

$$\text{sgn}_n(j_1, \cdots, j_n) = -\,\text{sgn}_n(i_1, \cdots, i_n)$$

Now consider the last claim. In computing $\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n)$ 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 $\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n)$ are those which are required to compute $\text{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n)$ multiplied by terms of the form $(n - i_j)$ which are nonnegative. It follows that

$$\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) = (-1)^{n-\theta}\,\text{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n)$$

It is obvious that if there are repeats in the list the function gives 0. ■

Lemma 3.3.2 *Every ordered list of distinct numbers from $\{1, 2, \cdots, n\}$ can be obtained from every other ordered list of distinct numbers 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 3.2 and 3.3, 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 3.3 gives both functions are equal to zero for that ordered list. ■

Definition 3.3.3 *When you have an ordered list of distinct numbers from* $\{1, 2, \cdots, n\}$, *say*

$$(i_1, \cdots, i_n),$$

this ordered list is called a permutation. The symbol for all such permutations is S_n. *The number* $\text{sgn}_n(i_1, \cdots, i_n)$ *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.

3.3.2 The Definition of the Determinant

Definition 3.3.4 *Let f be a real valued function which has the set of ordered lists of numbers from* $\{1, \cdots, n\}$ *as its domain. Define*

$$\sum_{(k_1, \cdots, k_n)} f(k_1 \cdots k_n)$$

to be the sum of all the $f(k_1 \cdots k_n)$ *for all possible choices of ordered lists* $(k_1, \cdots, k_n)$ *of numbers of* $\{1, \cdots, n\}$. *For example,*

$$\sum_{(k_1, k_2)} f(k_1, k_2) = f(1,2) + f(2,1) + f(1,1) + f(2,2).$$

Definition 3.3.5 *Let* $(a_{ij}) = A$ *denote an* $n \times n$ *matrix. The determinant of* A, *denoted by* $\det(A)$ *is defined by*

$$\det(A) \equiv \sum_{(k_1, \cdots, k_n)} \text{sgn}(k_1, \cdots, k_n)\, a_{1k_1} \cdots a_{nk_n}$$

where the sum is taken over all ordered lists of numbers from $\{1, \cdots, n\}$. *Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are,* $\text{sgn}(k_1, \cdots, k_n) = 0$ *and so that term contributes 0 to the sum.*

Let A be an $n \times n$ matrix $A = (a_{ij})$ and let $(r_1, \cdots, r_n)$ denote an ordered list of n numbers from $\{1, \cdots, n\}$. Let $A(r_1, \cdots, r_n)$ denote the matrix whose k^{th} row is the r_k row of the matrix A. Thus

$$\det(A(r_1, \cdots, r_n)) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \tag{3.5}$$

and $A(1, \cdots, n) = A$.

Proposition 3.3.6 *Let $(r_1, \cdots, r_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. Then*

$$\begin{aligned} \operatorname{sgn}(r_1, \cdots, r_n) \det(A) &= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} &(3.6) \\ &= \det(A(r_1, \cdots, r_n)). &(3.7) \end{aligned}$$

Proof: Let $(1, \cdots, n) = (1, \cdots, r, \cdots s, \cdots, n)$ so $r < s$.

$$\det(A(1, \cdots, r, \cdots, s, \cdots, n)) = \tag{3.8}$$

$$\sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n},$$

and renaming the variables, calling k_s, k_r and k_r, k_s, this equals

$$= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_s, \cdots, k_r, \cdots, k_n) a_{1k_1} \cdots a_{rk_s} \cdots a_{sk_r} \cdots a_{nk_n}$$

$$= \sum_{(k_1, \cdots, k_n)} -\operatorname{sgn}\left(k_1, \cdots, \overbrace{k_r, \cdots, k_s}^{\text{These got switched}}, \cdots, k_n\right) a_{1k_1} \cdots a_{sk_r} \cdots a_{rk_s} \cdots a_{nk_n}$$

$$= -\det(A(1, \cdots, s, \cdots, r, \cdots, n)). \tag{3.9}$$

Consequently,

$$\det(A(1, \cdots, s, \cdots, r, \cdots, n)) = -\det(A(1, \cdots, r, \cdots, s, \cdots, n)) = -\det(A)$$

Now letting $A(1, \cdots, s, \cdots, r, \cdots, n)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det(A(r_1, \cdots, r_n)) = (-1)^p \det(A)$$

where it took p switches to obtain$(r_1, \cdots, r_n)$ from $(1, \cdots, n)$. By Lemma 3.3.1, this implies

$$\det(A(r_1, \cdots, r_n)) = (-1)^p \det(A) = \operatorname{sgn}(r_1, \cdots, r_n) \det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1, \cdots, r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the same reasoning of 3.8 -3.9 shows that $\det(A(r_1, \cdots, r_n)) = 0$ but in this case, $\operatorname{sgn}(r_1, \cdots, r_n) = 0$ so the formula also holds even in case a number is repeated in the list of numbers from $1, 2, ..., n$. ■

Observation 3.3.7 *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.

3.3.3 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 3.3.8 *The following formula for* $\det(A)$ *is valid.*

$$\det(A) = \frac{1}{n!} \cdot \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(r_1, \cdots, r_n) \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}. \tag{3.10}$$

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 3.3.6, if the r_i are distinct,

$$\det(A) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(r_1, \cdots, r_n) \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.$$

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.)

$$n! \det(A) = \sum_{(r_1, \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(r_1, \cdots, r_n) \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T. ■

Corollary 3.3.9 *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 3.3.6 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 3.3.8 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.

$$\det(A) \equiv \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots (xa_{rk_i} + yb_{rk_i}) \cdots a_{nk_n}$$

$$= x \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots a_{rk_i} \cdots a_{nk_n}$$

$$+ y \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots b_{rk_i} \cdots a_{nk_n} \equiv x \det(A_1) + y \det(A_2).$$

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. ■

3.3.4 Basic Properties of the Determinant

Definition 3.3.10 *A vector,* $\mathbf{w}$*, is a linear combination of the vectors* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ *if there exist 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 3.3.11 *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. Say $\mathbf{a}_i = \sum_{j \neq i} c_j \mathbf{a}_j$. Then by Corollary 3.3.9, $\det(A) =$

$$\det \begin{pmatrix} \mathbf{a}_1 & \cdots & \sum_{j \neq i} c_j \mathbf{a}_j & \cdots & \mathbf{a}_n \end{pmatrix} = \sum_{j \neq i} c_j \det \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_j & \cdots & \mathbf{a}_n \end{pmatrix} = 0$$

because each of these determinants in the sum has two equal rows. ■

Recall the following definition of matrix multiplication.

Definition 3.3.12 *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 3.3.13 *Let* A *and* B *be* $n \times n$ *matrices. Then*

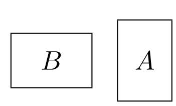
$$\det(AB) = \det(A) \det(B).$$

Proof: Let c_{ij} be the ij^{th} entry of AB. Then by Proposition 3.3.6, and the way we multiply matrices,

$$\begin{aligned} \det(AB) &= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) c_{1k_1} \cdots c_{nk_n} \\ &= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) \left(\sum_{r_1} a_{1r_1} b_{r_1 k_1} \right) \cdots \left(\sum_{r_n} a_{nr_n} b_{r_n k_n} \right) \end{aligned}$$

$$\begin{aligned} &= \sum_{(r_1 \cdots, r_n)} \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) b_{r_1 k_1} \cdots b_{r_n k_n} (a_{1r_1} \cdots a_{nr_n}) \\ &= \sum_{(r_1 \cdots, r_n)} \operatorname{sgn}(r_1 \cdots r_n) a_{1r_1} \cdots a_{nr_n} \det(B) = \det(A) \det(B). \blacksquare \end{aligned}$$

The Binet Cauchy formula is a generalization of the theorem which says the determinant of a product is the product of the determinants. The situation is illustrated in the following picture where A, B are matrices.



Theorem 3.3.14 *Let A be an $n \times m$ matrix with $n \geq m$ and let B be a $m \times n$ matrix. Also let A_i*

$$i = 1, \cdots, C(n,m)$$

be the $m \times m$ submatrices of A which are obtained by deleting $n-m$ rows and let B_i be the $m \times m$ submatrices of B which are obtained by deleting corresponding $n-m$ columns. Then

$$\det(BA) = \sum_{k=1}^{C(n,m)} \det(B_k)\det(A_k)$$

Proof: This follows from a computation. By Corollary 3.3.8 on Page 93, $\det(BA) =$

$$\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)(BA)_{i_1 j_1}(BA)_{i_2 j_2}\cdots(BA)_{i_m j_m}$$

$$\begin{gathered}\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot \\ \sum_{r_1=1}^{n} B_{i_1 r_1}A_{r_1 j_1}\sum_{r_2=1}^{n} B_{i_2 r_2}A_{r_2 j_2}\cdots\sum_{r_m=1}^{n} B_{i_m r_m}A_{r_m j_m}\end{gathered}$$

Now denote by I_k one of the subsets of $\{1, \cdots, n\}$ which has m elements. Thus there are $C(n,m)$ of these.

$$\begin{aligned}= & \sum_{k=1}^{C(n,m)}\sum_{\{r_1,\cdots,r_m\}=I_k}\frac{1}{m!}\sum_{(i_1\cdots i_m)}\sum_{(j_1\cdots j_m)} \operatorname{sgn}(i_1\cdots i_m)\operatorname{sgn}(j_1\cdots j_m)\cdot \\ & B_{i_1 r_1}A_{r_1 j_1}B_{i_2 r_2}A_{r_2 j_2}\cdots B_{i_m r_m}A_{r_m j_m}\end{aligned}$$

$$\begin{aligned}= & \sum_{k=1}^{C(n,m)}\sum_{\{r_1,\cdots,r_m\}=I_k}\frac{1}{m!}\sum_{(i_1\cdots i_m)} \operatorname{sgn}(i_1\cdots i_m)B_{i_1 r_1}B_{i_2 r_2}\cdots B_{i_m r_m}\cdot \\ & \sum_{(j_1\cdots j_m)} \operatorname{sgn}(j_1\cdots j_m)A_{r_1 j_1}A_{r_2 j_2}\cdots A_{r_m j_m}\end{aligned}$$

$$= \sum_{k=1}^{C(n,m)}\sum_{\{r_1,\cdots,r_m\}=I_k}\frac{1}{m!}\operatorname{sgn}(r_1\cdots r_m)^2\det(B_k)\det(A_k) = \sum_{k=1}^{C(n,m)}\det(B_k)\det(A_k)$$

since there are $m!$ ways of arranging the indices $\{r_1, \cdots, r_m\}$. ■

3.3.5 Expansion Using Cofactors

Lemma 3.3.15 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \text{ or } \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{3.11}$$

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 the earlier conventions used to prove Lemma 3.3.1, $\det(M)$ equals

$$\sum_{(k_1,\cdots,k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_1,\cdots,k_{\theta-1},\overset{\theta}{k_{\theta+1}},\cdots,\overset{n-1}{k_n}\right) m_{1k_1}\cdots m_{nk_n}$$

Now suppose the second case. 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 first case, use Corollary 3.3.8 to write

$$\det(M) = \det\left(M^T\right) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det\left(A^T\right) = a\det(A). \blacksquare$$

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 3.3.16 *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 3.3.17 *Let A be an $n\times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^n a_{ij}\operatorname{cof}(A)_{ij} = \sum_{i=1}^n a_{ij}\operatorname{cof}(A)_{ij}. \tag{3.12}$$

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 3.3.9,

$$\det(A) = \sum_{j=1}^n \det(B_j)$$

For example if

$$A=\begin{pmatrix} a & b & c \\ d & e & f \\ h & i & j \end{pmatrix}$$

and $i=2$, then

$$B_1=\begin{pmatrix} a & b & c \\ d & 0 & 0 \\ h & i & j \end{pmatrix}, B_2=\begin{pmatrix} a & b & c \\ 0 & e & 0 \\ h & i & j \end{pmatrix}, B_3=\begin{pmatrix} a & b & c \\ 0 & 0 & f \\ h & i & j \end{pmatrix}$$

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 3.3.6, 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 3.3.15,

$$\begin{aligned}\det(B_j) &= (-1)^{n-j}(-1)^{n-i}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) \\ &= (-1)^{i+j}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) = a_{ij}\operatorname{cof}(A)_{ij}.\end{aligned}$$

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,

$$\det(A)=\det\left(A^T\right)=\sum_{j=1}^{n} a_{ij}^T\operatorname{cof}\left(A^T\right)_{ij}=\sum_{j=1}^{n} a_{ji}\operatorname{cof}(A)_{ji}$$

which is the formula for expanding $\det(A)$ along the i^{th} column. ■

3.3.6 A Formula for the Inverse

Note that this gives an easy way to write a formula for the inverse of an $n\times n$ matrix. Recall the definition of the inverse of a matrix in Definition 2.1.22 on Page 48.

Theorem 3.3.18 *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 3.3.17 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 in the matrix A, replace the k^{th} column with the r^{th} column. This results in two equal columns if $r \neq k$ and no change if $r = k$. Then expand along the k^{th} column. This yields for $k \neq r$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = 0$$

because there are two equal columns by Corollary 3.3.9. Summarizing,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 3.3.17, 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 3.3.13,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

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 3.3.19 *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 3.3.13 implies $\det B \det A = 1$ and so $\det A \neq 0$. Therefore from Theorem 3.3.18, 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 3.3.18 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix A. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

In case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Definition 3.3.20 *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 3.3.17.

Corollary 3.3.21 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

3.3.7 Rank of a Matrix

Definition 3.3.22 *A submatrix of a matrix A is the rectangular array of numbers obtained by deleting some rows and columns of A. Let A be an $m \times n$ matrix. The* ***determinant rank*** *of the matrix equals r where r is the largest number such that some $r \times r$ submatrix of A has a non zero determinant. The* ***row rank*** *is defined to be the dimension of the span of the rows. The* ***column rank*** *is defined to be the dimension of the span of the columns.*

Theorem 3.3.23 *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 having ij^{th} entry 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

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} & a_{i_1 p} \\ \vdots & & \vdots & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} & a_{i_r p} \\ a_{l j_1} & \cdots & a_{l j_r} & a_{l p} \end{pmatrix}$$

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 $r \times r$ matrix in the upper left corner. 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 3.3.24 *The determinant rank equals the row rank.*

Proof: From Theorem 3.3.23, every row is in the span of r rows where r is the determinant rank. Therefore, the row rank (dimension of the span of the rows) is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, it follows from Theorem 3.3.23 that there exist p rows for $p < r \equiv$ determinant rank, such that the span of these p rows equals the row space. But then you could consider the $r \times r$ sub matrix which determines the determinant rank and it would follow that each of these rows would be in the span of the restrictions of the p rows just mentioned. By Theorem 2.6.4, the exchange theorem, the rows of this sub matrix would not be linearly independent and so some row is a linear combination of the others. By Corollary 3.3.11 the determinant would be 0, a contradiction. ■

Corollary 3.3.25 *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 3.3.24, 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 3.3.26 *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 3.3.23. 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

$$\mathbf{x}^T A = \mathbf{0}^T$$

where the $\mathbf{0}^T$ 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 3.3.18 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 3.3.27 *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.

3.3.8 Summary of Determinants

In all the following A, B are $n \times n$ matrices

1. $\det(A)$ is a number.
2. $\det(A)$ is linear in each row and in each column.

3. If you switch two rows or two columns, the determinant of the resulting matrix is -1 times the determinant of the unswitched matrix. (This and the previous one say

$$(\mathbf{a}_1 \cdots \mathbf{a}_n) \to \det (\mathbf{a}_1 \cdots \mathbf{a}_n)$$

is an alternating multilinear function or alternating tensor.

4. $\det (\mathbf{e}_1, \cdots, \mathbf{e}_n) = 1$.

5. $\det (AB) = \det (A) \det (B)$

6. $\det (A)$ can be expanded along any row or any column and the same result is obtained.

7. $\det (A) = \det \left(A^T\right)$

8. A^{-1} exists if and only if $\det (A) \neq 0$ and in this case

$$\left(A^{-1}\right)_{ij} = \frac{1}{\det (A)} \operatorname{cof} (A)_{ji} \tag{3.13}$$

9. Determinant rank, row rank and column rank are all the same number for any $m \times n$ matrix.

3.4 The Cayley Hamilton Theorem

Definition 3.4.1 *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.

Lemma 3.4.2 *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: Suppose some $A_i \neq 0$. Let p be the largest index of those which are non zero. Then multiply by λ^{-p}.

$$A_0 \lambda^{-p} + A_1 \lambda^{-p+1} + \cdots + A_{p-1} \lambda^{-1} + A_p = 0$$

Now let $\lambda \to \infty$. Thus $A_p = 0$ after all. Hence each $A_i = 0$. ■

With the lemma, here is a simple corollary.

[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.

Corollary 3.4.3 *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 3.4.4 *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 3.3.18

$$C(\lambda) = q(\lambda)(\lambda I - A)^{-1}.$$

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,

$$a_0 I + a_1 I\lambda + \cdots + I\lambda^n = C_0\lambda + C_1\lambda^2 + \cdots + C_{n-1}\lambda^n$$

$$-\left[C_0 A + C_1 A\lambda + \cdots + C_{n-1}A\lambda^{n-1}\right]$$

$$= -C_0 A + (C_0 - C_1 A)\lambda + (C_1 - C_2 A)\lambda^2 + \cdots + (C_{n-2} - C_{n-1}A)\lambda^{n-1} + C_{n-1}\lambda^n$$

Then, using Corollary 3.4.3, 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. ■

3.5 Block Multiplication of Matrices

Consider the following problem

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

You know how to do this. You get

$$\begin{pmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{pmatrix}.$$

Now what if instead of numbers, the entries, A, B, C, D, E, F, G are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that this is true.

Suppose A is a matrix of the form

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{3.14}$$

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:

$$\overbrace{\begin{pmatrix} \mathbf{0} & I_{s_i \times s_i} & \mathbf{0} \end{pmatrix}}^{s_i \times m} A \overbrace{\begin{pmatrix} \mathbf{0} \\ I_{p_j \times p_j} \\ \mathbf{0} \end{pmatrix}}^{n \times p_j}. \tag{3.15}$$

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

$$\begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_m \times p_m} \end{pmatrix}$$

where these little identity matrices don't overlap. A similar conclusion follows from consideration of the matrices $I_{s_i \times s_i}$. Note that in 3.15 the matrix on the right is a block column matrix for the above block diagonal matrix and the matrix on the left in 3.15 is a block row matrix taken from a similar block diagonal matrix consisting of the $I_{s_i \times s_i}$.

Next consider the question of multiplication of two block matrices. Let B, A be block matrices of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix}, \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{3.16}$$

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 3.5.1 *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 3.5.2 *Let B be a $q \times p$ block matrix as in 3.16 and let A be a $p \times n$ block matrix as in 3.16 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*

$$\sum_s B_{is}A_{sj}. \tag{3.17}$$

Proof: From 3.15

$$B_{is}A_{sj} = \begin{pmatrix} \mathbf{0} & I_{r_i \times r_i} & \mathbf{0} \end{pmatrix} B \begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j \times q_j} \\ \mathbf{0} \end{pmatrix}$$

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 3.5.1

$$\begin{pmatrix} \mathbf{0} \\ I_{p_s \times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{p_s \times p_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

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 3.5.3 *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, $q_M(t)$ denotes the characteristic polynomial, $\det(tI - M)$. The zeros of this polynomial will be shown later to be eigenvalues of the matrix M. First note that from block multiplication, for the following block matrices consisting of square blocks of an appropriate size,

$$\begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} \text{ so}$$

$$\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det(A) \det(C)$$

Theorem 3.5.4 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then*

$$q_{BA}(t) = t^{n-m} q_{AB}(t),$$

so the eigenvalues of BA and AB are the same including multiplicities except that BA has $n-m$ extra zero eigenvalues. Here $q_A(t)$ denotes the characteristic polynomial of the matrix A.

Proof: Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}$$

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

$$\begin{pmatrix} 0_{m\times m} & 0 \\ B & BA \end{pmatrix}, \begin{pmatrix} AB & 0 \\ B & 0_{n\times n} \end{pmatrix}$$

have the same characteristic polynomials. See Problem 8 on Page 88. Thus

$$\det \begin{pmatrix} tI_{m\times m} & 0 \\ -B & tI - BA \end{pmatrix} = \det \begin{pmatrix} tI - AB & 0 \\ -B & tI_{n\times n} \end{pmatrix} \tag{3.18}$$

Therefore,

$$t^m \det (tI - BA) = t^n \det (tI - AB) \tag{3.19}$$

and so $\det (tI - BA) = q_{BA}(t) = t^{n-m} \det (tI - AB) = t^{n-m} q_{AB}(t)$. ■

3.6 Exercises

1. Let $m < n$ and let A be an $m \times n$ matrix. Show that A is **not** one to one. **Hint:** Consider the $n \times n$ matrix A_1 which is of the form

$$A_1 \equiv \begin{pmatrix} A \\ 0 \end{pmatrix}$$

where the 0 denotes an $(n - m) \times n$ matrix of zeros. Thus $\det A_1 = 0$ and so A_1 is not one to one. Now observe that $A_1\mathbf{x}$ is the vector,

$$A_1\mathbf{x} = \begin{pmatrix} A\mathbf{x} \\ \mathbf{0} \end{pmatrix}$$

which equals zero if and only if $A\mathbf{x} = \mathbf{0}$.

2. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be vectors in $\mathbb{F}^n$ and let $M(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ denote the matrix whose i^{th} column equals $\mathbf{v}_i$. Define

$$d(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \det (M(\mathbf{v}_1, \cdots, \mathbf{v}_n)).$$

Prove that d is linear in each variable, (multilinear), that

$$d(\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_n) = -d(\mathbf{v}_1, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_n), \tag{3.20}$$

and

$$d(\mathbf{e}_1, \cdots, \mathbf{e}_n) = 1 \tag{3.21}$$

where here $\mathbf{e}_j$ is the vector in $\mathbb{F}^n$ which has a zero in every position except the j^{th} position in which it has a one.

3. Suppose $f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \to \mathbb{F}$ satisfies 3.20 and 3.21 and is linear in each variable. Show that $f = d$.

4. Show that if you replace a row (column) of an $n \times n$ matrix A with itself added to some multiple of another row (column) then the new matrix has the same determinant as the original one.

5. Use the result of Problem 4 to evaluate by hand the determinant

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

6. Find the inverse if it exists of the matrix

$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

7. 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 an 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,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ \vdots & \cdots & \vdots \\ y_1^{(n-2)}(x) & & y_n^{(n-2)}(x) \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

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.

8. 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 A is $\det(\lambda I - A)$.

9. Suppose the characteristic polynomial of an $n \times n$ matrix A is of the form
$$t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$$
and that $a_0 \neq 0$. Find a formula A^{-1} in terms of powers of the matrix A. Show that A^{-1} exists if and only if $a_0 \neq 0$. In fact, show that $a_0 = (-1)^n \det(A)$.

10. $\uparrow$Letting $p(t)$ denote the characteristic polynomial of A, show that $p_\varepsilon(t) \equiv p(t-\varepsilon)$ is the characteristic polynomial of $A + \varepsilon I$. Then show that if $\det(A) = 0$, it follows that $\det(A + \varepsilon I) \neq 0$ whenever $|\varepsilon|$ is sufficiently small.

11. In constitutive modeling of the stress and strain tensors, one sometimes considers sums of the form $\sum_{k=0}^{\infty} a_k A^k$ where A is a 3×3 matrix. Show using the Cayley Hamilton theorem that if such a thing makes any sense, you can always obtain it as a finite sum having no more than 3 terms.

12. Recall you can find the determinant from expanding along the j^{th} column.
$$\det(A) = \sum_i A_{ij} (\operatorname{cof}(A))_{ij}$$
Think of $\det(A)$ as a function of the entries, A_{ij}. Explain why the ij^{th} cofactor is really just
$$\frac{\partial \det(A)}{\partial A_{ij}}.$$

13. Let U be an open set in $\mathbb{R}^n$ and let $\mathbf{g}: U \to \mathbb{R}^n$ be such that all the first partial derivatives of all components of $\mathbf{g}$ exist and are continuous. Under these conditions form the matrix $D\mathbf{g}(\mathbf{x})$ given by
$$D\mathbf{g}(\mathbf{x})_{ij} \equiv \frac{\partial g_i(\mathbf{x})}{\partial x_j} \equiv g_{i,j}(\mathbf{x})$$
The best kept secret in calculus courses is that the linear transformation determined by this matrix $D\mathbf{g}(\mathbf{x})$ is called the derivative of $\mathbf{g}$ and is the correct generalization of the concept of derivative of a function of one variable. Suppose the second partial derivatives also exist and are continuous. Then show that $\sum_j (\operatorname{cof}(D\mathbf{g}))_{ij,j} = 0$. **Hint:** First explain why $\sum_i g_{i,k} \operatorname{cof}(D\mathbf{g})_{ij} = \delta_{jk} \det(D\mathbf{g})$. Next differentiate with respect to x_j and sum on j using the equality of mixed partial derivatives. Assume $\det(D\mathbf{g}) \neq 0$ to prove the identity in this special case. Then explain using Problem 10 why there exists a sequence $\varepsilon_k \to 0$ such that for $\mathbf{g}_{\varepsilon_k}(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x}) + \varepsilon_k \mathbf{x}$, $\det(D\mathbf{g}_{\varepsilon_k}) \neq 0$ and so the identity holds for $\mathbf{g}_{\varepsilon_k}$. Then take a limit to get the desired result in general. This is an extremely important identity which has surprising implications. One can build degree theory on it for example. It also leads to simple proofs of the Brouwer fixed point theorem from topology. See Evans [9] for example.

14. A determinant of the form
$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ a_0 & a_1 & \cdots & a_n \\ a_0^2 & a_1^2 & \cdots & a_n^2 \\ \vdots & \vdots & & \vdots \\ a_0^{n-1} & a_1^{n-1} & \cdots & a_n^{n-1} \\ a_0^n & a_1^n & \cdots & a_n^n \end{vmatrix}$$

is called a Vandermonde determinant. Show it equals $\prod_{0\leq i<j\leq n}(a_j - a_i)$. By this is meant to take the product of all terms of the form $(a_j - a_i)$ such that $j > i$. **Hint:** Show it works if $n = 1$ so you are looking at $\begin{vmatrix} 1 & 1 \\ a_0 & a_1 \end{vmatrix}$. Then suppose it holds for $n-1$ and consider the case n. Consider the polynomial in $t, p(t)$ which is obtained from the above by replacing the last column with the column $\begin{pmatrix} 1 & t & \cdots & t^n \end{pmatrix}^T$. Explain why $p(a_j) = 0$ for $i = 0, \cdots, n-1$. Explain why $p(t) = c\prod_{i=0}^{n-1}(t - a_i)$. Of course c is the coefficient of t^n. Find this coefficient from the above description of $p(t)$ and the induction hypothesis. Then plug in $t = a_n$ and observe you have the formula valid for n.

15. The example in this exercise was shown to me by Marc van Leeuwen and it helped to correct a misleading proof of the Cayley Hamilton theorem presented in this chapter. 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. The proof in question read as though it was using this incorrect argument. Let

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Show that for all $\lambda, (\lambda I + E_1)(\lambda I + E_2) = (\lambda^2 + \lambda) I = (\lambda I + E_2)(\lambda I + E_1)$. However, $(NI + E_1)(NI + E_2) \neq (NI + E_2)(NI + E_1)$. Explain why this can happen. In the proof of the Cayley-Hamilton theorem given in the chapter, show that the matrix A does commute with the matrices C_i in that argument. **Hint:** Multiply both sides out with N in place of λ. Does N commute with E_i?

16. Explain how 3.19 follows from 3.18. **Hint:** If you have two real or complex polynomials $p(t), q(t)$ of degree p and they are equal, for all $t \neq 0$, then by continuity, they are equal for all t. Also

$$\begin{pmatrix} tI & 0 \\ 0 & tI - BA \end{pmatrix} = \begin{pmatrix} tI & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & tI - BA \end{pmatrix}$$

thus the determinant of the one on the left equals $t^m \det(tI - BA)$.

17. Explain why the proof of the Cayley-Hamilton theorem given in this chapter cannot possibly hold for arbitrary fields of scalars.

18. Suppose A is $m \times n$ and B is $n \times m$. Letting I be the identity of the appropriate size, is it the case that $\det(I + AB) = \det(I + BA)$? Explain why or why not.

Chapter 4

Row Operations

4.1 Elementary Matrices

The elementary matrices result from doing a row operation to the identity matrix.

Definition 4.1.1 *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 4.1.2 *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. More generally, if $(i_1, i_2, \cdots, i_n)$ is a permutation, a matrix which has a 1 in the i_k position in row k and zero in every other position of that row is called a permutation matrix. Thus each permutation corresponds to a unique permutation matrix.*

As an example of why these elementary matrices are interesting, consider the following.

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b & c & d \\ x & y & z & w \\ f & g & h & i \end{pmatrix} = \begin{pmatrix} x & y & z & w \\ a & b & c & d \\ f & g & h & i \end{pmatrix}$$

A 3×4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. First consider the one which involves switching row i and row j where $i < j$. This matrix is of the form

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 0 & \cdots & 1 & & \\ & & \vdots & & \vdots & & \\ & & 1 & \cdots & 0 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix}$$

The two exceptional rows are shown. The i^{th} row was the j^{th} and the j^{th} row was the i^{th}

in the identity matrix. Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 0 & \cdots & 1 & & \\ & & \vdots & & \vdots & & \\ & & 1 & \cdots & 0 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_j \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix}$$

Now denote by P^{ij} the elementary matrix which comes from the identity from switching rows i and j. From what was just explained consider multiplication on the left by this elementary matrix.

$$P^{ij} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

From the way you multiply matrices this is a matrix which has the indicated columns.

$$\left(P^{ij} \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{n1} \end{pmatrix}, P^{ij} \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, P^{ij} \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \left(\begin{pmatrix} a_{11} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{n1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1p} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

This has established the following lemma.

Lemma 4.1.3 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows. Then*

$$P^{ij}A = B$$

where B is obtained from A by switching the i^{th} and the j^{th} rows.

As a consequence of the above lemma, if you have any permutation $(i_1, \cdots, i_n)$, it follows from Lemma 3.3.2 that the corresponding permutation matrix can be obtained by multiplying finitely many permutation matrices, each of which switch only two rows. Now every such permutation matrix in which only two rows are switched has determinant -1. Therefore, the determinant of the permutation matrix for $(i_1, \cdots, i_n)$ equals $(-1)^p$ where the given permutation can be obtained by making p switches. Now p is not unique. There are many ways to make switches and end up with a given permutation, but what this shows is that the total number of switches is either always odd or always even. That is, you could not obtain a given permutation by making $2m$ switches and $2k+1$ switches. A permutation is said to be even if p is even and odd if p is odd. This is an interesting result in abstract algebra which is obtained very easily from a consideration of elementary matrices and of course the theory of the determinant. Also, this shows that the composition of permutations corresponds to the product of the corresponding permutation matrices.

To see permutations considered more directly in the context of group theory, you should see a good abstract algebra book such as [18] or [14].

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c. The elementary matrix which results from applying this operation to the i^{th} row of the identity matrix is of the form

$$\begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & c & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & c & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ cv_i \\ \vdots \\ v_n \end{pmatrix}$$

Denote by $E(c,i)$ this elementary matrix which multiplies the i^{th} row of the identity by the nonzero constant, c. Then from what was just discussed and the way matrices are multiplied,

$$E(c,i)\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix having the columns indicated below.

$$= \left(E(c,i)\begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{n1} \end{pmatrix}, E(c,i)\begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, E(c,i)\begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & \vdots \\ ca_{i1} & ca_{i2} & \cdots & \cdots & ca_{ip} \\ \vdots & \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{np} \end{pmatrix}$$

This proves the following lemma.

Lemma 4.1.4 *Let $E(c,i)$ denote the elementary matrix corresponding to the row operation in which the i^{th} row is multiplied by the nonzero constant, c. Thus $E(c,i)$ involves multiplying the i^{th} row of the identity matrix by c. Then*

$$E(c,i)A = B$$

where B is obtained from A by multiplying the i^{th} row of A by c.

Finally consider the third of these row operations. Denote by $E(c \times i + j)$ the elementary matrix which replaces the j^{th} row with itself added to c times the i^{th} row added to it. In case $i < j$ this will be of the form

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & \ddots & & & \\ & & c & \cdots & 1 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & \ddots & & & \\ & & c & \cdots & 1 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ cv_i + v_j \\ \vdots \\ v_n \end{pmatrix}$$

Now from this and the way matrices are multiplied,

$$E(c \times i + j) \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix of the following form having the indicated columns.

$$\left(E(c \times i + j) \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n1} \end{pmatrix}, E(c \times i + j) \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots E(c \times i + j) \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{j2} + ca_{i1} & a_{j2} + ca_{i2} & \cdots & a_{jp} + ca_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

The case where $i > j$ is handled similarly. This proves the following lemma.

Lemma 4.1.5 *Let $E(c \times i + j)$ denote the elementary matrix obtained from I by replacing the j^{th} row with c times the i^{th} row added to it. Then*

$$E(c \times i + j) A = B$$

where B is obtained from A by replacing the j^{th} row of A with itself added to c times the i^{th} row of A.

The next theorem is the main result.

Theorem 4.1.6 *To perform any of the three row operations on a matrix A it suffices to do the row operation on the identity matrix obtaining an elementary matrix E and then take the product, EA. Furthermore, each elementary matrix is invertible and its inverse is an elementary matrix.*

Proof: The first part of this theorem has been proved in Lemmas 4.1.3 - 4.1.5. It only remains to verify the claim about the inverses. Consider first the elementary matrices corresponding to row operation of type three.

$$E(-c \times i + j) E(c \times i + j) = I$$

This follows because the first matrix takes c times row i in the identity and adds it to row j. When multiplied on the left by $E(-c \times i + j)$ it follows from the first part of this theorem that you take the i^{th} row of $E(c \times i + j)$ which coincides with the i^{th} row of I since that row was not changed, multiply it by $-c$ and add to the j^{th} row of $E(c \times i + j)$ which was the j^{th} row of I added to c times the i^{th} row of I. Thus $E(-c \times i + j)$ multiplied on the left, undoes the row operation which resulted in $E(c \times i + j)$. The same argument applied to the product

$$E(c \times i + j) E(-c \times i + j)$$

replacing c with $-c$ in the argument yields that this product is also equal to I. Therefore, $E(c \times i + j)^{-1} = E(-c \times i + j)$.

Similar reasoning shows that for $E(c, i)$ the elementary matrix which comes from multiplying the i^{th} row by the nonzero constant, c,

$$E(c, i)^{-1} = E\left(c^{-1}, i\right).$$

Finally, consider P^{ij} which involves switching the i^{th} and the j^{th} rows.

$$P^{ij} P^{ij} = I$$

because by the first part of this theorem, multiplying on the left by P^{ij} switches the i^{th} and j^{th} rows of P^{ij} which was obtained from switching the i^{th} and j^{th} rows of the identity. First you switch them to get P^{ij} and then you multiply on the left by P^{ij} which switches these rows again and restores the identity matrix. Thus $\left(P^{ij}\right)^{-1} = P^{ij}$. ■

4.2 The Rank of a Matrix

Recall the following definition of rank of a matrix.

Definition 4.2.1 *A submatrix of a matrix A is the rectangular array of numbers obtained by deleting some rows and columns of A. Let A be an $m \times n$ matrix. The* ***determinant rank*** *of the matrix equals r where r is the largest number such that some $r \times r$ submatrix of A has a non zero determinant. The* ***row rank*** *is defined to be the dimension of the span of the rows. The* ***column rank*** *is defined to be the dimension of the span of the columns. The rank of A is denoted as* $\text{rank}(A)$.

The following theorem is proved in the section on the theory of the determinant and is restated here for convenience.

Theorem 4.2.2 *Let A be an $m \times n$ matrix. Then the row rank, column rank and determinant rank are all the same.*

So how do you find the rank? It turns out that row operations are the key to the practical computation of the rank of a matrix.

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

Lemma 4.2.3 *Let B and A be two $m \times n$ matrices and suppose B results from a row operation applied to A. Then the k^{th} column of B is a linear combination of the $i_1, \cdots, i_r$ columns of B if and only if the k^{th} column of A is a linear combination of the $i_1, \cdots, i_r$ columns of A. Furthermore, the scalars in the linear combination are the same. (The linear relationship between the k^{th} column of A and the $i_1, \cdots, i_r$ columns of A is the same as the linear relationship between the k^{th} column of B and the $i_1, \cdots, i_r$ columns of B.)*

Proof: Let A equal the following matrix in which the $\mathbf{a}_k$ are the columns

$$\begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}$$

and let B equal the following matrix in which the columns are given by the $\mathbf{b}_k$

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{pmatrix}$$

Then by Theorem 4.1.6 on Page 116 $\mathbf{b}_k = E\mathbf{a}_k$ 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,

$$\mathbf{b}_k = E\mathbf{a}_k = \sum_{r \in S} c_r E\mathbf{a}_r = \sum_{r \in S} c_r \mathbf{b}_r. \blacksquare$$

Corollary 4.2.4 *Let A and B be two $m \times n$ matrices such that B is obtained by applying a row operation to A. Then the two matrices have the same rank.*

Proof: Lemma 4.2.3 says the linear relationships are the same between the columns of A and those of B. Therefore, the column rank of the two matrices is the same. $\blacksquare$

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

Example 4.2.5 *Find the rank of the following matrix and identify columns whose linear combinations yield all the other columns.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 3 & 7 & 8 & 6 & 6 \end{pmatrix} \tag{4.1}$$

Take (-1) times the first row and add to the second and then take (-3) times the first row and add to the third. This yields

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

By the above corollary, this matrix has the same rank as the first matrix. Now take (-1) times the second row and add to the third row and then -2 times the second added to the first yielding

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

At this point it is clear the rank is 2. This is because every column is in the span of the first two and these first two columns are linearly independent.

Example 4.2.6 *Find the rank of the following matrix and identify columns whose linear combinations yield all the other columns.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 6 & 0 & 2 \\ 3 & 6 & 8 & 6 & 6 \end{pmatrix} \tag{4.2}$$

Take (-1) times the first row and add to the second and then take (-3) times the first row and add to the last row. This yields

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

Now multiply the second row by $1/5$ and add 5 times it to the last row.

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

Add (-1) times the second row to the first.

$$\begin{pmatrix} 1 & 2 & 0 & 18/5 & 2 \\ 0 & 0 & 1 & -3/5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.3}$$

It is now clear the rank of this matrix is 2 because the first and third columns form a basis for the column space.

The matrix 4.3 is the row reduced echelon form for the matrix 4.2.

4.3 The Row Reduced Echelon Form

The following definition is for the row reduced echelon form of a matrix.

Definition 4.3.1 *Let* $\mathbf{e}_i$ *denote the column vector which has all zero entries except for the* i^{th} *slot which is one. An* $m \times n$ *matrix is said to be in row reduced echelon form if, in viewing 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,* $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$*.*

For example, here are some matrices which are in row reduced echelon form.

$$\begin{pmatrix} 0 & 1 & 3 & 0 & 3 \\ 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 & -11 & 0 \\ 0 & 1 & 4 & 4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Theorem 4.3.2 *Let A be an $m \times n$ matrix. Then A has a row reduced echelon form determined by a simple process.*

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

Definition 4.3.3 *The first pivot column of A is the first nonzero column of A. The next pivot column is the first column after this which is not a linear combination of the columns to its left. The third pivot column is the next column after this which is not a linear combination of those columns to its left, and so forth. Thus by Lemma 4.2.3 if a pivot column occurs as the j^{th} column from the left, it follows that in the row reduced echelon form there will be one of the $\mathbf{e}_k$ as the j^{th} column.*

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

Definition 4.3.4 *Two matrices are said to be **row equivalent** if one can be obtained from the other by a sequence of row operations.*

Since every row operation can be obtained by multiplication on the left by an elementary matrix and since each of these elementary matrices has an inverse which is also an elementary matrix, it follows that row equivalence is a similarity relation. Thus one can classify matrices according to which similarity class they are in. Later in the book, another more profound way of classifying matrices will be presented.

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 4.2.3 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 4.3.5 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are row equivalent to A, then $B = C$.*

Proof: Suppose B and C are both row reduced echelon forms for the matrix A. Then they clearly have the same zero columns since row operations leave zero columns unchanged. If B has the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occurring for the first time in the positions, $i_1, i_2, \cdots, i_r$, 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 4.2.3, 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 $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur respectively for the first time in columns $i_1, i_2, \cdots, i_r$ for C. Thus B, C have the same columns in these positions. By Lemma 4.2.3, the other columns in the two matrices are linear combinations, involving the **same scalars**, of the columns in the $i_1, \cdots, i_k$ position. Thus each column of B is identical to the corresponding column in C. ■

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.

The following corollary follows.

Corollary 4.3.6 *Let A be an $m \times n$ matrix and let R denote the row reduced echelon form obtained from A by row operations. Then there exists a sequence of elementary matrices, $E_1, \cdots, E_p$ such that*

$$(E_p E_{p-1} \cdots E_1) A = R.$$

Proof: This follows from the fact that row operations are equivalent to multiplication on the left by an elementary matrix. ■

Corollary 4.3.7 *Let A be an invertible $n \times n$ matrix. Then A equals a finite product of elementary matrices.*

Proof: Since A^{-1} is given to exist, it follows A must have rank n because by Theorem 3.3.18 $\det(A) \neq 0$ which says the determinant rank and hence the column rank of A is n and so the row reduced echelon form of A is I because the columns of A form a linearly independent set. Therefore, by Corollary 4.3.6 there is a sequence of elementary matrices, $E_1, \cdots, E_p$ such that

$$(E_p E_{p-1} \cdots E_1) A = I.$$

But now multiply on the left on both sides by E_p^{-1} then by E_{p-1}^{-1} and then by E_{p-2}^{-1} etc. until you get

$$A = E_1^{-1} E_2^{-1} \cdots E_{p-1}^{-1} E_p^{-1}$$

and by Theorem 4.1.6 each of these in this product is an elementary matrix. ■

Corollary 4.3.8 *The rank of a matrix equals the number of nonzero pivot columns. Furthermore, every column is contained in the span of the pivot columns.*

Proof: Write the row reduced echelon form for the matrix. From Corollary 4.2.4 this row reduced matrix has the same rank as the original matrix. Deleting all the zero rows and all the columns in the row reduced echelon form which do not correspond to a pivot column, yields an $r \times r$ identity submatrix in which r is the number of pivot columns. Thus the rank is at least r.

From Lemma 4.2.3 every column of A is a linear combination of the pivot columns since this is true by definition for the row reduced echelon form. Therefore, the rank is no more than r. ■

Here is a fundamental observation related to the above.

Corollary 4.3.9 *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.*

Proof: Since $m < n$, not all the columns of A can be pivot columns. That is, in the row reduced echelon form say $\mathbf{e}_i$ occurs for the first time at r_i where $r_1 < r_2 < \cdots < r_p$ where $p \leq m$. It follows since $m < n$, there exists some column in the row reduced echelon form which is a linear combination of the preceding columns. By Lemma 4.2.3 the same is true of the columns of A. ■

Definition 4.3.10 *Let A be an $m \times n$ matrix having rank, r. Then the nullity of A is defined to be $n - r$. Also define $\ker(A) \equiv \{\mathbf{x} \in \mathbb{F}^n : A\mathbf{x} = \mathbf{0}\}$. This is also denoted as $N(A)$.*

Observation 4.3.11 *Note that $\ker(A)$ is a subspace because if a, b are scalars and $\mathbf{x}, \mathbf{y}$ are vectors in $\ker(A)$, then*

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y} = \mathbf{0} + \mathbf{0} = \mathbf{0}$$

Recall that the dimension of the column space of a matrix equals its rank and since the column space is just $A(\mathbb{F}^n)$, the rank is just the dimension of $A(\mathbb{F}^n)$. The next theorem shows that the nullity equals the dimension of $\ker(A)$.

Theorem 4.3.12 *Let A be an $m \times n$ matrix. Then $\text{rank}(A) + \dim(\ker(A)) = n$.*

Proof: Since $\ker(A)$ is a subspace, there exists a basis for $\ker(A)$, $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$. Also let $\{A\mathbf{y}_1, \cdots, A\mathbf{y}_l\}$ be a basis for $A(\mathbb{F}^n)$. Let $\mathbf{u} \in \mathbb{F}^n$. Then there exist unique scalars c_i such that

$$A\mathbf{u} = \sum_{i=1}^{l} c_i A\mathbf{y}_i$$

It follows that

$$A\left(\mathbf{u} - \sum_{i=1}^{l} c_i \mathbf{y}_i\right) = \mathbf{0}$$

and so the vector in parenthesis is in $\ker(A)$. Thus there exist unique b_j such that

$$\mathbf{u} = \sum_{i=1}^{l} c_i \mathbf{y}_i + \sum_{j=1}^{k} b_j \mathbf{x}_j$$

Since $\mathbf{u}$ was arbitrary, this shows $\{\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{y}_1, \cdots, \mathbf{y}_l\}$ spans $\mathbb{F}^n$. If these vectors are independent, then they will form a basis and the claimed equation will be obtained. Suppose then that

$$\sum_{i=1}^{l} c_i \mathbf{y}_i + \sum_{j=1}^{k} b_j \mathbf{x}_j = \mathbf{0}$$

Apply A to both sides. This yields

$$\sum_{i=1}^{l} c_i A\mathbf{y}_i = \mathbf{0}$$

and so each $c_i = 0$. Then the independence of the $\mathbf{x}_j$ imply each $b_j = 0$. ■

4.4 Existence of Solutions to Linear Systems

Consider the linear system of equations,

$$A\mathbf{x} = \mathbf{b} \tag{4.4}$$

where A is an $m \times n$ matrix, $\mathbf{x}$ is a $n \times 1$ column vector, and $\mathbf{b}$ is an $m \times 1$ column vector. Suppose

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$$

where the $\mathbf{a}_k$ denote the columns of A. Then $\mathbf{x} = (x_1, \cdots, x_n)^T$ is a solution of the system 4.4, 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 $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$. This shows that there exists a solution to the system, 4.4 if and only if $\mathbf{b}$ is contained in $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$. In words, there is a solution to 4.4 if and only if $\mathbf{b}$ is in the column space of A. In terms of rank, the following proposition describes the situation.

Proposition 4.4.1 *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 4.4 if and only if*

$$\text{rank}\begin{pmatrix} A & | & \mathbf{b} \end{pmatrix} = \text{rank}(A). \tag{4.5}$$

Proof: Place $\begin{pmatrix} A & | & \mathbf{b} \end{pmatrix}$ and A in row reduced echelon form, respectively B and C. If the above condition on rank is true, then both B and C have the same number of nonzero rows. In particular, you cannot have a row of the form

$$\begin{pmatrix} 0 & \cdots & 0 & \bigstar \end{pmatrix}$$

where $\bigstar \neq 0$ in B. Therefore, there will exist a solution to the system 4.4.

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 4.5 is the same. ■

4.5 Fredholm Alternative

There is a very useful version of Proposition 4.4.1 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 4.5.1 *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

$$\ker\left(A^T\right) \equiv \left\{\mathbf{z} : A^T\mathbf{z} = \mathbf{0}\right\} = \left\{\mathbf{z} : \sum_{k=1}^{m} z_k\mathbf{a}_k = 0\right\}$$

Lemma 4.5.2 *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 inner 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 4.5.3 *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

$$\mathbf{x}^TA=\mathbf{0}, \text{then } \mathbf{x}^T\mathbf{b}=0.$$

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 $\text{rank}\left(\begin{array}{c|c} A & \mathbf{b}\end{array}\right)=\text{rank}\,(A)$. By Proposition 4.4.1, 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 4.5.2,

$$\mathbf{b}\cdot\mathbf{z}=A\mathbf{x}\cdot\mathbf{z}=\mathbf{x}\cdot A^T\mathbf{z}=\mathbf{x}\cdot\mathbf{0}=0 \blacksquare$$

This implies the following corollary which is also called the Fredholm alternative. The "alternative" becomes more clear in this corollary.

Corollary 4.5.4 *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 4.5.3. In other words, A maps $\mathbb{R}^n$ onto $\mathbb{R}^m$.

Conversely if A is onto, then by Theorem 4.5.3 every $\mathbf{b}\in\mathbb{R}^m$ is in $\ker\left(A^T\right)^\perp$ and so if $A^T\mathbf{x}=\mathbf{0}$, then $\mathbf{b}\cdot\mathbf{x}=0$ for every $\mathbf{b}$. In particular, this holds for $\mathbf{b}=\mathbf{x}$. Hence if $A^T\mathbf{x}=\mathbf{0}$, then $\mathbf{x}=\mathbf{0}$. $\blacksquare$

Here is an amusing example.

Example 4.5.5 *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 $A^T\mathbf{x}=\mathbf{0}$.

4.6 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 4.3.7.

4. Consider a permutation of $\{1, 2, \cdots, n\}$. This is an ordered list of numbers taken from this list with no repeats, $\{i_1, i_2, \cdots, i_n\}$. Define the permutation matrix
$$P(i_1, i_2, \cdots, i_n)$$
as the matrix which is obtained from the identity matrix by placing the j^{th} column of I as the i_j^{th} column of $P(i_1, i_2, \cdots, i_n)$. Thus the 1 in the i_j^{th} column of this permutation matrix occurs in the j^{th} slot. What does this permutation matrix do to the column vector $(1, 2, \cdots, n)^T$?

5. ↑Consider the 3×3 permutation matrices. List all of them and then determine the dimension of their span. Recall that you can consider an $m \times n$ matrix as something in $\mathbb{F}^{nm}$.

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

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

8. Find the rank and nullity of the following matrices. If the rank is r, identify r columns **in the original matrix** which have the property that every other column may be written as a linear combination of these.

   (a) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 3 & 2 & 12 & 1 & 6 & 8 \\ 0 & 1 & 1 & 5 & 0 & 2 & 3 \\ 0 & 2 & 1 & 7 & 0 & 3 & 4 \end{pmatrix}$

   (b) $\begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 6 & 0 & 5 & 4 \\ 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 2 & 1 & 4 & 0 & 3 & 2 \end{pmatrix}$

   (c) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$

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

10. Suppose A is an $m \times n$ matrix. Explain why the rank of A is always no larger than $\min(m, n)$.

11. 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.

12. Suppose A is an $m \times n$ matrix and that $m > n$. Show there exists $\mathbf{b} \in \mathbb{F}^m$ such that there is no solution to the equation
$$A\mathbf{x} = \mathbf{b}.$$

13. 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, $\mathbf{x} \neq \mathbf{0}$ 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.

14. Explain why an $n \times n$ matrix A is both one to one and onto if and only if its rank is n.

15. Suppose A is an $m \times n$ matrix and $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$ is a linearly independent set of vectors in $A(\mathbb{F}^n) \subseteq \mathbb{F}^m$. Suppose also that $A\mathbf{z}_i = \mathbf{w}_i$. Show that $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is also linearly independent.

16. Show $\operatorname{rank}(A + B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)$.

17. Suppose A is an $m \times n$ matrix, $m \geq n$ and the columns of A are independent. Suppose also that $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is a linearly independent set of vectors in $\mathbb{F}^n$. Show that $\{A\mathbf{z}_1, \cdots, A\mathbf{z}_k\}$ is linearly independent.

18. Suppose that 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
$$\ker(AB) \subseteq \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{z}_1, \cdots, \mathbf{z}_k).$$

19. Let $m < n$ and let A be an $m \times n$ matrix. Show that A is **not** one to one.

20. Let A be an $m \times n$ real matrix and let $\mathbf{b} \in \mathbb{R}^m$. Show there exists a solution, $\mathbf{x}$ to the system
$$A^T A\mathbf{x} = A^T \mathbf{b}$$
Next show that if $\mathbf{x}, \mathbf{x}_1$ are two solutions, then $A\mathbf{x} = A\mathbf{x}_1$. **Hint:** First show that $\left(A^T A\right)^T = A^T A$. Next show if $\mathbf{x} \in \ker\left(A^T A\right)$, then $A\mathbf{x} = \mathbf{0}$. Finally apply the Fredholm alternative. Show $A^T\mathbf{b} \in \ker(A^T A)^\perp$. This will give existence of a solution.

21. Show that in the context of Problem 20 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 is a solution to the least squares problem.

22. ↑Here is a point in $\mathbb{R}^4 : (1, 2, 3, 4)^T$. Find the point in span $\left(\begin{pmatrix} 1 \\ 0 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 3 \\ 2 \end{pmatrix}\right)$ which is closest to the given point.

23. ↑Here is a point in $\mathbb{R}^4 : (1, 2, 3, 4)^T$. Find the point on the plane described by $x + 2y - 4z + 4w = 0$ which is closest to the given point.

24. Suppose A, B are two invertible $n \times n$ matrices. Show there exists a sequence of row operations which when done to A yield B. **Hint:** Recall that every invertible matrix is a product of elementary matrices.

25. If A is invertible and $n \times n$ and B is $n \times p$, show that AB has the same null space as B and also the same rank as B.

26. Here are two matrices in row reduced echelon form
$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$
Does there exist a sequence of row operations which when done to A will yield B? Explain.

27. Is it true that an upper triagular matrix has rank equal to the number of nonzero entries down the main diagonal?

28. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}\}$ be vectors in $\mathbb{F}^n$. Describe a systematic way to obtain a vector $\mathbf{v}_n$ which is perpendicular to each of these vectors. **Hint:** You might consider something like this
$$\det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \\ v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & & \vdots \\ v_{(n-1)1} & v_{(n-1)2} & \cdots & v_{(n-1)n} \end{pmatrix}$$
where v_{ij} is the j^{th} entry of the vector $\mathbf{v}_i$. This is a lot like the cross product.

29. Let A be an $m \times n$ matrix. Then $\ker(A)$ is a subspace of $\mathbb{F}^n$. Is it true that every subspace of $\mathbb{F}^n$ is the kernel or null space of some matrix? Prove or disprove.

30. Let A be an $n \times n$ matrix and let P^{ij} be the permutation matrix which switches the i^{th} and j^{th} rows of the identity. Show that $P^{ij}AP^{ij}$ produces a matrix which is similar to A which switches the i^{th} and j^{th} entries on the main diagonal.

31. Recall the procedure for finding the inverse of a matrix on Page 49. It was shown that the procedure, when it works, finds the inverse of the matrix. Show that whenever the matrix has an inverse, the procedure works.

32. If $EA = B$ where E is invertible, show that A and B have the same linear relationships among their columns.

33. You could define column operations by analogy to row operations. That is, you switch two columns, multiply a column by a nonzero scalar, or add a scalar multiple of a column to another column. Let E be one of these column operations applied to the identity matrix. Show that AE produces the column operation on A which was used to define E.

Chapter 5

Some Factorizations

5.1 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. The L goes with "lower" and the U with "upper". It turns out many matrices can be written in this way and when this is possible, people get excited about slick ways of solving the system of equations, $A\mathbf{x} = \mathbf{y}$. The method lacks generality but is of interest just the same.

Example 5.1.1 *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 I mean above by saying the method lacks generality.

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 $i < j$.

5.2 Finding an LU Factorization

There is 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 and is called the multiplier method. It is due to Dolittle.

Example 5.2.1 *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 method for updating the matrix on the left. The 2 in the second entry of the first column is there because -2 times the first row of A added to the second row of A produced a 0. Now replace the third row in the matrix on the right by -1 times the first row added to the third. Thus the next step is

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

Finally, add the second row to the bottom row and make the following changes

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 0 & 0 & -11 \end{pmatrix}.$$

At this point, stop because the matrix on the right is upper triangular. An LU factorization is the above.

The justification for this gimmick will be given later.

Example 5.2.2 *Find an LU factorization for* $A = \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 3 & 1 & 3 & 2 \\ 1 & 0 & 1 & 1 & 2 \end{pmatrix}$.

This time everything is done at once for a whole column. This saves trouble. First multiply the first row by (-1) and then add to the last row. Next take (-2) times the first and add to the second and then (-2) times the first and add to the third.

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

This finishes the first column of L and the first column of U. Now take $-(1/4)$ times the second row in the matrix on the right and add to the third followed by $-(1/2)$ times the second added to the last.

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

This finishes the second column of L as well as the second column of U. Since the matrix on the right is upper triangular, stop. The LU 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.

5.3 Solving Linear Systems Using an LU Factorization

The reason people care about the LU factorization is it allows the quick solution of systems of equations. Here is an example.

Example 5.3.1 *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 an LU factorization and it turns out that an 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 $\mathbf{y} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$. Now you can find $\mathbf{x}$ by solving $U\mathbf{x} = \mathbf{y}$. Thus in this case,

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$$

which yields

$$\mathbf{x} = \begin{pmatrix} -\frac{3}{5} + \frac{7}{5}t \\ \frac{9}{5} - \frac{11}{5}t \\ t \\ -1 \end{pmatrix}, t \in \mathbb{R}.$$

Work this out by hand and you will see the advantage of working only with triangular matrices.

It may seem like a trivial thing but it is used because it cuts down on the number of operations involved in finding a solution to a system of equations enough that it makes a difference for large systems.

5.4 The PLU Factorization

As indicated above, some matrices don't have an LU factorization. Here is an example.

$$M = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} \tag{5.1}$$

In this case, there is another factorization which is useful called a PLU factorization. Here P is a permutation matrix.

Example 5.4.1 *Find a PLU factorization for the above matrix in 5.1.*

Proceed as before trying to find the row echelon form of the matrix. First add -1 times the first row to the second row and then add -4 times the first to the third. This yields

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & 0 & 0 & -2 \\ 0 & -5 & -11 & -7 \end{pmatrix}$$

There is no way to do only row operations involving replacing a row with itself added to a multiple of another row to the second matrix in such a way as to obtain an upper triangular matrix. Therefore, consider M with the bottom two rows switched.

$$M' = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix}.$$

Now try again with this matrix. First take -1 times the first row and add to the bottom row and then take -4 times the first row and add to the second row. This yields

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

The second matrix is upper triangular and so the LU factorization of the matrix M' is

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

Thus $M' = PM = LU$ where L and U are given above. Therefore, $M = P^2 M = PLU$ 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 5.4.2 *Use a 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,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Then multiplying both sides by P gives

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

and so

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Now $U\mathbf{x} = \mathbf{y}$ and so it only remains to solve

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

which yields

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} + \frac{7}{5}t \\ \frac{9}{10} - \frac{11}{5}t \\ t \\ -\frac{1}{2} \end{pmatrix} : t \in \mathbb{R}.$$

5.5 Justification for the Multiplier Method

Why does the multiplier method work for finding an 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. Furthermore, in the multiplier method for finding the LU factorization, we zero out the elements below the pivot entry in first column and then the next and so on when scanning from the left. In terms of elementary matrices, this means the row operations used to reduce A to upper triangular form correspond to multiplication on the left by lower triangular matrices having all ones down the main diagonal and the sequence of elementary matrices which row reduces A has the property that in scanning the list of elementary matrices from the right to the left, this list consists of several matrices which involve only changes from the identity in the first column, then several which involve only changes from the identity in the second column and so forth. More precisely, $E_p \cdots E_1 A = U$ where U is upper triangular, E_k having all zeros below the main diagonal except for a single column. Therefore, $A = \overbrace{E_1^{-1} \cdots E_{p-1}^{-1} E_p^{-1}}^{\text{Will be } L} U$. You multiply the inverses in the reverse order. Now each of the E_i^{-1} is also lower triangular with 1 down the main diagonal. Therefore their product has this property. Recall also that if E_i equals the identity matrix except for having an a in a single column somewhere below the main diagonal, E_i^{-1} is obtained by replacing the a in E_i with $-a$, thus explaining why we replace with -1 times the multiplier in computing L. In the case where A is a $3 \times m$ matrix, $E_1^{-1} \cdots E_{p-1}^{-1} E_p^{-1}$ is of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}.$$

Note that scanning from left to right, the first two in the product involve changes in the identity only in the first column while in the third matrix, the change is only in the second. If the entries in the first column had been zeroed out in a different order, the following would have resulted.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}$$

However, it is important to be working from the left to the right, one column at a time.

A similar observation holds in any dimension. Multiplying the elementary matrices which involve a change only in the j^{th} column you obtain A equal to an upper triangular, $n \times m$ matrix U which is multiplied by a sequence of lower triangular matrices on its left which is of the following form, in which the a_{ij} are negatives of multipliers used in row reducing to an upper triangular matrix.

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{11} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ a_{1,n-1} & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & a_{2,n-2} & \cdots & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & a_{n,n-1} & 1 \end{pmatrix}$$

From the matrix multiplication, this product equals

$$\begin{pmatrix} 1 & & & \\ a_{11} & 1 & & \\ \vdots & & \ddots & \\ a_{1,n-1} & \cdots & a_{n,n-1} & 1 \end{pmatrix}$$

Notice how the end result of the matrix multiplication made no change in the a_{ij}. It just filled in the empty spaces with the a_{ij} which occurred in one of the matrices in the product. This is why, in computing L, it is sufficient to begin with the left column and work column by column toward the right, replacing entries with the negative of the multiplier used in the row operation which produces a zero in that entry.

5.6 Existence for the PLU Factorization

Here I will consider an invertible $n \times n$ matrix and show that such a matrix always has a PLU factorization. More general matrices could also be considered but this is all I will present.

Let A be such an invertible matrix and consider the first column of A. If $A_{11} \neq 0$, use this to zero out everything below it. The entry A_{11} is called the pivot. Thus in this case there is a lower triangular matrix L_1 which has all ones on the diagonal such that

$$L_1 P_1 A = \begin{pmatrix} * & * \\ \mathbf{0} & A_1 \end{pmatrix} \tag{5.2}$$

Here $P_1 = I$. In case $A_{11} = 0$, let r be such that $A_{r1} \neq 0$ and r is the first entry for which this happens. In this case, let P_1 be the permutation matrix which switches the first row and the r^{th} row. Then as before, there exists a lower triangular matrix L_1 which has all ones on the diagonal such that 5.2 holds in this case also. In the first column, this L_1 has zeros between the first row and the r^{th} row.

Go to A_1. Following the same procedure as above, there exists a lower triangular matrix and permutation matrix L_2', P_2' such that

$$L_2' P_2' A_1 = \begin{pmatrix} * & * \\ \mathbf{0} & A_2 \end{pmatrix}$$

Let

$$L_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & L_2' \end{pmatrix}, \ P_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & P_2' \end{pmatrix}$$

Then using block multiplication, Theorem 3.5.2,

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & L_2' \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & P_2' \end{pmatrix} \begin{pmatrix} * & * \\ \mathbf{0} & A_1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & L_2' \end{pmatrix} \begin{pmatrix} * & * \\ \mathbf{0} & P_2' A_1 \end{pmatrix} = \begin{pmatrix} * & * \\ \mathbf{0} & L_2' P_2' A_1 \end{pmatrix}$$

$$\begin{pmatrix} * & \cdots & * \\ 0 & * & * \\ \mathbf{0} & \mathbf{0} & A_2 \end{pmatrix} = L_2 P_2 L_1 P_1 A$$

and L_2 has all the subdiagonal entries equal to 0 except possibly some nonzero entries in the second column starting with position r_2 where P_2 switches rows r_2 and 2. Continuing this way, it follows there are lower triangular matrices L_j having all ones down the diagonal and permutation matrices P_i which switch only two rows such that

$$L_{n-1} P_{n-1} L_{n-2} P_{n-2} L_{n-3} \cdots L_2 P_2 L_1 P_1 A = U \tag{5.3}$$

where U is upper triangular. The matrix L_j has all zeros below the main diagonal except for the j^{th} column and even in this column it has zeros between position j and r_j where P_j switches rows j and r_j. Of course in the case where no switching is necessary, you could get all nonzero entries below the main diagonal in the j^{th} column for L_j.

The fact that L_j is the identity except for the j^{th} column means that each P_k for $k > j$ almost commutes with L_j. Say P_k switches the k^{th} and the q^{th} rows for $q \geq k > j$. When you place P_k on the right of L_j it just switches the k^{th} and the q^{th} columns and leaves the j^{th} column unchanged. Therefore, the same result as placing P_k on the left of L_j can be obtained by placing P_k on the right of L_j and modifying L_j by switching the k^{th} and the q^{th} entries in the j^{th} column. (Note this could possibly interchange a 0 for something nonzero.) It follows from 5.3 there exists P, the product of permutation matrices, $P = P_{n-1} \cdots P_1$

each of which switches two rows, and L a lower triangular matrix having all ones on the main diagonal, $L = L'_{n-1} \cdots L'_2 L'_1$, where the L'_j are obtained as just described by moving a succession of P_k from the left to the right of L_j and modifying the j^{th} column as indicated, such that $LPA = U$. Then $A = P^T L^{-1} U$.

It is customary to write this more simply as

$$A = PLU$$

where L is an upper triangular matrix having all ones on the diagonal and P is a permutation matrix consisting of $P_1 \cdots P_{n-1}$ as described above. This proves the following theorem.

Theorem 5.6.1 *Let A be any invertible $n \times n$ matrix. Then there exists a permutation matrix P and a lower triangular matrix L having all ones on the main diagonal and an upper triangular matrix U such that $A = PLU$.*

5.7 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 5.7.1 *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 5.7.2 *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.*

Consider the following problem.

Problem 5.7.3 *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.

$$\begin{aligned}Q(\mathbf{x}-\mathbf{y}) &= (\mathbf{x}-\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \\ &= (\mathbf{x}-\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}|\mathbf{x}-\mathbf{y}|^2=\mathbf{y}-\mathbf{x}\end{aligned}$$

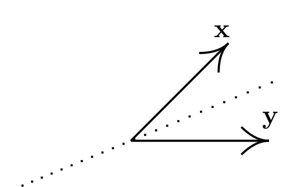
$$\begin{aligned}Q(\mathbf{x}+\mathbf{y}) &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}+\mathbf{y}) \\ &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}((\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}+\mathbf{y})) \\ &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}\left(|\mathbf{x}|^2-|\mathbf{y}|^2\right)=\mathbf{x}+\mathbf{y}\end{aligned}$$

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* $2Q\mathbf{y}=2\mathbf{x}$.

A picture of the geometric significance follows.



The orthogonal matrix Q reflects across the dotted line taking $\mathbf{x}$ to $\mathbf{y}$ and $\mathbf{y}$ to $\mathbf{x}$.

Definition 5.7.4 *Let* A *be an* $m\times n$ *matrix. Then a QR factorization of* A *consists of two matrices,* Q *orthogonal and* R *upper triangular (right triangular) having all the entries on the main diagonal nonnegative 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

$$Q_1 A = \begin{pmatrix} |\mathbf{a}_1| & * \\ \mathbf{0} & A_2 \end{pmatrix}$$

where A_2 is a $m-1 \times n-1$ matrix. Now find in the same way as was just done a $m-1 \times m-1$ matrix $\widehat{Q}_2$ such that

$$\widehat{Q}_2 A_2 = \begin{pmatrix} * & * \\ \mathbf{0} & A_3 \end{pmatrix}$$

Let

$$Q_2 \equiv \begin{pmatrix} 1 & 0 \\ \mathbf{0} & \widehat{Q}_2 \end{pmatrix}.$$

Then

$$Q_2 Q_1 A = \begin{pmatrix} 1 & 0 \\ \mathbf{0} & \widehat{Q}_2 \end{pmatrix} \begin{pmatrix} |\mathbf{a}_1| & * \\ \mathbf{0} & A_2 \end{pmatrix}$$

$$= \begin{pmatrix} |\mathbf{a}_1| & * & * \\ \vdots & * & * \\ 0 & \mathbf{0} & A_3 \end{pmatrix}$$

Continuing this way until the result is upper triangular, you get a sequence of orthogonal matrices $Q_p Q_{p-1} \cdots Q_1$ such that

$$Q_p Q_{p-1} \cdots Q_1 A = R \tag{5.4}$$

where R is upper triangular.

Now if Q_1 and Q_2 are orthogonal, then from properties of matrix multiplication,

$$Q_1 Q_2 (Q_1 Q_2)^T = Q_1 Q_2 Q_2^T Q_1^T = Q_1 I Q_1^T = I$$

and similarly

$$(Q_1 Q_2)^T Q_1 Q_2 = I.$$

Thus the product of orthogonal matrices is orthogonal. Also the transpose of an orthogonal matrix is orthogonal directly from the definition. Therefore, from 5.4

$$A = (Q_p Q_{p-1} \cdots Q_1)^T R \equiv QR.$$

This proves the following theorem.

Theorem 5.7.5 *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 on the main diagonal such that*

$$A = QR$$

and this factorization can be accomplished in a systematic manner.

►►

5.8 Exercises

1. Find a LU factorization of $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix}$.

2. Find a LU factorization of $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 1 \\ 5 & 0 & 1 & 3 \end{pmatrix}$.

3. Find a PLU factorization of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix}$.

4. Find a PLU factorization of $\begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 1 \\ 1 & 2 & 1 & 3 & 2 \end{pmatrix}$.

5. Find a PLU factorization of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 4 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

6. Is there only one LU factorization for a given matrix? **Hint:** Consider the equation
$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

7. Here is a matrix and an LU factorization of it.
$$A = \begin{pmatrix} 1 & 2 & 5 & 0 \\ 1 & 1 & 4 & 9 \\ 0 & 1 & 2 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 5 & 0 \\ 0 & -1 & -1 & 9 \\ 0 & 0 & 1 & 14 \end{pmatrix}$$
Use this factorization to solve the system of equations
$$A\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

8. Find a QR factorization for the matrix
$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & -2 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

9. Find a QR factorization for the matrix
$$\begin{pmatrix} 1 & 2 & 1 & 0 \\ 3 & 0 & 1 & 1 \\ 1 & 0 & 2 & 1 \end{pmatrix}$$

10. If you had a QR factorization, $A = QR$, describe how you could use it to solve the equation $A\mathbf{x} = \mathbf{b}$.

11. If Q is an orthogonal matrix, show the columns are an orthonormal set. That is show that for
$$Q = \begin{pmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{pmatrix}$$
it follows that $\mathbf{q}_i \cdot \mathbf{q}_j = \delta_{ij}$. Also show that any orthonormal set of vectors is linearly independent.

12. Show you can't expect uniqueness for QR factorizations. Consider
$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$
and verify this equals
$$\begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2}\sqrt{2} & 0 & \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & 0 & -\frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 0 & 0 & \sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
and also
$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$
Using Definition 5.7.4, can it be concluded that if A is an invertible matrix it will follow there is only one QR factorization?

13. Suppose $\{\mathbf{a}_1, \cdots, \mathbf{a}_n\}$ are linearly independent vectors in $\mathbb{R}^n$ and let
$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$$
Form a QR factorization for A.
$$\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}$$

Show that for each $k \leq n$,

$$\operatorname{span}(\mathbf{a}_1, \cdots, \mathbf{a}_k) = \operatorname{span}(\mathbf{q}_1, \cdots, \mathbf{q}_k)$$

Prove that every subspace of $\mathbb{R}^n$ has an orthonormal basis. The procedure just described is similar to the Gram Schmidt procedure which will be presented later.

14. Suppose $Q_n R_n$ converges to an orthogonal matrix Q where Q_n is orthogonal and R_n is upper triangular having all positive entries on the diagonal. Show that then Q_n converges to Q and R_n converges to the identity.

Chapter 6

Spectral Theory

Spectral Theory refers to the study of eigenvalues and eigenvectors of a matrix. It is of fundamental importance in many areas. Row operations will no longer be such a useful tool in this subject.

6.1 Eigenvalues and Eigenvectors of a Matrix

The field of scalars in spectral theory is best taken to equal $\mathbb{C}$ although I will sometimes refer to it as $\mathbb{F}$ when it could be either $\mathbb{C}$ or $\mathbb{R}$.

Definition 6.1.1 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a nonzero vector for which*

$$M\mathbf{x} = \lambda \mathbf{x} \tag{6.1}$$

for some scalar, λ. Then $\mathbf{x}$ is called an eigenvector and λ is called an eigenvalue (characteristic value) of the matrix M.

$$\boxed{\textbf{\textit{Eigenvectors are never equal to zero!}}}$$

The set of all eigenvalues of an $n \times n$ matrix M, is denoted by $\sigma(M)$ and is referred to as the spectrum of M.

Eigenvectors are vectors which are shrunk, stretched or reflected upon multiplication by a matrix. How can they be identified? Suppose $\mathbf{x}$ satisfies 6.1. Then

$$(\lambda I - M)\,\mathbf{x} = \mathbf{0}$$

for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the matrix $M - \lambda I$ cannot have an inverse and so by Theorem 3.3.18

$$\det(\lambda I - M) = 0. \tag{6.2}$$

In other words, λ must be a zero of the characteristic polynomial. Since M is an $n \times n$ matrix, it follows from the theorem on expanding a matrix by its cofactor that this is a polynomial equation of degree n. As such, it has a solution, $\lambda \in \mathbb{C}$. Is it actually an eigenvalue? The answer is yes and this follows from Theorem 3.3.26 on Page 100. Since $\det(\lambda I - M) = 0$ the matrix $\lambda I - M$ cannot be one to one and so there exists a nonzero vector, $\mathbf{x}$ such that $(\lambda I - M)\,\mathbf{x} = \mathbf{0}$. This proves the following corollary.

Corollary 6.1.2 *Let M be an $n \times n$ matrix and* $\det(M - \lambda I) = 0$. *Then there exists $\mathbf{x} \in \mathbb{C}^n$ such that $(M - \lambda I)\,\mathbf{x} = \mathbf{0}$.*

As an example, consider the following.

Example 6.1.3 *Find the eigenvalues and eigenvectors for the matrix*

$$A = \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}.$$

You first need to identify the eigenvalues. Recall this requires the solution of the equation

$$\det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\right) = 0$$

When you expand this determinant, you find the equation is

$$(\lambda - 5)\left(\lambda^2 - 20\lambda + 100\right) = 0$$

and so the eigenvalues are

$$5, 10, 10.$$

I have listed 10 twice because it is a zero of multiplicity two due to

$$\lambda^2 - 20\lambda + 100 = (\lambda - 10)^2.$$

Having found the eigenvalues, it only remains to find the eigenvectors. First find the eigenvectors for $\lambda = 5$. As explained above, this requires you to solve the equation,

$$\left(5\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

That is you need to find the solution to

$$\begin{pmatrix} 0 & 10 & 5 \\ -2 & -9 & -2 \\ 4 & 8 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

By now this is an old problem. You set up the augmented matrix and row reduce to get the solution. Thus the matrix you must row reduce is

$$\begin{pmatrix} 0 & 10 & 5 & 0 \\ -2 & -9 & -2 & 0 \\ 4 & 8 & -1 & 0 \end{pmatrix}. \tag{6.3}$$

The reduced row echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{5}{4} & 0 \\ 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so the solution is any vector of the form

$$\begin{pmatrix} \frac{5}{4}z \\ \frac{-1}{2}z \\ z \end{pmatrix} = z\begin{pmatrix} \frac{5}{4} \\ \frac{-1}{2} \\ 1 \end{pmatrix}$$

where $z \in \mathbb{F}$. You would obtain the same collection of vectors if you replaced z with $4z$. Thus a simpler description for the solutions to this system of equations whose augmented matrix is in 6.3 is

$$z\begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} \tag{6.4}$$

where $z \in \mathbb{F}$. Now you need to remember that you can't take $z = 0$ because this would result in the zero vector and

$$\boxed{\textbf{\underline{Eigenvectors} \underline{are} \underline{never} \underline{equal} \underline{to} \underline{zero}!}}$$

Other than this value, every other choice of z in 6.4 results in an eigenvector. It is a good idea to check your work! To do so, I will take the original matrix and multiply by this vector and see if I get 5 times this vector.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 25 \\ -10 \\ 20 \end{pmatrix} = 5 \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix}$$

so it appears this is correct. Always check your work on these problems if you care about getting the answer right.

The variable, z is called a free variable or sometimes a parameter. The set of vectors in 6.4 is called the eigenspace and it equals $\ker(\lambda I - A)$. You should observe that in this case the eigenspace has dimension 1 because there is one vector which spans the eigenspace. In general, you obtain the solution from the row echelon form and the number of different free variables gives you the dimension of the eigenspace. Just remember that not every vector in the eigenspace is an eigenvector. The vector, $\mathbf{0}$ is not an eigenvector although it is in the eigenspace because

$$\boxed{\textbf{\underline{Eigenvectors are never equal to zero!}}}$$

Next consider the eigenvectors for $\lambda = 10$. These vectors are solutions to the equation,

$$\left(10 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

That is you must find the solutions to

$$\begin{pmatrix} 5 & 10 & 5 \\ -2 & -4 & -2 \\ 4 & 8 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

which reduces to consideration of the augmented matrix

$$\begin{pmatrix} 5 & 10 & 5 & 0 \\ -2 & -4 & -2 & 0 \\ 4 & 8 & 4 & 0 \end{pmatrix}$$

The row reduced echelon form for this matrix is

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

and so the eigenvectors are of the form

$$\begin{pmatrix} -2y - z \\ y \\ z \end{pmatrix} = y \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

You can't pick z and y both equal to zero because this would result in the zero vector and

$$\boxed{\textbf{\underline{Eigenvectors are never equal to zero!}}}$$

However, every other choice of z and y does result in an eigenvector for the eigenvalue $\lambda = 10$. As in the case for $\lambda = 5$ you should check your work if you care about getting it right.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -10 \\ 0 \\ 10 \end{pmatrix} = 10 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

so it worked. The other vector will also work. Check it.

The above example shows how to find eigenvectors and eigenvalues algebraically. You may have noticed it is a bit long. Sometimes students try to first row reduce the matrix before looking for eigenvalues. This is a terrible idea because row operations destroy the value of the eigenvalues. The eigenvalue problem is really not about row operations. A general rule to remember about the eigenvalue problem is this.

If it is not long and hard it is usually wrong!

The eigenvalue problem is the hardest problem in algebra and people still do research on ways to find eigenvalues. Now if you are so fortunate as to find the eigenvalues as in the above example, then finding the eigenvectors does reduce to row operations and this part of the problem is easy. However, finding the eigenvalues is anything but easy because for an $n \times n$ matrix, it involves solving a polynomial equation of degree n and none of us are very good at doing this. If you only find a good approximation to the eigenvalue, it won't work. It either is or is not an eigenvalue and if it is not, the only solution to the equation, $(\lambda I - M)\,\mathbf{x} = \mathbf{0}$ will be the zero solution as explained above and

Eigenvectors are never equal to zero!

Here is another example.

Example 6.1.4 *Let*

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

First find the eigenvalues.

$$\det\left(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & 2 & -2 \\ 1 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}\right) = 0$$

This is $\lambda^3 - 6\lambda^2 + 8\lambda = 0$ and the solutions are 0, 2, and 4.

0 **Can be an Eigenvalue!**

Now find the eigenvectors. For $\lambda = 0$ the augmented matrix for finding the solutions is

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

and the row reduced echelon form is

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

Therefore, the eigenvectors are of the form

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

where $z \neq 0$.

Next find the eigenvectors for $\lambda = 2$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

and the row reduced echelon form is

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

and so the eigenvectors are of the form

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

where $z \neq 0$.

Finally find the eigenvectors for $\lambda = 4$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

and the row reduced echelon form is

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

Therefore, the eigenvectors are of the form

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

where $y \neq 0$.

Example 6.1.5 *Let*

$$A = \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}.$$

Find the eigenvectors and eigenvalues.

In this case the eigenvalues are $3, 6, 6$ where I have listed 6 twice because it is a zero of algebraic multiplicity two, the characteristic equation being

$$(\lambda - 3)(\lambda - 6)^2 = 0.$$

It remains to find the eigenvectors for these eigenvalues. First consider the eigenvectors for $\lambda = 3$. You must solve

$$\left(3\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Using routine row operations, the eigenvectors are nonzero vectors of the form

$$\begin{pmatrix} z \\ -z \\ z \end{pmatrix} = z\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

Next consider the eigenvectors for $\lambda = 6$. This requires you to solve

$$\left(6\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and using the usual procedures yields the eigenvectors for $\lambda = 6$ are of the form

$$z\begin{pmatrix} -\frac{1}{8} \\ -\frac{1}{4} \\ 1 \end{pmatrix}$$

or written more simply,

$$z\begin{pmatrix} -1 \\ -2 \\ 8 \end{pmatrix}$$

where $z \in \mathbb{F}$.

Note that in this example the eigenspace for the eigenvalue $\lambda = 6$ is of dimension 1 because there is only one parameter which can be chosen. However, this eigenvalue is of multiplicity two as a root to the characteristic equation.

Definition 6.1.6 *If A is an $n \times n$ matrix with the property that some eigenvalue has algebraic multiplicity as a root of the characteristic equation which is greater than the dimension of the eigenspace associated with this eigenvalue, then the matrix is called defective.*

There may be repeated roots to the characteristic equation, 6.2 and it is not known whether the dimension of the eigenspace equals the multiplicity of the eigenvalue. However, the following theorem is available.

Theorem 6.1.7 *Suppose* $M\mathbf{v}_i = \lambda_i \mathbf{v}_i, i = 1, \cdots, r$, $\mathbf{v}_i \neq 0$, *and that if* $i \neq j$, *then* $\lambda_i \neq \lambda_j$. *Then the set of eigenvectors,* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ *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

$$\sum_{j=1}^{r} c_j \mathbf{w}_j = \mathbf{0} \tag{6.5}$$

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 6.5.

$$\sum_{j=1}^{r} c_j \mu_j \mathbf{w}_j = \mathbf{0}. \tag{6.6}$$

Next pick $\mu_k \neq 0$ and multiply both sides of 6.5 by μ_k. Such a μ_k exists because $r > 1$. Thus

$$\sum_{j=1}^{r} c_j \mu_k \mathbf{w}_j = \mathbf{0} \tag{6.7}$$

Subtract the sum in 6.7 from the sum in 6.6 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. ■

In words, this theorem says that eigenvectors associated with distinct eigenvalues are linearly independent.

Sometimes you have to consider eigenvalues which are complex numbers. This occurs in differential equations for example. You do these problems exactly the same way as you do the ones in which the eigenvalues are real. Here is an example.

Example 6.1.8 *Find the eigenvalues and eigenvectors of the matrix*

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

You need to find the eigenvalues. Solve

$$\det \left(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix} \right) = 0.$$

This reduces to $(\lambda-1)\left(\lambda^2-4\lambda+5\right)=0$. The solutions are $\lambda=1, \lambda=2+i, \lambda=2-i$.

There is nothing new about finding the eigenvectors for $\lambda=1$ so consider the eigenvalue $\lambda=2+i$. You need to solve

$$\left((2+i)\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}-\begin{pmatrix}1&0&0\\0&2&-1\\0&1&2\end{pmatrix}\right)\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

In other words, you must consider the augmented matrix

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

for the solution. Divide the top row by $(1+i)$ and then take $-i$ times the second row and add to the bottom. This yields

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

Now multiply the second row by $-i$ to obtain

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

Therefore, the eigenvectors are of the form

$$z\begin{pmatrix}0\\i\\1\end{pmatrix}.$$

You should find the eigenvectors for $\lambda=2-i$. These are

$$z\begin{pmatrix}0\\-i\\1\end{pmatrix}.$$

As usual, if you want to get it right you had better check it.

$$\begin{pmatrix}1&0&0\\0&2&-1\\0&1&2\end{pmatrix}\begin{pmatrix}0\\-i\\1\end{pmatrix}=\begin{pmatrix}0\\-1-2i\\2-i\end{pmatrix}=(2-i)\begin{pmatrix}0\\-i\\1\end{pmatrix}$$

so it worked.

6.2 Some Applications of Eigenvalues and Eigenvectors

Recall that $n\times n$ matrices can be considered as linear transformations. If F is a 3×3 real matrix having positive determinant, it can be shown that $F=RU$ where R is a rotation matrix and U is a symmetric real matrix having positive eigenvalues. An application of this wonderful result, known to mathematicians as the right polar decomposition, is to continuum mechanics where a chunk of material is identified with a set of points in three dimensional space.

The linear transformation, F in this context is called the deformation gradient and it describes the local deformation of the material. Thus it is possible to consider this deformation in terms of two processes, one which distorts the material and the other which just rotates it. It is the matrix U which is responsible for stretching and compressing. This is why in continuum mechanics, the stress is often taken to depend on U which is known in this context as the right Cauchy Green strain tensor. This process of writing a matrix as a product of two such matrices, one of which preserves distance and the other which distorts is also important in applications to geometric measure theory an interesting field of study in mathematics and to the study of quadratic forms which occur in many applications such as statistics. Here I am emphasizing the application to mechanics in which the eigenvectors of U determine the principle directions, those directions in which the material is stretched or compressed to the maximum extent.

Example 6.2.1 *Find the principle directions determined by the matrix*

$$\begin{pmatrix} \frac{29}{11} & \frac{6}{11} & \frac{6}{11} \\ \frac{6}{11} & \frac{41}{44} & \frac{19}{44} \\ \frac{6}{11} & \frac{19}{44} & \frac{41}{44} \end{pmatrix}$$

The eigenvalues are $3, 1,$ *and* $\frac{1}{2}$.

It is nice to be given the eigenvalues. The largest eigenvalue is 3 which means that in the direction determined by the eigenvector associated with 3 the stretch is three times as large. The smallest eigenvalue is $1/2$ and so in the direction determined by the eigenvector for $1/2$ the material is compressed, becoming locally half as long. It remains to find these directions. First consider the eigenvector for 3. It is necessary to solve

$$\left(3\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{29}{11} & \frac{6}{11} & \frac{6}{11} \\ \frac{6}{11} & \frac{41}{44} & \frac{19}{44} \\ \frac{6}{11} & \frac{19}{44} & \frac{41}{44} \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Thus the augmented matrix for this system of equations is

$$\begin{pmatrix} \frac{4}{11} & -\frac{6}{11} & -\frac{6}{11} & 0 \\ -\frac{6}{11} & \frac{91}{44} & -\frac{19}{44} & 0 \\ -\frac{6}{11} & -\frac{19}{44} & \frac{91}{44} & 0 \end{pmatrix}$$

The row reduced echelon form is

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

and so the principle direction for the eigenvalue 3 in which the material is stretched to the maximum extent is

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

A direction vector in this direction is

$$\begin{pmatrix} 3/\sqrt{11} \\ 1/\sqrt{11} \\ 1/\sqrt{11} \end{pmatrix}.$$

You should show that the direction in which the material is compressed the most is in the direction

$$\begin{pmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

Note this is meaningful information which you would have a hard time finding without the theory of eigenvectors and eigenvalues.

Another application is to the problem of finding solutions to systems of differential equations. It turns out that vibrating systems involving masses and springs can be studied in the form

$$\mathbf{x}'' = A\mathbf{x} \tag{6.8}$$

where A is a real symmetric $n \times n$ matrix which has nonpositive eigenvalues. This is analogous to the case of the scalar equation for undamped oscillation, $x'' + \omega^2 x = 0$. The main difference is that here the scalar ω^2 is replaced with the matrix $-A$. Consider the problem of finding solutions to 6.8. You look for a solution which is in the form

$$\mathbf{x}(t) = \mathbf{v}e^{\lambda t} \tag{6.9}$$

and substitute this into 6.8. Thus

$$\mathbf{x}'' = \mathbf{v}\lambda^2 e^{\lambda t} = e^{\lambda t} A\mathbf{v}$$

and so

$$\lambda^2 \mathbf{v} = A\mathbf{v}.$$

Therefore, λ^2 needs to be an eigenvalue of A and $\mathbf{v}$ needs to be an eigenvector. Since A has nonpositive eigenvalues, $\lambda^2 = -a^2$ and so $\lambda = \pm ia$ where $-a^2$ is an eigenvalue of A. Corresponding to this you obtain solutions of the form

$$\mathbf{x}(t) = \mathbf{v}\cos(at), \mathbf{v}\sin(at).$$

Note these solutions oscillate because of the $\cos(at)$ and $\sin(at)$ in the solutions. Here is an example.

Example 6.2.2 *Find oscillatory solutions to the system of differential equations,* $\mathbf{x}'' = A\mathbf{x}$ *where*

$$A = \begin{pmatrix} -\frac{5}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{13}{6} & \frac{5}{6} \\ -\frac{1}{3} & \frac{5}{6} & -\frac{13}{6} \end{pmatrix}.$$

The eigenvalues are $-1, -2,$ *and* -3.

According to the above, you can find solutions by looking for the eigenvectors. Consider the eigenvectors for -3. The augmented matrix for finding the eigenvectors is

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

and its row echelon form is

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

Therefore, the eigenvectors are of the form

$$\mathbf{v} = z \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

It follows

$$\begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \cos\left(\sqrt{3}t\right), \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \sin\left(\sqrt{3}t\right)$$

are both solutions to the system of differential equations. You can find other oscillatory solutions in the same way by considering the other eigenvalues. You might try checking these answers to verify they work.

This is just a special case of a procedure used in differential equations to obtain closed form solutions to systems of differential equations using linear algebra. The overall philosophy is to take one of the easiest problems in analysis and change it into the eigenvalue problem which is the most difficult problem in algebra. However, when it works, it gives precise solutions in terms of known functions.

6.3 Exercises

1. 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.

2. If A is an $n \times n$ matrix and c is a nonzero constant, compare the eigenvalues of A and cA.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} -7 & -12 & 30 \\ -3 & -7 & 15 \\ -3 & -6 & 14 \end{pmatrix}$.

10. 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.

11. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 3 & -2 & -1 \\ 0 & 5 & 1 \\ 0 & 2 & 4 \end{pmatrix}$.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}$.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. Here is a matrix.
$$\begin{pmatrix} 1 & a & 0 & 0 \\ 0 & 1 & b & 0 \\ 0 & 0 & 2 & c \\ 0 & 0 & 0 & 2 \end{pmatrix}$$
Find values of a, b, c for which the matrix is defective and values of a, b, c for which it is nondefective.

26. Here is a matrix.
$$\begin{pmatrix} a & 1 & 0 \\ 0 & b & 1 \\ 0 & 0 & c \end{pmatrix}$$
where a, b, c are numbers. Show this is sometimes defective depending on the choice of a, b, c. What is an easy case which will ensure it is not defective?

27. 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 $a - ib$.

28. Recall an $n \times n$ matrix is said to be symmetric if it has all real entries and if $A = A^T$. Show the eigenvalues of a real symmetric matrix are real and for each eigenvalue, it has a real eigenvector.

29. 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.

30. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?

31. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.

32. Suppose A is an $n \times n$ matrix having all real eigenvalues which are distinct. Show there exists S such that $S^{-1}AS = D$, a diagonal matrix. If
$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$
define e^D by
$$e^D \equiv \begin{pmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n} \end{pmatrix}$$
and define
$$e^A \equiv Se^D S^{-1}.$$
Next show that if A is as just described, so is tA where t is a real number and the eigenvalues of At are $t\lambda_k$. If you differentiate a matrix of functions entry by entry so that for the ij^{th} entry of $A'(t)$ you get $a'_{ij}(t)$ where $a_{ij}(t)$ is the ij^{th} entry of $A(t)$, show
$$\frac{d}{dt}\left(e^{At}\right) = Ae^{At}$$
Next show $\det\left(e^{At}\right) \neq 0$. This is called the matrix exponential. Note I have only defined it for the case where the eigenvalues of A are real, but the same procedure will work even for complex eigenvalues. All you have to do is to define what is meant by e^{a+ib}.

33. Find the principle directions determined by the matrix $\begin{pmatrix} \frac{7}{12} & -\frac{1}{4} & \frac{1}{6} \\ -\frac{1}{4} & \frac{7}{12} & -\frac{1}{6} \\ \frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \end{pmatrix}$. The eigenvalues are $\frac{1}{3}, 1$, and $\frac{1}{2}$ listed according to multiplicity.

34. Find the principle directions determined by the matrix

$\begin{pmatrix} \frac{5}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{7}{6} & \frac{1}{6} \\ -\frac{1}{3} & \frac{1}{6} & \frac{7}{6} \end{pmatrix}$ The eigenvalues are $1, 2$, and 1. What is the physical interpretation of the repeated eigenvalue?

35. Find oscillatory solutions to the system of differential equations, $\mathbf{x}'' = A\mathbf{x}$ where $A = \begin{pmatrix} -3 & -1 & -1 \\ -1 & -2 & 0 \\ -1 & 0 & -2 \end{pmatrix}$ The eigenvalues are $-1, -4$, and -2.

36. Let A and B be $n \times n$ matrices and let the columns of B be
$$\mathbf{b}_1, \cdots, \mathbf{b}_n$$
and the rows of A are
$$\mathbf{a}_1^T, \cdots, \mathbf{a}_n^T.$$
Show the columns of AB are
$$A\mathbf{b}_1 \cdots A\mathbf{b}_n$$
and the rows of AB are
$$\mathbf{a}_1^T B \cdots \mathbf{a}_n^T B.$$

37. 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.

38. 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 said to be diagonalizable.

39. Show that a $n \times n$ matrix M is diagonalizable if and only if $\mathbb{F}^n$ has a basis of eigenvectors. **Hint:** The first part is done in Problem 38. 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.

40. Let
$$A = \begin{pmatrix} \boxed{\begin{matrix} 1 & 2 \\ 3 & 4 \end{matrix}} & \boxed{\begin{matrix} 2 \\ 0 \end{matrix}} \\ \boxed{\begin{matrix} 0 & 1 \end{matrix}} & \boxed{3} \end{pmatrix}$$
and let
$$B = \begin{pmatrix} \boxed{\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}} \\ \boxed{\begin{matrix} 2 & 1 \end{matrix}} \end{pmatrix}$$

Multiply AB verifying the block multiplication formula. Here $A_{11} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_{12} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, $A_{21} = \begin{pmatrix} 0 & 1 \end{pmatrix}$ and $A_{22} = (3)$.

41. Suppose A, B are $n \times n$ matrices and λ is a nonzero eigenvalue of AB. Show that then it is also an eigenvalue of BA. **Hint:** Use the definition of what it means for λ to be an eigenvalue. That is,
$$AB\mathbf{x} = \lambda \mathbf{x}$$
where $\mathbf{x} \neq \mathbf{0}$. Maybe you should multiply both sides by B.

42. Using the above problem show that if A, B are $n \times n$ matrices, it is not possible that $AB - BA = aI$ for any $a \neq 0$. **Hint:** First show that if A is a matrix, then the eigenvalues of $A - aI$ are $\lambda - a$ where λ is an eigenvalue of A.

43. Consider the following matrix.
$$C = \begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & 0 & & -a_1 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & -a_{n-1} \end{pmatrix}$$
Show $\det(\lambda I - C) = a_0 + \lambda a_1 + \cdots a_{n-1}\lambda^{n-1} + \lambda^n$. This matrix is called a companion matrix for the given polynomial.

44. A discreet dynamical system is a relation of the following form in which $\mathbf{x}(k)$ is a $n \times 1$ vector and A is a $n \times n$ square matrix.
$$\mathbf{x}(k+1) = A\mathbf{x}(k), \ \mathbf{x}(0) = \mathbf{x}_0$$
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.

45. 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,
$$A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I = 0.$$

46. Suppose an $n \times n$ nondefective matrix A has only 1 and -1 as eigenvalues. Find A^{12}.

47. 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?

48. 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 44 to find a formula for x_n.

49. 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 $a_0 = (-1)^n \det(A)$.

6.4 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.

Lemma 6.4.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for $\mathbb{F}^n$. Then there exists an orthonormal basis for $\mathbb{F}^n$, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1, \cdots, \mathbf{x}_k) =$* span$(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.

Proof: Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for $\mathbb{F}^n$. Let $\mathbf{u}_1 \equiv \mathbf{x}_1 / |\mathbf{x}_1|$. Thus for $k = 1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen with $(\mathbf{u}_j \cdot \mathbf{u}_l) = \delta_{jl}$ and $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then define
$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\,\mathbf{u}_j}{\left|\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\,\mathbf{u}_j\right|}, \tag{6.10}$$
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, $\mathbf{x}_{k+1} \in \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving 6.10 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 $l \leq k$, then for $c = 1/\left|\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j\right|$,

$$(\mathbf{u}_{k+1} \cdot \mathbf{u}_l) = C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)(\mathbf{u}_j \cdot \mathbf{u}_l)\right) =$$

$$C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\delta_{lj}\right) = C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l)\right) = 0.$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^{n}$, generated in this way are therefore an orthonormal basis because each vector has unit length. ■

The process by which these vectors were generated is called the Gram Schmidt process. Here is a fundamental definition.

Definition 6.4.2 *An $n \times n$ matrix U, is unitary if $UU^* = I = U^*U$ where U^* is defined to be the transpose of the conjugate of U.*

Proposition 6.4.3 *An $n \times n$ matrix is unitary if and only if the columns (rows) are an orthonormal set.*

Proof: This follows right away from the way we multiply matrices. If U is an $n \times n$ complex matrix, then

$$(U^*U)_{ij} = \mathbf{u}_i^* \mathbf{u}_j = \overline{(\mathbf{u}_i, \mathbf{u}_j)}$$

and the matrix is unitary if and only if this equals δ_{ij} if and only if the columns are orthonormal.

Note that if U is unitary, then so is U^T. This is because

$$\left(U^T\right)^* U^T \equiv \overline{\left(U^T\right)^T} U^T = \left(U\left(\overline{U^T}\right)\right)^T = (UU^*)^T = I^T = I$$

Thus an $n \times n$ matrix is unitary if and only if the rows are an orthonormal set. ■

Theorem 6.4.4 *Let A be an $n \times n$ matrix. Then there exists a unitary matrix U such that*

$$U^*AU = T, \tag{6.11}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal listed according to multiplicity as roots of the characteristic equation.

Proof: The theorem is clearly true if A is a 1×1 matrix. Just let $U = 1$ the 1×1 matrix which has 1 down the main diagonal and zeros elsewhere. Suppose it is true for $(n-1) \times (n-1)$ matrices and let A be an $n \times n$ matrix. Then let $\mathbf{v}_1$ be a unit eigenvector for A . Then there exists λ_1 such that

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1,\ |\mathbf{v}_1| = 1.$$

Extend $\{\mathbf{v}_1\}$ to a basis and then use Lemma 6.4.1 to obtain $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, an orthonormal basis in $\mathbb{F}^n$. Let U_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then from the above, it follows U_0 is unitary. Then $U_0^* A U_0$ is of the form

$$B \equiv \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & A_1 \end{pmatrix}$$

where A_1 is an $n-1 \times n-1$ matrix. The above matrix B has the same eigenvalues as A. Also note in case of an eigenvalue μ for B,

$$\mu \begin{pmatrix} a \\ \mathbf{x} \end{pmatrix} = B \begin{pmatrix} a \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} * \\ A_1 \mathbf{x} \end{pmatrix}$$

so $\mathbf{x}$ is an eigenvector for A_1 with the same eigenvalue μ. 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

$$U_1 \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}$$

This is a unitary matrix and

$$U_1^* U_0^* A U_0 U_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & A_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & T_{n-1} \end{pmatrix} \equiv T$$

where T is upper triangular. Then let $U = U_0 U_1$. Since $(U_0 U_1)^* = U_1^* U_0^*$, it follows 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 are the diagonal entries listed according to algebraic multiplicity, these are also the eigenvalues of A listed according to multiplicity. ■

Corollary 6.4.5 *Let A be a real $n \times n$ matrix having only real eigenvalues. Then there exists a real orthogonal matrix Q and an upper triangular matrix T such that*

$$Q^T A Q = T$$

and furthermore, if the eigenvalues of A are listed in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

Q can be chosen such that T is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

Proof: Repeat the above argument but pick a real eigenvector for the first step which corresponds to λ_1 as just described. Then use induction as above. Simply replace the word "unitary" with the word "orthogonal". ■

As a simple consequence of the above theorem, here is an interesting lemma.

Lemma 6.4.6 *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

$$\det(A) = \prod_k \det(P_k).$$

Also, the eigenvalues of A consist of the union of the eigenvalues of the P_j.

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 it follows that for

$$U \equiv \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s \end{pmatrix}, \; U^* = \begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s^* \end{pmatrix}$$

and also

$$\begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s^* \end{pmatrix} \begin{pmatrix} P_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_s \end{pmatrix} \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s \end{pmatrix} = \begin{pmatrix} T_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_s \end{pmatrix}.$$

Therefore, since the determinant of an upper triangular matrix is the product of the diagonal entries,

$$\det(A) = \prod_k \det(T_k) = \prod_k \det(P_k).$$

From the above formula, the eigenvalues of A consist of the eigenvalues of the upper triangular matrices T_k, and each T_k has the same eigenvalues as P_k. ■

What if A is a real matrix and you only want to consider real unitary matrices?

Theorem 6.4.7 *Let A be a real $n \times n$ matrix. Then there exists a real unitary (orthogonal) matrix Q and a matrix T of the form*

$$T = \begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix} \tag{6.12}$$

where P_i equals either a real 1×1 matrix or P_i equals a real 2×2 matrix having as its eigenvalues a conjugate pair of eigenvalues of A such that $Q^T A Q = T$. The matrix T is called the real Schur form of the matrix A. Recall that a real unitary matrix is also called an orthogonal matrix.

Proof: Suppose

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \; |\mathbf{v}_1| = 1$$

where λ_1 is real. Then let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis of vectors in $\mathbb{R}^n$. Let Q_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is a real $n-1 \times n-1$ matrix. This is just like the proof of Theorem 6.4.4 up to this point.

Now consider the case where $\lambda_1 = \alpha + i\beta$ where $\beta \neq 0$. It follows since A is real that $\mathbf{v}_1 = \mathbf{z}_1 + i\mathbf{w}_1$ and that $\overline{\mathbf{v}}_1 = \mathbf{z}_1 - i\mathbf{w}_1$ is an eigenvector for the eigenvalue $\alpha - i\beta$. Here $\mathbf{z}_1$ and $\mathbf{w}_1$ are real vectors. Since $\overline{\mathbf{v}}_1$ and $\mathbf{v}_1$ are eigenvectors corresponding to distinct eigenvalues, they form a linearly independent set. From this it follows that $\{\mathbf{z}_1, \mathbf{w}_1\}$ is an independent set of vectors in $\mathbb{C}^n$, hence in $\mathbb{R}^n$. Indeed,$\{\mathbf{v}_1, \overline{\mathbf{v}}_1\}$ is an independent set and also $\operatorname{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1) = \operatorname{span}(\mathbf{z}_1, \mathbf{w}_1)$. Now using the Gram Schmidt theorem in $\mathbb{R}^n$, there exists $\{\mathbf{u}_1, \mathbf{u}_2\}$, an orthonormal set of real vectors such that $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2) = \operatorname{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1)$. For example,

$$\mathbf{u}_1 = \mathbf{z}_1 / |\mathbf{z}_1|, \ \mathbf{u}_2 = \frac{|\mathbf{z}_1|^2 \mathbf{w}_1 - (\mathbf{w}_1 \cdot \mathbf{z}_1) \mathbf{z}_1}{\left| |\mathbf{z}_1|^2 \mathbf{w}_1 - (\mathbf{w}_1 \cdot \mathbf{z}_1) \mathbf{z}_1 \right|}$$

Let $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ be an orthonormal basis in $\mathbb{R}^n$ and let Q_0 be a unitary matrix whose i^{th} column is $\mathbf{u}_i$ so Q_0 is a real orthogonal matrix. Then $A\mathbf{u}_j$ are both in $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$ for $j = 1, 2$ and so $\mathbf{u}_k^T A \mathbf{u}_j = 0$ whenever $k \geq 3$. It follows that $Q_0^* A Q_0$ is of the form

$$Q_0^* A Q_0 = \begin{pmatrix} * & * & \cdots & * \\ * & * & & \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} P_1 & * \\ 0 & A_1 \end{pmatrix}$$

where A_1 is now an $n-2 \times n-2$ matrix and P_1 is a 2×2 matrix. Now this is similar to A and so two of its eigenvalues are $\alpha + i\beta$ and $\alpha - i\beta$.

Now find $\widetilde{Q}_1$ an $n-2 \times n-2$ matrix to put A_1 in an appropriate form as above and come up with A_2 either an $n-4 \times n-4$ matrix or an $n-3 \times n-3$ matrix. Then the only other difference is to let

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & \widetilde{Q}_1 & \\ 0 & 0 & & & \end{pmatrix}$$

thus putting a 2×2 identity matrix in the upper left corner rather than a one. Repeating this process with the above modification for the case of a complex eigenvalue leads eventually to 6.12 where Q is the product of real unitary matrices Q_i above. When the block P_i is 2×2, its eigenvalues are a conjugate pair of eigenvalues of A and if it is 1×1 it is a real eigenvalue of A.

Here is why this last claim is true

$$\lambda I - T = \begin{pmatrix} \lambda I_1 - P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & \lambda I_r - P_r \end{pmatrix}$$

where I_k is the 2×2 identity matrix in the case that P_k is 2×2 and is the number 1 in the case where P_k is a 1×1 matrix. Now by Lemma 6.4.6,

$$\det(\lambda I - T) = \prod_{k=1}^{r} \det(\lambda I_k - P_k).$$

Therefore, λ is an eigenvalue of T if and only if it is an eigenvalue of some P_k. This proves the theorem since the eigenvalues of T are the same as those of A including multiplicity because they have the same characteristic polynomial due to the similarity of A and T. ■

Of course there is a similar conclusion which says that the blocks can be ordered according to order of the size of the eigenvalues.

Corollary 6.4.8 *Let A be a real $n \times n$ matrix. Then there exists a real orthogonal matrix Q and an upper triangular matrix T such that*

$$Q^T A Q = T = \begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix}$$

where P_i equals either a real 1×1 matrix or P_i equals a real 2×2 matrix having as its eigenvalues a conjugate pair of eigenvalues of A. If P_k corresponds to the two eigenvalues $\alpha_k \pm i\beta_k \equiv \sigma(P_k)$, Q can be chosen such that

$$|\sigma(P_1)| \geq |\sigma(P_2)| \geq \cdots$$

where

$$|\sigma(P_k)| \equiv \sqrt{\alpha_k^2 + \beta_k^2}$$

The blocks, P_k can be arranged in any other order also.

Definition 6.4.9 *When a linear transformation A, mapping a linear space V to V has a basis of eigenvectors, the linear transformation is called non defective. Otherwise it is called defective. An $n \times n$ matrix A, is called normal if $AA^* = A^*A$. An important class of normal matrices is that of the Hermitian or self adjoint matrices. An $n \times n$ matrix A is self adjoint or Hermitian if $A = A^*$.*

You can check that an example of a normal matrix which is neither symmetric nor Hermitian is $\begin{pmatrix} 6i & -(1+i)\sqrt{2} \\ (1-i)\sqrt{2} & 6i \end{pmatrix}$.

The next lemma is the basis for concluding that every normal matrix is unitarily similar to a diagonal matrix.

Lemma 6.4.10 *If T is upper triangular and normal, then T is a diagonal matrix.*

Proof: This is obviously true if T is 1×1. In fact, it can't help being diagonal in this case. Suppose then that the lemma is true for $(n-1) \times (n-1)$ matrices and let T be an upper triangular normal $n \times n$ matrix. Thus T is of the form

$$T = \begin{pmatrix} t_{11} & \mathbf{a}^* \\ \mathbf{0} & T_1 \end{pmatrix}, \ T^* = \begin{pmatrix} \overline{t_{11}} & \mathbf{0}^T \\ \mathbf{a} & T_1^* \end{pmatrix}$$

Then

$$\begin{aligned} TT^* &= \begin{pmatrix} t_{11} & \mathbf{a}^* \\ \mathbf{0} & T_1 \end{pmatrix} \begin{pmatrix} \overline{t_{11}} & \mathbf{0}^T \\ \mathbf{a} & T_1^* \end{pmatrix} = \begin{pmatrix} |t_{11}|^2 + \mathbf{a}^*\mathbf{a} & \mathbf{a}^* T_1^* \\ T_1 \mathbf{a} & T_1 T_1^* \end{pmatrix} \\ T^*T &= \begin{pmatrix} \overline{t_{11}} & \mathbf{0}^T \\ \mathbf{a} & T_1^* \end{pmatrix} \begin{pmatrix} t_{11} & \mathbf{a}^* \\ \mathbf{0} & T_1 \end{pmatrix} = \begin{pmatrix} |t_{11}|^2 & \overline{t_{11}}\mathbf{a}^* \\ \mathbf{a} t_{11} & \mathbf{a}\mathbf{a}^* + T_1^* T_1 \end{pmatrix} \end{aligned}$$

Since these two matrices are equal, it follows $\mathbf{a} = \mathbf{0}$. But now it follows that $T_1^* T_1 = T_1 T_1^*$ and so by induction T_1 is a diagonal matrix D_1. Therefore,

$$T = \begin{pmatrix} t_{11} & \mathbf{0}^T \\ \mathbf{0} & D_1 \end{pmatrix}$$

a diagonal matrix.

Now here is a proof which doesn't involve block multiplication. Since T is normal, $T^*T = TT^*$. Writing this in terms of components and using the description of the adjoint as the transpose of the conjugate, yields the following for the ik^{th} entry of $T^*T = TT^*$.

$$\overbrace{\sum_j t_{ij} t_{jk}^* = \sum_j t_{ij} \overline{t_{kj}}}^{TT^*} = \overbrace{\sum_j t_{ij}^* t_{jk} = \sum_j \overline{t_{ji}} t_{jk}}^{T^*T}.$$

Now use the fact that T is upper triangular and let $i = k = 1$ to obtain the following from the above.

$$\sum_j |t_{1j}|^2 = \sum_j |t_{j1}|^2 = |t_{11}|^2$$

You see, $t_{j1} = 0$ unless $j = 1$ due to the assumption that T is upper triangular. This shows T is of the form

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

Now do the same thing only this time take $i = k = 2$ and use the result just established. Thus, from the above,

$$\sum_j |t_{2j}|^2 = \sum_j |t_{j2}|^2 = |t_{22}|^2 ,$$

showing that $t_{2j} = 0$ if $j > 2$ which means T has the form

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

Next let $i = k = 3$ and obtain that T looks like a diagonal matrix in so far as the first 3 rows and columns are concerned. Continuing in this way, it follows T is a diagonal matrix. ■

Theorem 6.4.11 *Let A be a normal matrix. Then there exists a unitary matrix U such that U^*AU is a diagonal matrix. Also if A is normal and U is unitary, then U^*AU is also normal.*

Proof: From Theorem 6.4.4 there exists a unitary matrix U such that U^*AU equals an upper triangular matrix. The theorem is now proved if it is shown that the property of being normal is preserved under unitary similarity transformations. That is, verify that if A is normal and if $B = U^*AU$, then B is also normal. But this is easy.

$$\begin{aligned} B^*B &= U^*A^*UU^*AU = U^*A^*AU \\ &= U^*AA^*U = U^*AUU^*A^*U = BB^*. \end{aligned}$$

Therefore, U^*AU is a normal and upper triangular matrix and by Lemma 6.4.10 it must be a diagonal matrix. ■

The converse is also true. See Problem 9 below.

Corollary 6.4.12 *If A is Hermitian, then all the eigenvalues of A are real and there exists an orthonormal basis of eigenvectors. Also there exists a unitary U such that $U^*AU = D$, a diagonal matrix whose diagonal is comprised of the eigenvalues of A. The columns of U are the corresponding eigenvectors. By permuting the columns of U one can cause the diagonal entries of D to occur in any desired order.*

Proof: Since A is normal, there exists unitary, U such that $U^*AU = D$, a diagonal matrix whose diagonal entries are the eigenvalues of A. Therefore, $D^* = U^*A^*U = U^*AU = D$ showing D is real.

Finally, let

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$$

where the $\mathbf{u}_i$ denote the columns of U and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

The equation, $U^*AU = D$ implies

$$\begin{aligned} AU &= \begin{pmatrix} A\mathbf{u}_1 & A\mathbf{u}_2 & \cdots & A\mathbf{u}_n \end{pmatrix} \\ &= UD = \begin{pmatrix} \lambda_1\mathbf{u}_1 & \lambda_2\mathbf{u}_2 & \cdots & \lambda_n\mathbf{u}_n \end{pmatrix} \end{aligned} \tag{6.13}$$

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

$$\delta_{ij} = \mathbf{u}_i^*\mathbf{u}_j \equiv \mathbf{u}_j \cdot \mathbf{u}_i.$$

This proves the corollary because it shows the vectors $\{\mathbf{u}_i\}$ are orthonormal. Therefore, they form a basis because every orthonormal set of vectors is linearly independent. It follows from 6.13 that one can achieve any order for the λ_i by permuting the columns of U. ■

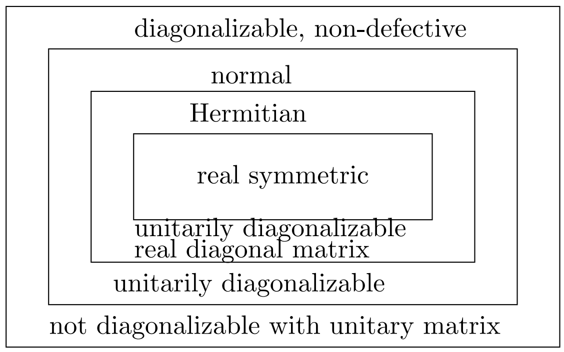
Corollary 6.4.13 *If A is a real symmetric matrix, then A is Hermitian and there exists a real unitary (orthogonal) matrix U such that $U^TAU = D$ where D is a diagonal matrix whose diagonal entries are the eigenvalues of A. By arranging the columns of U the diagonal entries of D can be made to appear in any order.*

Proof: It is clear that $A = A^* = A^T$. Thus A is real and all eigenvalues are real and it is Hermitian. Now by Corollary 6.4.5, there is an orthogonal matrix U such that $U^TAU = T$. Since A is normal, so is T by Theorem 6.4.11. Hence by Lemma 6.4.10 T is a diagonal matrix. Then it follows the diagonal entries are the eigenvalues of A and the columns of U

are the corresponding eigenvectors. Permuting these columns, one can cause the eigenvalues to appear in any order on the diagonal. ■

The converse for the above theorems about normal and Hermitian matrices is also true. That is, the Hermitian matrices, $(A = A^*)$ are exactly those for which there is a unitary U such that U^*AU is a real diagonal matrix. The normal matrices are exactly those for which there is a unitary U such that U^*AU is a diagonal matrix, maybe not real.

To summarize these types of matrices which have just been discussed, here is a little diagram.



6.5 Trace and Determinant

The determinant has already been discussed. It is also clear that if $A = S^{-1}BS$ so that A, B are similar, then

$$\begin{aligned} \det(A) &= \det\left(S^{-1}\right)\det(S)\det(B) = \det\left(S^{-1}S\right)\det(B) \\ &= \det(I)\det(B) = \det(B) \end{aligned}$$

The **trace** is defined in the following definition.

Definition 6.5.1 *Let A be an $n \times n$ matrix whose ij^{th} entry is denoted as a_{ij}. Then*

$$\operatorname{trace}(A) \equiv \sum_i a_{ii}$$

In other words it is the sum of the entries down the main diagonal.

Theorem 6.5.2 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix. Then*

$$\operatorname{trace}(AB) = \operatorname{trace}(BA).$$

Also if $B = S^{-1}AS$ so that A, B are similar, then

$$\operatorname{trace}(A) = \operatorname{trace}(B).$$

Proof:

$$\operatorname{trace}(AB) \equiv \sum_i \left(\sum_k A_{ik}B_{ki}\right) = \sum_k \sum_i B_{ki}A_{ik} = \operatorname{trace}(BA)$$

Therefore,

$$\operatorname{trace}(B) = \operatorname{trace}\left(S^{-1}AS\right) = \operatorname{trace}\left(ASS^{-1}\right) = \operatorname{trace}(A). \blacksquare$$

Theorem 6.5.3 *Let A be an $n\times n$ matrix. Then* $\operatorname{trace}(A)$ *equals the sum of the eigenvalues of A and* $\det(A)$ *equals the product of the eigenvalues of A.*

This is proved using Schur's theorem and is in Problem 17 below. Another important property of the trace is in the following theorem.

6.6 Quadratic Forms

Definition 6.6.1 *A quadratic form in three dimensions is an expression of the form*

$$\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.14}$$

where A is a 3×3 symmetric matrix. In higher dimensions the idea is the same except you use a larger symmetric matrix in place of A. In two dimensions A is a 2×2 matrix.

For example, consider

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.15}$$

which equals $3x^2 - 8xy + 2xz - 8yz + 3z^2$. This is very awkward because of the mixed terms such as $-8xy$. The idea is to pick different axes such that if x, y, z are taken with respect to these axes, the quadratic form is much simpler. In other words, look for new variables, x', y', and z' and a unitary matrix U such that

$$U \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.16}$$

and if you write the quadratic form in terms of the primed variables, there will be no mixed terms. Any symmetric real matrix is Hermitian and is therefore normal. From Corollary 6.4.13, it follows there exists a real unitary matrix U, (an orthogonal matrix) such that $U^T AU = D$ a diagonal matrix. Thus in the quadratic form, 6.14

$$\begin{aligned} \begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} x' & y' & z' \end{pmatrix} U^T AU \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \\ &= \begin{pmatrix} x' & y' & z' \end{pmatrix} D \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \end{aligned}$$

and in terms of these new variables, the quadratic form becomes

$$\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2$$

where $D = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$. Similar considerations apply equally well in any other dimension. For the given example,

$$\begin{pmatrix} -\frac{1}{2}\sqrt{2} & 0 & \frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6} & \frac{1}{3}\sqrt{6} & \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \cdot$$

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

and so if the new variables are given by

$$\begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

it follows that in terms of the new variables the quadratic form is $2(x')^2 - 4(y')^2 + 8(z')^2$. You can work other examples the same way.

6.7 Second Derivative Test

Under certain conditions the **mixed partial derivatives** will always be equal. This astonishing fact was first observed by Euler around 1734. It is also called Clairaut's theorem.

Theorem 6.7.1 *Suppose $f : U \subseteq \mathbb{F}^2 \to \mathbb{R}$ where U is an open set on which f_x, f_y, f_{xy} and f_{yx} exist. Then if f_{xy} and f_{yx} are continuous at the point $(x, y) \in U$, it follows*

$$f_{xy}(x, y) = f_{yx}(x, y).$$

Proof: Since U is open, there exists $r > 0$ such that $B((x, y), r) \subseteq U$. Now let $|t|, |s| < r/2, t, s$ real numbers and consider

$$\Delta(s, t) \equiv \frac{1}{st}\{\overbrace{f(x+t, y+s) - f(x+t, y)}^{h(t)} - \overbrace{(f(x, y+s) - f(x, y))}^{h(0)}\}. \tag{6.17}$$

Note that $(x + t, y + s) \in U$ because

$$\begin{aligned} |(x+t, y+s) - (x, y)| &= |(t, s)| = (t^2 + s^2)^{1/2} \\ &\leq \left(\frac{r^2}{4} + \frac{r^2}{4}\right)^{1/2} = \frac{r}{\sqrt{2}} < r. \end{aligned}$$

As implied above, $h(t) \equiv f(x + t, y + s) - f(x + t, y)$. Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

$$\begin{aligned} \Delta(s, t) &= \frac{1}{st}(h(t) - h(0)) = \frac{1}{st}h'(\alpha t)t \\ &= \frac{1}{s}(f_x(x + \alpha t, y + s) - f_x(x + \alpha t, y)) \end{aligned}$$

for some $\alpha \in (0, 1)$. Applying the mean value theorem again,

$$\Delta(s, t) = f_{xy}(x + \alpha t, y + \beta s)$$

where $\alpha, \beta \in (0,1)$.

If the terms $f(x+t,y)$ and $f(x,y+s)$ are interchanged in 6.17, $\Delta(s,t)$ is unchanged and the above argument shows there exist $\gamma, \delta \in (0,1)$ such that

$$\Delta(s,t) = f_{yx}(x+\gamma t, y+\delta s).$$

Letting $(s,t) \to (0,0)$ and using the continuity of f_{xy} and f_{yx} at (x,y),

$$\lim_{(s,t)\to(0,0)} \Delta(s,t) = f_{xy}(x,y) = f_{yx}(x,y). \blacksquare$$

The following is obtained from the above by simply fixing all the variables except for the two of interest.

Corollary 6.7.2 *Suppose U is an open subset of $\mathbb{F}^n$ and $f : U \to \mathbb{R}$ has the property that for two indices, k, l, f_{x_k}, $f_{x_l}, f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.*

Thus the theorem asserts that the mixed partial derivatives are equal at $\mathbf{x}$ if they are defined near $\mathbf{x}$ and continuous at $\mathbf{x}$.

Now recall the Taylor formula with the Lagrange form of the remainder. What follows is a proof of this important result based on the mean value theorem or Rolle's theorem.

Theorem 6.7.3 *Suppose f has $n+1$ derivatives on an interval, (a,b) and let $c \in (a,b)$. Then if $x \in (a,b)$, there exists ξ between c and x such that*

$$f(x) = f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}.$$

(In this formula, the symbol $\sum_{k=1}^{0} a_k$ will denote the number 0.)

Proof: It can be assumed $x \neq c$ because if $x = c$ there is nothing to show. Then there exists K such that

$$f(x) - \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + K(x-c)^{n+1} \right) = 0 \tag{6.18}$$

In fact,

$$K = \frac{-f(x) + \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k \right)}{(x-c)^{n+1}}.$$

Now define $F(t)$ for t in the closed interval determined by x and c by

$$F(t) \equiv f(x) - \left(f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(t)}{k!}(x-t)^k + K(x-t)^{n+1} \right).$$

The c in 6.18 got replaced by t.

Therefore, $F(c) = 0$ by the way K was chosen and also $F(x) = 0$. By the mean value theorem or Rolle's theorem, there exists ξ between x and c such that $F'(\xi) = 0$. Therefore,

$$\begin{aligned}
0 &= f'(\xi) + \sum_{k=1}^{n} \frac{f^{(k+1)}(\xi)}{k!}(x-\xi)^k - \sum_{k=1}^{n} \frac{f^{(k)}(\xi)}{(k-1)!}(x-\xi)^{k-1} - K(n+1)(x-\xi)^n \\
&= f'(\xi) + \sum_{k=1}^{n} \frac{f^{(k+1)}(\xi)}{k!}(x-\xi)^k - \sum_{k=0}^{n-1} \frac{f^{(k+1)}(\xi)}{k!}(x-\xi)^k - K(n+1)(x-\xi)^n \\
&= f'(\xi) + \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n - f'(\xi) - K(n+1)(x-\xi)^n \\
&= \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n - K(n+1)(x-\xi)^n
\end{aligned}$$

Then therefore,

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!} \blacksquare$$

The following is a special case and is what will be used.

Theorem 6.7.4 *Let $h : (-\delta, 1+\delta) \to \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in [0,1]$ such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Now let $f : U \to \mathbb{R}$ where $U \subseteq \mathbb{R}^n$ and suppose $f \in C^m(U)$. Let $\mathbf{x} \in U$ and let $r > 0$ be such that

$$B(\mathbf{x},r) \subseteq U.$$

Then for $||\mathbf{v}|| < r$, consider

$$f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) \equiv h(t)$$

for $t \in [0,1]$. Then by the chain rule,

$$h'(t) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x}+t\mathbf{v}) v_k, \ h''(t) = \sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_k}(\mathbf{x}+t\mathbf{v}) v_k v_j \blacksquare$$

Then from the Taylor formula stopping at the second derivative, the following theorem can be obtained.

Theorem 6.7.5 *Let $f : U \to \mathbb{R}$ and let $f \in C^2(U)$. Then if*

$$B(\mathbf{x},r) \subseteq U,$$

and $||\mathbf{v}|| < r$, there exists $t \in (0,1)$ such that.

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x}) v_k + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_k}(\mathbf{x}+t\mathbf{v}) v_k v_j \tag{6.19}$$

Definition 6.7.6 *Define the following matrix.*

$$H_{ij}(\mathbf{x}+t\mathbf{v}) \equiv \frac{\partial^2 f(\mathbf{x}+t\mathbf{v})}{\partial x_j \partial x_i}.$$

It is called the Hessian matrix. From Corollary 6.7.2, this is a symmetric matrix. Then in terms of this matrix, 6.19 can be written as

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}) v_k + \frac{1}{2}\mathbf{v}^T H(\mathbf{x}+t\mathbf{v})\mathbf{v}$$

Then this implies $f(\mathbf{x}+\mathbf{v}) =$

$$f(\mathbf{x}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x})\, v_k + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T \left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right). \tag{6.20}$$

Using the above formula, here is the second derivative test.

Theorem 6.7.7 *In the above situation, suppose $f_{x_j}(\mathbf{x}) = 0$ for each x_j. Then if $H(\mathbf{x})$ has all positive eigenvalues, $\mathbf{x}$ is a local minimum for f. If $H(\mathbf{x})$ has all negative eigenvalues, then $\mathbf{x}$ is a local maximum. If $H(\mathbf{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\mathbf{x}$, while if $H(\mathbf{x})$ has a negative eigenvalue, there exists a direction in which $H(\mathbf{x})$ has a local maximum at $\mathbf{x}$.*

Proof: Since $f_{x_j}(\mathbf{x}) = 0$ for each x_j, formula 6.20 implies

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T \left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)$$

where $H(\mathbf{x})$ is a symmetric matrix. Thus, by Corollary 6.4.12 $H(\mathbf{x})$ has all real eigenvalues. Suppose first that $H(\mathbf{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$. Then $H(\mathbf{x})$ has an orthonormal basis of eigenvectors, $\{\mathbf{v}_i\}_{i=1}^n$ and if $\mathbf{u}$ is an arbitrary vector, $\mathbf{u} = \sum_{j=1}^n u_j \mathbf{v}_j$ where $u_j = \mathbf{u} \cdot \mathbf{v}_j$. Thus

$$\mathbf{u}^T H(\mathbf{x})\,\mathbf{u} = \left(\sum_{k=1}^{n} u_k \mathbf{v}_k^T\right) H(\mathbf{x}) \left(\sum_{j=1}^{n} u_j \mathbf{v}_j\right) = \sum_{j=1}^{n} u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^{n} u_j^2 = \delta^2 |\mathbf{u}|^2.$$

From 6.20 and the continuity of H, if $\mathbf{v}$ is small enough,

$$f(\mathbf{x}+\mathbf{v}) \geq f(\mathbf{x}) + \frac{1}{2}\delta^2 |\mathbf{v}|^2 - \frac{1}{4}\delta^2 |\mathbf{v}|^2 = f(\mathbf{x}) + \frac{\delta^2}{4}|\mathbf{v}|^2.$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H(\mathbf{x})$ has a positive eigenvalue λ^2. Then let $\mathbf{v}$ be an eigenvector for this eigenvalue. From 6.20,

$$f(\mathbf{x}+t\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}t^2\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}t^2\left(\mathbf{v}^T \left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)$$

which implies

$$\begin{aligned} f(\mathbf{x}+t\mathbf{v}) &= f(\mathbf{x}) + \frac{1}{2}t^2\lambda^2 |\mathbf{v}|^2 + \frac{1}{2}t^2\left(\mathbf{v}^T \left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \\ &\geq f(\mathbf{x}) + \frac{1}{4}t^2\lambda^2 |\mathbf{v}|^2 \end{aligned}$$

whenever t is small enough. Thus in the direction $\mathbf{v}$ the function has a local minimum at $\mathbf{x}$. The assertion about the local maximum in some direction follows similarly. ■

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1(x,y) = x^4 + y^2,\ f_2(x,y) = -x^4 + y^2.$$

Then $Df_i(0,0) = \mathbf{0}$ and for both functions, the Hessian matrix evaluated at $(0,0)$ equals

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

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.

6.8 The Estimation of Eigenvalues

There are ways to estimate the eigenvalues for matrices. The most famous is known as Gerschgorin's theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

Theorem 6.8.1 *Let A be an $n \times n$ matrix. Consider the n Gerschgorin discs defined as*

$$D_i \equiv \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

Then every eigenvalue is contained in some Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the i^{th} row which are off the main diagonal and form the disc centered at a_{ii} having this radius. The union of these discs contains $\sigma(A)$.

Proof: Suppose $A\mathbf{x} = \lambda \mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$. Then for $A = (a_{ij})$, 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 6.8.2 *Here is a matrix. Estimate its eigenvalues.*

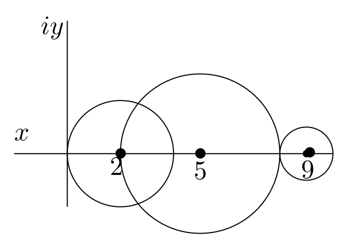
$$\begin{pmatrix} 2 & 1 & 1 \\ 3 & 5 & 0 \\ 0 & 1 & 9 \end{pmatrix}$$

According to Gerschgorin's theorem the eigenvalues are contained in the disks

$$D_1 = \{\lambda \in \mathbb{C} : |\lambda - 2| \leq 2\}, D_2 = \{\lambda \in \mathbb{C} : |\lambda - 5| \leq 3\},$$

$$D_3 = \{\lambda \in \mathbb{C} : |\lambda - 9| \leq 1\}$$

It is important to observe that these disks are in the complex plane. In general this is the case. If you want to find eigenvalues they will be complex numbers.



So what are the values of the eigenvalues? In this case they are real. You can compute them by graphing the characteristic polynomial, $\lambda^3 - 16\lambda^2 + 70\lambda - 66$ and then zooming in on the zeros. If you do this you find the solution is $\{\lambda = 1.2953\}, \{\lambda = 5.5905\}$, $\{\lambda = 9.1142\}$. Of course these are only approximations and so this information is useless for finding eigenvectors. However, in many applications, it is the size of the eigenvalues which is important and so these numerical values would be helpful for such applications. In this case, you might think there is no real reason for Gerschgorin's theorem. Why not just compute the characteristic equation and graph and zoom? This is fine up to a point, but what if the matrix was huge? Then it might be hard to find the characteristic polynomial. Remember the difficulties in expanding a big matrix along a row or column. Also, what if the eigenvalues were complex? You don't see these by following this procedure. However, Gerschgorin's theorem will at least estimate them.

6.9 Advanced Theorems

More can be said but this requires some theory from complex variables[1]. The following is a fundamental theorem about counting zeros.

Theorem 6.9.1 *Let U be a region and let $\gamma : [a,b] \to U$ be closed, continuous, bounded variation, and the winding number, $n(\gamma, z) = 0$ for all $z \notin U$. Suppose also that f is analytic on U having zeros $a_1, \cdots, a_m$ where the zeros are repeated according to multiplicity, and suppose that none of these zeros are on $\gamma([a,b])$. Then*

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz = \sum_{k=1}^m n(\gamma, a_k).$$

Proof: It is given that $f(z) = \prod_{j=1}^m (z-a_j)\, g(z)$ where $g(z) \neq 0$ on U. Hence using the product rule,

$$\frac{f'(z)}{f(z)} = \sum_{j=1}^m \frac{1}{z-a_j} + \frac{g'(z)}{g(z)}$$

where $\frac{g'(z)}{g(z)}$ is analytic on U and so

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz = \sum_{j=1}^m n(\gamma, a_j) + \frac{1}{2\pi i}\int_\gamma \frac{g'(z)}{g(z)}dz = \sum_{j=1}^m n(\gamma, a_j). \blacksquare$$

Now let A be an $n \times n$ matrix. Recall that the eigenvalues of A are given by the zeros of the polynomial, $p_A(z) = \det(zI - A)$ where I is the $n \times n$ identity. You can argue that small changes in A will produce small changes in $p_A(z)$ and $p_A'(z)$. Let γ_k denote a very small closed circle which winds around z_k, one of the eigenvalues of A, in the counter clockwise direction so that $n(\gamma_k, z_k) = 1$. This circle is to enclose only z_k and is to have no other eigenvalue on it. Then apply Theorem 6.9.1. According to this theorem

$$\frac{1}{2\pi i}\int_\gamma \frac{p_A'(z)}{p_A(z)}dz$$

is always an integer equal to the multiplicity of z_k as a root of $p_A(t)$. Therefore, small changes in A result in no change to the above contour integral because it must be an integer and small changes in A result in small changes in the integral. Therefore whenever B is close enough to A, the two matrices have the same number of zeros inside γ_k, the zeros being counted according to multiplicity. By making the radius of the small circle equal to ε where ε is less than the minimum distance between any two distinct eigenvalues of A, this shows that if B is close enough to A, every eigenvalue of B is closer than ε to some eigenvalue of A. $\blacksquare$

[1] If you haven't studied the theory of a complex variable, you should skip this section because you won't understand any of it.

Theorem 6.9.2 *If λ is an eigenvalue of A, then if all the entries of B are close enough to the corresponding entries of A, some eigenvalue of B will be within ε of λ.*

Consider the situation that $A(t)$ is an $n \times n$ matrix and that $t \to A(t)$ is continuous for $t \in [0,1]$.

Lemma 6.9.3 *Let $\lambda(t) \in \sigma(A(t))$ for $t < 1$ and let $\Sigma_t = \cup_{s \geq t} \sigma(A(s))$. Also let K_t be the connected component of $\lambda(t)$ in Σ_t. Then there exists $\eta > 0$ such that $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$.*

Proof: Denote by $D(\lambda(t), \delta)$ the disc centered at $\lambda(t)$ having radius $\delta > 0$, with other occurrences of this notation being defined similarly. Thus

$$D(\lambda(t), \delta) \equiv \{z \in \mathbb{C} : |\lambda(t) - z| \leq \delta\}.$$

Suppose $\delta > 0$ is small enough that $\lambda(t)$ is the only element of $\sigma(A(t))$ contained in $D(\lambda(t), \delta)$ and that $p_{A(t)}$ has no zeroes on the boundary of this disc. Then by continuity, and the above discussion and theorem, there exists $\eta > 0, t + \eta < 1$, such that for $s \in [t, t+\eta]$, $p_{A(s)}$ also has no zeroes on the boundary of this disc and $A(s)$ has the same number of eigenvalues, counted according to multiplicity, in the disc as $A(t)$. Thus $\sigma(A(s)) \cap D(\lambda(t), \delta) \neq \emptyset$ for all $s \in [t, t+\eta]$. Now let

$$H = \bigcup_{s \in [t, t+\eta]} \sigma(A(s)) \cap D(\lambda(t), \delta).$$

It will be shown that H is connected. Suppose not. Then $H = P \cup Q$ where P, Q are separated and $\lambda(t) \in P$. Let $s_0 \equiv \inf\{s : \lambda(s) \in Q \text{ for some } \lambda(s) \in \sigma(A(s))\}$. There exists $\lambda(s_0) \in \sigma(A(s_0)) \cap D(\lambda(t), \delta)$. If $\lambda(s_0) \notin Q$, then from the above discussion there are $\lambda(s) \in \sigma(A(s)) \cap Q$ for $s > s_0$ arbitrarily close to $\lambda(s_0)$. Therefore, $\lambda(s_0) \in Q$ which shows that $s_0 > t$ because $\lambda(t)$ is the only element of $\sigma(A(t))$ in $D(\lambda(t), \delta)$ and $\lambda(t) \in P$. Now let $s_n \uparrow s_0$. Then $\lambda(s_n) \in P$ for any $\lambda(s_n) \in \sigma(A(s_n)) \cap D(\lambda(t), \delta)$ and also it follows from the above discussion that for some choice of $s_n \to s_0$, $\lambda(s_n) \to \lambda(s_0)$ which contradicts P and Q separated and nonempty. Since P is nonempty, this shows $Q = \emptyset$. Therefore, H is connected as claimed. But $K_t \supseteq H$ and so $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$. ■

Theorem 6.9.4 *Suppose $A(t)$ is an $n \times n$ matrix and that $t \to A(t)$ is continuous for $t \in [0,1]$. Let $\lambda(0) \in \sigma(A(0))$ and define $\Sigma \equiv \cup_{t \in [0,1]} \sigma(A(t))$. Let $K_{\lambda(0)} = K_0$ denote the connected component of $\lambda(0)$ in Σ. Then $K_0 \cap \sigma(A(t)) \neq \emptyset$ for all $t \in [0,1]$.*

Proof: Let $S \equiv \{t \in [0,1] : K_0 \cap \sigma(A(s)) \neq \emptyset \text{ for all } s \in [0,t]\}$. Then $0 \in S$. Let $t_0 = \sup(S)$. Say $\sigma(A(t_0)) = \lambda_1(t_0), \cdots, \lambda_r(t_0)$.

Claim: At least one of these is a limit point of K_0 and consequently must be in K_0 which shows that S has a last point. Why is this claim true? Let $s_n \uparrow t_0$ so $s_n \in S$. Now let the discs, $D(\lambda_i(t_0), \delta), i = 1, \cdots, r$ be disjoint with $p_{A(t_0)}$ having no zeroes on γ_i the boundary of $D(\lambda_i(t_0), \delta)$. Then for n large enough it follows from Theorem 6.9.1 and the discussion following it that $\sigma(A(s_n))$ is contained in $\cup_{i=1}^r D(\lambda_i(t_0), \delta)$. It follows that $K_0 \cap (\sigma(A(t_0)) + D(0, \delta)) \neq \emptyset$ for all δ small enough. This requires at least one of the $\lambda_i(t_0)$ to be in $\overline{K_0}$. Therefore, $t_0 \in S$ and S has a last point.

Now by Lemma 6.9.3, if $t_0 < 1$, then $K_0 \cup K_t$ would be a strictly larger connected set containing $\lambda(0)$. (The reason this would be strictly larger is that $K_0 \cap \sigma(A(s)) = \emptyset$ for some $s \in (t, t+\eta)$ while $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$.) Therefore, $t_0 = 1$. ■

Corollary 6.9.5 *Suppose one of the Gerschgorin discs, D_i is disjoint from the union of the others. Then D_i contains an eigenvalue of A. Also, if there are n disjoint Gerschgorin discs, then each one contains an eigenvalue of A.*

Proof: Denote by $A(t)$ the matrix $\left(a_{ij}^t\right)$ where if $i \neq j$, $a_{ij}^t = t a_{ij}$ and $a_{ii}^t = a_{ii}$. Thus to get $A(t)$ multiply all non diagonal terms by t. Let $t \in [0,1]$. Then $A(0) = \text{diag}(a_{11}, \cdots, a_{nn})$ and $A(1) = A$. Furthermore, the map, $t \to A(t)$ is continuous. Denote by D_j^t the Gerschgorin disc obtained from the j^{th} row for the matrix $A(t)$. Then it is clear that $D_j^t \subseteq D_j$ the j^{th} Gerschgorin disc for A. It follows a_{ii} is the eigenvalue for $A(0)$ which is contained in the disc, consisting of the single point a_{ii} which is contained in D_i. Letting K be the connected component in Σ for Σ defined in Theorem 6.9.4 which is determined by a_{ii}, Gerschgorin's theorem implies that $K \cap \sigma(A(t)) \subseteq \cup_{j=1}^n D_j^t \subseteq \cup_{j=1}^n D_j = D_i \cup (\cup_{j \neq i} D_j)$ and also, since K is connected, there are not points of K in both D_i and $(\cup_{j \neq i} D_j)$. Since at least one point of K is in D_i,(a_{ii}), it follows all of K must be contained in D_i. Now by Theorem 6.9.4 this shows there are points of $K \cap \sigma(A)$ in D_i. The last assertion follows immediately. ■

This can be improved even more. This involves the following lemma.

Lemma 6.9.6 *In the situation of Theorem 6.9.4 suppose $\lambda(0) = K_0 \cap \sigma(A(0))$ and that $\lambda(0)$ is a simple root of the characteristic equation of $A(0)$. Then for all $t \in [0,1]$,*

$$\sigma(A(t)) \cap K_0 = \lambda(t)$$

where $\lambda(t)$ is a simple root of the characteristic equation of $A(t)$.

Proof: Let

$$S \equiv \{t \in [0,1] : K_0 \cap \sigma(A(s)) = \lambda(s), \text{ a simple eigenvalue for all } s \in [0,t]\}.$$

Then $0 \in S$ so it is nonempty. Let $t_0 = \sup(S)$ and suppose $\lambda_1 \neq \lambda_2$ are two elements of $\sigma(A(t_0)) \cap K_0$. Then choosing $\eta > 0$ small enough, and letting D_i be disjoint discs containing λ_i respectively, similar arguments to those of Lemma 6.9.3 can be used to conclude

$$H_i \equiv \cup_{s \in [t_0 - \eta, t_0]} \sigma(A(s)) \cap D_i$$

is a connected and nonempty set for $i = 1, 2$ which would require that $H_i \subseteq K_0$. But then there would be two different eigenvalues of $A(s)$ contained in K_0, contrary to the definition of t_0. Therefore, there is at most one eigenvalue $\lambda(t_0) \in K_0 \cap \sigma(A(t_0))$. Could it be a repeated root of the characteristic equation? Suppose $\lambda(t_0)$ is a repeated root of the characteristic equation. As before, choose a small disc, D centered at $\lambda(t_0)$ and η small enough that

$$H \equiv \cup_{s \in [t_0 - \eta, t_0]} \sigma(A(s)) \cap D$$

is a nonempty connected set containing either multiple eigenvalues of $A(s)$ or else a single repeated root to the characteristic equation of $A(s)$. But since H is connected and contains $\lambda(t_0)$ it must be contained in K_0 which contradicts the condition for $s \in S$ for all these $s \in [t_0 - \eta, t_0]$. Therefore, $t_0 \in S$ as hoped. If $t_0 < 1$, there exists a small disc centered at $\lambda(t_0)$ and $\eta > 0$ such that for all $s \in [t_0, t_0 + \eta]$, $A(s)$ has only simple eigenvalues in D and the only eigenvalues of $A(s)$ which could be in K_0 are in D. (This last assertion follows from noting that $\lambda(t_0)$ is the only eigenvalue of $A(t_0)$ in K_0 and so the others are at a positive distance from K_0. For s close enough to t_0, the eigenvalues of $A(s)$ are either close to these eigenvalues of $A(t_0)$ at a positive distance from K_0 or they are close to the eigenvalue $\lambda(t_0)$ in which case it can be assumed they are in D.) But this shows that t_0 is not really an upper bound to S. Therefore, $t_0 = 1$ and the lemma is proved. ■

With this lemma, the conclusion of the above corollary can be sharpened.

Corollary 6.9.7 *Suppose one of the Gerschgorin discs, D_i is disjoint from the union of the others. Then D_i contains exactly one eigenvalue of A and this eigenvalue is a simple root to the characteristic polynomial of A.*

Proof: In the proof of Corollary 6.9.5, note that a_{ii} is a simple root of $A(0)$ since otherwise the i^{th} Gerschgorin disc would not be disjoint from the others. Also, K, the connected component determined by a_{ii} must be contained in D_i because it is connected and by Gerschgorin's theorem above, $K \cap \sigma(A(t))$ must be contained in the union of the Gerschgorin discs. Since all the other eigenvalues of $A(0)$, the a_{jj}, are outside D_i, it follows that $K \cap \sigma(A(0)) = a_{ii}$. Therefore, by Lemma 6.9.6, $K \cap \sigma(A(1)) = K \cap \sigma(A)$ consists of a single simple eigenvalue. ■

Example 6.9.8 *Consider the matrix*

$$\begin{pmatrix} 5 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

The Gerschgorin discs are $D(5,1), D(1,2)$, and $D(0,1)$. Observe $D(5,1)$ is disjoint from the other discs. Therefore, there should be an eigenvalue in $D(5,1)$. The actual eigenvalues are not easy to find. They are the roots of the characteristic equation, $t^3 - 6t^2 + 3t + 5 = 0$. The numerical values of these are $-.669\,66, 1.423\,1$, and $5.246\,55$, verifying the predictions of Gerschgorin's theorem.

6.10 Exercises

1. Explain why it is typically impossible to compute the upper triangular matrix whose existence is guaranteed by Schur's theorem.

2. Now recall the QR factorization of Theorem 5.7.5 on Page 139. The QR algorithm is a technique which does compute the upper triangular matrix in Schur's theorem. There is much more to the QR algorithm than will be presented here. In fact, what I am about to show you is not the way it is done in practice. One first obtains what is called a Hessenburg matrix for which the algorithm will work better. However, the idea is as follows. Start with A an $n \times n$ matrix having real eigenvalues. Form $A = QR$ where Q is orthogonal and R is upper triangular. (Right triangular.) This can be done using the technique of Theorem 5.7.5 using Householder matrices. Next take $A_1 \equiv RQ$. Show that $A = QA_1Q^T$. In other words these two matrices, A, A_1 are similar. Explain why they have the same eigenvalues. Continue by letting A_1 play the role of A. Thus the algorithm is of the form $A_n = QR_n$ and $A_{n+1} = R_{n+1}Q$. Explain why $A = Q_nA_nQ_n^T$ for some Q_n orthogonal. Thus A_n is a sequence of matrices each similar to A. The remarkable thing is that often these matrices converge to an upper triangular matrix T and $A = QTQ^T$ for some orthogonal matrix, the limit of the Q_n where the limit means the entries converge. Then the process computes the upper triangular Schur form of the matrix A. Thus the eigenvalues of A appear on the diagonal of T. You will see approximately what these are as the process continues.

3. $\uparrow$Try the QR algorithm on
$$\begin{pmatrix} -1 & -2 \\ 6 & 6 \end{pmatrix}$$
which has eigenvalues 3 and 2. I suggest you use a computer algebra system to do the computations.

4. $\uparrow$Now try the QR algorithm on
$$\begin{pmatrix} 0 & -1 \\ 2 & 0 \end{pmatrix}$$

Show that the algorithm cannot converge for this example. **Hint:** Try a few iterations of the algorithm. Use a computer algebra system if you like.

5. ↑Show the two matrices $A \equiv \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix}$ and $B \equiv \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}$ are similar; that is there exists a matrix S such that $A = S^{-1}BS$ but there is no orthogonal matrix Q such that $Q^T BQ = A$. Show the QR algorithm does converge for the matrix B although it fails to do so for A.

6. Let F be an $m \times n$ matrix. Show that F^*F has all real eigenvalues and furthermore, they are all nonnegative.

7. If A is a real $n \times n$ matrix and λ is a complex eigenvalue $\lambda = a + ib, b \neq 0$, of A having eigenvector $\mathbf{z} + i\mathbf{w}$, show that $\mathbf{w} \neq \mathbf{0}$.

8. Suppose $A = Q^T DQ$ where Q is an orthogonal matrix and all the matrices are real. Also D is a diagonal matrix. Show that A must be symmetric.

9. Suppose A is an $n \times n$ matrix and there exists a unitary matrix U such that
$$A = U^* DU$$
where D is a diagonal matrix. Explain why A must be normal.

10. If A is Hermitian, show that $\det(A)$ must be real.

11. Show that every unitary matrix preserves distance. That is, if U is unitary,
$$|U\mathbf{x}| = |\mathbf{x}| .$$

12. Show that if a matrix does preserve distances, then it must be unitary.

13. ↑Show that a complex normal matrix A is unitary if and only if its eigenvalues have magnitude equal to 1.

14. Suppose A is an $n \times n$ matrix which is diagonally dominant. Recall this means
$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|$$
show A^{-1} must exist.

15. Give some disks in the complex plane whose union contains all the eigenvalues of the matrix
$$\begin{pmatrix} 1+2i & 4 & 2 \\ 0 & i & 3 \\ 5 & 6 & 7 \end{pmatrix}$$

16. Show a square matrix is invertible if and only if it has no zero eigenvalues.

17. Using Schur's theorem, show the trace of an $n \times n$ matrix equals the sum of the eigenvalues and the determinant of an $n \times n$ matrix is the product of the eigenvalues.

18. Using Schur's theorem, show that if A is any complex $n \times n$ matrix having eigenvalues $\{\lambda_i\}$ listed according to multiplicity, then $\sum_{i,j} |A_{ij}|^2 \geq \sum_{i=1}^n |\lambda_i|^2$. Show that equality holds if and only if A is normal. This inequality is called Schur's inequality. [20]

19. Here is a matrix.
$$\begin{pmatrix} 1234 & 6 & 5 & 3 \\ 0 & -654 & 9 & 123 \\ 98 & 123 & 10,000 & 11 \\ 56 & 78 & 98 & 400 \end{pmatrix}$$
I know this matrix has an inverse before doing any computations. How do I know?

20. Show the critical points of the following function are
$$(0,-3,0)\,,(2,-3,0)\,,\text{and }\left(1,-3,-\frac{1}{3}\right)$$
and classify them as local minima, local maxima or saddle points.
$f(x,y,z) = -\frac{3}{2}x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2}z^2$.

21. Here is a function of three variables.
$$f(x,y,z) = 13x^2 + 2xy + 8xz + 13y^2 + 8yz + 10z^2$$
change the variables so that in the new variables there are no mixed terms, terms involving xy, yz etc. Two eigenvalues are 12 and 18.

22. Here is a function of three variables.
$$f(x,y,z) = 2x^2 - 4x + 2 + 9yx - 9y - 3zx + 3z + 5y^2 - 9zy - 7z^2$$
change the variables so that in the new variables there are no mixed terms, terms involving xy, yz etc. The eigenvalues of the matrix which you will work with are $-\frac{17}{2}, \frac{19}{2}, -1$.

23. Here is a function of three variables.
$$f(x,y,z) = -x^2 + 2xy + 2xz - y^2 + 2yz - z^2 + x$$
change the variables so that in the new variables there are no mixed terms, terms involving xy, yz etc.

24. Show the critical points of the function,
$$f(x,y,z) = -2yx^2 - 6yx - 4zx^2 - 12zx + y^2 + 2yz.$$
are points of the form,
$$(x,y,z) = \left(t, 2t^2 + 6t, -t^2 - 3t\right)$$
for $t \in \mathbb{R}$ and classify them as local minima, local maxima or saddle points.

25. Show the critical points of the function
$$f(x,y,z) = \frac{1}{2}x^4 - 4x^3 + 8x^2 - 3zx^2 + 12zx + 2y^2 + 4y + 2 + \frac{1}{2}z^2.$$
are $(0,-1,0)\,,(4,-1,0)\,,$ and $(2,-1,-12)$ and classify them as local minima, local maxima or saddle points.

26. Let $f(x,y) = 3x^4 - 24x^2 + 48 - yx^2 + 4y$. Find and classify the critical points using the second derivative test.

27. Let $f(x,y) = 3x^4 - 5x^2 + 2 - y^2x^2 + y^2$. Find and classify the critical points using the second derivative test.

28. Let $f(x,y) = 5x^4 - 7x^2 - 2 - 3y^2x^2 + 11y^2 - 4y^4$. Find and classify the critical points using the second derivative test.

29. Let $f(x,y,z) = -2x^4 - 3yx^2 + 3x^2 + 5x^2z + 3y^2 - 6y + 3 - 3zy + 3z + z^2$. Find and classify the critical points using the second derivative test.

30. Let $f(x,y,z) = 3yx^2 - 3x^2 - x^2z - y^2 + 2y - 1 + 3zy - 3z - 3z^2$. Find and classify the critical points using the second derivative test.

31. Let Q be orthogonal. Find the possible values of $\det(Q)$.

32. Let U be unitary. Find the possible values of $\det(U)$.

33. If a matrix is nonzero can it have only zero for eigenvalues?

34. A matrix A is called nilpotent if $A^k = 0$ for some positive integer k. Suppose A is a nilpotent matrix. Show it has only 0 for an eigenvalue.

35. If A is a nonzero nilpotent matrix, show it must be defective.

36. Suppose A is a nondefective $n \times n$ matrix and its eigenvalues are all either 0 or 1. Show $A^2 = A$. Could you say anything interesting if the eigenvalues were all either 0,1,or -1? By DeMoivre's theorem, an n^{th} root of unity is of the form

$$\left(\cos\left(\frac{2k\pi}{n}\right) + i\sin\left(\frac{2k\pi}{n}\right)\right)$$

Could you generalize the sort of thing just described to get $A^n = A$? **Hint:** Since A is nondefective, there exists S such that $S^{-1}AS = D$ where D is a diagonal matrix.

37. This and the following problems will present most of a differential equations course. Most of the explanations are given. You fill in any details needed. 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 $y = y_0e^{a(t-t_0)}$. Next suppose

$$y' = (a+ib)\,y,\ y(t_0) = y_0 \tag{6.21}$$

where $y(t) = u(t) + iv(t)$. Show there exists a unique solution and it is given by $y(t) =$

$$y_0e^{a(t-t_0)}\left(\cos b(t-t_0) + i\sin b(t-t_0)\right) \equiv e^{(a+ib)(t-t_0)}y_0. \tag{6.22}$$

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

$$y' = (a + ib)\, y,\ y(t_0) = 0$$

and verify this requires $y(t) = 0$. To do this, note

$$\overline{y}' = (a - ib)\, \overline{y},\ \overline{y}(t_0) = 0$$

and that $|y|^2 (t_0) = 0$ and

$$\frac{d}{dt}|y(t)|^2 = y'(t)\,\overline{y}(t) + \overline{y}'(t)\, y(t)$$

$$= (a + ib)\, y(t)\,\overline{y}(t) + (a - ib)\,\overline{y}(t)\, y(t) = 2a\,|y(t)|^2.$$

Thus from the first part $|y(t)|^2 = 0e^{-2at} = 0$. Finally observe by a simple computation that 6.21 is solved by 6.22. 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.

38. 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

$$\mathbf{y}' = T\mathbf{y}, \mathbf{y}(t_0) = Q^{-1}\mathbf{x}_0 \tag{6.23}$$

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 6.23.

39. Now suppose $\Phi(t)$ is an $n \times n$ matrix of the form

$$\Phi(t) = \begin{pmatrix} \mathbf{x}_1(t) & \cdots & \mathbf{x}_n(t) \end{pmatrix} \tag{6.24}$$

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 6.24. Also explain why if $\mathbf{c} \in \mathbb{F}^n, \mathbf{y}(t) \equiv \Phi(t)\,\mathbf{c}$ solves the equation $\mathbf{y}'(t) = A\mathbf{y}(t)$.

40. In the above problem, consider the question whether all solutions to
$$\mathbf{x}' = A\mathbf{x} \tag{6.25}$$
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 6.10.1 *Suppose $\Phi(t)$ is an $n \times n$ matrix which satisfies $\Phi'(t) = A\Phi(t)$. Then the general solution to 6.25 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 the equation $\mathbf{c} = \Phi(t_0)\mathbf{x}$. By the existence part of Problem 38 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.

41. 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
$$\mathbf{x}' = A\mathbf{x} + \mathbf{f},\ \mathbf{x}(t_0) = \mathbf{x}_0 \tag{6.26}$$
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: Uniquenss 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 $\mathbf{u}' = A\mathbf{u},\ u(t_0) = \mathbf{0}$. Then use Problem 38. 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.

42. Show there exists a special Φ such that $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$, and suppose $\Phi(t)^{-1}$ 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 6.26 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.

43. You can see more on this problem and the next one in the latest version of Horn and Johnson, [17]. Two $n \times n$ matrices A, B are said to be congruent if there is an invertible P such that

$$B = PAP^*$$

Let A be a Hermitian matrix. Thus it has all real eigenvalues. Let n_+ be the number of positive eigenvalues, n_-, the number of negative eigenvalues and n_0 the number of zero eigenvalues. For k a positive integer, let I_k denote the $k \times k$ identity matrix and O_k the $k \times k$ zero matrix. Then the inertia matrix of A is the following block diagonal $n \times n$ matrix.

$$\begin{pmatrix} I_{n_+} & & \\ & I_{n_-} & \\ & & O_{n_0} \end{pmatrix}$$

Show that A is congruent to its inertia matrix. Next show that congruence is an equivalence relation on the set of Hermitian matrices. Finally, show that if two Hermitian matrices have the same inertia matrix, then they must be congruent. **Hint:** First recall that there is a unitary matrix, U such that

$$U^*AU = \begin{pmatrix} D_{n_+} & & \\ & D_{n_-} & \\ & & O_{n_0} \end{pmatrix}$$

where the D_{n_+} is a diagonal matrix having the positive eigenvalues of A, D_{n_-} being defined similarly. Now let $\left|D_{n_-}\right|$ denote the diagonal matrix which replaces each entry of D_{n_-} with its absolute value. Consider the two diagonal matrices

$$D = D^* = \begin{pmatrix} D_{n_+}^{-1/2} & & \\ & \left|D_{n_-}\right|^{-1/2} & \\ & & I_{n_0} \end{pmatrix}$$

Now consider D^*U^*AUD.

44. Show that if A, B are two congruent Hermitian matrices, then they have the same inertia matrix. **Hint:** Let $A = SBS^*$ where S is invertible. Show that A, B have the same rank and this implies that they are each unitarily similar to a diagonal matrix which has the same number of zero entries on the main diagonal. Therefore, letting V_A be the span of the eigenvectors associated with positive eigenvalues of A and V_B being defined similarly, it suffices to show that these have the same dimensions. Show that $(A\mathbf{x}, \mathbf{x}) > 0$ for all $\mathbf{x} \in V_A$. Next consider S^*V_A. For $\mathbf{x} \in V_A$, explain why

$$\begin{aligned}(BS^*\mathbf{x}, S^*\mathbf{x}) &= \left(S^{-1}A\left(S^*\right)^{-1}S^*\mathbf{x}, S^*\mathbf{x}\right) \\ &= \left(S^{-1}A\mathbf{x}, S^*\mathbf{x}\right) = \left(A\mathbf{x}, \left(S^{-1}\right)^* S^*\mathbf{x}\right) = (A\mathbf{x}, \mathbf{x}) > 0\end{aligned}$$

Next explain why this shows that S^*V_A is a subspace of V_B and so the dimension of V_B is at least as large as the dimension of V_A. Hence there are at least as many positive eigenvalues for B as there are for A. Switching A, B you can turn the inequality around. Thus the two have the same inertia matrix.

45. Let A be an $m \times n$ matrix. Then if you unraveled it, you could consider it as a vector in $\mathbb{C}^{nm}$. The Frobenius inner product on the vector space of $m \times n$ matrices is defined as

$$(A, B) \equiv \operatorname{trace}(AB^*)$$

Show that this really does satisfy the axioms of an inner product space and that it also amounts to nothing more than considering $m \times n$ matrices as vectors in $\mathbb{C}^{nm}$.

46. ↑Consider the $n \times n$ unitary matrices. Show that whenever U is such a matrix, it follows that

$$|U|_{\mathbb{C}^{nn}} = \sqrt{n}$$

Next explain why if $\{U_k\}$ is any sequence of unitary matrices, there exists a subsequence $\{U_{k_m}\}_{m=1}^{\infty}$ such that $\lim_{m\to\infty} U_{k_m} = U$ where U is unitary. Here the limit takes place in the sense that the entries of U_{k_m} converge to the corresponding entries of U.

47. ↑Let A, B be two $n \times n$ matrices. Denote by $\sigma(A)$ the set of eigenvalues of A. Define

$$\operatorname{dist}(\sigma(A), \sigma(B)) = \max_{\lambda \in \sigma(A)} \min\{|\lambda - \mu| : \mu \in \sigma(B)\}$$

Explain why $\operatorname{dist}(\sigma(A), \sigma(B))$ is small if and only if every eigenvalue of A is close to some eigenvalue of B. Now prove the following theorem using the above problem and Schur's theorem. This theorem says roughly that if A is close to B then the eigenvalues of A are close to those of B in the sense that every eigenvalue of A is close to an eigenvalue of B.

Theorem 6.10.2 *Suppose* $\lim_{k\to\infty} A_k = A$. *Then*

$$\lim_{k\to\infty} \operatorname{dist}(\sigma(A_k), \sigma(A)) = 0$$

48. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a 2×2 matrix which is not a multiple of the identity. Show that A is similar to a 2×2 matrix which has at least one diagonal entry equal to 0.

Hint: First note that there exists a vector $\mathbf{a}$ such that $A\mathbf{a}$ is not a multiple of $\mathbf{a}$. Then consider

$$B = \begin{pmatrix} \mathbf{a} & A\mathbf{a} \end{pmatrix}^{-1} A \begin{pmatrix} \mathbf{a} & A\mathbf{a} \end{pmatrix}$$

Show B has a zero on the main diagonal.

49. $\uparrow$ Let A be a complex $n \times n$ matrix which has trace equal to 0. Show that A is similar to a matrix which has all zeros on the main diagonal. **Hint:** Use Problem 30 on Page 128 to argue that you can say that a given matrix is similar to one which has the diagonal entries permuted in any order desired. Then use the above problem and block multiplication to show that if the A has k nonzero entries, then it is similar to a matrix which has $k-1$ nonzero entries. Finally, when A is similar to one which has at most one nonzero entry, this one must also be zero because of the condition on the trace.

50. $\uparrow$An $n \times n$ matrix X is a comutator if there are $n \times n$ matrices A, B such that $X = AB - BA$. Show that the trace of any comutator is 0. Next show that if a complex matrix X has trace equal to 0, then it is in fact a comutator. **Hint:** Use the above problem to show that it suffices to consider X having all zero entries on the main diagonal. Then define

$$A = \begin{pmatrix} 1 & & & 0 \\ & 2 & & \\ & & \ddots & \\ 0 & & & n \end{pmatrix}, \ B_{ij} = \begin{cases} \frac{X_{ij}}{i-j} \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}$$

6.11 Cauchy's Interlacing Theorem for Eigenvalues

Recall that every Hermitian matrix has all real eigenvalues. The Cauchy interlacing theorem compares the location of the eigenvalues of a Hermitian matrix with the eigenvalues of a principal submatrix. It is an extremely interesting theorem.

Theorem 6.11.1 *Let A be a Hermitian $n \times n$ matrix and let*

$$A = \begin{pmatrix} a & \mathbf{y}^* \\ \mathbf{y} & B \end{pmatrix}$$

where B is $(n-1) \times (n-1)$. Let the eigenvalues of B be $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{n-1}$. Then if the eigenvalues of A are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, it follows that $\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-1} \leq \lambda_n$.

Proof: First note that B is Hermitian because

$$A^* = \begin{pmatrix} \overline{a} & \mathbf{y}^* \\ \mathbf{y} & B^* \end{pmatrix} = A = \begin{pmatrix} a & \mathbf{y}^* \\ \mathbf{y} & B \end{pmatrix}$$

It is easiest to consider the case where strict inequality holds for the eigenvalues for B so first is an outline of reducing to this case.

There exists U unitary, depending on B such that $U^*BU = D$ where

$$D = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_{n-1} \end{pmatrix}$$

Now let $\{\varepsilon_k\}$ be a decreasing sequence of very small positive numbers converging to 0 and let B_k be defined by

$$U^* B_k U = D_k, \quad D_k \equiv \begin{pmatrix} \mu_1 + \varepsilon_k & & & 0 \\ & \mu_2 + 2\varepsilon_k & & \\ & & \ddots & \\ 0 & & & \mu_{n-1} + (n-1)\,\varepsilon_k \end{pmatrix}$$

where U is the above unitary matrix. Thus the eigenvalues of $B_k, \hat{\mu}_1 < \cdots < \hat{\mu}_{n-1}$ are strictly increasing and $\hat{\mu}_j \equiv \mu_j + j\varepsilon_k$. Let A_k be given by

$$A_k = \begin{pmatrix} a & \mathbf{y}^* \\ \mathbf{y} & B_k \end{pmatrix}$$

Then

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U^* \end{pmatrix} A_k \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U \end{pmatrix} &= \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U^* \end{pmatrix} \begin{pmatrix} a & \mathbf{y}^* \\ \mathbf{y} & B_k \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U \end{pmatrix} \\ &= \begin{pmatrix} a & \mathbf{y}^* \\ U^*\mathbf{y} & U^* B_k \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & U \end{pmatrix} = \begin{pmatrix} a & \mathbf{y}^* U \\ U^*\mathbf{y} & D_k \end{pmatrix} \end{aligned}$$

We can replace $\mathbf{y}$ in the statement of the theorem with $\mathbf{y}_k$ such that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{y}$ but $\mathbf{z}_k \equiv U^*\mathbf{y}_k$ has the property that each component of $\mathbf{z}_k$ is nonzero. This will probably take place automatically but if not, make the change. This makes a change in A_k but still $\lim_{k\to\infty} A_k = A$. The main part of this argument which follows has to do with fixed k.

Expanding $\det(\lambda I - A_k)$ along the top row, the characteristic polynomial for A_k is then

$$q(\lambda) = (\lambda - a) \prod_{i=1}^{n-1} (\lambda - \hat{\mu}_i) - \sum_{i=2}^{n-1} |z_i|^2 (\lambda - \hat{\mu}_1) \cdots \widehat{(\lambda - \hat{\mu}_i)} \cdots (\lambda - \hat{\mu}_{n-1}) \tag{6.27}$$

where $\widehat{(\lambda - \hat{\mu}_i)}$ indicates that this factor is omitted from the product $\prod_{i=1}^{n-1} (\lambda - \hat{\mu}_i)$. To see why this is so, consider the case where B_k is 3×3. In this case, you would have

$$\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & U^* \end{pmatrix} (\lambda I - A_k) \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & U \end{pmatrix} = \begin{pmatrix} \lambda - a & \overline{z}_1 & \overline{z}_2 & \overline{z}_3 \\ z_1 & \lambda - \hat{\mu}_1 & 0 & 0 \\ z_2 & 0 & \lambda - \hat{\mu}_2 & 0 \\ z_3 & 0 & 0 & \lambda - \hat{\mu}_3 \end{pmatrix}$$

In general, you would have an $n \times n$ matrix on the right with the same appearance. Then expanding as indicated, the determinant is

$$\begin{gathered} (\lambda - a) \prod_{i=1}^{3} (\lambda - \hat{\mu}_i) - \overline{z}_1 \det \begin{pmatrix} z_1 & 0 & 0 \\ z_2 & \lambda - \hat{\mu}_2 & 0 \\ z_3 & 0 & \lambda - \hat{\mu}_3 \end{pmatrix} \\ + \overline{z}_2 \det \begin{pmatrix} z_1 & \lambda - \hat{\mu}_1 & 0 \\ z_2 & 0 & 0 \\ z_3 & 0 & \lambda - \hat{\mu}_3 \end{pmatrix} - \overline{z}_3 \det \begin{pmatrix} z_1 & \lambda - \hat{\mu}_1 & 0 \\ z_2 & 0 & \lambda - \hat{\mu}_2 \\ z_3 & 0 & 0 \end{pmatrix} \end{gathered}$$

$$= (\lambda - a)\prod_{i=1}^{3}(\lambda - \hat{\mu}_i) - \left(\begin{array}{c} |z_1|^2 (\lambda - \hat{\mu}_2)(\lambda - \hat{\mu}_3) + |z_2|^2 (\lambda - \hat{\mu}_1)(\lambda - \hat{\mu}_3) \\ + |z_3|^2 (\lambda - \hat{\mu}_1)(\lambda - \hat{\mu}_2) \end{array} \right)$$

Notice how, when you expand the 3×3 determinants along the first column, you have only one non-zero term and the sign is adjusted to give the above claim. Clearly, it works the same for any size matrix. Since the $\hat{\mu}_i$ are strictly increasing in i, it follows from 6.27 that $q(\hat{\mu}_i) q(\hat{\mu}_{i+1}) \leq 0$. However, since each $|z_i| \neq 0$, none of the $q(\hat{\mu}_i)$ can equal 0 and so $q(\hat{\mu}_i) q(\hat{\mu}_{i+1}) < 0$. Hence, from the intermediate value theorem of calculus, there is a root of $q(\lambda)$ in each of the disjoint open intervals $(\hat{\mu}_i, \hat{\mu}_{i+1})$. There are $n-2$ of these intervals and so this accounts for $n-2$ roots of $q(\lambda)$.

$$q(\lambda) = (\lambda - a)\prod_{i=1}^{n-1}(\lambda - \hat{\mu}_i) - \sum_{i=2}^{n-1} |z_i|^2 (\lambda - \hat{\mu}_1) \cdots \widehat{(\lambda - \hat{\mu}_i)} \cdots (\lambda - \hat{\mu}_{n-1})$$

What of $q(\hat{\mu}_1)$? Its sign is the same as $(-1)^{n-3}$ and also $q(\hat{\mu}_{n-1}) < 0$. Therefore, there is a root to $q(\lambda)$ which is larger than $\hat{\mu}_{n-1}$. Indeed, $\lim_{\lambda \to \infty} q(\lambda) = \infty$ so there exists a root of $q(\lambda)$ strictly larger than $\hat{\mu}_{n-1}$. This accounts for $n-1$ roots of $q(\lambda)$. Now consider $q(\hat{\mu}_1)$. Suppose first that n is odd. Then you have $q(\hat{\mu}_1) > 0$. Hence, there is a root of $q(\lambda)$ which is no larger than $\hat{\mu}_1$ because in this case, $\lim_{\lambda \to -\infty} q(\lambda) = -\infty$. If n is even, then $q(\hat{\mu}_1) < 0$ and so there is a root of $q(\lambda)$ which is smaller than $\hat{\mu}_1$ because in this case, $\lim_{\lambda \to -\infty} q(\lambda) = \infty$. This accounts for all roots of $q(\lambda)$. Hence, if the roots of $q(\lambda)$ are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, it follows that

$$\lambda_1 < \hat{\mu}_1 < \lambda_2 < \hat{\mu}_2 < \cdots < \hat{\mu}_{n-1} < \lambda_n$$

To get the complete result, simply take the limit as $k \to \infty$. Then $\lim_{k \to \infty} \hat{\mu}_k = \mu_k$ and $A_k \to A$ and so the eigenvalues of A_k converge to the corresponding eigenvalues of A (See Problem 47 on Page 184), and so, passing to the limit, gives the desired result in which it may be necessary to replace $<$ with $\leq$. ■

Definition 6.11.2 *Let A be an $n \times n$ matrix. An $(n-r) \times (n-r)$ matrix is called a principal submatrix of A if it is obtained by deleting from A the rows $i_1, i_2, \cdots, i_r$ and the columns $i_1, i_2, \cdots, i_r$.*

Now the Cauchy interlacing theorem is really the following corollary.

Corollary 6.11.3 *Let A be an $n \times n$ Hermitian matrix and let B be an $(n-1) \times (n-1)$ principal submatrix. Then the interlacing inequality holds $\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{n-1} \leq \lambda_n$ where the μ_i are the eigenvalues of B listed in increasing order and the λ_i are the eigenvalues of A listed in increasing order.*

Proof: Suppose B is obtained from A by deleting the i^{th} row and the i^{th} column. Then let P be the permutation matrix which switches the i^{th} row with the first row. It is an orthogonal matrix and so its inverse is its transpose. The transpose switches the i^{th} column with the first column. See Problem 33 on Page 128. Thus $PAP^T = \begin{pmatrix} a & \mathbf{y}^* \\ \mathbf{y} & B \end{pmatrix}$ and it follows that the result of the multiplication is indeed as shown, a Hermitian matrix because P, P^T are orthogonal matrices. Now the conclusion of the corollary follows from Theorem 6.11.1. ■

Chapter 7

Vector Spaces and Fields

7.1 Vector Space Axioms

It is time to consider the idea of a vector space.

Definition 7.1.1 *A vector space is an Abelian group of "vectors" satisfying the axioms of an Abelian group,*

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

the commutative law of addition,

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

the associative law for addition,

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

the existence of an additive identity,

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

the existence of an additive inverse, along with a field of "scalars", $\mathbb{F}$ which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)

$$\alpha(\mathbf{v}+\mathbf{w})=\alpha\mathbf{v}+\alpha\mathbf{w}, \tag{7.1}$$

$$(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}, \tag{7.2}$$

$$\alpha(\beta\mathbf{v})=\alpha\beta(\mathbf{v}), \tag{7.3}$$

$$1\mathbf{v}=\mathbf{v}. \tag{7.4}$$

The field of scalars is usually $\mathbb{R}$ or $\mathbb{C}$ and the vector space will be called real or complex depending on whether the field is $\mathbb{R}$ or $\mathbb{C}$. However, other fields are also possible. For example, one could use the field of rational numbers or even the field of the integers mod p for p a prime. A vector space is also called a linear space.

For example, $\mathbb{R}^n$ with the usual conventions is an example of a real vector space and $\mathbb{C}^n$ is an example of a complex vector space. Up to now, the discussion has been for $\mathbb{R}^n$ or $\mathbb{C}^n$ and all that is taking place is an increase in generality and abstraction.

There are many examples of vector spaces.

Example 7.1.2 *Let Ω be a nonempty set and let V consist of all functions defined on Ω which have values in some field $\mathbb{F}$. The vector operations are defined as follows.*

$$\begin{aligned}(f+g)(x) &= f(x)+g(x)\\ (\alpha f)(x) &= \alpha f(x)\end{aligned}$$

Then it is routine to verify that V with these operations is a vector space.

Note that $\mathbb{F}^n$ actually fits in to this framework. You consider the set Ω to be $\{1,2,\cdots,n\}$ and then the mappings from Ω to $\mathbb{F}$ give the elements of $\mathbb{F}^n$. Thus a typical vector can be considered as a function.

Example 7.1.3 *Generalize the above example by letting V denote all functions defined on Ω which have values in a vector space W which has field of scalars $\mathbb{F}$. The definitions of scalar multiplication and vector addition are identical to those of the above example.*

7.2 Subspaces and Bases

7.2.1 Basic Definitions

Definition 7.2.1 *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 subset, $W \subseteq V$ *is said to be a subspace if it is also a vector space with the same field of scalars. Thus* $W \subseteq V$ *for* W *nonempty is a subspace if* $ax + by \in W$ *whenever* $a, b \in \mathbb{F}$ *and* $x, y \in W$. *The span of a set of vectors as just described is an example of a subspace.*

Example 7.2.2 *Consider the real valued functions defined on an interval* $[a, b]$. *A subspace is the set of continuous real valued functions defined on the interval. Another subspace is the set of polynomials of degree no more than 4.*

Definition 7.2.3 *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*

$$\operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_n) = V$$

and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is linearly independent. The set of vectors is linearly dependent if it is not linearly independent.*

7.2.2 A Fundamental Theorem

The next theorem is called the exchange theorem. It is very important that you understand this theorem. It is so important that I have given several proofs of it. Some amount to the same thing, just worded differently.

Theorem 7.2.4 *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: Define $\operatorname{span}\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{7.5}$$

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

Say $c_k \neq 0$. Then solve 7.5 for $\mathbf{y}_k$ and obtain

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

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

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

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

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

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

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

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{7.6}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (7.6) 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 2: Let

$$\mathbf{x}_k = \sum_{j=1}^{s} a_{jk} \mathbf{y}_j$$

If $r > s$, then the matrix $A = (a_{jk})$ has more columns than rows. By Corollary 4.3.9 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

$$\sum_{k=1}^{r} c_k \mathbf{x}_k = \sum_{k=1}^{r} c_k \sum_{j=1}^{s} a_{jk} \mathbf{y}_j = \sum_{j=1}^{s} \left(\sum_{k=1}^{r} c_k a_{jk} \right) \mathbf{y}_j = \mathbf{0}$$

which contradicts the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent. Hence $r \leq s$.

Proof 3: 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

$$\operatorname{span}(\mathbf{y}_1, \cdots, \mathbf{y}_s) = \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_m, \mathbf{z}_1, \cdots, \mathbf{z}_j)$$

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 them in terms of $\mathbf{x}_{m+1}, \mathbf{x}_m, \cdots, \mathbf{x}_1$, and the other $\mathbf{z}_k$. Therefore, there exists

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{j-1}\} \subseteq \{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$$

such that

$$\operatorname{span}(\mathbf{y}_1, \cdots, \mathbf{y}_s) = \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_{m+1}, \mathbf{z}_1, \cdots, \mathbf{z}_{j-1})$$

contradicting the choice of j. Hence $j = 0$ and

$$\operatorname{span}(\mathbf{y}_1, \cdots, \mathbf{y}_s) = \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_s)$$

It follows that

$$\mathbf{x}_{s+1} \in \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_s)$$

contrary to the assumption the $\mathbf{x}_k$ are linearly independent. Therefore, $r \leq s$ as claimed. ■

Corollary 7.2.5 *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 7.2.4, $m \leq n$ and $n \leq m$. ■

Definition 7.2.6 *A vector space V is of dimension n if it has a basis consisting of n vectors. This is well defined thanks to Corollary 7.2.5. It is always assumed here that $n < \infty$ and in this case, such a vector space is said to be finite dimensional.*

Example 7.2.7 *Consider the polynomials defined on $\mathbb{R}$ of degree no more than 3, denoted here as P_3. Then show that a basis for P_3 is $\{1, x, x^2, x^3\}$. Here x^k symbolizes the function $x \mapsto x^k$.*

It is obvious that the span of the given vectors yields P_3. Why is this set of vectors linearly independent? Suppose

$$c_0 + c_1 x + c_2 x^2 + c_3 x^3 = 0$$

where 0 is the zero function which maps everything to 0. Then you could differentiate three times and obtain the following equations

$$\begin{aligned} c_1 + 2c_2 x + 3c_3 x^2 &= 0 \\ 2c_2 + 6c_3 x &= 0 \\ 6c_3 &= 0 \end{aligned}$$

Now this implies $c_3 = 0$. Then from the equations above the bottom one, you find in succession that $c_2 = 0, c_1 = 0, c_0 = 0$.

There is a somewhat interesting theorem about linear independence of smooth functions (those having plenty of derivatives) which I will show now. It is often used in differential equations.

Definition 7.2.8 *Let $f_1, \cdots, f_n$ be smooth functions defined on an interval $[a, b]$. The Wronskian of these functions is defined as follows.*

$$W(f_1, \cdots, f_n)(x) \equiv \begin{vmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}$$

Note that to get from one row to the next, you just differentiate everything in that row. The notation $f^{(k)}(x)$ denotes the k^{th} derivative.

With this definition, the following is the theorem. The interesting theorem involving the Wronskian has to do with the situation where the functions are solutions of a differential equation. Then much more can be said and it is much more interesting than the following theorem.

Theorem 7.2.9 *Let $\{f_1, \cdots, f_n\}$ be smooth functions defined on $[a, b]$. Then they are linearly independent if there exists some point $t \in [a, b]$ where $W(f_1, \cdots, f_n)(t) \neq 0$.*

Proof: Form the linear combination of these vectors (functions) and suppose it equals 0. Thus

$$a_1 f_1 + a_2 f_2 + \cdots + a_n f_n = 0$$

The question you must answer is whether this requires each a_j to equal zero. If they all must equal 0, then this means these vectors (functions) are independent. This is what it means to be linearly independent.

Differentiate the above equation $n - 1$ times yielding the equations

$$\begin{pmatrix} a_1 f_1 + a_2 f_2 + \cdots + a_n f_n = 0 \\ a_1 f_1' + a_2 f_2' + \cdots + a_n f_n' = 0 \\ \vdots \\ a_1 f_1^{(n-1)} + a_2 f_2^{(n-1)} + \cdots + a_n f_n^{(n-1)} = 0 \end{pmatrix}$$

Now plug in t. Then the above yields

$$\begin{pmatrix} f_1(t) & f_2(t) & \cdots & f_n(t) \\ f_1'(t) & f_2'(t) & \cdots & f_n'(t) \\ \vdots & \vdots & & \vdots \\ f_1^{(n-1)}(t) & f_2^{(n-1)}(t) & \cdots & f_n^{(n-1)}(t) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Since the determinant of the matrix on the left is assumed to be nonzero, it follows this matrix has an inverse and so the only solution to the above system of equations is to have each $a_k = 0$. ■

Here is a useful lemma.

Lemma 7.2.10 *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 $\sum_{i=1}^{k} c_i \mathbf{u}_i + d\mathbf{v} = 0$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$,

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

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

Given a spanning set, you can delete vectors till you end up with a basis. Given a linearly independent set, you can add vectors till you get a basis. This is what the following theorem is about, weeding and planting.

Theorem 7.2.11 *If $V = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ then some subset of $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is a basis for V. Also, if $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\} \subseteq V$ is linearly independent and the vector space is finite dimensional, then the set, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$, can be enlarged to obtain a basis of V.*

Proof: Let

$$S = \{E \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\} \text{ such that } \text{span}(E) = V\}.$$

For $E \in S$, let $|E|$ denote the number of elements of E. Let

$$m \equiv \min\{|E| \text{ such that } E \in S\}.$$

Thus there exist vectors

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_m\} \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$$

such that

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_m) = V$$

and m is as small as possible for this to happen. If this set is linearly independent, it follows it is a basis for V and the theorem is proved. On the other hand, if the set is not linearly independent, then there exist scalars

$$c_1, \cdots, c_m$$

such that

$$\mathbf{0} = \sum_{i=1}^{m} c_i \mathbf{v}_i$$

and not all the c_i are equal to zero. Suppose $c_k \neq 0$. Then the vector, $\mathbf{v}_k$ may be solved for in terms of the other vectors. Consequently,

$$V = \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_m)$$

contradicting the definition of m. This proves the first part of the theorem.

To obtain the second part, begin with $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ and suppose a basis for V is

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}.$$

If

$$\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k) = V,$$

then $k = n$. If not, there exists a vector,

$$\mathbf{u}_{k+1} \notin \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k).$$

Then by Lemma 7.2.10, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is also linearly independent. Continue adding vectors in this way until n linearly independent vectors have been obtained. Then

$$\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_n) = V$$

because if it did not do so, there would exist $\mathbf{u}_{n+1}$ as just described and $\{\mathbf{u}_1, \cdots, \mathbf{u}_{n+1}\}$ would be a linearly independent set of vectors having $n+1$ elements even though $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis. This would contradict Theorem 7.2.4. Therefore, this list is a basis. ■

7.2.3 The Basis of a Subspace

Every subspace of a finite dimensional vector space is a span of some vectors and in fact it has a basis. This is the content of the next theorem.

Theorem 7.2.12 *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 7.2.10 $\{\mathbf{v}_1, \mathbf{v}_2\}$ 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 7.2.4. ∎

7.3 Lots of Fields

7.3.1 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 7.3.1 *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. A polynomial of degree n is monic if the coefficient of λ^n is 1. In any case, this coefficient is called the leading coefficient.*

Example 7.3.2 *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. It is Lemma 1.10.10 on Page 25. We state it here for convenience.

Lemma 7.3.3 *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.

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 7.3.4 *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. In other words, it is not possible to write $f(\lambda) = a(\lambda) b(\lambda)$ where each of $a(\lambda), b(\lambda)$ have degree less than the degree of $f(\lambda)$.*

Proposition 7.3.5 *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 7.3.6 *Let $\psi(\lambda)$ be the greatest common divisor 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 7.3.3

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

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

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

$$\begin{aligned} &= \left(\phi_1(\lambda) - \sum_{i=1}^{p} r_i(\lambda) \phi_i(\lambda) l(\lambda)\right) a \\ &= \left((1 - r_1(\lambda)) \phi_1(\lambda) + \sum_{i=2}^{p} (-r_i(\lambda) l(\lambda)) \phi_i(\lambda)\right) a \end{aligned}$$

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 7.3.7 *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 7.3.8 *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 7.3.7, 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,

$$\begin{aligned} \psi(\lambda)\, n(\lambda) &\equiv \psi(\lambda)\, l(\lambda) \prod_{i=1}^{p} n_i(\lambda)^{k_i} = \prod_{i=1}^{p} \left(n_i(\lambda)\, \phi_i(\lambda)\right)^{k_i} \\ &= \prod_{i=1}^{p} \left(1 - \psi(\lambda)\, m_i(\lambda)\right)^{k_i} = 1 + g(\lambda)\, \psi(\lambda) \end{aligned}$$

for a polynomial $g(\lambda)$. Thus

$$1 = \psi(\lambda)\left(n(\lambda) - g(\lambda)\right)$$

which is impossible because $\psi(\lambda)$ is not equal to 1. ■

Now here is a simple lemma about canceling monic polynomials.

Lemma 7.3.9 *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 7.3.10 *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 7.3.8, each $\psi_i(\lambda)$ equals one of the $\phi_j(\lambda)$. By the above cancellation lemma, Lemma 7.3.9, 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 $a = b$. $\blacksquare$

The following corollary will be well used. This corollary seems rather believable but does require a proof.

Corollary 7.3.11 *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 7.3.10, 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. ■

7.3.2 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 7.3.12 *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 7.3.13 *$[x]$ 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 7.3.14 *Let $\sim$ be an equivalence relation defined on a set, S and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.*

Definition 7.3.15 *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 7.3.16 *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

$$a(x) - c(x) = a(x) - b(x) + b(x) - c(x)$$

$$= k(x)p(x) + l(x)p(x) = (l(x) + k(x))p(x)$$

and so $a(x) \sim c(x)$ and this shows the transitive law. ■

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 7.3.17 *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, and its leading coefficient is 1. For the similarity relation defined in Definition 7.3.15, 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 $a(x)$.*

$$[a(x)] + [b(x)] \equiv [a(x) + b(x)]$$

$$[a(x)][b(x)] \equiv [a(x)b(x)]$$

This collection of equivalence classes is sometimes denoted by $\mathbb{F}[x]/(p(x))$.

Proposition 7.3.18 *In the situation of Definition 7.3.17 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 $\mathbb{F}$ can be considered 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

$$[a(x)] = [a'(x)], [b(x)] = [b'(x)]$$

It is necessary to show

$$[a(x) + b(x)] = [a'(x) + b'(x)]$$

$$[a(x)b(x)] = [a'(x)b'(x)]$$

Consider the second of the two.

$$\begin{aligned} & a'(x)\, b'(x) - a(x)\, b(x) \\ = \;& a'(x)\, b'(x) - a(x)\, b'(x) + a(x)\, b'(x) - a(x)\, b(x) \\ = \;& b'(x)\, (a'(x) - a(x)) + a(x)\, (b'(x) - b(x)) \end{aligned}$$

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. ■

From the above theorem, it makes perfect sense to 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.

Remark 7.3.19 *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 7.3.20 *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 $\mathbb{Q}[x] / (x^2 - 2)$ is of the form $a + b[x]$ where $a, b \in \mathbb{Q}$ and $[x]^2 - 2 = 0$. Thus one can regard $[x]$ as $\sqrt{2}$. $\mathbb{Q}[x] / (x^2 - 2)$ is of the form $a + b\sqrt{2}$.*

In the above example, $[x^2 + x]$ is not zero because it is not a multiple of $x^2 - 2$. What is $[x^2 + x]^{-1}$? You know that the two polynomials are relatively prime and so there exists $n(x), m(x)$ such that

$$1 = n(x)\left(x^2 - 2\right) + m(x)\left(x^2 + x\right)$$

Thus $[m(x)] = [x^2 + x]^{-1}$. How could you find these polynomials? First of all, it suffices to consider only $n(x)$ and $m(x)$ having degree less than 2.

$$1 = (ax + b)\left(x^2 - 2\right) + (cx + d)\left(x^2 + x\right)$$

$$1 = ax^3 - 2b + bx^2 + cx^2 + cx^3 + dx^2 - 2ax + dx$$

Now you solve the resulting system of equations.

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

Then the desired inverse is $\left[-\frac{1}{2}x + 1\right]$. To check,

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

Thus $\left[-\frac{1}{2}x + 1\right]\left[x^2 + x\right] - [1] = [0]$.

The above is an example of something general described in the following definition.

Definition 7.3.21 *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 7.3.22 *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 $\{k_j\}_{j=1}^m$ 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 $\{k_j l_i\}$ is a spanning set. If

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ji} k_j l_i = 0$$

Then, since the l_i are independent, it follows that

$$\sum_{j=1}^{m} f_{ji} k_j = 0$$

and since $\{k_j\}$ is independent, each $f_{ji} = 0$ for each j for a given arbitrary i. Therefore, $\{k_j l_i\}$ is a basis. ■

Note that if $p(x)$ were not irreducible, then you could find a field extension $\mathbb{G}$ containing a root of $p(x)$ such that $[\mathbb{G} : \mathbb{F}] \leq n$. You could do this by working with an irreducible factor of $p(x)$.

Theorem 7.3.23 *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 $\mathbb{G}$ and $\{z_1, \cdots, z_n\}$ contained in $\mathbb{G}$, 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 7.3.18, 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 7.3.18, there is a larger field $\mathbb{G}_1$ such that $p(x)$ has a root a_1 in $\mathbb{G}_1$ and $[\mathbb{G}_1 : \mathbb{F}] \leq n$. Then

$$p(x) = (x - a_1) q(x)$$

Continue this way until the polynomial equals the product of linear factors. Then by Proposition 7.3.22 applied multiple times, $[\mathbb{G} : \mathbb{F}] \leq n!$. ■

Example 7.3.24 *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 7.3.25 *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 7.3.26 *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 $(ai+b)^{-1} = \frac{b-ai}{a^2+b^2}$.

7.3.3 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 7.3.27 *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 $p(a) = 0$.*

The next theorem is on the uniqueness of the minimal polynomial.

Theorem 7.3.28 *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 a 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 with smallest degree such that $q(a) = 0$. Then

$$q(x) = p(x) l(x) + r(x)$$

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 $q(x)$ has smallest degree, $l(x) = 1$ showing that $p(x) = q(x)$. ■

Definition 7.3.29 *For a an algebraic number, let $\deg(a)$ denote the degree of the minimal polynomial of a.*

Also, here is another definition.

Definition 7.3.30 *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 7.3.31 *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 of a 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. Hence $q(a) = r(a)$. 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 polynomials 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 7.2.4 this proves the first claim.

Finally consider the last claim. Let $g(a_1, \cdots, a_m)$ be a polynomial in $\{a_1, \cdots, a_m\}$ 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 7.3.32 *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 34 on Page 211 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 7.3.33 *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 7.3.31 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.

7.3.4 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 7.3.34 *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.

For a proof, you can see my book "Linear Algebra and Analysis".

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]

7.4 Exercises

1. 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.

2. 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.

[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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. Suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is a set of vectors from $\mathbb{F}^n$. Show that $\mathbf{0}$ is in $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k)$.

10. 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.

11. 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.

12. Consider the vectors of the form
$$\left\{ \begin{pmatrix} 2t + u + 1 \\ t + 3u \\ t + s + v \\ u \end{pmatrix} : s, t, u, v \in \mathbb{R} \right\}.$$
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.

13. Let V denote the set of functions defined on $[0, 1]$. Vector addition is defined as $(f + g)(x) \equiv f(x) + g(x)$ and scalar multiplication is defined as $(\alpha f)(x) \equiv \alpha(f(x))$. Verify V is a vector space. What is its dimension, finite or infinite? Justify your answer.

14. Let V denote the set of polynomial functions defined on $[0,1]$. Vector addition is defined as $(f+g)(x) \equiv f(x)+g(x)$ and scalar multiplication is defined as $(\alpha f)(x) \equiv \alpha(f(x))$. Verify V is a vector space. What is its dimension, finite or infinite? Justify your answer.

15. Let V be the set of polynomials defined on $\mathbb{R}$ having degree no more than 4. Give a basis for this vector space.

16. Let the vectors be of the form $a+b\sqrt{2}$ where a,b are rational numbers and let the field of scalars be $\mathbb{F}=\mathbb{Q}$, the rational numbers. Show directly this is a vector space. What is its dimension? What is a basis for this vector space?

17. Let V be a vector space with field of scalars $\mathbb{F}$ and suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for V. Now let W also be a vector space with field of scalars $\mathbb{F}$. Let $L : \{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \to W$ be a function such that $L\mathbf{v}_j = \mathbf{w}_j$. Explain how L can be extended to a linear transformation mapping V to W in a unique way.

18. 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.

19. If you have 6 vectors in $\mathbb{F}^5$, is it possible they are linearly independent? Explain.

20. 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.

21. Suppose V and W both have dimension equal to 7 and they are subspaces of a vector space of dimension 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.

22. 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.

23. If $\mathbf{b}\neq\mathbf{0}$, can the solution set of $A\mathbf{x}=\mathbf{b}$ be a plane through the origin? Explain.

24. 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.

25. 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.

26. 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.

27. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.

28. 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, $\operatorname{rank}(A)=4$ and $\operatorname{rank}(A|\mathbf{b})=4$. **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 $\text{rank}(A|\mathbf{b}) = 2$.

29. 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.

30. 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 that this would require the column rank to be less than n.

31. Explain why an $n \times n$ matrix A is both one to one and onto if and only if its rank is n.

32. If you have not done this already, here it is again. It is a very important result of Sylvester. Even if you have done it, a review is a good idea. 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

$$\ker(AB) \subseteq \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{z}_1, \cdots, \mathbf{z}_k).$$

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

$$\mathbf{x} - \sum_{i=1}^k \mathbf{z}_i \in \ker(B).$$

33. Recall that every positive integer can be factored into a product of primes in a unique way. Show there must be infinitely many primes. **Hint:** Show that if you have any finite set of primes and you multiply them and then add 1, the result cannot be divisible by any of the primes in your finite set. This idea in the hint is due to Euclid who lived about 300 B.C.

34. 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.

$$[a] + [b] \equiv [a + b]$$

$$[a][b] \equiv [ab]$$

Show these operations are well defined. That is, if $[a] = [a']$ and $[b] = [b']$, then $[a] + [b] = [a'] + [b']$ with a similar conclusion holding for multiplication. Thus for addition you need to verify $[a + b] = [a' + b']$ and for multiplication you need to verify $[ab] = [a'b']$. For example, if $p = 5$ you have $[3] = [8]$ and $[2] = [7]$. Is $[2 \times 3] = [8 \times 7]$? Is $[2 + 3] = [8 + 7]$? Clearly so in this example because when you subtract, the result is divisible by 5. So why is this so in general? Now verify that $\{[0], [1], \cdots, [p - 1]\}$ 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? Of course you could now consider field extensions based on these fields.

35. Suppose the field of scalars is $\mathbb{Z}_2$ described above. Show that

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

Thus the identity is a comutator. Compare this with Problem 50 on Page 185.

36. Suppose V is a vector space with field of scalars $\mathbb{F}$. Let $T \in \mathcal{L}(V, W)$, the space of linear transformations mapping V onto W where W is another vector space. Define an equivalence relation on V as follows. $\mathbf{v} \sim \mathbf{w}$ means $\mathbf{v} - \mathbf{w} \in \ker(T)$. Recall that $\ker(T) \equiv \{\mathbf{v} : T\mathbf{v} = \mathbf{0}\}$. Show this is an equivalence relation. Now for $[\mathbf{v}]$ an equivalence class define $T'[\mathbf{v}] \equiv T\mathbf{v}$. Show this is well defined. Also show that with the operations

$$\begin{aligned} [\mathbf{v}] + [\mathbf{w}] &\equiv [\mathbf{v} + \mathbf{w}] \\ \alpha[\mathbf{v}] &\equiv [\alpha\mathbf{v}] \end{aligned}$$

this set of equivalence classes, denoted by $V/\ker(T)$ is a vector space. Show next that $T' : V/\ker(T) \to W$ is one to one, linear, and onto. This new vector space, $V/\ker(T)$ is called a quotient space. Show its dimension equals the difference between the dimension of V and the dimension of $\ker(T)$.

37. Let V be an n dimensional vector space and let W be a subspace. Generalize the above problem to define and give properties of V/W. What is its dimension? What is a basis?

38. If $\mathbb{F}$ and $\mathbb{G}$ are two fields and $\mathbb{F} \subseteq \mathbb{G}$, can you consider $\mathbb{G}$ as a vector space with field of scalars $\mathbb{F}$? Explain.

39. Let $\mathbb{A}$ denote the real roots of polynomials in $\mathbb{Q}[x]$. Show $\mathbb{A}$ can be considered a vector space with field of scalars $\mathbb{Q}$. What is the dimension of this vector space, finite or infinite?

40. As mentioned, for distinct algebraic numbers α_i, the complex numbers $\{e^{\alpha_i}\}_{i=1}^n$ are linearly independent over the field of scalars $\mathbb{A}$ where $\mathbb{A}$ denotes the algebraic numbers, those which are roots of a polynomial having integer (rational) coefficients. What is the dimension of the vector space $\mathbb{C}$ with field of scalars $\mathbb{A}$, finite or infinite? If the field of scalars were $\mathbb{C}$ instead of $\mathbb{A}$, would this change? What if the field of scalars were $\mathbb{R}$?

41. Suppose $\mathbb{F}$ is a countable field and let $\mathbb{A}$ be the algebraic numbers, those numbers in $\mathbb{G}$ which are roots of a polynomial in $\mathbb{F}[x]$. Show $\mathbb{A}$ is also countable.

42. This problem is on partial fractions. Suppose you have

$$R(x) = \frac{p(x)}{q_1(x)\cdots q_m(x)}, \text{ degree of } p(x) < \text{ degree of denominator.}$$

where the polynomials $q_i(x)$ are relatively prime and all the polynomials $p(x)$ and $q_i(x)$ have coefficients in a field of scalars $\mathbb{F}$. Thus there exist polynomials $a_i(x)$ having coefficients in $\mathbb{F}$ such that

$$1 = \sum_{i=1}^m a_i(x) q_i(x)$$

Explain why

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

Now continue doing this on each term in the above sum till finally you obtain an expression of the form

$$\sum_{i=1}^m \frac{b_i(x)}{q_i(x)}$$

Using the Euclidean algorithm for polynomials, explain why the above is of the form

$$M(x) + \sum_{i=1}^m \frac{r_i(x)}{q_i(x)}$$

where the degree of each $r_i(x)$ is less than the degree of $q_i(x)$ and $M(x)$ is a polynomial. Now argue that $M(x) = 0$. From this explain why the usual partial fractions expansion of calculus must be true. You can use the fact that every polynomial having real coefficients factors into a product of irreducible quadratic polynomials and linear polynomials having real coefficients. This follows from the fundamental theorem of algebra in the appendix.

43. Suppose $\{f_1, \cdots, f_n\}$ is an independent set of smooth functions defined on some interval (a, b). Now let A be an invertible $n \times n$ matrix. Define new functions $\{g_1, \cdots, g_n\}$ as follows.

$$\begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} = A \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$$

Is it the case that $\{g_1, \cdots, g_n\}$ is also independent? Explain why.

44. A number is transcendental if it is not the root of any nonzero polynomial with rational coefficients. As mentioned, there are many known transcendental numbers. Suppose α is a real transcendental number. Show that $\{1, \alpha, \alpha^2, \cdots\}$ is a linearly independent set of real numbers if the field of scalars is the rational numbers.

Chapter 8

Linear Transformations

8.1 Matrix Multiplication as a Linear Transformation

Definition 8.1.1 *Let V and W be two finite dimensional vector spaces. A function, L which maps V to W is called a linear transformation and written $L \in \mathcal{L}(V,W)$ if for all scalars α and β, and vectors v,w,*

$$L(\alpha v+\beta w) = \alpha L(v) + \beta L(w).$$

An example of a linear transformation is familiar matrix multiplication. Let $A = (a_{ij})$ be an $m \times n$ matrix. Then an example of a linear transformation $L : \mathbb{F}^n \to \mathbb{F}^m$ is given by

$$(L\mathbf{v})_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$

Here

$$\mathbf{v} \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{F}^n.$$

8.2 $\mathcal{L}(V,W)$ as a Vector Space

Definition 8.2.1 *Given $L, M \in \mathcal{L}(V,W)$ define a new element of $\mathcal{L}(V,W)$, denoted by $L+M$ according to the rule*[1]

$$(L+M)v \equiv Lv + Mv.$$

For α a scalar and $L \in \mathcal{L}(V,W)$, define $\alpha L \in \mathcal{L}(V,W)$ by

$$\alpha L(v) \equiv \alpha(Lv).$$

You should verify that all the axioms of a vector space hold for $\mathcal{L}(V,W)$ with the above definitions of vector addition and scalar multiplication. What about the dimension of $\mathcal{L}(V,W)$?

Before answering this question, here is a useful lemma. It gives a way to define linear transformations and a way to tell when two of them are equal.

Lemma 8.2.2 *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 $\mathcal{L}(V,W)$. 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 if $\sum_{k=1}^{n} a_k v_k$ is an arbitrary vector of V,

$$L\left(\sum_{k=1}^{n} a_k v_k\right) = \sum_{k=1}^{n} a_k L v_k = \sum_{k=1}^{n} a_k M v_k = M\left(\sum_{k=1}^{n} a_k v_k\right)$$

[1] Note that this is the standard way of defining the sum of two functions.

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.

Theorem 8.2.3 *Let V and W be finite dimensional linear spaces of dimension n and m respectively Then* $\dim(\mathcal{L}(V,W)) = mn$.

Proof: Let two sets of bases be

$$\{v_1, \cdots, v_n\} \text{ and } \{w_1, \cdots, w_m\}$$

for V and W respectively. Using Lemma 8.2.2, let $w_i v_j \in \mathcal{L}(V,W)$ be the linear transformation defined on the basis, $\{v_1, \cdots, v_n\}$, by

$$w_i v_k (v_j) \equiv w_i \delta_{jk}$$

where $\delta_{ik} = 1$ if $i = k$ and 0 if $i \neq k$. I will show that $L \in \mathcal{L}(V,W)$ is a linear combination of these special linear transformations called dyadics.

Then let $L \in \mathcal{L}(V,W)$. Since $\{w_1, \cdots, w_m\}$ is a basis, there exist constants, d_{jk} such that

$$Lv_r = \sum_{j=1}^{m} d_{jr} w_j$$

Now consider the following sum of dyadics.

$$\sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i$$

Apply this to v_r. This yields

$$\sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i (v_r) = \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j \delta_{ir} = \sum_{j=1}^{m} d_{jr} w_i = Lv_r$$

Therefore, $L = \sum_{j=1}^{m} \sum_{i=1}^{n} d_{ji} w_j v_i$ showing the span of the dyadics is all of $\mathcal{L}(V,W)$.

Now consider whether these dyadics form a linearly independent set. Suppose

$$\sum_{i,k} d_{ik} w_i v_k = \mathbf{0}.$$

Are all the scalars d_{ik} equal to 0?

$$\mathbf{0} = \sum_{i,k} d_{ik} w_i v_k (v_l) = \sum_{i=1}^{m} d_{il} w_i$$

and so, since $\{w_1, \cdots, w_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows that the dimension of $\mathcal{L}(V,W)$ is mn as claimed because there are m choices for the w_i and n choices for the v_j. ■

8.3 The Matrix of a Linear Transformation

Definition 8.3.1 *In Theorem 8.2.3, the matrix of the linear transformation $L \in \mathcal{L}(V, W)$ with respect to the ordered bases $\beta \equiv \{v_1, \cdots, v_n\}$ for V and $\gamma \equiv \{w_1, \cdots, w_m\}$ for W is defined to be $[L]$ where $[L]_{ij} = d_{ij}$. Thus this matrix is defined by $L = \sum_{i,j} [L]_{ij} w_i v_i$. When it is desired to feature the bases β, γ, this matrix will be denoted as $[L]_{\gamma\beta}$. When there is only one basis β, this is denoted as $[L]_\beta$.*

If V is an n dimensional vector space and $\beta = \{v_1, \cdots, v_n\}$ is a basis for V, there exists a linear map

$$q_\beta : \mathbb{F}^n \to V$$

defined as

$$q_\beta(\mathbf{a}) \equiv \sum_{i=1}^n a_i v_i$$

where

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \sum_{i=1}^n a_i \mathbf{e}_i,$$

for $\mathbf{e}_i$ the standard basis vectors for $\mathbb{F}^n$ consisting of $\begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^T$. Thus the 1 is in the i^{th} position and the other entries are 0. Conversely, if $q : \mathbb{F}^n \to V$ is one to one, onto, and linear, it must be of the form just described. Just let $v_i \equiv q(\mathbf{e}_i)$.

It is clear that q defined in this way, is one to one, onto, and linear. For $v \in V$, $q_\beta^{-1}(v)$ is a vector in $\mathbb{F}^n$ called the component vector of v with respect to the basis $\{v_1, \cdots, v_n\}$.

Proposition 8.3.2 *The matrix of a linear transformation with respect to ordered bases β, γ as described above is characterized by the requirement that multiplication of the components of v by $[L]_{\gamma\beta}$ gives the components of Lv.*

Proof: This happens because by definition, if $v = \sum_i x_i v_i$, then

$$Lv = \sum_i x_i L v_i \equiv \sum_i \sum_j [L]_{ji} x_i w_j = \sum_j \sum_i [L]_{ji} x_i w_j$$

and so the j^{th} component of Lv is $\sum_i [L]_{ji} x_i$, the j^{th} component of the matrix times the component vector of v. Could there be some other matrix which will do this? No, because if such a matrix is M, then for any $\mathbf{x}$, it follows from what was just shown that $[L]\mathbf{x} = M\mathbf{x}$. Hence $[L] = M$. ■

The above proposition shows that the following diagram determines the matrix of a linear transformation. Here q_β and q_γ are the maps defined above with reference to the ordered bases, $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_m\}$ respectively.

$$\begin{array}{ccccc} & & L & & \\ \beta = \{v_1, \cdots, v_n\} & V & \rightarrow & W & \{w_1, \cdots, w_m\} = \gamma \\ & q_\beta \uparrow & \circ & \uparrow q_\gamma & \\ & \mathbb{F}^n & \rightarrow & \mathbb{F}^m & \\ & & [L]_{\gamma\beta} & & \end{array} \tag{8.1}$$

In terms of this diagram, the matrix $[L]_{\gamma\beta}$ is the matrix chosen to make the diagram "commute". It may help to write the description of $[L]_{\gamma\beta}$ in the form

$$\begin{pmatrix} Lv_1 & \cdots & Lv_n \end{pmatrix} = \begin{pmatrix} w_1 & \cdots & w_m \end{pmatrix} [L]_{\gamma\beta} \tag{8.2}$$

with the understanding that you do the multiplications in a formal manner just as you would if everything were numbers. If this helps, use it. If it does not help, ignore it.

Example 8.3.3 *Let*

$$V \equiv \{ \text{ polynomials of degree 3 or less}\},$$

$$W \equiv \{ \text{ polynomials of degree 2 or less}\},$$

and $L \equiv D$ *where* D *is the differentiation operator. A basis for* V *is* $\beta = \{1, x, x^2, x^3\}$ *and a basis for* W *is* $\gamma = \{1, x, x^2\}$.

What is the matrix of this linear transformation with respect to this basis? Using 8.2,

$$\begin{pmatrix} 0 & 1 & 2x & 3x^2 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 \end{pmatrix} [D]_{\gamma\beta} .$$

It follows from this that the first column of $[D]_{\gamma\beta}$ is

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

The next three columns of $[D]_{\gamma\beta}$ are

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

and so

$$[D]_{\gamma\beta} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Now consider the important case where $V = \mathbb{F}^n$, $W = \mathbb{F}^m$, and the basis chosen is the standard basis of vectors $\mathbf{e}_i$ described above.

$$\beta = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}, \ \gamma = \{\mathbf{e}_1, \cdots, \mathbf{e}_m\}$$

Let L be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$ and let A be the matrix of the transformation with respect to these bases. In this case the coordinate maps q_β and q_γ are simply the identity maps on $\mathbb{F}^n$ and $\mathbb{F}^m$ respectively, and can be accomplished by simply multiplying by the appropriate sized identity matrix. The requirement that A is the matrix of the transformation amounts to

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

What about the situation where different pairs of bases are chosen for V and W? How are the two matrices with respect to these choices related? Consider the following diagram which illustrates the situation.

$$\begin{array}{ccc} \mathbb{F}^n & \overset{A_2}{\rightarrow} & \mathbb{F}^m \\ q_{\beta_2} \downarrow & \circ & q_{\gamma_2} \downarrow \\ V & \overset{L}{\rightarrow} & W \\ q_{\beta_1} \uparrow & \circ & q_{\gamma_1} \uparrow \\ \mathbb{F}^n & \overset{A_1}{\rightarrow} & \mathbb{F}^m \end{array}$$

In this diagram q_{β_i} and q_{γ_i} are coordinate maps as described above. From the diagram,

$$q_{\gamma_1}^{-1} q_{\gamma_2} A_2 q_{\beta_2}^{-1} q_{\beta_1} = A_1,$$

where $q_{\beta_2}^{-1} q_{\beta_1}$ and $q_{\gamma_1}^{-1} q_{\gamma_2}$ are one to one, onto, and linear maps which may be accomplished by multiplication by a square matrix. Thus there exist matrices P, Q such that $P : \mathbb{F}^n \to \mathbb{F}^n$ and $Q : \mathbb{F}^m \to \mathbb{F}^m$ are invertible and

$$PA_2Q = A_1.$$

Example 8.3.4 *Let $\beta \equiv \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\gamma \equiv \{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ be two bases for V. Let L be the linear transformation which maps $\mathbf{v}_i$ to $\mathbf{w}_i$. Find $[L]_{\gamma\beta}$. In case $V = \mathbb{F}^n$ and letting $\delta = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$, the usual basis for $\mathbb{F}^n$, find $[L]_\delta$.*

Letting δ_{ij} be the symbol which equals 1 if $i = j$ and 0 if $i \neq j$, it follows that $L = \sum_{i,j} \delta_{ij} \mathbf{w}_i \mathbf{v}_j$ and so $[L]_{\gamma\beta} = I$ the identity matrix. For the second part, you must have

$$\left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{array} \right) = \left(\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{array} \right) [L]_\delta$$

and so

$$[L]_\delta = \left(\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{array} \right)^{-1} \left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{array} \right)$$

where $\left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{array} \right)$ is the $n \times n$ matrix having i^{th} column equal to $\mathbf{w}_i$.

Definition 8.3.5 *In the special case where $V = W$ and only one basis is used for $V = W$, this becomes*

$$q_{\beta_1}^{-1} q_{\beta_2} A_2 q_{\beta_2}^{-1} q_{\beta_1} = A_1.$$

Letting S be the matrix of the linear transformation $q_{\beta_2}^{-1} q_{\beta_1}$ with respect to the standard basis vectors in $\mathbb{F}^n$,

$$S^{-1} A_2 S = A_1. \tag{8.3}$$

When this occurs, A_1 is said to be similar to A_2 and $A \to S^{-1}AS$ is called a similarity transformation.

Recall the following.

Definition 8.3.6 *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 8.3.7 *$[x]$ 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.*

Also recall the notion of equivalence classes.

Theorem 8.3.8 *Let $\sim$ be an equivalence class defined on a set S and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.*

Theorem 8.3.9 *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 $L \in \mathcal{L}(V, V)$ and bases $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_n\}$ such that A is the matrix of L with respect to $\{v_1, \cdots, v_n\}$ and B is the matrix of L with respect to $\{w_1, \cdots, w_n\}$.

Proof: $A \sim A$ because $S = I$ works in the definition. If $A \sim B$, then $B \sim A$, because

$$A = S^{-1}BS$$

implies $B = SAS^{-1}$. If $A \sim B$ and $B \sim C$, then

$$A = S^{-1}BS, \ B = T^{-1}CT$$

and so

$$A = S^{-1}T^{-1}CTS = (TS)^{-1} CTS$$

which implies $A \sim C$. This verifies the first part of the conclusion.

Now let V be an n dimensional vector space, $A \sim B$ so $A = S^{-1}BS$ and pick a basis for V,

$$\beta \equiv \{v_1, \cdots, v_n\}.$$

Define $L \in \mathcal{L}(V, V)$ by

$$Lv_i \equiv \sum_j a_{ji} v_j$$

where $A = (a_{ij})$. Thus A is the matrix of the linear transformation L. Consider the diagram

$$\begin{array}{ccc} \mathbb{F}^n & \underset{\rightarrow}{B} & \mathbb{F}^n \\ q_\gamma \downarrow & \circ & q_\gamma \downarrow \\ V & \underset{\rightarrow}{L} & V \\ q_\beta \uparrow & \circ & q_\beta \uparrow \\ \mathbb{F}^n & \underset{\rightarrow}{A} & \mathbb{F}^n \end{array}$$

where q_γ is chosen to make the diagram commute. Thus we need $S = q_\gamma^{-1} q_\beta$ which requires

$$q_\gamma = q_\beta S^{-1}$$

Then it follows that B is the matrix of L with respect to the basis

$$\{q_\gamma \mathbf{e}_1, \cdots, q_\gamma \mathbf{e}_n\} \equiv \{w_1, \cdots, w_n\}.$$

That is, A and B are matrices of the same linear transformation L. Conversely, suppose whenever V is an n dimensional vector space, there exists $L \in \mathcal{L}(V, V)$ and bases $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_n\}$ such that A is the matrix of L with respect to $\{v_1, \cdots, v_n\}$ and B is the matrix of L with respect to $\{w_1, \cdots, w_n\}$. Then it was shown above that $A \sim B$. ■

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 next proposition considers this.

Proposition 8.3.10 *Let A be an $m \times n$ matrix and let L be the linear transformation which is defined by*

$$L\left(\sum_{k=1}^n x_k \mathbf{e}_k\right) \equiv \sum_{k=1}^n (A\mathbf{e}_k) x_k \equiv \sum_{i=1}^m \sum_{k=1}^n A_{ik} x_k \mathbf{e}_i$$

In simple language, to find $L\mathbf{x}$, you multiply on the left of $\mathbf{x}$ by A. (A is the matrix of L with respect to the standard basis.) Then the matrix M of this linear transformation with respect to the bases $\beta = \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ for $\mathbb{F}^n$ and $\gamma = \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ 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.

$$\begin{array}{ccc} & L & \\ \mathbb{F}^n & \to & \mathbb{F}^m \\ q_\beta \uparrow & \circ & \uparrow q_\gamma \\ \mathbb{F}^n & \to & \mathbb{F}^m \\ & M & \end{array}$$

Here the coordinate maps are defined in the usual way. Thus

$$q_\beta \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T \equiv \sum_{i=1}^n x_i \mathbf{u}_i.$$

Therefore, q_β can be considered the same as multiplication of a vector in $\mathbb{F}^n$ on the left by the matrix $\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix}$. Similar considerations apply to q_γ. Thus it is desired to have the following for an arbitrary $\mathbf{x} \in \mathbb{F}^n$.

$$A \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{pmatrix} M\mathbf{x}$$

Therefore, the conclusion of the proposition follows. ■

In the special case where $m = n$ and $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is an orthonormal basis and you want M, the matrix of L with respect to this new orthonormal basis, it follows from the above that

$$M = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{pmatrix}^* A \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} = U^* A U$$

where U is a unitary matrix. Thus matrices with respect to two orthonormal bases are unitarily similar.

Definition 8.3.11 *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.*

The following theorem is of great significance.

Theorem 8.3.12 *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 8.3.11 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 $\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$ and let $S^{-1} \equiv \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix}$ where

$$\mathbf{u}_i^T \mathbf{v}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}.$$

S^{-1} exists because S has rank n. Then from block multiplication,

$$S^{-1}AS = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (A\mathbf{v}_1 \cdots A\mathbf{v}_n) = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (\lambda_1 \mathbf{v}_1 \cdots \lambda_n \mathbf{v}_n)$$

$$= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.$$

Next suppose A is diagonalizable so $S^{-1}AS = D \equiv \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$. Then the columns of S form a basis because S^{-1} is given to exist. It only remains to verify that these columns of S are eigenvectors. But letting $S = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$, $AS = SD$ and so $\begin{pmatrix} A\mathbf{v}_1 & \cdots & A\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 \mathbf{v}_1 & \cdots & \lambda_n \mathbf{v}_n \end{pmatrix}$ 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 8.3.13 *Let $L \in \mathcal{L}(V, V)$ where V is an n dimensional vector space and let A be the matrix of this linear transformation with respect to a basis on V. Then it is possible to define*

$$\det(L) \equiv \det(A).$$

Proof: Each choice of basis for V determines a matrix for L with respect to the basis. If A and B are two such matrices, it follows from Theorem 8.3.9 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

$$\det(A) = \det(B) \ \blacksquare$$

Definition 8.3.14 *Let $A \in \mathcal{L}(X,Y)$ where X and Y are finite dimensional vector spaces. Define* $\operatorname{rank}(A)$ *to equal the dimension of $A(X)$.*

The following theorem explains how the rank of A is related to the rank of the matrix of A.

Theorem 8.3.15 *Let $A \in \mathcal{L}(X,Y)$. Then* $\operatorname{rank}(A) = \operatorname{rank}(M)$ *where M is the matrix of A taken with respect to a pair of bases for the vector spaces X, and Y.*

Proof: Recall the diagram which describes what is meant by the matrix of A. Here the two bases are as indicated.

$$\begin{array}{ccccc} \beta = \{v_1, \cdots, v_n\} & X & \underset{\to}{A} & Y & \{w_1, \cdots, w_m\} = \gamma \\ & q_\beta \uparrow & \circ & \uparrow q_\gamma & \\ & \mathbb{F}^n & \underset{\to}{M} & \mathbb{F}^m & \end{array}$$

Let $\{Ax_1, \cdots, Ax_r\}$ be a basis for AX. Thus

$$\left\{q_\gamma M q_\beta^{-1} x_1, \cdots, q_\gamma M q_\beta^{-1} x_r\right\}$$

is a basis for AX. It follows that

$$\left\{M q_X^{-1} x_1, \cdots, M q_X^{-1} x_r\right\}$$

is linearly independent and so $\operatorname{rank}(A) \leq \operatorname{rank}(M)$. However, one could interchange the roles of M and A in the above argument and thereby turn the inequality around. ■

The following result is a summary of many concepts.

Theorem 8.3.16 *Let $L \in \mathcal{L}(V,V)$ where V is a finite dimensional vector space. Then the following are equivalent.*

1. *L is one to one.*
2. *L maps a basis to a basis.*
3. *L is onto.*
4. $\det(L) \neq 0$
5. *If $Lv = 0$ then $v = 0$.*

Proof: Suppose first L is one to one and let $\beta = \{v_i\}_{i=1}^n$ be a basis. Then if $\sum_{i=1}^n c_i L v_i = 0$ it follows $L\left(\sum_{i=1}^n c_i v_i\right) = 0$ which means that since $L(0) = 0$, and L is one to one, it must be the case that $\sum_{i=1}^n c_i v_i = 0$. Since $\{v_i\}$ is a basis, each $c_i = 0$ which shows $\{Lv_i\}$ is a linearly independent set. Since there are n of these, it must be that this is a basis.

Now suppose 2.). Then letting $\{v_i\}$ be a basis, and $y \in V$, it follows from part 2.) that there are constants, $\{c_i\}$ such that $y = \sum_{i=1}^n c_i L v_i = L\left(\sum_{i=1}^n c_i v_i\right)$. Thus L is onto. It has been shown that 2.) implies 3.).

Now suppose 3.). Then the operation consisting of multiplication by the matrix of $L, [L]$, 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 3.3.23 this equals the determinant rank and so $\det([L]) \equiv \det(L) \neq 0$.

Now assume 4.) If $Lv = 0$ for some $v \neq 0$, it follows that $[L]\,\mathbf{x} = 0$ for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the columns of $[L]$ are linearly dependent and so by Theorem 3.3.23, $\det([L]) = \det(L) = 0$ contrary to 4.). Therefore, 4.) implies 5.).

Now suppose 5.) and suppose $Lv = Lw$. Then $L(v - w) = 0$ and so by 5.), $v - w = 0$ showing that L is one to one. ■

Also it is important to note that composition of linear transformations corresponds to multiplication of the matrices. Consider the following diagram in which $[A]_{\gamma\beta}$ denotes the matrix of A relative to the bases γ on Y and β on X, $[B]_{\delta\gamma}$ defined similarly.

$$\begin{array}{ccccc} X & \underrightarrow{A} & Y & \underrightarrow{B} & Z \\ q_\beta \uparrow & \circ & \uparrow q_\gamma & \circ & \uparrow q_\delta \\ \mathbb{F}^n & \underrightarrow{[A]_{\gamma\beta}} & \mathbb{F}^m & \underrightarrow{[B]_{\delta\gamma}} & \mathbb{F}^p \end{array}$$

where A and B are two linear transformations, $A \in \mathcal{L}(X, Y)$ and $B \in \mathcal{L}(Y, Z)$. Then $B \circ A \in \mathcal{L}(X, Z)$ and so it has a matrix with respect to bases given on X and Z, the coordinate maps for these bases being q_β and q_δ respectively. Then

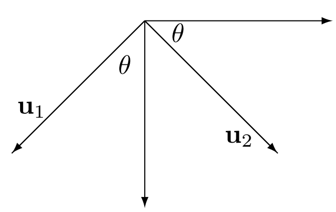
$$B \circ A = q_\delta\,[B]_{\delta\gamma}\,q_\gamma^{-1} q_\gamma\,[A]_{\gamma\beta}\,q_\beta^{-1} = q_\delta\,[B]_{\delta\gamma}\,[A]_{\gamma\beta}\,q_\beta^{-1}.$$

But this shows that $[B]_{\delta\gamma}\,[A]_{\gamma\beta}$ plays the role of $[B \circ A]_{\delta\beta}$, the matrix of $B \circ A$. Hence the matrix of $B \circ A$ equals the product of the two matrices $[A]_{\gamma\beta}$ and $[B]_{\delta\gamma}$. Of course it is interesting to note that although $[B \circ A]_{\delta\beta}$ must be unique, the matrices, $[A]_{\gamma\beta}$ and $[B]_{\delta\gamma}$ are not unique because they depend on γ, the basis chosen for Y.

Theorem 8.3.17 *The matrix of the composition of linear transformations equals the product of the matrices of these linear transformations.*

8.3.1 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

$$T(a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3) = aT\mathbf{u}_1 + bT\mathbf{u}_2 + cT\mathbf{u}_3$$

$$= (a\cos\theta - b\sin\theta)\,\mathbf{u}_1 + (a\sin\theta + b\cos\theta)\,\mathbf{u}_2 + c\mathbf{u}_3.$$

Thus in terms of the basis $\gamma \equiv \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, the matrix of this transformation is

$$[T]_\gamma \equiv \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

I want to obtain the matrix of the transformation in terms of the usual basis $\beta \equiv \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ because it is in terms of this basis that we usually deal with vectors. From Proposition 8.3.10, if $[T]_\beta$ is this matrix,

$$\begin{aligned} &\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ = &\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}^{-1} [T]_\beta \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} \end{aligned}$$

and so you can solve for $[T]_\beta$ if you know the $\mathbf{u}_i$.

Recall why this is so.

$$\begin{array}{ccc} \mathbb{R}^3 & \xrightarrow{[T]_\gamma} & \mathbb{R}^3 \\ q_\gamma \downarrow & \circ & q_\gamma \downarrow \\ \mathbb{R}^3 & \xrightarrow{T} & \mathbb{R}^3 \\ I \uparrow & \circ & I \uparrow \\ \mathbb{R}^3 & \xrightarrow{[T]_\beta} & \mathbb{R}^3 \end{array}$$

The map q_γ is accomplished by a multiplication on the left by $\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}$. Thus

$$[T]_\beta = q_\gamma [T]_\gamma q_\gamma^{-1} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} [T]_\gamma \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}^{-1}.$$

Suppose the unit vector $\mathbf{u}_3$ about which the counterclockwise rotation takes place is (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 orthonormal 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 counter clockwise 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

$$= \begin{pmatrix} a^2 + \left(1-a^2\right)\cos\theta & ab\left(1-\cos\theta\right) - c\sin\theta & ac\left(1-\cos\theta\right) + b\sin\theta \\ ab\left(1-\cos\theta\right) + c\sin\theta & b^2 + \left(1-b^2\right)\cos\theta & bc\left(1-\cos\theta\right) - a\sin\theta \\ ac\left(1-\cos\theta\right) - b\sin\theta & bc\left(1-\cos\theta\right) + a\sin\theta & c^2 + \left(1-c^2\right)\cos\theta \end{pmatrix}. \tag{8.4}$$

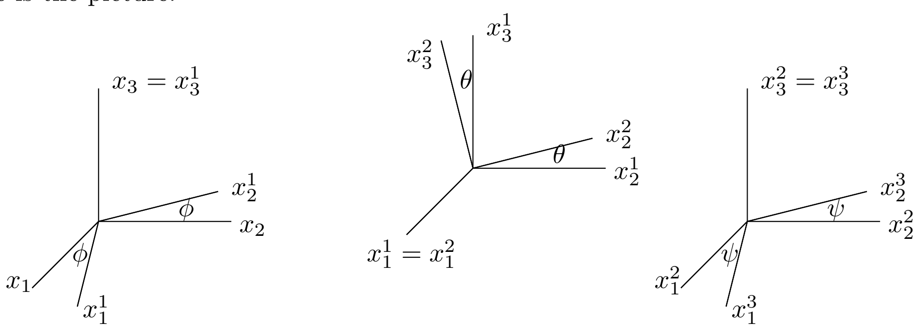
With this, it is clear how to rotate clockwise about the unit vector, (a, b, c). Just rotate counter clockwise through an angle of $-\theta$. Thus the matrix for this clockwise rotation is just

$$= \begin{pmatrix} a^2 + \left(1-a^2\right)\cos\theta & ab\left(1-\cos\theta\right) + c\sin\theta & ac\left(1-\cos\theta\right) - b\sin\theta \\ ab\left(1-\cos\theta\right) - c\sin\theta & b^2 + \left(1-b^2\right)\cos\theta & bc\left(1-\cos\theta\right) + a\sin\theta \\ ac\left(1-\cos\theta\right) + b\sin\theta & bc\left(1-\cos\theta\right) - a\sin\theta & c^2 + \left(1-c^2\right)\cos\theta \end{pmatrix}.$$

In deriving 8.4 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 8.4 reduces to 2.24.

8.3.2 The Euler Angles

An important application of the above theory is to the Euler angles, important in the mechanics of rotating bodies. Lagrange studied these things back in the 1700's. To describe the Euler angles consider the following picture in which x_1, x_2 and x_3 are the usual coordinate axes fixed in space and the axes labeled with a superscript denote other coordinate axes. Here is the picture.



We obtain ϕ by rotating counter clockwise about the fixed x_3 axis. Thus this rotation has the matrix

$$\begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \equiv M_1\left(\phi\right)$$

Next rotate counter clockwise about the x_1^1 axis which results from the first rotation through an angle of θ. Thus it is desired to rotate counter clockwise through an angle θ about the unit vector

$$\begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\phi \\ \sin\phi \\ 0 \end{pmatrix}.$$

Therefore, in 8.4, $a = \cos\phi, b = \sin\phi$, and $c = 0$. It follows the matrix of this transformation with respect to the usual basis is

$$\begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta & \cos\phi\sin\phi\,(1-\cos\theta) & \sin\phi\sin\theta \\ \cos\phi\sin\phi\,(1-\cos\theta) & \sin^2\phi + \cos^2\phi\cos\theta & -\cos\phi\sin\theta \\ -\sin\phi\sin\theta & \cos\phi\sin\theta & \cos\theta \end{pmatrix} \equiv M_2(\phi,\theta)$$

Finally, we rotate counter clockwise about the positive x_3^2 axis by ψ. The vector in the positive x_3^1 axis is the same as the vector in the fixed x_3 axis. Thus the unit vector in the positive direction of the x_3^2 axis is

$$\begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta & \cos\phi\sin\phi\,(1-\cos\theta) & \sin\phi\sin\theta \\ \cos\phi\sin\phi\,(1-\cos\theta) & \sin^2\phi + \cos^2\phi\cos\theta & -\cos\phi\sin\theta \\ -\sin\phi\sin\theta & \cos\phi\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta \\ \cos\phi\sin\phi\,(1-\cos\theta) \\ -\sin\phi\sin\theta \end{pmatrix} = \begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta \\ \cos\phi\sin\phi\,(1-\cos\theta) \\ -\sin\phi\sin\theta \end{pmatrix}$$

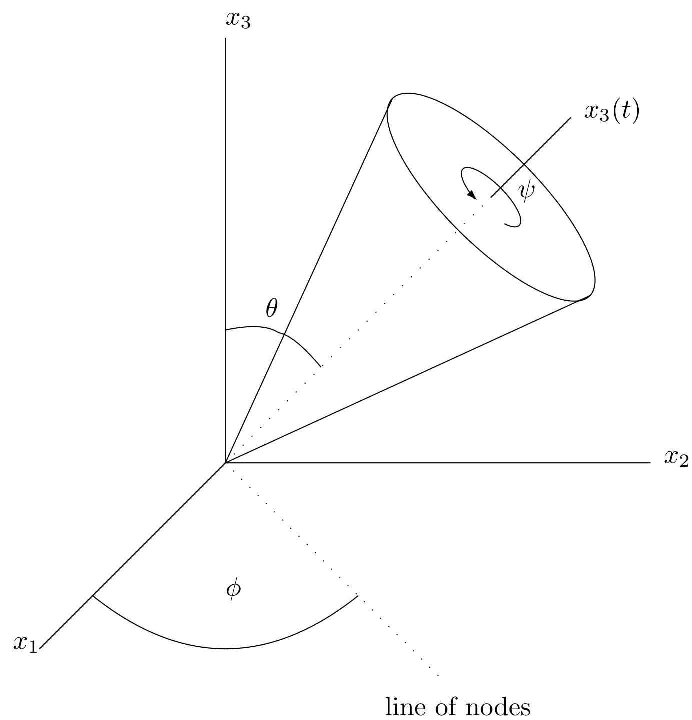
and it is desired to rotate counter clockwise through an angle of ψ about this vector. Thus, in this case,

$$a = \cos^2\phi + \sin^2\phi\cos\theta, b = \cos\phi\sin\phi\,(1-\cos\theta)\,, c = -\sin\phi\sin\theta.$$

and you could substitute in to the formula of Theorem 8.4 and obtain a matrix which represents the linear transformation obtained by rotating counter clockwise about the positive x_3^2 axis, $M_3(\phi,\theta,\psi)$. Then what would be the matrix with respect to the usual basis for the linear transformation which is obtained as a composition of the three just described? By Theorem 8.3.17, this matrix equals the product of these three,

$$M_3(\phi,\theta,\psi)\,M_2(\phi,\theta)\,M_1(\phi)\,.$$

I leave the details to you. There are procedures due to Lagrange which will allow you to write differential equations for the Euler angles in a rotating body. To give an idea how these angles apply, consider the following picture.



This is as far as I will go on this topic. The point is, it is possible to give a systematic description in terms of matrix multiplication of a very elaborate geometrical description of a composition of linear transformations. You see from the picture it is possible to describe the motion of the spinning top shown in terms of these Euler angles.

8.4 Eigenvalues and Eigenvectors of Linear Transformations

Let V be a finite dimensional vector space. For example, it could be a subspace of $\mathbb{C}^n$or $\mathbb{R}^n$. Also suppose $A \in \mathcal{L}(V, V)$.

Definition 8.4.1 *The characteristic polynomial of A is defined as $q(\lambda) \equiv \det(\lambda I - A)$. The zeros of $q(\lambda)$ in $\mathbb{F}$ are called the eigenvalues of A.*

Lemma 8.4.2 *When λ is an eigenvalue of A which is also in $\mathbb{F}$, the field of scalars, then there exists $v \neq 0$ such that $Av = \lambda v$.*

Proof: This follows from Theorem 8.3.16. Since $\lambda \in \mathbb{F}$,

$$\lambda I - A \in \mathcal{L}(V, V)$$

and since it has zero determinant, it is not one to one. ■

The following lemma gives the existence of something called the minimal polynomial.

Lemma 8.4.3 *Let $A \in \mathcal{L}(V, V)$ where V is a finite dimensional vector space of dimension n with arbitrary field of scalars. Then there exists a unique polynomial of the form*

$$p(\lambda) = \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$$

such that $p(A) = 0$ and m is as small as possible for this to occur.

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 8.2.3 the set is linearly dependent. Thus there exist constants, $c_i \in \mathbb{F}$ 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, one example is $q(\lambda) \equiv c_0 + \sum_{k=1}^{n^2} c_k \lambda^k$. 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 the one which has the smallest degree m. This is called the minimal polynomial and will be denoted here by $p(\lambda)$. If there were two minimal polynomials, the one just found and another,

$$\lambda^m + d_{m-1}\lambda^{m-1} + \cdots + d_1\lambda + d_0.$$

Then subtracting these would give the following polynomial,

$$\widetilde{q}(\lambda) = (d_{m-1} - c_{m-1})\lambda^{m-1} + \cdots + (d_1 - c_1)\lambda + d_0 - c_0$$

Since $\widetilde{q}(A) = 0$, this requires each $d_k = c_k$ since otherwise you could divide by $d_k - c_k$ where k is the largest one which is nonzero. Thus the choice of m would be contradicted. ■

Theorem 8.4.4 *Let V be a nonzero finite dimensional vector space of dimension n with the field of scalars equal to $\mathbb{F}$. Suppose $A \in \mathcal{L}(V, V)$ and for $p(\lambda)$ the minimal polynomial defined above, let $\mu \in \mathbb{F}$ be a zero of this polynomial. Then there exists $v \neq 0, v \in V$ such that*

$$Av = \mu v.$$

If $\mathbb{F} = \mathbb{C}$, then A always has an eigenvector and eigenvalue. Furthermore, if $\{\lambda_1, \cdots, \lambda_m\}$ are the zeros of $p(\lambda)$ in $\mathbb{F}$, these are exactly the eigenvalues of A for which there exists an eigenvector in V.

Proof: Suppose first μ is a zero of $p(\lambda)$. Since $p(\mu) = 0$, it follows

$$p(\lambda) = (\lambda - \mu) k(\lambda)$$

where $k(\lambda)$ is a polynomial having coefficients in $\mathbb{F}$. Since p has minimal degree, $k(A) \neq 0$ and so there exists a vector, $u \neq 0$ such that $k(A) u \equiv v \neq 0$. But then

$$(A - \mu I) v = (A - \mu I) k(A)(u) = \mathbf{0}.$$

The next claim about the existence of an eigenvalue follows from the fundamental theorem of algebra and what was just shown.

It has been shown that every zero of $p(\lambda)$ is an eigenvalue which has an eigenvector in V. Now suppose μ is an eigenvalue which has an eigenvector in V so that $Av = \mu v$ for some $v \in V, v \neq 0$. Does it follow μ is a zero of $p(\lambda)$?

$$\mathbf{0} = p(A) v = p(\mu) v$$

and so μ is indeed a zero of $p(\lambda)$. ■

In summary, the theorem says that the eigenvalues which have eigenvectors in V are exactly the zeros of the minimal polynomial which are in the field of scalars $\mathbb{F}$.

8.5 Exercises

1. If A, B, and C are each $n \times n$ matrices and ABC is invertible, why are each of A, B, and C invertible?

2. 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.

3. Explain why $A\mathbf{x} = \mathbf{0}$ always has a solution whenever A is a linear transformation.

4. Review problem: Suppose $\det(A - \lambda I) = 0$. Show using Theorem 3.1.15 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

5. How does the minimal polynomial of an algebraic number relate to the minimal polynomial of a linear transformation? Can an algebraic number be thought of as a linear transformation? How?

6. Recall the fact from algebra that if $p(\lambda)$ and $q(\lambda)$ are polynomials, then there exists $l(\lambda)$, a polynomial such that
$$q(\lambda) = p(\lambda) l(\lambda) + r(\lambda)$$
where the degree of $r(\lambda)$ is less than the degree of $p(\lambda)$ or else $r(\lambda) = 0$. With this in mind, why must the minimal polynomial always divide the characteristic polynomial? That is, why does there always exist a polynomial $l(\lambda)$ such that $p(\lambda) l(\lambda) = q(\lambda)$? Can you give conditions which imply the minimal polynomial equals the characteristic polynomial? Go ahead and use the Cayley Hamilton theorem.

7. In the following examples, a linear transformation, T is given by specifying its action on a basis β. Find its matrix with respect to this basis.

   (a) $T\begin{pmatrix} 1 \\ 2 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 1\begin{pmatrix} -1 \\ 1 \end{pmatrix}, T\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

   (b) $T\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 2\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 1\begin{pmatrix} -1 \\ 1 \end{pmatrix}, T\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

   (c) $T\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \end{pmatrix} + 1\begin{pmatrix} 1 \\ 0 \end{pmatrix}, T\begin{pmatrix} 1 \\ 2 \end{pmatrix} = 1\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

8. Let $\beta = \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ be a basis for $\mathbb{F}^n$ and let $T : \mathbb{F}^n \to \mathbb{F}^n$ be defined as follows.
$$T\left(\sum_{k=1}^{n} a_k \mathbf{u}_k\right) = \sum_{k=1}^{n} a_k b_k \mathbf{u}_k$$
First show that T is a linear transformation. Next show that the matrix of T with respect to this basis, $[T]_\beta$ is
$$\begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_n \end{pmatrix}$$
Show that the above definition is equivalent to simply specifying T on the basis vectors of β by
$$T(\mathbf{u}_k) = b_k \mathbf{u}_k.$$

9. $\uparrow$In the situation of the above problem, let $\gamma = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$ be the standard basis for $\mathbb{F}^n$ where $\mathbf{e}_k$ is the vector which has 1 in the k^{th} entry and zeros elsewhere. Show that $[T]_\gamma =$
$$\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} [T]_\beta \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix}^{-1} \tag{8.5}$$

10. $\uparrow$Generalize the above problem to the situation where T is given by specifying its action on the vectors of a basis $\beta = \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ as follows.
$$T\mathbf{u}_k = \sum_{j=1}^{n} a_{jk}\mathbf{u}_j.$$
Letting $A = (a_{ij})$, verify that for $\gamma = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$, 8.5 still holds and that $[T]_\beta = A$.

11. Let P_3 denote the set of real polynomials of degree no more than 3, defined on an interval $[a, b]$. Show that P_3 is a subspace of the vector space of all functions defined on this interval. Show that a basis for P_3 is $\{1, x, x^2, x^3\}$. Now let D denote the differentiation operator which sends a function to its derivative. Show D is a linear transformation which sends P_3 to P_3. Find the matrix of this linear transformation with respect to the given basis.

12. Generalize the above problem to P_n, the space of polynomials of degree no more than n with basis $\{1, x, \cdots, x^n\}$.

13. In the situation of the above problem, let the linear transformation be $T = D^2 + 1$, defined as $Tf = f'' + f$. Find the matrix of this linear transformation with respect to the given basis $\{1, x, \cdots, x^n\}$. Write it down for $n = 4$.

14. In calculus, the following situation is encountered. There exists a vector valued function $\mathbf{f} : U \to \mathbb{R}^m$ where U is an open subset of $\mathbb{R}^n$. Such a function is said to have a derivative or to be differentiable at $\mathbf{x} \in U$ if there exists a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ such that
$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{|\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) - T\mathbf{v}|}{|\mathbf{v}|} = 0.$$
First show that this linear transformation, if it exists, must be unique. Next show that for $\beta = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$, , the standard basis, the k^{th} column of $[T]_\beta$ is
$$\frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}).$$
Actually, the result of this problem is a well kept secret. People typically don't see this in calculus. It is seen for the first time in advanced calculus if then.

15. Recall that A is similar to B if there exists a matrix P such that $A = P^{-1}BP$. Show that if A and B are similar, then they have the same determinant. Give an example of two matrices which are not similar but have the same determinant.

16. Suppose $A \in \mathcal{L}(V, W)$ where $\dim(V) > \dim(W)$. Show $\ker(A) \neq \{\mathbf{0}\}$. That is, show there exist nonzero vectors $\mathbf{v} \in V$ such that $A\mathbf{v} = \mathbf{0}$.

17. A vector $\mathbf{v}$ is in the convex hull of a nonempty set S if there are finitely many vectors of $S, \{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ and nonnegative scalars $\{t_1, \cdots, t_m\}$ such that
$$\mathbf{v} = \sum_{k=1}^{m} t_k\mathbf{v}_k, \ \sum_{k=1}^{m} t_k = 1.$$

Such a linear combination is called a convex combination. Suppose now that $S \subseteq V$, a vector space of dimension n. Show that if $\mathbf{v} = \sum_{k=1}^{m} t_k \mathbf{v}_k$ is a vector in the convex hull for $m > n+1$, then there exist other scalars $\{t_k'\}$ such that

$$\mathbf{v} = \sum_{k=1}^{m-1} t_k' \mathbf{v}_k.$$

Thus every vector in the convex hull of S can be obtained as a convex combination of at most $n+1$ points of S. This incredible result is in Rudin [24]. **Hint:** Consider $L : \mathbb{R}^m \to V \times \mathbb{R}$ defined by

$$L(\mathbf{a}) \equiv \left(\sum_{k=1}^{m} a_k \mathbf{v}_k, \sum_{k=1}^{m} a_k \right)$$

Explain why $\ker(L) \neq \{\mathbf{0}\}$. Next, letting $\mathbf{a} \in \ker(L) \setminus \{\mathbf{0}\}$ and $\lambda \in \mathbb{R}$, note that $\lambda \mathbf{a} \in \ker(L)$. Thus for all $\lambda \in \mathbb{R}$,

$$\mathbf{v} = \sum_{k=1}^{m} (t_k + \lambda a_k) \mathbf{v}_k.$$

Now vary λ till some $t_k + \lambda a_k = 0$ for some $a_k \neq 0$.

18. For those who know about compactness, use Problem 17 to show that if $S \subseteq \mathbb{R}^n$ and S is compact, then so is its convex hull.

19. 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.

20. Let A be an $n \times n$ matrix of elements of $\mathbb{F}$. There are two cases. In the first case, $\mathbb{F}$ contains a splitting field of $p_A(\lambda)$ so that $p(\lambda)$ factors into a product of linear polynomials having coefficients in $\mathbb{F}$. It is the second case which is of interest here where $p_A(\lambda)$ does not factor into linear factors having coefficients in $\mathbb{F}$. Let $\mathbb{G}$ be a splitting field of $p_A(\lambda)$ and let $q_A(\lambda)$ be the minimal polynomial of A with respect to the field $\mathbb{G}$. Explain why $q_A(\lambda)$ must divide $p_A(\lambda)$. Now why must $q_A(\lambda)$ factor completely into linear factors?

21. In Lemma 8.2.2 verify that L is linear.

Chapter 9

Canonical Forms

9.1 A Theorem of Sylvester, Direct Sums

The notation is defined as follows.

Definition 9.1.1 *Let* $L \in \mathcal{L}(V, W)$. *Then* $\ker(L) \equiv \{v \in V : Lv = 0\}$.

Lemma 9.1.2 *Whenever* $L \in \mathcal{L}(V, W)$, $\ker(L)$ *is a subspace.*

Proof: If a, b are scalars and v,w are in $\ker(L)$, then

$$L(av + bw) = aL(v) + bL(w) = 0 + 0 = 0 \blacksquare$$

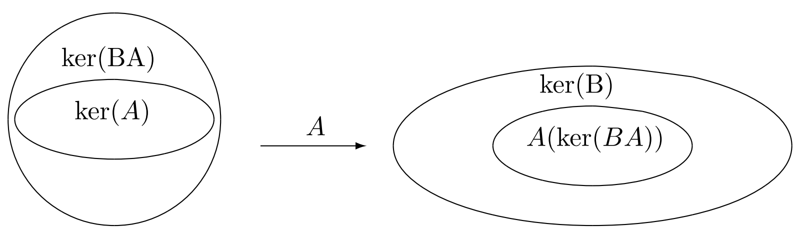
Suppose now that $A \in \mathcal{L}(V, W)$ and $B \in \mathcal{L}(W, U)$ where V, W, U are all finite dimensional vector spaces. Then it is interesting to consider $\ker(BA)$. The following theorem of Sylvester is a very useful and important result.

Theorem 9.1.3 *Let* $A \in \mathcal{L}(V, W)$ *and* $B \in \mathcal{L}(W, U)$ *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*

$$\dim(\ker(BA)) \leq \dim(\ker(B)) + \dim(\ker(A)).$$

Equality holds if and only if $A(\ker(BA)) = \ker(B)$.

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 $Az = \sum_{i=1}^{m} a_i Ay_i$ 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.)

$$\dim(\ker(BA)) \leq n + m \leq \dim(\ker(A)) + \dim(\ker(B))$$

Now $\{x_1, \cdots, x_n, y_1, \cdots, y_m\}$ is linearly independent because if

$$\sum_i a_i x_i + \sum_j b_j y_j = 0$$

then you could do A to both sides and conclude that $\sum_j b_j A y_j = 0$ which requires that each $b_j = 0$. Then it follows that each $a_i = 0$ also because it implies $\sum_i a_i x_i = 0$. Thus

$$\{x_1, \cdots, x_n, y_1, \cdots, y_m\}$$

is a basis for $\ker(BA)$. Then $A(\ker(BA)) = \ker(B)$ if and only if $m = \dim(\ker(B))$ if and only if

$$\dim(\ker(BA)) = m + n = \dim(\ker(B)) + \dim(\ker(A)). \blacksquare$$

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 $\mathcal{L}(V, V)$. Then

$$\dim\left(\ker \prod_{i=1}^{l} L_i\right) \leq \sum_{i=1}^{l} \dim(\ker L_i).$$

Definition 9.1.4 *Let $\{V_i\}_{i=1}^{r}$ be subspaces of V. Then*

$$\sum_{i=1}^{r} V_i \equiv V_1 + \cdots + V_r$$

denotes all sums of the form $\sum_{i=1}^{r} v_i$ where $v_i \in V_i$. If whenever

$$\sum_{i=1}^{r} v_i = 0, v_i \in V_i, \tag{9.1}$$

it follows that $v_i = 0$ 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.

Now here is a useful lemma which is likely already understood.

Lemma 9.1.5 *Let $L \in \mathcal{L}(V, W)$ where V, W are n dimensional vector spaces. Then L is one to one, if and only if L is also onto. In fact, if $\{v_1, \cdots, v_n\}$ is a basis, then so is $\{Lv_1, \cdots, Lv_n\}$.*

Proof: Let $\{v_1, \cdots, v_n\}$ be a basis for V. Then I claim that $\{Lv_1, \cdots, Lv_n\}$ is a basis for W. First of all, I show $\{Lv_1, \cdots, Lv_n\}$ is linearly independent. Suppose

$$\sum_{k=1}^{n} c_k L v_k = 0.$$

Then

$$L\left(\sum_{k=1}^{n} c_k v_k\right) = 0$$

and since L is one to one, it follows

$$\sum_{k=1}^{n} c_k v_k = 0$$

which implies each $c_k = 0$. Therefore, $\{Lv_1, \cdots, Lv_n\}$ is linearly independent. If there exists w not in the span of these vectors, then by Lemma 7.2.10, $\{Lv_1, \cdots, Lv_n, w\}$ would be independent and this contradicts the exchange theorem, Theorem 7.2.4 because it would be a linearly independent set having more vectors than the spanning set $\{v_1, \cdots, v_n\}$.

Conversely, suppose L is onto. Then there exists a basis for W which is of the form $\{Lv_1, \cdots, Lv_n\}$. It follows that $\{v_1, \cdots, v_n\}$ is linearly independent. Hence it is a basis for V by similar reasoning to the above. Then if $Lx = 0$, it follows that there are scalars c_i such that $x = \sum_i c_i v_i$ and consequently $0 = Lx = \sum_i c_i L v_i$. Therefore, each $c_i = 0$ and so $x = 0$ also. Thus L is one to one. ■

Lemma 9.1.6 *If $V = V_1 \oplus \cdots \oplus V_r$ and if $\beta_i = \left\{v_1^i, \cdots, v_{m_i}^i\right\}$ is a basis for V_i, then a basis for V is $\{\beta_1, \cdots, \beta_r\}$. Thus*

$$\dim(V) = \sum_{i=1}^{r} \dim(V_i).$$

Proof: Suppose $\sum_{i=1}^{r} \sum_{j=1}^{m_i} c_{ij} v_j^i = 0$. then since it is a direct sum, it follows for each i,

$$\sum_{j=1}^{m_i} c_{ij} v_j^i = 0$$

and now since $\left\{v_1^i, \cdots, v_{m_i}^i\right\}$ is a basis, each $c_{ij} = 0$. ■

Here is a fundamental lemma.

Lemma 9.1.7 *Let L_i be in $\mathcal{L}(V, V)$ and suppose for $i \neq j$, $L_i L_j = L_j L_i$ 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.

Proof : Note that since the operators commute, $L_j : \ker(L_i) \to \ker(L_i)$. Here is why. If $L_i y = 0$ so that $y \in \ker(L_i)$, then

$$L_i L_j y = L_j L_i y = L_j 0 = 0$$

and so $L_j : \ker(L_i) \mapsto \ker(L_i)$. Next observe that it is obvious that, since the operators commute,

$$\sum_{i=1}^{p} \ker(L_p) \subseteq \ker\left(\prod_{i=1}^{p} L_i\right)$$

Next, why is $\sum_i \ker(L_p) = \ker(L_1) \oplus \cdots \oplus \ker(L_p)$? Suppose

$$\sum_{i=1}^{p} v_i = 0, \ v_i \in \ker(L_i),$$

but some $v_i \neq 0$. Then do $\prod_{j \neq i} L_j$ to both sides. Since the linear transformations commute, this results in

$$\prod_{j \neq i} L_j(v_i) = 0$$

which contradicts the assumption that these L_j are one to one on $\ker(L_i)$ and the observation that they map $\ker(L_i)$ to $\ker(L_i)$. Thus if

$$\sum_i v_i = 0,\ v_i \in \ker(L_i)$$

then each $v_i = 0$. 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 9.1.6,

$$\begin{aligned}\sum_{i=1}^{p} \dim(\ker(L_i)) &= \dim(\ker(L_1) \oplus + \cdots + \oplus \ker(L_p)) \\ &\leq \dim\left(\ker\left(\prod_{i=1}^{p} L_i\right)\right) \leq \sum_{i=1}^{p} \dim(\ker(L_i))\end{aligned}$$

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 v is not in the span of this basis, then 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

$$\ker(L_1) \oplus + \cdots + \oplus \ker(L_p) = \ker\left(\prod_{i=1}^{p} L_i\right) \blacksquare$$

9.2 Direct Sums, Block Diagonal Matrices

Let V be a finite dimensional vector space with field of scalars $\mathbb{F}$. Here I will make no assumption on $\mathbb{F}$. Also suppose $A \in \mathcal{L}(V, V)$.

Recall Lemma 8.4.3 which gives the existence of the minimal polynomial for a linear transformation A. This is the monic polynomial p which has smallest possible degree such that $p(A) = 0$. It is stated again for convenience.

Lemma 9.2.1 *Let $A \in \mathcal{L}(V, V)$ where V is a finite dimensional vector space of dimension n with field of scalars $\mathbb{F}$. Then there exists a unique monic polynomial of the form*

$$p(\lambda) = \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$$

such that $p(A) = 0$ and m is as small as possible for this to occur.

Now it is time to consider the notion of a direct sum of subspaces. Recall you can always assert the existence of a factorization of the minimal polynomial into a product of irreducible polynomials. This fact will now be used to show how to obtain such a direct sum of subspaces.

Definition 9.2.2 *For $A \in \mathcal{L}(V, V)$ where $\dim(V) = n$, suppose the minimal polynomial is*

$$p(\lambda) = \prod_{k=1}^{q} (\phi_k(\lambda))^{r_k}$$

where the polynomials ϕ_k have coefficients in $\mathbb{F}$ and are irreducible. Now define the generalized eigenspaces

$$V_k \equiv \ker\left(\left(\phi_k\left(A\right)\right)^{r_k}\right)$$

Note that if one of these polynomials $(\phi_k(\lambda))^{r_k}$ is a monic linear polynomial, then the generalized eigenspace would be an eigenspace.

Theorem 9.2.3 *In the context of Definition 9.2.2,*

$$V = V_1 \oplus \cdots \oplus V_q \tag{9.2}$$

and each V_k is A invariant, meaning $A\left(V_k\right) \subseteq V_k$. $\phi_l\left(A\right)$ is one to one on each V_k for $k \neq l$. If $\beta_i = \left\{v_1^i, \cdots, v_{m_i}^i\right\}$ is a basis for V_i, then $\left\{\beta_1, \beta_2, \cdots, \beta_q\right\}$ is a basis for V.

Proof: It is clear V_k is a subspace which is A invariant because A commutes with $\phi_k\left(A\right)^{m_k}$. It is clear the operators $\phi_k\left(A\right)^{r_k}$ commute. Thus if $v \in V_k$,

$$\phi_k\left(A\right)^{r_k}\phi_l\left(A\right)^{r_l}v = \phi_l\left(A\right)^{r_l}\phi_k\left(A\right)^{r_k}v = \phi_l\left(A\right)^{r_l}0 = 0$$

and so $\phi_l\left(A\right)^{r_l} : V_k \to V_k$.

I claim $\phi_l\left(A\right)$ is one to one on V_k whenever $k \neq l$. The two polynomials $\phi_l\left(\lambda\right)$ and $\phi_k\left(\lambda\right)^{r_k}$ are relatively prime so there exist polynomials $m\left(\lambda\right), n\left(\lambda\right)$ such that

$$m\left(\lambda\right)\phi_l\left(\lambda\right) + n\left(\lambda\right)\phi_k\left(\lambda\right)^{r_k} = 1$$

It follows that the sum of all coefficients of λ raised to a positive power are zero and the constant term on the left is 1. Therefore, using the convention $A^0 = I$ it follows

$$m\left(A\right)\phi_l\left(A\right) + n\left(A\right)\phi_k\left(A\right)^{r_k} = I$$

If $v \in V_k$, then from the above,

$$m\left(A\right)\phi_l\left(A\right)v + n\left(A\right)\phi_k\left(A\right)^{r_k}v = v$$

Since v is in V_k, it follows by definition,

$$m\left(A\right)\phi_l\left(A\right)v = v$$

and so $\phi_l\left(A\right)v \neq 0$ unless $v = 0$. Thus $\phi_l\left(A\right)$ and hence $\phi_l\left(A\right)^{r_l}$ is one to one on V_k for every $k \neq l$. By Lemma 9.1.7 and the fact that $\ker\left(\prod_{k=1}^q \phi_k\left(\lambda\right)^{r_k}\right) = V$, 9.2 is obtained. The claim about the bases follows from Lemma 9.1.6. ■

You could consider the restriction of A to V_k. It turns out that this restriction has minimal polynomial equal to $\phi_k\left(\lambda\right)^{m_k}$.

Corollary 9.2.4 *Let the minimal polynomial of A be $p\left(\lambda\right) = \prod_{k=1}^q \phi_k\left(\lambda\right)^{m_k}$ where each ϕ_k is irreducible. Let $V_k = \ker\left(\phi\left(A\right)^{m_k}\right)$. Then*

$$V_1 \oplus \cdots \oplus V_q = V$$

and letting A_k denote the restriction of A to V_k, it follows the minimal polynomial of A_k is $\phi_k\left(\lambda\right)^{m_k}$.

Proof: Recall the direct sum, $V_1 \oplus \cdots \oplus V_q = V$ where $V_k = \ker\left(\phi_k\left(A\right)^{m_k}\right)$ for $p\left(\lambda\right) = \prod_{k=1}^q \phi_k\left(\lambda\right)^{m_k}$ the minimal polynomial for A where the $\phi_k\left(\lambda\right)$ are all irreducible. Thus each V_k is invariant with respect to A. What is the minimal polynomial of A_k, the restriction of A to V_k? First note that $\phi_k\left(A_k\right)^{m_k}\left(V_k\right) = \{0\}$ by definition. Thus if $\eta\left(\lambda\right)$ is the minimal polynomial for A_k then it must divide $\phi_k(\lambda)^{m_k}$ and so by Corollary 7.3.11 $\eta(\lambda) = \phi_k(\lambda)^{r_k}$ where $r_k \leq m_k$. Could $r_k < m_k$? No, this is not possible because then $p(\lambda)$ would fail to be the minimal polynomial for A. You could substitute for the term $\phi_k(\lambda)^{m_k}$ in the factorization of $p(\lambda)$ with $\phi_k(\lambda)^{r_k}$ and the resulting polynomial p' would satisfy $p'(A) = 0$. Here is why. From Theorem 9.2.3, a typical $x \in V$ is of the form

$$\sum_{i=1}^{q} v_i,\ v_i \in V_i$$

Then since all the factors commute,

$$p'(A)\left(\sum_{i=1}^{q} v_i\right) = \prod_{i\neq k}^{q} \phi_i(A)^{m_i}\,\phi_k(A)^{r_k}\left(\sum_{i=1}^{q} v_i\right)$$

For $j \neq k$

$$\prod_{i\neq k}^{q} \phi_i(A)^{m_i}\,\phi_k(A)^{r_k}\,v_j = \prod_{i\neq k,j}^{q} \phi_i(A)^{m_i}\,\phi_k(A)^{r_k}\,\phi_j(A)^{m_j}\,v_j = 0$$

If $j = k$,

$$\prod_{i\neq k}^{q} \phi_i(A)^{m_i}\,\phi_k(A)^{r_k}\,v_k = 0$$

which shows $p'(\lambda)$ is a monic polynomial having smaller degree than $p(\lambda)$ such that $p'(A) = 0$. Thus the minimal polynomial for A_k is $\phi_k(\lambda)^{m_k}$ as claimed. ■

How does Theorem 9.2.3 relate to matrices?

Theorem 9.2.5 *Suppose V is a vector space with field of scalars $\mathbb{F}$ and $A \in \mathcal{L}(V,V)$. Suppose also*

$$V = V_1 \oplus \cdots \oplus V_q$$

where each V_k is A invariant. ($AV_k \subseteq V_k$) Also let β_k be an ordered basis for V_k and let A_k denote the restriction of A to V_k. Letting M^k denote the matrix of A_k with respect to this basis, it follows the matrix of A with respect to the basis $\{\beta_1, \cdots, \beta_q\}$ is

$$\begin{pmatrix} M^1 & & 0 \\ & \ddots & \\ 0 & & M^q \end{pmatrix}$$

Proof: Let β denote the ordered basis $\{\beta_1, \cdots, \beta_q\}$, $|\beta_k|$ being the number of vectors in β_k. Let $q_k : \mathbb{F}^{|\beta_k|} \to V_k$ be the usual map such that the following diagram commutes.

$$\begin{array}{ccc} & A_k & \\ V_k & \rightarrow & V_k \\ q_k \uparrow & \circ & \uparrow q_k \\ \mathbb{F}^{|\beta_k|} & \rightarrow & \mathbb{F}^{|\beta_k|} \\ & M^k & \end{array}$$

Thus $A_k q_k = q_k M^k$. Then if q is the map from $\mathbb{F}^n$ to V corresponding to the ordered basis β just described,

$$q\begin{pmatrix} \mathbf{0} & \cdots & \mathbf{x} & \cdots & \mathbf{0} \end{pmatrix}^T = q_k\mathbf{x},$$

where $\mathbf{x}$ occupies the positions between $\sum_{i=1}^{k-1}|\beta_i|+1$ and $\sum_{i=1}^{k}|\beta_i|$. Then M will be the matrix of A with respect to β if and only if a similar diagram to the above commutes. Thus it is required that $Aq = qM$. However, from the description of q just made, and the invariance of each V_k,

$$Aq\begin{pmatrix}\mathbf{0}\\ \vdots\\ \mathbf{x}\\ \vdots\\ \mathbf{0}\end{pmatrix} = A_k q_k \mathbf{x} = q_k M^k \mathbf{x} = q\begin{pmatrix} M^1 & & & & 0\\ & \ddots & & & \\ & & M^k & & \\ & & & \ddots & \\ 0 & & & & M^q\end{pmatrix}\begin{pmatrix}\mathbf{0}\\ \vdots\\ \mathbf{x}\\ \vdots\\ \mathbf{0}\end{pmatrix}$$

It follows that the above block diagonal matrix is the matrix of A with respect to the given ordered basis. ■

An examination of the proof of the above theorem yields the following corollary.

Corollary 9.2.6 *If any β_k in the above consists of eigenvectors, then M^k is a diagonal matrix having the corresponding eigenvalues down the diagonal.*

It follows that it would be interesting to consider special bases for the vector spaces in the direct sum. This leads to the Jordan form or more generally other canonical forms such as the rational canonical form.

9.3 Cyclic Sets

It was shown above that for $A \in \mathcal{L}(V,V)$ for V a finite dimensional vector space over the field of scalars $\mathbb{F}$, there exists a direct sum decomposition

$$V = V_1 \oplus \cdots \oplus V_q$$

where

$$V_k = \ker\left(\phi_k(A)^{m_k}\right)$$

and $\phi_k(\lambda)$ is an irreducible polynomial. Here the minimal polynomial of A was

$$\prod_{k=1}^{q}\phi_k(\lambda)^{m_k}$$

Next I will consider the problem of finding a basis for V_k such that the matrix of A restricted to V_k assumes various forms.

Definition 9.3.1 *Letting $x \neq 0$ denote by β_x the vectors $\left\{x, Ax, A^2x, \cdots, A^{m-1}x\right\}$ where m is the smallest such that $A^m x \in \operatorname{span}\left(x, \cdots, A^{m-1}x\right)$. This is called an A cyclic set. The vectors which result are also called a Krylov sequence. For such a sequence of vectors, $|\beta_x| \equiv m$.*

The first thing to notice is that such a Krylov sequence is always linearly independent.

Lemma 9.3.2 *Let $\beta_x = \left\{x, Ax, A^2x, \cdots, A^{m-1}x\right\}, x \neq 0$ where m is the smallest such that $A^m x \in \operatorname{span}\left(x, \cdots, A^{m-1}x\right)$. Then $\left\{x, Ax, A^2x, \cdots, A^{m-1}x\right\}$ is linearly independent.*

Proof: Suppose that there are scalars a_k, not all zero such that

$$\sum_{k=0}^{m-1} a_k A^k x = 0$$

Then letting a_r be the last nonzero scalar in the sum, you can divide by a_r and solve for $A^r x$ as a linear combination of the $A^j x$ for $j < r \leq m-1$ contrary to the definition of m. ■

Now here is a nice lemma which has been pretty much discussed earlier.

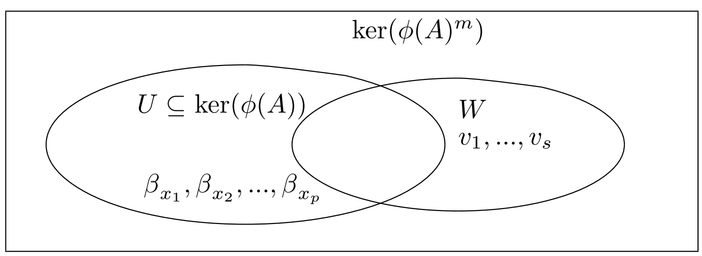
Lemma 9.3.3 *Suppose W is a subspace of V where V is a finite dimensional vector space and $L \in \mathcal{L}(V,V)$ and suppose $LW = LV$. Then $V = W + \ker(L)$.*

Proof: Let a basis for $LV = LW$ be $\{Lw_1, \cdots, Lw_m\}, w_i \in W$. Then let $y \in V$. Thus $Ly = \sum_{i=1}^m c_i L w_i$ and so

$$L\left(\overbrace{y - \sum_{i=1}^m c_i w_i}^{=z}\right) \equiv Lz = 0$$

It follows that $z \in \ker(L)$ and so $y = \sum_{i=1}^m c_i w_i + z \in W + \ker(L)$. ■

For more on the next lemma and the following theorem, see Hofman and Kunze [15]. I am following the presentation in Friedberg Insel and Spence [10]. See also Herstein [14] for a different approach to canonical forms. To help organize the ideas in the lemma, here is a diagram.



Lemma 9.3.4 *Let W be an A invariant $(AW \subseteq W)$ subspace of $\ker(\phi(A)^m)$ for m a positive integer where $\phi(\lambda)$ is an irreducible monic polynomial of degree d. Let U be an A invariant subspace of $\ker(\phi(A))$.*

If $\{v_1, \cdots, v_s\}$ is a basis for W then if $x \in U \setminus W$,

$$\{v_1, \cdots, v_s, \beta_x\}$$

is linearly independent.

There exist vectors $x_1, \cdots, x_p$ each in U such that

$$\left\{v_1, \cdots, v_s, \beta_{x_1}, \cdots, \beta_{x_p}\right\}$$

is a basis for

$$U + W.$$

Also, if $x \in \ker(\phi(A)^m)$, $|\beta_x| = kd$ where $k \leq m$. Here $|\beta_x|$ is the length of β_x, the degree of the monic polynomial $\eta(\lambda)$ satisfying $\eta(A)x = 0$ with $\eta(\lambda)$ having smallest possible degree.

Proof: Claim: If $x \in \ker \phi(A)$, and $|\beta_x|$ denotes the length of β_x, then $|\beta_x| = d$ the degree of the irreducible polynomial $\phi(\lambda)$ and so

$$\beta_x = \left\{x, Ax, A^2x, \cdots, A^{d-1}x\right\}$$

also $\text{span}(\beta_x)$ is A invariant, $A(\text{span}(\beta_x)) \subseteq \text{span}(\beta_x)$.

Proof of the claim: Let $m = |\beta_x|$. That is, there exists monic $\eta(\lambda)$ of degree m and $\eta(A)x = 0$ with m is as small as possible for this to happen. Then from the usual process of division of polynomials, there exist $l(\lambda), r(\lambda)$ such that $r(\lambda) = 0$ or else has smaller degree than that of $\eta(\lambda)$ such that

$$\phi(\lambda) = \eta(\lambda) l(\lambda) + r(\lambda)$$

If $\deg(r(\lambda)) < \deg(\eta(\lambda))$, then the equation implies $0 = \phi(A)x = r(A)x$ and so m was incorrectly chosen. Hence $r(\lambda) = 0$ and so if $l(\lambda) \neq 1$, then $\eta(\lambda)$ divides $\phi(\lambda)$ contrary to the assumption that $\phi(\lambda)$ is irreducible. Hence $l(\lambda) = 1$ and $\eta(\lambda) = \phi(\lambda)$. The claim about $\text{span}(\beta_x)$ is obvious because $A^d x \in \text{span}(\beta_x)$. This shows the claim.

Suppose now $x \in U \setminus W$ where $U \subseteq \ker(\phi(A))$. Consider

$$\{v_1, \cdots, v_s, \beta_x\}.$$

Is this set of vectors independent? Suppose

$$\sum_{i=1}^{s} a_i v_i + \sum_{j=1}^{d} d_j A^{j-1} x = 0.$$

If $z \equiv \sum_{j=1}^{d} d_j A^{j-1} x$, then $z \in W \cap \text{span}\left(x, Ax, \cdots, A^{d-1}x\right)$. Then also for each $m \leq d-1$,

$$A^m z \in W \cap \text{span}\left(x, Ax, \cdots, A^{d-1}x\right)$$

because $W, \text{span}\left(x, Ax, \cdots, A^{d-1}x\right)$ are A invariant. Therefore,

$$\begin{aligned} \text{span}\left(z, Az, \cdots, A^{d-1}z\right) &\subseteq W \cap \text{span}\left(x, Ax, \cdots, A^{d-1}x\right) \\ &\subseteq \text{span}\left(x, Ax, \cdots, A^{d-1}x\right) \end{aligned} \tag{9.3}$$

Suppose $z \neq 0$. Then from the Lemma 9.3.2 above, $\left\{z, Az, \cdots, A^{d-1}z\right\}$ must be linearly independent. Therefore,

$$d = \dim\left(\text{span}\left(z, Az, \cdots, A^{d-1}z\right)\right) \leq \dim\left(W \cap \text{span}\left(x, Ax, \cdots, A^{d-1}x\right)\right)$$

$$\leq \dim\left(\text{span}\left(x, Ax, \cdots, A^{d-1}x\right)\right) = d$$

Thus

$$W \cap \text{span}\left(x, Ax, \cdots, A^{d-1}x\right) = \text{span}\left(x, Ax, \cdots, A^{d-1}x\right)$$

which would require $x \in W$ but this is assumed not to take place. Hence $z = 0$ and so the linear independence of the $\{v_1, \cdots, v_s\}$ implies each $a_i = 0$. Then the linear independence of $\left\{x, Ax, \cdots, A^{d-1}x\right\}$, which follows from Lemma 9.3.2, shows each $d_j = 0$. Thus $\left\{v_1, \cdots, v_s, x, Ax, \cdots, A^{d-1}x\right\}$ is linearly independent as claimed.

Let $x \in U \setminus W \subseteq \ker(\phi(A))$. Then it was just shown that $\{v_1, \cdots, v_s, \beta_x\}$ is linearly independent. Let W_1 be given by

$$y \in \text{span}(v_1, \cdots, v_s, \beta_x) \equiv W_1$$

Then W_1 is A invariant. If W_1 equals $U+W$, then you are done. If not, let W_1 play the role of W and pick $x_1 \in U \setminus W_1$ and repeat the argument. Continue till

$$\operatorname{span}\left(v_1, \cdots, v_s, \beta_{x_1}, \cdots, \beta_{x_n}\right) = U + W$$

The process stops because $\ker\left(\phi\left(A\right)^m\right)$ is finite dimensional.

Finally, letting $x \in \ker\left(\phi\left(A\right)^m\right)$, there is a monic polynomial $\eta\left(\lambda\right)$ such that $\eta\left(A\right)x = 0$ and $\eta\left(\lambda\right)$ is of smallest possible degree, which degree equals $\left|\beta_x\right|$. Then

$$\phi\left(\lambda\right)^m = \eta\left(\lambda\right) l\left(\lambda\right) + r\left(\lambda\right)$$

If $\deg\left(r\left(\lambda\right)\right) < \deg\left(\eta\left(\lambda\right)\right)$, then $r\left(A\right)x = 0$ and $\eta\left(\lambda\right)$ was incorrectly chosen. Hence $r\left(\lambda\right) = 0$ and so $\eta\left(\lambda\right)$ must divide $\phi\left(\lambda\right)^m$. Hence by Corollary 7.3.11 $\eta\left(\lambda\right) = \phi\left(\lambda\right)^k$ where $k \leq m$. Thus $\left|\beta_x\right| = kd = \deg\left(\eta\left(\lambda\right)\right)$. ■

With this preparation, here is the main result about a basis V where $A \in \mathcal{L}\left(V,V\right)$ and the minimal polynomial for A is $\phi\left(A\right)^m$ for $\phi\left(\lambda\right)$ irreducible an irreducible monic polynomial. There is a very interesting generalization of this theorem in [15] which pertains to the existence of complementary subspaces. For an outline of this generalization, see Problem 9 on Page 292.

Theorem 9.3.5 *Suppose $A \in \mathcal{L}\left(V,V\right)$ for V some finite dimensional vector space. Then for each $k \in \mathbb{N}$, there exists a cyclic basis for $\ker\left(\phi\left(A\right)^k\right)$ which is one of the form $\beta = \left\{\beta_{x_1}, \cdots, \beta_{x_p}\right\}$ or $\ker\left(\phi\left(A\right)^k\right) = \{0\}$. Note that if $\ker\left(\phi\left(A\right)\right) \neq \{0\}$, then the same is true for all $\ker\left(\phi\left(A\right)^k\right), k \in \mathbb{N}$.*

Proof: If $k = 1$, you can use Lemma 9.3.4 and let $W = \{0\}$ and $U = \ker\left(\phi\left(A\right)\right)$ to obtain the cyclic basis. Suppose then that the theorem is true for $m-1, m-1 \geq 1$ meaning that for any finite dimensional vector space V and $A \in \mathcal{L}\left(V,V\right)$, $\ker\left(\phi\left(A\right)^k\right)$ has a cyclic basis for all $k \leq m-1$. Consider a new vector space $\phi\left(A\right)\ker\left(\phi\left(A\right)^m\right) \equiv \hat{V}$ in place of V and the restriction of A to $\hat{V}$ which we will call $\hat{A}$. Then $\hat{A} \in \mathcal{L}\left(\hat{V},\hat{V}\right)$. It follows $\phi\left(A\right)^{m-1}\left(\phi\left(A\right)\ker\left(\phi\left(A\right)^m\right)\right) = \phi\left(A\right)^{m-1}\hat{V} = 0$ and since $\phi\left(\lambda\right)$ is irreducible, the minimum polynomial of $\hat{A}$ on $\hat{V}$ is $\phi\left(\hat{A}\right)^k$ for some $k \leq m-1$. Thus $\ker\left(\phi\left(\hat{A}\right)^k\right) \equiv \left\{v \in \hat{V} : \phi\left(\hat{A}\right)^k v = 0\right\}$. Since $k \leq m-1$ the cyclic basis in $\hat{V}$ exists by induction. If $k = 0$, then you would have $\hat{V} = \{0\}$ and $\{0\} = \phi\left(A\right)\ker\left(\phi\left(A\right)^m\right) \supseteq \ker\left(\phi\left(A\right)\right)$ so nothing is of any interest because all of these spaces are $\{0\}$.

Let the cyclic basis for $\hat{V} \equiv \phi\left(A\right)\ker\left(\phi\left(A\right)^m\right)$ be

$$\left\{\beta_{x_1}, \cdots, \beta_{x_p}\right\},$$

$x_i \in \phi\left(A\right)\ker\left(\phi\left(A\right)^m\right)$. Let $x_i = \phi\left(A\right)y_i, y_i \in \ker\left(\phi\left(A\right)^m\right)$. Consider $\left\{\beta_{y_1}, \cdots, \beta_{y_p}\right\}$, $y_i \in \ker\left(\phi\left(A\right)^m\right)$. Are these vectors independent? Suppose

$$0 = \sum_{i=1}^{p}\sum_{j=1}^{\left|\beta_{y_i}\right|} a_{ij}A^{j-1}y_i \equiv \sum_{i=1}^{p} f_i\left(A\right)y_i \tag{9.4}$$

If the sum involved x_i in place of y_i, then something could be said because $\left\{\beta_{x_1}, \cdots, \beta_{x_p}\right\}$ is a basis.

Do $\phi(A)$ to both sides to obtain

$$0=\sum_{i=1}^{p}\sum_{j=1}^{\left|\beta_{y_i}\right|}a_{ij}A^{j-1}x_i\equiv\sum_{i=1}^{p}f_i\left(\hat{A}\right)x_i$$

Now $f_i\left(\hat{A}\right)x_i=0$ for each i since $f_i\left(\hat{A}\right)x_i\in\operatorname{span}\left(\beta_{x_i}\right)$ and as just mentioned,

$$\left\{\beta_{x_1},\cdots,\beta_{x_p}\right\}$$

is a basis. Let $\eta_i(\lambda)$ be the monic polynomial of smallest degree such that $\eta_i\left(\hat{A}\right)x_i=0$. Then

$$f_i(\lambda)=\eta_i(\lambda)\,l(\lambda)+r(\lambda)$$

where $r(\lambda)=0$ or else it has smaller degree than $\eta_i(\lambda)$. However, the equation then shows that $r\left(\hat{A}\right)x_i=0$ which would contradict the choice of $\eta_i(\lambda)$. Thus $r(\lambda)=0$ and $\eta_i(\lambda)$ divides $f_i(\lambda)$. Also, $\phi\left(\hat{A}\right)^{m-1}x_i=\phi\left(\hat{A}\right)^{m-1}\phi(A)\,y_i=0$ and so $\eta_i(\lambda)$ must divide $\phi(\lambda)^{m-1}$. From Corollary 7.3.11, it follows that, since $\phi(\lambda)$ is irreducible, $\eta_i(\lambda)=\phi(\lambda)^r$ for some $r\leq m-1$. Thus $\phi(\lambda)$ divides $\eta_i(\lambda)$ which divides $f_i(\lambda)$. Hence $f_i(\lambda)=\phi(\lambda)\,g_i(\lambda)$! Now

$$0=\sum_{i=1}^{p}f_i(A)\,y_i=\sum_{i=1}^{p}g_i(A)\,\phi(A)\,y_i=\sum_{i=1}^{p}g_i\left(\hat{A}\right)x_i.$$

By the same reasoning just given, since $g_i\left(\hat{A}\right)x_i\in\operatorname{span}\left(\beta_{x_i}\right)$, it follows that each

$$g_i\left(\hat{A}\right)x_i=0.$$

Therefore,

$$f_i(A)\,y_i=g_i\left(\hat{A}\right)\phi(A)\,y_i=g_i\left(\hat{A}\right)x_i=0.$$

Therefore,

$$f_i(A)\,y_i=\sum_{j=1}^{\left|\beta_{y_j}\right|}a_{ij}A^{j-1}y_i=0$$

and by independence of the β_{y_i}, this implies $a_{ij}=0$ for each j for each i.

Next, it follows from the definition that

$$\phi(A)\left(\ker\left(\phi(A)^m\right)\right)=\operatorname{span}\left(\beta_{x_1},\cdots,\beta_{x_p}\right).$$

Now

$$W\equiv\operatorname{span}\left(\beta_{y_1},\cdots,\beta_{y_p}\right)\subseteq\ker\left(\phi(A)^m\right)$$

because each $y_i\in\ker\left(\phi(A)^m\right)$. Then from the above description of $\left\{\beta_{x_1},\cdots,\beta_{x_p}\right\}$ as a cyclic basis for $\phi(A)\left(\ker\left(\phi(A)^m\right)\right)$,

$$\begin{aligned}\phi(A)\left(\ker\left(\phi(A)^m\right)\right)&=\operatorname{span}\left(\beta_{x_1},\cdots,\beta_{x_p}\right)\subseteq\phi(A)\operatorname{span}\left(\beta_{y_1},\cdots,\beta_{y_p}\right)\\&\equiv\phi(A)(W)\subseteq\phi(A)\ker\left(\phi(A)^m\right)\end{aligned}$$

To see the first inclusion,

$$A^r x_q = A^r \phi(A) y_q = \phi(A) A^r y_q \in \phi(A) \operatorname{span}\left(\beta_{y_q}\right) \subseteq \phi(A) \operatorname{span}\left(\beta_{y_1}, \cdots, \beta_{y_p}\right)$$

It follows from Lemma 9.3.3 that $\ker(\phi(A)^m) = W + \ker(\phi(A))$. From Lemma 9.3.4 $W + \ker(\phi(A))$ has a basis of the form $\left\{\beta_{y_1}, \cdots, \beta_{y_p}, \beta_{z_1}, \cdots, \beta_{z_s}\right\}$. ■

9.4 Nilpotent Transformations

Definition 9.4.1 *Let V be a vector space over the field of scalars $\mathbb{F}$. Then $N \in \mathcal{L}(V, V)$ is called nilpotent if for some m, it follows that $N^m = 0$.*

The following lemma contains some significant observations about nilpotent transformations.

Lemma 9.4.2 *Suppose $N^k x \neq 0$. Then $\left\{x, Nx, \cdots, N^k x\right\}$ is linearly independent. Also, the minimal polynomial of N is λ^m where m is the first such that $N^m = 0$.*

Proof: Suppose $\sum_{i=0}^k c_i N^i x = 0$ where not all $c_i = 0$. There exists l such that $k \leq l < m$ and $N^{l+1} x = 0$ but $N^l x \neq 0$. Then multiply both sides by N^l to conclude that $c_0 = 0$. Next multiply both sides by N^{l-1} to conclude that $c_1 = 0$ and continue this way to obtain that all the $c_i = 0$.

Next consider the claim that λ^m is the minimal polynomial. If $p(\lambda)$ is the minimal polynomial, then by the division algorithm,

$$\lambda^m = p(\lambda) l(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than that of $p(\lambda)$ or else $r(\lambda) = 0$. The above implies $0 = 0 + r(N)$ contrary to $p(\lambda)$ being minimal. Hence $r(\lambda) = 0$ and so $p(\lambda)$ divides λ^m. Hence $p(\lambda) = \lambda^k$ for $k \leq m$. But if $k < m$, this would contradict the definition of m as being the smallest such that $N^m = 0$. ■

For such a nilpotent transformation, let $\left\{\beta_{x_1}, \cdots, \beta_{x_q}\right\}$ be a basis for $\ker(N^m) = V$ where these β_{x_i} are cyclic. This basis exists thanks to Theorem 9.3.5. Note that you can have $|\beta_x| < m$ because it is possible for $N^k x = 0$ without $N^k = 0$. Thus

$$V = \operatorname{span}\left(\beta_{x_1}\right) \oplus \cdots \oplus \operatorname{span}\left(\beta_{x_q}\right),$$

each of these subspaces in the above direct sum being N invariant. For x one of the x_k, consider β_x given by

$$x, Nx, N^2 x, \cdots, N^{r-1} x$$

where $N^r x$ is in the span of the above vectors. Then by the above lemma, $N^r x = 0$.

By Theorem 9.2.5, the matrix of N with respect to the above basis is the block diagonal matrix

$$\begin{pmatrix} M^1 & & 0 \\ & \ddots & \\ 0 & & M^q \end{pmatrix}$$

where M^k denotes the matrix of N restricted to $\operatorname{span}\left(\beta_{x_k}\right)$. In computing this matrix, I will order β_{x_k} as follows:

$$\left(N^{r_k - 1} x_k, \cdots, x_k\right)$$

Also the cyclic sets $\beta_{x_1}, \beta_{x_2}, \cdots, \beta_{x_q}$ will be ordered according to length, the length of β_{x_i} being at least as large as the length of $\beta_{x_{i+1}}$, $\left|\beta_{x_k}\right| \equiv r_k$. Then since $N^{r_k} x_k = 0$, it is now easy to find M^k. Using the procedure mentioned above for determining the matrix of a linear transformation,

$$\begin{pmatrix} 0 & N^{r_k-1}x_k & \cdots & Nx_k \end{pmatrix} =$$

$$\begin{pmatrix} N^{r_k-1}x_k & N^{r_k-2}x_k & \cdots & x_k \end{pmatrix} \begin{pmatrix} 0 & 1 & & 0 \\ 0 & 0 & \ddots & \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Thus the matrix M_k is the $r_k \times r_k$ matrix which has ones down the super diagonal and zeros elsewhere. The following convenient notation will be used.

Definition 9.4.3 *$J_k(\alpha)$ 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.

Then with this definition and the above discussion, the following proposition has been proved.

Proposition 9.4.4 *Let $N \in \mathcal{L}(W, W)$ be nilpotent,*

$$N^m = 0$$

for some $m \in \mathbb{N}$. Here W is a p dimensional vector space with field of scalars $\mathbb{F}$. Then there exists a basis for W such that the matrix of N with respect to this basis is of the form

$$J = \begin{pmatrix} J_{r_1}(0) & & & 0 \\ & J_{r_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{r_s}(0) \end{pmatrix} \tag{9.5}$$

where $r_1 \geq r_2 \geq \cdots \geq r_s \geq 1$ and $\sum_{i=1}^{s} r_i = p$. In the above, the $J_{r_j}(0)$ is called a Jordan block of size $r_j \times r_j$ with 0 down the main diagonal.

Observation 9.4.5 *Observe that $J_r(0)^r = 0$ but $J_r(0)^{r-1} \neq 0$.*

In fact, the matrix of the above proposition is unique.

Corollary 9.4.6 *Let J, J' both be matrices of the nilpotent linear transformation $N \in \mathcal{L}(W, W)$ which are of the form described in Proposition 9.4.4. Then $J = J'$. In fact, if the rank of J^k equals the rank of J'^k for all nonnegative integers k, then $J = J'$.*

Proof: Since J and J' are similar, it follows that for each k an integer, J^k and J'^k are similar. Hence, for each k, these matrices have the same rank. Now suppose $J \neq J'$. Note first that

$$J_r(0)^r = 0,\ J_r(0)^{r-1} \neq 0.$$

Denote the blocks of J as $J_{r_k}(0)$ and the blocks of J' as $J_{r'_k}(0)$. Let k be the first such that $J_{r_k}(0) \neq J_{r'_k}(0)$. Suppose that $r_k > r'_k$. By block multiplication and the above observation, it follows that the two matrices J^{r_k-1} and J'^{r_k-1} are respectively of the forms

$$\begin{pmatrix} M_{r_1} & & & & & 0 \\ & \ddots & & & & \\ & & M_{r_k} & & & \\ & & & * & & \\ & & & & \ddots & \\ 0 & & & & & * \end{pmatrix}, \begin{pmatrix} M_{r'_1} & & & & & 0 \\ & \ddots & & & & \\ & & M_{r'_k} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{pmatrix}$$

where $M_{r_j} = M_{r'_j}$ for $j \leq k-1$ but $M_{r'_k}$ is a zero $r'_k \times r'_k$ matrix while M_{r_k} is a larger matrix which is not equal to 0. For example, M_{r_k} could look like

$$M_{r_k} = \begin{pmatrix} 0 & \cdots & 1 \\ & \ddots & \vdots \\ 0 & & 0 \end{pmatrix}$$

Thus there are more pivot columns in J^{r_k-1} than in $(J')^{r_k-1}$, contradicting the requirement that J^k and J'^k have the same rank. ■

9.5 The Jordan Canonical Form

The Jordan canonical form has to do with the case where the minimal polynomial of $A \in \mathcal{L}(V,V)$ splits. Thus there exist λ_k in the field of scalars such that the minimal polynomial of A is of the form

$$p(\lambda) = \prod_{k=1}^{r} (\lambda - \lambda_k)^{m_k}$$

Recall the following which follows from Theorem 8.4.4.

Proposition 9.5.1 *Let the minimal polynomial of $A \in \mathcal{L}(V,V)$ be given by*

$$p(\lambda) = \prod_{k=1}^{r} (\lambda - \lambda_k)^{m_k}$$

Then the eigenvalues of A are $\{\lambda_1, \cdots, \lambda_r\}$.

It follows from Corollary 9.2.3 that

$$\begin{aligned} V &= \ker(A - \lambda_1 I)^{m_1} \oplus \cdots \oplus \ker(A - \lambda_r I)^{m_r} \\ &\equiv V_1 \oplus \cdots \oplus V_r \end{aligned}$$

where I denotes the identity linear transformation. Without loss of generality, let the dimensions of the V_k be decreasing from left to right. These V_k are called the generalized eigenspaces.

It follows from the definition of V_k that $(A - \lambda_k I)$ is nilpotent on V_k and clearly each V_k is A invariant. Therefore from Proposition 9.4.4, and letting A_k denote the restriction of A to V_k, there exists an ordered basis for V_k, β_k such that with respect to this basis, the matrix of $(A_k - \lambda_k I)$ is of the form given in that proposition, denoted here by J^k. What is the matrix of A_k with respect to β_k? Letting $\{b_1, \cdots, b_r\} = \beta_k$,

$$A_k b_j = (A_k - \lambda_k I)\, b_j + \lambda_k I b_j \equiv \sum_s J^k_{sj} b_s + \sum_s \lambda_k \delta_{sj} b_s = \sum_s \left(J^k_{sj} + \lambda_k \delta_{sj}\right) b_s$$

and so the matrix of A_k with respect to this basis is $J^k + \lambda_k I$ where I is the identity matrix.

Therefore, with respect to the ordered basis $\{\beta_1, \cdots, \beta_r\}$ the matrix of A is in Jordan canonical form. This means the matrix is of the form

$$\begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_r) \end{pmatrix} \tag{9.6}$$

where $J(\lambda_k)$ is an $m_k \times m_k$ matrix of the form

$$\begin{pmatrix} J_{k_1}(\lambda_k) & & & 0 \\ & J_{k_2}(\lambda_k) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\lambda_k) \end{pmatrix} \tag{9.7}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = m_k$. Here $J_k(\lambda)$ is a $k \times k$ Jordan block of the form

$$\begin{pmatrix} \lambda & 1 & & 0 \\ 0 & \lambda & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 0 & \lambda \end{pmatrix} \tag{9.8}$$

This proves the existence part of the following fundamental theorem.

Note that if any of the β_k consists of eigenvectors, then the corresponding Jordan block will consist of a diagonal matrix having λ_k down the main diagonal. This corresponds to $m_k = 1$. The vectors which are in $\ker (A - \lambda_k I)^{m_k}$ which are not in $\ker (A - \lambda_k I)$ are called generalized eigenvectors.

The following is the main result on the Jordan canonical form.

Theorem 9.5.2 *Let V be an n dimensional vector space with field of scalars $\mathbb{C}$ or some other field such that the minimal polynomial of $A \in \mathcal{L}(V, V)$ completely factors into powers of linear factors. Then there exists a unique Jordan canonical form for A as described in 9.6 - 9.8, where uniqueness is in the sense that any two have the same number and size of Jordan blocks.*

Proof: It only remains to verify uniqueness. Suppose there are two, J and J'. Then these are matrices of A with respect to possibly different bases and so they are similar. Therefore, they have the same minimal polynomials and the generalized eigenspaces have the same dimension. Thus the size of the matrices $J(\lambda_k)$ and $J'(\lambda_k)$ defined by the dimension of these generalized eigenspaces, also corresponding to the algebraic multiplicity of λ_k, must be the same. Therefore, they comprise the same set of positive integers. Thus listing the eigenvalues in the same order, corresponding blocks $J(\lambda_k), J'(\lambda_k)$ are the same size.

It remains to show that $J(\lambda_k)$ and $J'(\lambda_k)$ are not just the same size but also are the same up to order of the Jordan blocks running down their respective diagonals. It is only necessary to worry about the number and size of the Jordan blocks making up $J(\lambda_k)$ and $J'(\lambda_k)$. Since J, J' are similar, so are $J - \lambda_k I$ and $J' - \lambda_k I$.

Thus the following two matrices are similar

$$A \equiv \begin{pmatrix} J(\lambda_1) - \lambda_k I & & & & 0 \\ & \ddots & & & \\ & & J(\lambda_k) - \lambda_k I & & \\ & & & \ddots & \\ 0 & & & & J(\lambda_r) - \lambda_k I \end{pmatrix}$$

$$B \equiv \begin{pmatrix} J'(\lambda_1) - \lambda_k I & & & & 0 \\ & \ddots & & & \\ & & J'(\lambda_k) - \lambda_k I & & \\ & & & \ddots & \\ 0 & & & & J'(\lambda_r) - \lambda_k I \end{pmatrix}$$

and consequently, $\operatorname{rank}(A^k) = \operatorname{rank}(B^k)$ for all $k \in \mathbb{N}$. Also, both $J(\lambda_j) - \lambda_k I$ and $J'(\lambda_j) - \lambda_k I$ are one to one for every $\lambda_j \neq \lambda_k$. Since all the blocks in both of these matrices are one to one except the blocks $J'(\lambda_k) - \lambda_k I$, $J(\lambda_k) - \lambda_k I$, it follows that this requires the two sequences of numbers $\{\operatorname{rank}((J(\lambda_k) - \lambda_k I)^m)\}_{m=1}^{\infty}$ and $\left\{\operatorname{rank}\left((J'(\lambda_k) - \lambda_k I)^m\right)\right\}_{m=1}^{\infty}$ must be the same.

Then

$$J(\lambda_k) - \lambda_k I \equiv \begin{pmatrix} J_{k_1}(0) & & & 0 \\ & J_{k_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(0) \end{pmatrix}$$

and a similar formula holds for $J'(\lambda_k)$

$$J'(\lambda_k) - \lambda_k I \equiv \begin{pmatrix} J_{l_1}(0) & & & 0 \\ & J_{l_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{l_p}(0) \end{pmatrix}$$

and it is required to verify that $p = r$ and that the same blocks occur in both. Without loss of generality, let the blocks be arranged according to size with the largest on upper left corner falling to smallest in lower right. Now the desired conclusion follows from Corollary 9.4.6. ■

Note that if any of the generalized eigenspaces $\ker(A - \lambda_k I)^{m_k}$ has a basis of eigenvectors, then it would be possible to use this basis and obtain a diagonal matrix in the block corresponding to λ_k. By uniqueness, this is **the** block corresponding to the eigenvalue λ_k. Thus when this happens, the block in the Jordan canonical form corresponding to λ_k is just the diagonal matrix having λ_k down the diagonal and there are **no generalized eigenvectors**.

The Jordan canonical form is very significant when you try to understand powers of a matrix. There exists an $n \times n$ matrix S[1] such that

$$A = S^{-1}JS.$$

Therefore, $A^2 = S^{-1}JSS^{-1}JS = S^{-1}J^2S$ and continuing this way, it follows

$$A^k = S^{-1}J^kS.$$

where J is given in the above corollary. Consider J^k. By block multiplication,

$$J^k = \begin{pmatrix} J_1^k & & 0 \\ & \ddots & \\ 0 & & J_r^k \end{pmatrix}.$$

The matrix J_s is an $m_s \times m_s$ matrix which is of the form

$$J_s = D + N$$

for D a multiple of the identity and N an upper triangular matrix with zeros down the main diagonal. Thus $N^{m_s} = 0$. Now since D is just a multiple of the identity, it follows that $DN = ND$. Therefore, the usual binomial theorem may be applied and this yields the following equations for $k \geq m_s$.

$$\begin{aligned} J_s^k &= (D+N)^k = \sum_{j=0}^{k} \binom{k}{j} D^{k-j}N^j \\ &= \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j}N^j, \end{aligned} \tag{9.9}$$

the third equation holding because $N^{m_s} = 0$. Thus J_s^k is of the form

$$J_s^k = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}.$$

Lemma 9.5.3 *Suppose J is of the form J_s, a Jordan block where the constant α, on the main diagonal is less than one in absolute value. Then*

$$\lim_{k\to\infty} \left(J^k\right)_{ij} = 0.$$

Proof: From 9.9, it follows that for large k, and $j \leq m_s$,

$$\binom{k}{j} \leq \frac{k(k-1)\cdots(k-m_s+1)}{m_s!}.$$

Therefore, letting C be the largest value of $\left|\left(N^j\right)_{pq}\right|$ for $0 \leq j \leq m_s$,

$$\left|\left(J^k\right)_{pq}\right| \leq m_s C \left(\frac{k(k-1)\cdots(k-m_s+1)}{m_s!}\right) |\alpha|^{k-m_s}$$

[1] The S here is written as S^{-1} in the corollary.

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. ■

9.6 Exercises

1. In the discussion of Nilpotent transformations, it was asserted that if two $n \times n$ matrices A, B are similar, then A^k is also similar to B^k. Why is this so? If two matrices are similar, why must they have the same rank?

2. If A, B are both invertible, then they are both row equivalent to the identity matrix. Are they necessarily similar? Explain.

3. Suppose you have two nilpotent matrices A, B and A^k and B^k both have the same rank for all $k \geq 1$. Does it follow that A, B are similar? What if it is not known that A, B are nilpotent? Does it follow then?

4. When we say a polynomial equals zero, we mean that all the coefficients equal 0. If we assign a different meaning to it which says that a polynomial $p(\lambda)$ equals zero when it is the zero function, ($p(\lambda) = 0$ for every $\lambda \in \mathbb{F}$.) does this amount to the same thing? Is there any difference in the two definitions for ordinary fields like $\mathbb{Q}$? **Hint:** Consider for the field of scalars $\mathbb{Z}_2$, the integers mod 2 and consider $p(\lambda) = \lambda^2 + \lambda$.

5. Let $A \in \mathcal{L}(V, V)$ where V is a finite dimensional vector space with field of scalars $\mathbb{F}$. Let $p(\lambda)$ be the minimal polynomial and suppose $\phi(\lambda)$ is any nonzero polynomial such that $\phi(A)$ is not one to one and $\phi(\lambda)$ has smallest possible degree such that $\phi(A)$ is nonzero and not one to one. Show $\phi(\lambda)$ must divide $p(\lambda)$.

6. Let $A \in \mathcal{L}(V, V)$ where V is a finite dimensional vector space with field of scalars $\mathbb{F}$. Let $p(\lambda)$ be the minimal polynomial and suppose $\phi(\lambda)$ is an irreducible polynomial with the property that $\phi(A)x = 0$ for some specific $x \neq 0$. Show that $\phi(\lambda)$ must divide $p(\lambda)$. **Hint:** First write $p(\lambda) = \phi(\lambda) g(\lambda) + r(\lambda)$ where $r(\lambda)$ is either 0 or has degree smaller than the degree of $\phi(\lambda)$. If $r(\lambda) = 0$ you are done. Suppose it is not 0. Let $\eta(\lambda)$ be the monic polynomial of smallest degree with the property that $\eta(A)x = 0$. Now use the Euclidean algorithm to divide $\phi(\lambda)$ by $\eta(\lambda)$. Contradict the irreducibility of $\phi(\lambda)$.

7. Suppose A is a linear transformation and let the characteristic polynomial be
$$\det(\lambda I - A) = \prod_{j=1}^{q} \phi_j(\lambda)^{n_j}$$
where the $\phi_j(\lambda)$ are irreducible. Explain using Corollary 7.3.11 why the irreducible factors of the minimal polynomial are $\phi_j(\lambda)$ and why the minimal polynomial is of the form $\prod_{j=1}^{q} \phi_j(\lambda)^{r_j}$ where $r_j \leq n_j$. You can use the Cayley Hamilton theorem if you like.

8. Let
$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

Find the minimal polynomial for A.

9. Suppose A is an $n \times n$ matrix and let $\mathbf{v}$ be a vector. Consider the A cyclic set of vectors $\left\{\mathbf{v}, A\mathbf{v}, \cdots, A^{m-1}\mathbf{v}\right\}$ where this is an independent set of vectors but $A^m\mathbf{v}$ is a linear combination of the preceding vectors in the list. Show how to obtain a monic polynomial of smallest degree, m, $\phi_{\mathbf{v}}(\lambda)$ such that
$$\phi_{\mathbf{v}}(A)\,\mathbf{v} = \mathbf{0}$$
Now let $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ be a basis and let $\phi(\lambda)$ be the least common multiple of the $\phi_{\mathbf{w}_k}(\lambda)$. Explain why this must be the minimal polynomial of A. Give a reasonably easy algorithm for computing $\phi_{\mathbf{v}}(\lambda)$.

10. Here is a matrix.
$$\begin{pmatrix} -7 & -1 & -1 \\ -21 & -3 & -3 \\ 70 & 10 & 10 \end{pmatrix}$$
Using the process of Problem 9 find the minimal polynomial of this matrix. It turns out the characteristic polynomial is λ^3.

11. Find the minimal polynomial for
$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ -3 & 2 & 1 \end{pmatrix}$$
by the above technique. Is what you found also the characteristic polynomial?

12. Let A be an $n \times n$ matrix with field of scalars $\mathbb{C}$. Letting λ be an eigenvalue, show the dimension of the eigenspace equals the number of Jordan blocks in the Jordan canonical form which are associated with λ. Recall the eigenspace is $\ker(\lambda I - A)$.

13. For any $n \times n$ matrix, why is the dimension of the eigenspace always less than or equal to the algebraic multiplicity of the eigenvalue as a root of the characteristic equation? **Hint:** Note the algebraic multiplicity is the size of the appropriate block in the Jordan form.

14. Give an example of two nilpotent matrices which are not similar but have the same minimal polynomial if possible.

15. Use the existence of the Jordan canonical form for a linear transformation whose minimal polynomial factors completely to give a proof of the Cayley Hamilton theorem which is valid for any field of scalars. **Hint:** First assume the minimal polynomial factors completely into linear factors. If this does not happen, consider a splitting field of the minimal polynomial. Then consider the minimal polynomial with respect to this larger field. How will the two minimal polynomials be related? Show the minimal polynomial always divides the characteristic polynomial.

16. Here is a matrix. Find its Jordan canonical form by directly finding the eigenvectors and generalized eigenvectors based on these to find a basis which will yield the Jordan form. The eigenvalues are 1 and 2.
$$\begin{pmatrix} -3 & -2 & 5 & 3 \\ -1 & 0 & 1 & 2 \\ -4 & -3 & 6 & 4 \\ -1 & -1 & 1 & 3 \end{pmatrix}$$

Why is it typically impossible to find the Jordan canonical form?

17. People like to consider the solutions of first order linear systems of equations which are of the form
$$\mathbf{x}'(t) = A\mathbf{x}(t)$$
where here A is an $n \times n$ matrix. From the theorem on the Jordan canonical form, there exist S and S^{-1} such that $A = SJS^{-1}$ where J is a Jordan form. Define $\mathbf{y}(t) \equiv S^{-1}\mathbf{x}(t)$. Show $\mathbf{y}' = J\mathbf{y}$. Now suppose $\Psi(t)$ is an $n \times n$ matrix whose columns are solutions of the above differential equation. Thus
$$\Psi' = A\Psi$$
Now let Φ be defined by $S\Phi S^{-1} = \Psi$. Show
$$\Phi' = J\Phi.$$

18. In the above Problem show that
$$\det(\Psi)' = \operatorname{trace}(A)\det(\Psi)$$
and so
$$\det(\Psi(t)) = Ce^{\operatorname{trace}(A)t}$$
This is called Abel's formula and $\det(\Psi(t))$ is called the Wronskian. **Hint:** Show it suffices to consider
$$\Phi' = J\Phi$$
and establish the formula for Φ. Next let
$$\Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}$$
where the ϕ_j are the rows of Φ. Then explain why
$$\det(\Phi)' = \sum_{i=1}^{n} \det(\Phi_i) \tag{9.10}$$
where Φ_i is the same as Φ except the i^{th} row is replaced with ϕ_i' instead of the row ϕ_i. Now from the form of J,
$$\Phi' = D\Phi + N\Phi$$
where N has all nonzero entries above the main diagonal. Explain why
$$\phi_i'(t) = \lambda_i \phi_i(t) + a_i \phi_{i+1}(t)$$
Now use this in the formula for the derivative of the Wronskian given in 9.10 and use properties of determinants to obtain
$$\det(\Phi)' = \sum_{i=1}^{n} \lambda_i \det(\Phi).$$
Obtain Abel's formula
$$\det(\Phi) = Ce^{\operatorname{trace}(A)t}$$
and so the Wronskian $\det \Phi$ either vanishes identically or never.

19. Let A be an $n \times n$ matrix and let J be its Jordan canonical form. Recall J is a block diagonal matrix having blocks $J_k(\lambda)$ down the diagonal. Each of these blocks is of the form
$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}$$
Now for $\varepsilon > 0$ given, let the diagonal matrix D_ε be given by
$$D_\varepsilon = \begin{pmatrix} 1 & & & 0 \\ & \varepsilon & & \\ & & \ddots & \\ 0 & & & \varepsilon^{k-1} \end{pmatrix}$$
Show that $D_\varepsilon^{-1} J_k(\lambda) D_\varepsilon$ has the same form as $J_k(\lambda)$ but instead of ones down the super diagonal, there is ε down the super diagonal. That is $J_k(\lambda)$ is replaced with
$$\begin{pmatrix} \lambda & \varepsilon & & 0 \\ & \lambda & \ddots & \\ & & \ddots & \varepsilon \\ 0 & & & \lambda \end{pmatrix}$$
Now show that for A an $n \times n$ matrix, it is similar to one which is just like the Jordan canonical form except instead of the blocks having 1 down the super diagonal, it has ε.

20. Let A be in $\mathcal{L}(V,V)$ and suppose that $A^p x \neq 0$ for some $x \neq 0$. Show that $A^p e_k \neq 0$ for some $e_k \in \{e_1, \cdots, e_n\}$, a basis for V. If you have a matrix which is nilpotent, ($A^m = 0$ for some m) will it always be possible to find its Jordan form? Describe how to do it if this is the case. **Hint:** First explain why all the eigenvalues are 0. Then consider the way the Jordan form for nilpotent transformations was constructed in the above.

21. Suppose A is an $n \times n$ matrix and that it has n distinct eigenvalues. How do the minimal polynomial and characteristic polynomials compare? Determine other conditions based on the Jordan Canonical form which will cause the minimal and characteristic polynomials to be different.

22. Suppose A is a 3×3 matrix and it has at least two distinct eigenvalues. Is it possible that the minimal polynomial is different than the characteristic polynomial?

23. If A is an $n \times n$ matrix of entries from a field of scalars and if the minimal polynomial of A splits over this field of scalars, does it follow that the characteristic polynomial of A also splits? Explain why or why not.

24. Show that if two $n \times n$ matrices A, B are similar, then they have the same minimal polynomial and also that if this minimal polynomial is of the form $p(\lambda) = \prod_{i=1}^{s} \phi_i(\lambda)^{r_i}$ where the $\phi_i(\lambda)$ are irreducible and monic, then $\ker(\phi_i(A)^{r_i})$ and $\ker(\phi_i(B)^{r_i})$ have the same dimension. Why is this so? This was what was responsible for the blocks corresponding to an eigenvalue being of the same size.

25. Show that a given complex $n \times n$ matrix is non defective (diagonalizable) if and only if the minimal polynomial has no repeated roots.

26. Describe a straight forward way to determine the minimal polynomial of an $n \times n$ matrix using row operations. Next show that if $p(\lambda)$ and $p'(\lambda)$ are relatively prime, then $p(\lambda)$ has no repeated roots. With the above problem, explain how this gives a way to determine whether a matrix is non defective.

27. In Theorem 9.3.5 show that each cyclic set β_x is associated with a monic polynomial $\eta_x(\lambda)$ such that $\eta_x(A)(x) = 0$ and this polynomial has smallest possible degree such that this happens. Show that the cyclic sets β_{x_i} can be arranged such that $\eta_{x_{i+1}}(\lambda) / \eta_{x_i}(\lambda)$.

28. Show that if A is a complex $n \times n$ matrix, then A and A^T are similar. **Hint:** Consider a Jordan block. Note that

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

29. Let A be a linear transformation defined on a finite dimensional vector space V. Let the minimal polynomial be $\prod_{i=1}^{q} \phi_i(\lambda)^{m_i}$ and let $\left(\beta_{v_1^i}^i, \cdots, \beta_{v_{r_i}^i}^i\right)$ be the cyclic sets such that $\left\{\beta_{v_1^i}^i, \cdots, \beta_{v_{r_i}^i}^i\right\}$ is a basis for $\ker(\phi_i(A)^{m_i})$. Let $v = \sum_i \sum_j v_j^i$. Now let $q(\lambda)$ be any polynomial and suppose that

$$q(A)v = 0$$

Show that it follows $q(A) = 0$. **Hint:** First consider the special case where a basis for V is $\{x, Ax, \cdots, A^{n-1}x\}$ and $q(A)x = 0$.

9.7 The Rational Canonical Form*

Here one has the minimal polynomial in the form $\prod_{k=1}^{q} \phi(\lambda)^{m_k}$ where $\phi(\lambda)$ is an irreducible monic polynomial. It is not necessarily the case that $\phi(\lambda)$ is a linear factor. Thus this case is completely general and includes the situation where the field is arbitrary. In particular, it includes the case where the field of scalars is, for example, the rational numbers. This may be partly why it is called the rational canonical form. As you know, the rational numbers are notorious for not having roots to polynomial equations which have integer or rational coefficients.

This canonical form is due to Frobenius. I am following the presentation given in [10] and there are more details given in this reference. Another good source which has additional results is [15].

Here is a definition of the concept of a companion matrix.

Definition 9.7.1 *Let*

$$q(\lambda) = a_0 + a_1\lambda + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n$$

be a monic polynomial. The companion matrix of $q(\lambda)$, *denoted as* $C(q(\lambda))$ *is the matrix*

$$\begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & 0 & & -a_1 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & -a_{n-1} \end{pmatrix}$$

Proposition 9.7.2 *Let $q(\lambda)$ be a polynomial and let $C(q(\lambda))$ be its companion matrix. Then $q(C(q(\lambda))) = 0$.*

Proof: Write C instead of $C(q(\lambda))$ for short. Note that

$$C\mathbf{e}_1 = \mathbf{e}_2, C\mathbf{e}_2 = \mathbf{e}_3, \cdots, C\mathbf{e}_{n-1} = \mathbf{e}_n$$

Thus

$$\mathbf{e}_k = C^{k-1}\mathbf{e}_1,\ k = 1, \cdots, n \tag{9.11}$$

and so it follows

$$\left\{\mathbf{e}_1, C\mathbf{e}_1, C^2\mathbf{e}_1, \cdots, C^{n-1}\mathbf{e}_1\right\} \tag{9.12}$$

are linearly independent. Hence these form a basis for $\mathbb{F}^n$. Now note that $C\mathbf{e}_n$ is given by

$$C\mathbf{e}_n = -a_0\mathbf{e}_1 - a_1\mathbf{e}_2 - \cdots - \mathbf{a}_{n-1}\mathbf{e}_n$$

and from 9.11 this implies

$$C^n\mathbf{e}_1 = -a_0\mathbf{e}_1 - a_1C\mathbf{e}_1 - \cdots - \mathbf{a}_{n-1}C^{n-1}\mathbf{e}_1$$

and so $q(C)\mathbf{e}_1 = \mathbf{0}$. Now since 9.12 is a basis, every vector of $\mathbb{F}^n$ is of the form $k(C)\mathbf{e}_1$ for some polynomial $k(\lambda)$. Therefore, if $\mathbf{v} \in \mathbb{F}^n$,

$$q(C)\mathbf{v} = q(C)k(C)\mathbf{e}_1 = k(C)q(C)\mathbf{e}_1 = \mathbf{0}$$

which shows $q(C) = 0$. ■

The following theorem is on the existence of the rational canonical form.

Theorem 9.7.3 *Let $A \in \mathcal{L}(V,V)$ where V is a vector space with field of scalars $\mathbb{F}$ and minimal polynomial $\prod_{i=1}^{q}\phi_i(\lambda)^{m_i}$ where each $\phi_i(\lambda)$ is irreducible and monic. Letting $V_k \equiv \ker\left(\phi_k(\lambda)^{m_k}\right)$, it follows*

$$V = V_1 \oplus \cdots \oplus V_q$$

where each V_k is A invariant. Letting B_k denote a basis for V_k and M^k the matrix of the restriction of A to V_k, it follows that the matrix of A with respect to the basis $\{B_1, \cdots, B_q\}$ is the block diagonal matrix of the form

$$\begin{pmatrix} M^1 & & 0 \\ & \ddots & \\ 0 & & M^q \end{pmatrix} \tag{9.13}$$

If B_k is given as $\left\{\beta_{v_1}, \cdots, \beta_{v_s}\right\}$ as described in Theorem 9.3.5 where each β_{v_j} is an A cyclic set of vectors, then the matrix M^k is of the form

$$M^k = \begin{pmatrix} C\left(\phi_k(\lambda)^{r_1}\right) & & 0 \\ & \ddots & \\ 0 & & C\left(\phi_k(\lambda)^{r_s}\right) \end{pmatrix} \tag{9.14}$$

where the A cyclic sets of vectors may be arranged in order such that the positive integers r_j satisfy $r_1 \geq \cdots \geq r_s$ and $C\left(\phi_k(\lambda)^{r_j}\right)$ is the companion matrix of the polynomial $\phi_k(\lambda)^{r_j}$.

Proof: By Theorem 9.2.5 the matrix of A with respect to $\{B_1, \cdots, B_q\}$ is of the form given in 9.13. Now by Theorem 9.3.5 the basis B_k may be chosen in the form $\{\beta_{v_1}, \cdots, \beta_{v_s}\}$ where each β_{v_k} is an A cyclic set of vectors and also it can be assumed the lengths of these β_{v_k} are decreasing. Thus

$$V_k = \operatorname{span}\left(\beta_{v_1}\right) \oplus \cdots \oplus \operatorname{span}\left(\beta_{v_s}\right)$$

and it only remains to consider the matrix of A restricted to $\operatorname{span}\left(\beta_{v_k}\right)$. Then you can apply Theorem 9.2.5 to get the result in 9.14. Say

$$\beta_{v_k} = v_k, Av_k, \cdots, A^{d-1}v_k$$

where $\eta(A) v_k = 0$ and the degree of $\eta(\lambda)$ is d, the smallest degree such that this is so, η being a monic polynomial. Then $\eta(\lambda)$ must divide $\phi_k(\lambda)^{m_k}$. By Corollary 7.3.11, $\eta(\lambda) = \phi_k(\lambda)^{r_k}$ where $r_k \leq m_k$. It remains to consider the matrix of A restricted to $\operatorname{span}\left(\beta_{v_k}\right)$. Say

$$\eta(\lambda) = \phi_k(\lambda)^{r_k} = a_0 + a_1\lambda + \cdots + a_{d-1}\lambda^{d-1} + \lambda^d$$

Thus, since $\eta(A) v_k = 0$,

$$A^d v_k = -a_0 v_k - a_1 A v_k - \cdots - a_{d-1} A^{d-1} v_k$$

Recall the formalism for finding the matrix of A restricted to this invariant subspace.

$$\begin{pmatrix} Av_k & A^2v_k & A^3v_k & \cdots & -a_0v_k - a_1Av_k - \cdots - a_{d-1}A^{d-1}v_k \end{pmatrix} =$$

$$\begin{pmatrix} v_k & Av_k & A^2v_k & \cdots & A^{d-1}v_k \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & \cdots & -a_0 \\ 1 & 0 & & & -a_1 \\ 0 & 1 & \ddots & & \vdots \\ & \ddots & \ddots & 0 & -a_{d-2} \\ 0 & & 0 & 1 & -a_{d-1} \end{pmatrix}$$

Thus the matrix of the transformation is the above. This is the companion matrix of $\phi_k(\lambda)^{r_k} = \eta(\lambda)$. In other words, $C = C\left(\phi_k(\lambda)^{r_k}\right)$ and so M^k has the form claimed in the theorem. ■

9.8 Uniqueness

Given $A \in \mathcal{L}(V, V)$ where V is a vector space having field of scalars $\mathbb{F}$, the above shows there exists a rational canonical form for A. Could A have more than one rational canonical form? Recall the definition of an A cyclic set. For convenience, here it is again.

Definition 9.8.1 *Letting $x \neq 0$ denote by β_x the vectors $\{x, Ax, A^2x, \cdots, A^{m-1}x\}$ where m is the smallest such that $A^m x \in \operatorname{span}\left(x, \cdots, A^{m-1}x\right)$.*

The following proposition ties these A cyclic sets to polynomials. It is just a review of ideas used above to prove existence.

Proposition 9.8.2 *Let $x \neq 0$ and consider $\{x, Ax, A^2x, \cdots, A^{m-1}x\}$. Then this is an A cyclic set if and only if there exists a monic polynomial $\eta(\lambda)$ such that $\eta(A) x = 0$ and among all such polynomials $\psi(\lambda)$ satisfying $\psi(A) x = 0$, $\eta(\lambda)$ has the smallest degree. If $V = \ker\left(\phi(\lambda)^m\right)$ where $\phi(\lambda)$ is monic and irreducible, then for some positive integer $p \leq m, \eta(\lambda) = \phi(\lambda)^p$.*

The following is the main consideration for proving uniqueness. It will depend on what was already shown for the Jordan canonical form. This will apply to the nilpotent matrix $\phi(A)$.

Lemma 9.8.3 *Let V be a vector space and $A \in \mathcal{L}(V,V)$ has minimal polynomial $\phi(\lambda)^m$ where $\phi(\lambda)$ is irreducible and has degree d. Let the basis for V consist of $\{\beta_{v_1}, \cdots, \beta_{v_s}\}$ where β_{v_k} is A cyclic as described above and the rational canonical form for A is the matrix taken with respect to this basis. Then letting $|\beta_{v_k}|$ denote the number of vectors in β_{v_k}, it follows there is only one possible set of numbers $|\beta_{v_k}|$.*

Proof: Say β_{v_j} is associated with the polynomial $\phi(\lambda)^{p_j}$. Thus, as described above $\left|\beta_{v_j}\right|$ equals $p_j d$. Consider the following table which comes from the A cyclic set

$$\{v_j, Av_j, \cdots, A^{d-1}v_j, \cdots, A^{p_j d-1}v_j\}$$

α_0^j	α_1^j	α_2^j	$\cdots$	α_{d-1}^j
v_j	Av_j	A^2v_j	$\cdots$	$A^{d-1}v_j$
$\phi(A)v_j$	$\phi(A)Av_j$	$\phi(A)A^2v_j$	$\cdots$	$\phi(A)A^{d-1}v_j$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\phi(A)^{p_j-1}v_j$	$\phi(A)^{p_j-1}Av_j$	$\phi(A)^{p_j-1}A^2v_j$	$\cdots$	$\phi(A)^{p_j-1}A^{d-1}v_j$

In the above, α_k^j signifies the vectors below it in the k^{th} column. None of these vectors below the top row are equal to 0 because the degree of $\phi(\lambda)^{p_j-1}\lambda^{d-1}$ is $dp_j - 1$, which is less than $p_j d$ and the smallest degree of a nonzero polynomial sending v_j to 0 is $p_j d$. Also, each of these vectors is in the span of β_{v_j} and there are dp_j of them, just as there are dp_j vectors in β_{v_j}.

Claim: The vectors $\{\alpha_0^j, \cdots, \alpha_{d-1}^j\}$ are linearly independent.

Proof of claim: Suppose

$$\sum_{i=0}^{d-1}\sum_{k=0}^{p_j-1} c_{ik}\phi(A)^k A^i v_j = 0$$

Then multiplying both sides by $\phi(A)^{p_j-1}$ this yields

$$\sum_{i=0}^{d-1} c_{i0}\phi(A)^{p_j-1} A^i v_j = 0$$

this is because if $k \geq 1$, you have a typical term of the form

$$c_{ik}\phi(A)^{p_j-1}\phi(A)^k A^i v_j = A^i\phi(A)^{k-1} c_{ik}\phi(A)^{p_j} v_j = 0$$

Now if any of the c_{i0} is nonzero this would imply there exists a polynomial having degree smaller than $p_j d$ which sends v_j to 0. In fact, the polynomial would have degree $d-1+p_j-1$. Since this does not happen, it follows each $c_{i0} = 0$. Thus

$$\sum_{i=0}^{d-1}\sum_{k=1}^{p_j-1} c_{ik}\phi(A)^k A^i v_j = 0$$

Now multiply both sides by $\phi(A)^{p_j-2}$ and do a similar argument to assert that $c_{i1}=0$ for each i. Continuing this way, all the $c_{ik}=0$ and this proves the claim.

Thus the vectors $\left\{\alpha_0^j,\cdots,\alpha_{d-1}^j\right\}$ are linearly independent and there are $p_jd=\left|\beta_{v_j}\right|$ of them. Therefore, they form a basis for span $\left(\beta_{v_j}\right)$. Also note that if you list the columns in reverse order starting from the bottom and going toward the top, the vectors $\left\{\alpha_0^j,\cdots,\alpha_{d-1}^j\right\}$ yield Jordan blocks in the matrix of $\phi(A)$. Hence, considering all these vectors $\left\{\alpha_0^j,\cdots,\alpha_{d-1}^j\right\}_{j=1}^s$, each listed in the reverse order, the matrix of $\phi(A)$ with respect to this basis of V is in Jordan canonical form. See Proposition 9.4.4 and Theorem 9.5.2 on existence and uniqueness for the Jordan form. This Jordan form is unique up to order of the blocks. For a given j $\left\{\alpha_0^j,\cdots,\alpha_{d-1}^j\right\}$ yields d Jordan blocks of size p_j for $\phi(A)$. The size and number of Jordan blocks of $\phi(A)$ depends only on $\phi(A)$, hence only on A. Once A is determined, $\phi(A)$ is determined and hence the number and size of Jordan blocks is determined, so the exponents p_j are determined and this shows the lengths of the β_{v_j}, p_jd are also determined. ■

Note that if the p_j are known, then so is the rational canonical form because it comes from blocks which are companion matrices of the polynomials $\phi(\lambda)^{p_j}$. Now here is the main result.

Theorem 9.8.4 *Let V be a vector space having field of scalars $\mathbb{F}$ and let $A\in\mathcal{L}(V,V)$. Then the rational canonical form of A is unique up to order of the blocks.*

Proof: Let the minimal polynomial of A be $\prod_{k=1}^q\phi_k(\lambda)^{m_k}$. Then recall from Corollary 9.2.3

$$V=V_1\oplus\cdots\oplus V_q$$

where $V_k=\ker\left(\phi_k(A)^{m_k}\right)$. Also recall from Corollary 9.2.4 that the minimal polynomial of the restriction of A to V_k is $\phi_k(\lambda)^{m_k}$. Now apply Lemma 9.8.3 to A restricted to V_k. ■

In the case where two $n\times n$ matrices M,N are similar, recall this is equivalent to the two being matrices of the same linear transformation taken with respect to two different bases. Hence each are similar to the same rational canonical form.

Example 9.8.5 *Here is a matrix.*

$$A=\begin{pmatrix}5&-2&1\\2&10&-2\\9&0&9\end{pmatrix}$$

Find a similarity transformation which will produce the rational canonical form for A.

The minimal polynomial is $\lambda^3-24\lambda^2+180\lambda-432$. Why? This factors as

$$(\lambda-6)^2(\lambda-12)$$

Thus $\mathbb{Q}^3$ is the direct sum of $\ker\left((A-6I)^2\right)$ and $\ker(A-12I)$. Consider the first of these. You see easily that this is

$$y\begin{pmatrix}1\\1\\0\end{pmatrix}+z\begin{pmatrix}-1\\0\\1\end{pmatrix},y,z\in\mathbb{Q}.$$

What about the length of A cyclic sets? It turns out it doesn't matter much. You can start with either of these and get a cycle of length 2. Lets pick the second one. This leads to the cycle

$$\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -4 \\ -4 \\ 0 \end{pmatrix} = A \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -12 \\ -48 \\ -36 \end{pmatrix} = A^2 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

where the last of the three is a linear combination of the first two. Take the first two as the first two columns of S. To get the third, you need a cycle of length 1 corresponding to $\ker(A - 12I)$. This yields the eigenvector $\begin{pmatrix} 1 & -2 & 3 \end{pmatrix}^T$. Thus

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

Now using Proposition 8.3.10, the Rational canonical form for A should be

$$\begin{pmatrix} -1 & -4 & 1 \\ 0 & -4 & -2 \\ 1 & 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 5 & -2 & 1 \\ 2 & 10 & -2 \\ 9 & 0 & 9 \end{pmatrix} \begin{pmatrix} -1 & -4 & 1 \\ 0 & -4 & -2 \\ 1 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & -36 & 0 \\ 1 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix}$$

Example 9.8.6 *Here is a matrix.*

$$A = \begin{pmatrix} 12 & -3 & -19 & -14 & 8 \\ -4 & 1 & 1 & 6 & -4 \\ 4 & 5 & 5 & -2 & 4 \\ 0 & -5 & -5 & 2 & 0 \\ -4 & 3 & 11 & 6 & 0 \end{pmatrix}$$

Find a basis such that if S is the matrix which has these vectors as columns $S^{-1}AS$ is in rational canonical form assuming the field of scalars is $\mathbb{Q}$.

First it is necessary to find the minimal polynomial. Of course you can find the characteristic polynomial and then take away factors till you find the minimal polynomial. However, there is a much better way which is described in the exercises. Leaving out this detail, the minimal polynomial is

$$\lambda^3 - 12\lambda^2 + 64\lambda - 128$$

This polynomial factors as

$$(\lambda - 4)\left(\lambda^2 - 8\lambda + 32\right) \equiv \phi_1(\lambda)\,\phi_2(\lambda)$$

where the second factor is irreducible over $\mathbb{Q}$. Consider $\phi_2(\lambda)$ first. Messy computations yield

$$\ker(\phi_2(A)) = a \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + b \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + c \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + d \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Now start with one of these basis vectors and look for an A cycle. Picking the first one, you obtain the cycle

$$\begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -15 \\ 5 \\ 1 \\ -5 \\ 7 \end{pmatrix}$$

because the next vector involving A^2 yields a vector which is in the span of the above two. You check this by making the vectors the columns of a matrix and finding the row reduced echelon form. Clearly this cycle does not span $\ker(\phi_2(A))$, so look for another cycle. Begin with a vector which is not in the span of these two. The last one works well. Thus another A cycle is

$$\begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -16 \\ 4 \\ -4 \\ 0 \\ 8 \end{pmatrix}$$

It follows a basis for $\ker(\phi_2(A))$ is

$$\left\{ \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -16 \\ 4 \\ -4 \\ 0 \\ 8 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -15 \\ 5 \\ 1 \\ -5 \\ 7 \end{pmatrix} \right\}$$

Finally consider a cycle coming from $\ker(\phi_1(A))$. This amounts to nothing more than finding an eigenvector for A corresponding to the eigenvalue 4. An eigenvector is

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

Now the desired matrix for the similarity transformation is

$$S \equiv \begin{pmatrix} -2 & -16 & -1 & -15 & -1 \\ 0 & 4 & 1 & 5 & 0 \\ 0 & -4 & 0 & 1 & 0 \\ 0 & 0 & 0 & -5 & 0 \\ 1 & 8 & 0 & 7 & 1 \end{pmatrix}$$

Then doing the computations, you get

$$S^{-1}AS = \begin{pmatrix} 0 & -32 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & -32 & 0 \\ 0 & 0 & 1 & 8 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}$$

and you see this is in rational canonical form, the two 2×2 blocks being companion matrices for the polynomial $\lambda^2 - 8\lambda + 32$ and the 1×1 block being a companion matrix for $\lambda - 4$. Note that you could have written this without finding a similarity transformation to produce it. This follows from the above theory which gave the existence of the rational canonical form.

Obviously there is a lot more which could be considered about rational canonical forms. Just begin with a strange field and start investigating what can be said. One can also derive more systematic methods for finding the rational canonical form. The advantage of this is you don't need to find the eigenvalues in order to compute the rational canonical form and it can often be computed for this reason, unlike the Jordan form. The uniqueness of this rational canonical form can be used to determine whether two matrices consisting of entries in some field are similar.

9.9 Exercises

1. Suppose A is a linear transformation and let the characteristic polynomial be
$$\det(\lambda I - A) = \prod_{j=1}^{q} \phi_j(\lambda)^{n_j}$$
where the $\phi_j(\lambda)$ are irreducible. Explain using Corollary 7.3.11 why the irreducible factors of the minimal polynomial are $\phi_j(\lambda)$ and why the minimal polynomial is of the form $\prod_{j=1}^{q} \phi_j(\lambda)^{r_j}$ where $r_j \leq n_j$. You can use the Cayley Hamilton theorem if you like.

2. Find the minimal polynomial for
$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ -3 & 2 & 1 \end{pmatrix}$$
by the above technique assuming the field of scalars is the rational numbers. Is what you found also the characteristic polynomial?

3. Show, using the rational root theorem, the minimal polynomial for A in the above problem is irreducible with respect to $\mathbb{Q}$. Letting the field of scalars be $\mathbb{Q}$ find the rational canonical form and a similarity transformation which will produce it.

4. Letting the field of scalars be $\mathbb{Q}$, find the rational canonical form for the matrix
$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ 2 & 3 & 0 & 2 \\ 1 & 3 & 2 & 4 \\ 1 & 2 & 1 & 2 \end{pmatrix}$$

5. Let $A : \mathbb{Q}^3 \to \mathbb{Q}^3$ be linear. Suppose the minimal polynomial is $(\lambda - 2)\left(\lambda^2 + 2\lambda + 7\right)$. Find the rational canonical form. Can you give generalizations of this rather simple problem to other situations?

6. Find the rational canonical form with respect to the field of scalars equal to $\mathbb{Q}$ for the matrix
$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}$$
Observe that this particular matrix is already a companion matrix of $\lambda^3 - \lambda^2 + \lambda - 1$. Then find the rational canonical form if the field of scalars equals $\mathbb{C}$ or $\mathbb{Q} + i\mathbb{Q}$.

7. Let $q(\lambda)$ be a polynomial and C its companion matrix. Show the characteristic and minimal polynomial of C are the same and both equal $q(\lambda)$.

8. $\uparrow$Use the existence of the rational canonical form to give a proof of the Cayley Hamilton theorem valid for any field, even fields like the integers mod p for p a prime. The earlier proof based on determinants was fine for fields like $\mathbb{Q}$ or $\mathbb{R}$ where you could let $\lambda \to \infty$ but it is not clear the same result holds in general.

9. Suppose you have two $n \times n$ matrices A, B whose entries are in a field $\mathbb{F}$ and suppose $\mathbb{G}$ is an extension of $\mathbb{F}$. For example, you could have $\mathbb{F} = \mathbb{Q}$ and $\mathbb{G} = \mathbb{C}$. Suppose A and B are similar with respect to the field $\mathbb{G}$. Can it be concluded that they are similar with respect to the field $\mathbb{F}$? **Hint:** First show that the two have the same minimal polynomial over $\mathbb{F}$. Next consider the proof of Lemma 9.8.3 and show that they have the same rational canonical form with respect to $\mathbb{F}$.

Chapter 10

Markov Processes

10.1 Regular Markov Matrices

The existence of the Jordan form is the basis for the proof of limit theorems for certain kinds of matrices called Markov matrices.

Definition 10.1.1 *An $n \times n$ matrix $A = (a_{ij})$, is a Markov matrix if $a_{ij} \geq 0$ for all i, j and*

$$\sum_i a_{ij} = 1.$$

It may also be called a stochastic matrix or a transition matrix. A Markov or stochastic matrix is called regular if some power of A has all entries strictly positive. A vector $\mathbf{v} \in \mathbb{R}^n$, is a steady state if $A\mathbf{v} = \mathbf{v}$.

Lemma 10.1.2 *The property of being a stochastic matrix is preserved by taking products. It is also true if the sum is of the form $\sum_j a_{ij} = 1$.*

Proof: Suppose the sum over a row equals 1 for A and B. Then letting the entries be denoted by (a_{ij}) and (b_{ij}) respectively and the entries of AB by (c_{ij}),

$$\sum_i c_{ij} = \sum_i \sum_k a_{ik} b_{kj} = \sum_k \sum_i a_{ik} b_{kj} = \sum_k b_{kj} = 1$$

It is obvious that when the product is taken, if each $a_{ij}, b_{ij} \geq 0$, then the same will be true of sums of products of these numbers. Similar reasoning works for the assumption that $\sum_j a_{ij} = 1$. ■

The following theorem is convenient for showing the existence of limits.

Theorem 10.1.3 *Let A be a real $p \times p$ matrix having the properties*

1. *$a_{ij} \geq 0$*
2. *Either $\sum_{i=1}^p a_{ij} = 1$ or $\sum_{j=1}^p a_{ij} = 1$.*
3. *The distinct eigenvalues of A are $\{1, \lambda_2, \ldots, \lambda_m\}$ where each $|\lambda_j| < 1$.*

Then $\lim_{n\to\infty} A^n = A_\infty$ exists in the sense that $\lim_{n\to\infty} a_{ij}^n = a_{ij}^\infty$, the ij^{th} entry A_∞. ***Here a_{ij}^n denotes the ij^{th} entry of A^n.*** *Also, if $\lambda = 1$ has algebraic multiplicity r, then the Jordan block corresponding to $\lambda = 1$ is just the $r \times r$ identity.*

Proof. By the existence of the Jordan form for A, it follows that there exists an invertible matrix P such that

$$P^{-1}AP = \begin{pmatrix} I+N & & & \\ & J_{r_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{r_m}(\lambda_m) \end{pmatrix} = J$$

where I is $r \times r$ for r the multiplicity of the eigenvalue 1 and N is a nilpotent matrix for which $N^r = 0$. I will show that because of Condition 2, $N = 0$.

First of all,

$$J_{r_i}(\lambda_i) = \lambda_i I + N_i$$

where N_i satisfies $N_i^{r_i} = 0$ for some $r_i > 0$. It is clear that $N_i (\lambda_i I) = (\lambda_i I) N$ and so

$$(J_{r_i}(\lambda_i))^n = \sum_{k=0}^{n} \binom{n}{k} N^k \lambda_i^{n-k} = \sum_{k=0}^{r} \binom{n}{k} N^k \lambda_i^{n-k}$$

which converges to 0 due to the assumption that $|\lambda_i| < 1$. There are finitely many terms and a typical one is a matrix whose entries are no larger than an expression of the form

$$|\lambda_i|^{n-k} C_k n(n-1)\cdots(n-k+1) \leq C_k |\lambda_i|^{n-k} n^k$$

which converges to 0 because, by the root test, the series $\sum_{n=1}^{\infty} |\lambda_i|^{n-k} n^k$ converges. Thus for each $i = 2, \ldots, p$,

$$\lim_{n\to\infty} (J_{r_i}(\lambda_i))^n = 0.$$

By Condition 2, if a_{ij}^n denotes the ij^{th} entry of A^n, then either

$$\sum_{i=1}^{p} a_{ij}^n = 1 \text{ or } \sum_{j=1}^{p} a_{ij}^n = 1,\ a_{ij}^n \geq 0.$$

This follows from Lemma 10.1.2. It is obvious each $a_{ij}^n \geq 0$, and so the entries of A^n must be bounded independent of n.

It follows easily from

$$\overbrace{P^{-1}APP^{-1}APP^{-1}AP\cdots P^{-1}AP}^{n \text{ times}} = P^{-1}A^nP$$

that

$$P^{-1}A^nP = J^n \tag{10.1}$$

Hence J^n must also have bounded entries as $n \to \infty$. However, this requirement is incompatible with an assumption that $N \neq 0$.

If $N \neq 0$, then $N^s \neq 0$ but $N^{s+1} = 0$ for some $1 \leq s \leq r$. Then

$$(I+N)^n = I + \sum_{k=1}^{s} \binom{n}{k} N^k$$

One of the entries of N^s is nonzero by the definition of s. Let this entry be n_{ij}^s. Then this implies that one of the entries of $(I+N)^n$ is of the form $\binom{n}{s} n_{ij}^s$. This entry dominates the ij^{th} entries of $\binom{n}{k} N^k$ for all $k < s$ because

$$\lim_{n\to\infty} \binom{n}{s} / \binom{n}{k} = \infty$$

Therefore, the entries of $(I+N)^n$ cannot all be bounded. From block multiplication,

$$P^{-1}A^nP = \begin{pmatrix} (I+N)^n & & & \\ & (J_{r_2}(\lambda_2))^n & & \\ & & \ddots & \\ & & & (J_{r_m}(\lambda_m))^n \end{pmatrix}$$

and this is a contradiction because entries are bounded on the left and unbounded on the right.

Since $N = 0$, the above equation implies $\lim_{n\to\infty} A^n$ exists and equals

$$P\begin{pmatrix} I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}P^{-1} \blacksquare$$

Are there examples which will cause the eigenvalue condition of this theorem to hold? The following lemma gives such a condition. It turns out that if $a_{ij} > 0$, not just ≥ 0, then the eigenvalue condition of the above theorem is valid.

Lemma 10.1.4 *Suppose $A = (a_{ij})$ is a stochastic matrix. Then $\lambda = 1$ is an eigenvalue. If $a_{ij} > 0$ for all i, j, then if μ is an eigenvalue of A, either $|\mu| < 1$ or $\mu = 1$.*

Proof: First consider the claim that 1 is an eigenvalue. By definition,

$$\sum_i 1a_{ij} = 1$$

and so $A^T\mathbf{v} = \mathbf{v}$ where $\mathbf{v} = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T$. Since A, A^T have the same eigenvalues, this shows 1 is an eigenvalue. Suppose then that μ is an eigenvalue. Is $|\mu| < 1$ or $\mu = 1$? Let $\mathbf{v}$ be an eigenvector for A^T and let $|v_i|$ be the largest of the $|v_j|$.

$$\mu v_i = \sum_j a_{ji} v_j$$

and now multiply both sides by $\overline{\mu v_i}$ to obtain

$$\begin{aligned} |\mu|^2 |v_i|^2 &= \sum_j a_{ji} v_j \overline{\mu v_i} = \sum_j a_{ji} \operatorname{Re}\left(v_j \overline{\mu v_i}\right) \\ &\leq \sum_j a_{ji} |v_i|^2 |\mu| = |\mu| |v_i|^2 \end{aligned}$$

Therefore, $|\mu| \leq 1$. If $|\mu| = 1$, then equality must hold in the above, and so $v_j \overline{v_i \mu}$ must be real and nonnegative for each j. In particular, this holds for $j = i$ which shows $\overline{\mu}$ is real and nonnegative. Thus, in this case, $\mu = 1$ because $\bar{\mu} = \mu$ is nonnegative and equal to 1. The only other case is where $|\mu| < 1$. $\blacksquare$

Lemma 10.1.5 *Let A be any Markov matrix and let $\mathbf{v}$ be a vector having all its components non negative with $\sum_i v_i = c$. Then if $\mathbf{w} = A\mathbf{v}$, it follows that $w_i \geq 0$ for all i and $\sum_i w_i = c$.*

Proof: From the definition of $\mathbf{w}$,

$$w_i \equiv \sum_j a_{ij} v_j \geq 0.$$

Also

$$\sum_i w_i = \sum_i \sum_j a_{ij} v_j = \sum_j \sum_i a_{ij} v_j = \sum_j v_j = c. \blacksquare$$

The following theorem about limits is now easy to obtain.

Theorem 10.1.6 *Suppose A is a Markov matrix in which $a_{ij} > 0$ for all i, j and suppose $\mathbf{w}$ is a vector. Then for each i,*

$$\lim_{k \to \infty} \left(A^k \mathbf{w}\right)_i = v_i$$

where $A\mathbf{v} = \mathbf{v}$. In words, $A^k \mathbf{w}$ always converges to a steady state. In addition to this, if the vector $\mathbf{w}$ satisfies $w_i \geq 0$ for all i and $\sum_i w_i = c$, then the vector $\mathbf{v}$ will also satisfy the conditions, $v_i \geq 0$, $\sum_i v_i = c$.

Proof: By Lemma 10.1.4, since each $a_{ij} > 0$, the eigenvalues are either 1 or have absolute value less than 1. Therefore, the claimed limit exists by Theorem 10.1.3. The assertion that the components are nonnegative and sum to c follows from Lemma 10.1.5. That $A\mathbf{v} = \mathbf{v}$ follows from

$$\mathbf{v} = \lim_{n \to \infty} A^n \mathbf{w} = \lim_{n \to \infty} A^{n+1} \mathbf{w} = A \lim_{n \to \infty} A^n \mathbf{w} = A\mathbf{v}. \blacksquare$$

It is not hard to generalize the conclusion of this theorem to regular Markov processes.

Corollary 10.1.7 *Suppose A is a regular Markov matrix, one for which the entries of A^k are all positive for some k, and suppose $\mathbf{w}$ is a vector. Then for each i,*

$$\lim_{n \to \infty} \left(A^n \mathbf{w}\right)_i = v_i$$

where $A\mathbf{v} = \mathbf{v}$. In words, $A^n \mathbf{w}$ always converges to a steady state. In addition to this, if the vector $\mathbf{w}$ satisfies $w_i \geq 0$ for all i and $\sum_i w_i = c$, Then the vector $\mathbf{v}$ will also satisfy the conditions $v_i \geq 0$, $\sum_i v_i = c$.

Proof: Let the entries of A^k be all positive for some k. Now suppose that $a_{ij} \geq 0$ for all i, j and $A = (a_{ij})$ is a Markov matrix. Then if $B = (b_{ij})$ is a Markov matrix with $b_{ij} > 0$ for all ij, it follows that BA is a Markov matrix which has strictly positive entries. This is because the ij^{th} entry of BA is

$$\sum_k b_{ik} a_{kj} > 0,$$

Thus, from Lemma 10.1.4, A^k has an eigenvalue equal to 1 for all k sufficiently large, and all the other eigenvalues have absolute value strictly less than 1. The same must be true of A. If $\mathbf{v} \neq \mathbf{0}$ and $A\mathbf{v} = \lambda \mathbf{v}$ and $|\lambda| = 1$, then $A^k \mathbf{v} = \lambda^k \mathbf{v}$ and so, by Lemma 10.1.4, $\lambda^m = 1$ if $m \geq k$. Thus

$$1 = \lambda^{k+1} = \lambda^k \lambda = \lambda$$

By Theorem 10.1.3, $\lim_{n \to \infty} A^n \mathbf{w}$ exists. The rest follows as in Theorem 10.1.6. $\blacksquare$

10.2 Migration Matrices

Definition 10.2.1 *Let n locations be denoted by the numbers $1, 2, \cdots, n$. Also suppose it is the case that each year a_{ij} denotes the proportion of residents in location j which move to location i. Also suppose no one escapes or emigrates from without these n locations. This last assumption requires $\sum_i a_{ij} = 1$. Thus (a_{ij}) is a Markov matrix referred to as a migration matrix.*

If $\mathbf{v} = (x_1, \cdots, x_n)^T$ where x_i is the population of location i at a given instant, you obtain the population of location i one year later by computing $\sum_j a_{ij} x_j = (A\mathbf{v})_i$. Therefore, the population of location i after k years is $\left(A^k \mathbf{v}\right)_i$. Furthermore, Corollary 10.1.7 can be used to predict in the case where A is regular what the long time population will be for the given locations.

As an example of the above, consider the case where $n = 3$ and the migration matrix is of the form

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}.$$

Now

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}^2 = \begin{pmatrix} .38 & .02 & .15 \\ .28 & .64 & .02 \\ .34 & .34 & .83 \end{pmatrix}$$

and so the Markov matrix is regular. Therefore, $\left(A^k \mathbf{v}\right)_i$ will converge to the i^{th} component of a steady state. It follows the steady state can be obtained from solving the system

$$\begin{aligned} .6x + .1z &= x \\ .2x + .8y &= y \\ .2x + .2y + .9z &= z \end{aligned}$$

along with the stipulation that the sum of x, y, and z must equal the constant value present at the beginning of the process. The solution to this system is

$$\{y = x, z = 4x, x = x\}.$$

If the total population at the beginning is 150,000, then you solve the following system

$$y = x,\ z = 4x,\ x + y + z = 150000$$

whose solution is easily seen to be $\{x = 25\,000, z = 100\,000, y = 25\,000\}$. Thus, after a long time there would be about four times as many people in the third location as in either of the other two.

10.3 Absorbing States

There is a different kind of Markov process containing so called absorbing states which result in transition matrices which are not regular. However, Theorem 10.1.3 may still apply. One such example is the Gambler's ruin problem. There is a total amount of money denoted by b. The Gambler starts with an amount $j > 0$ and gambles till he either loses everything or gains everything. He does this by playing a game in which he wins with probability p and loses with probability q. When he wins, the amount of money he has increases by 1 and when he loses, the amount of money he has decreases by 1. Thus the states are the integers from 0 to b. Let p_{ij} denote the probability that the gambler has i at the end of a game given that he had j at the beginning. Let p_{ij}^n denote the probability that the gambler has i after n games given that he had j initially. Thus

$$p_{ij}^{n+1} = \sum_k p_{ik} p_{kj}^n,$$

and so p_{ij}^n is the ij^{th} entry of P^n where P is the transition matrix. The above description indicates that this transition probability matrix is of the form

$$P = \begin{pmatrix} 1 & q & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & 0 \\ 0 & p & \ddots & q & \vdots \\ \vdots & & \ddots & 0 & 0 \\ 0 & \cdots & 0 & p & 1 \end{pmatrix} \tag{10.2}$$

The absorbing states are 0 and b. In the first, the gambler has lost everything and hence has nothing else to gamble, so the process stops. In the second, he has won everything and there is nothing else to gain, so again the process stops.

Consider the eigenvalues of this matrix.

Lemma 10.3.1 *Let $p, q > 0$ and $p + q = 1$. Then the eigenvalues of*

$$\begin{pmatrix} 0 & q & 0 & \cdots & 0 \\ p & 0 & q & \cdots & 0 \\ 0 & p & 0 & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & q \\ 0 & \vdots & 0 & p & 0 \end{pmatrix}$$

have absolute value less than 1.

Proof: By Gerschgorin's theorem, (See Page 173) if λ is an eigenvalue, then $|\lambda| \leq 1$. Now suppose $\mathbf{v}$ is an eigenvector for λ. Then

$$A\mathbf{v} = \begin{pmatrix} qv_2 \\ pv_1 + qv_3 \\ \vdots \\ pv_{n-2} + qv_n \\ pv_{n-1} \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{pmatrix}.$$

Suppose $|\lambda| = 1$. Let v_k be the first nonzero entry. Then

$$qv_{k+1} = \lambda v_k$$

and so $|v_{k+1}| > |v_k|$. If $\{|v_j|\}_{j=k}^m$ is increasing for some $m > k$, then

$$p\,|v_{m-1}| + q\,|v_m| \geq |pv_{m-2} + qv_m| = |\lambda v_{m-1}| = |v_{m-1}|$$

and so $q\,|v_m| \geq q\,|v_{m-1}|$. Thus by induction, the sequence is increasing. Hence $|v_n| \geq |v_{n-1}| > 0$. However, the last line states that $p\,|v_{n-1}| = |v_n|$ which requires that $|v_{n-1}| > |v_n|$, a contradiction. ■

Now consider the eigenvalues of 10.2. For P given there,

$$P - \lambda I = \begin{pmatrix} 1-\lambda & q & 0 & \cdots & 0 \\ 0 & -\lambda & \ddots & & 0 \\ 0 & p & \ddots & q & \vdots \\ \vdots & & \ddots & -\lambda & 0 \\ 0 & \cdots & 0 & p & 1-\lambda \end{pmatrix}$$

and so, expanding the determinant of the matrix along the first column and then along the last column yields

$$(1-\lambda)^2 \det \begin{pmatrix} -\lambda & q & & \\ p & \ddots & \ddots & \\ & \ddots & -\lambda & q \\ & & p & -\lambda \end{pmatrix}.$$

The roots of the polynomial after $(1-\lambda)^2$ have absolute value less than 1 because they are just the eigenvalues of a matrix of the sort in Lemma 10.3.1. It follows that the conditions of Theorem 10.1.3 apply and therefore, $\lim_{n\to\infty} P^n$ exists. ■

Of course, the above transition matrix, models many other kinds of problems. It is called a Markov process with two absorbing states, sometimes a random walk with two absorbing states.

It is interesting to find the probability that the gambler loses all his money. This is given by $\lim_{n\to\infty} p_{0j}^n$.From the transition matrix for the gambler's ruin problem, it follows that

$$\begin{aligned} p_{0j}^n &= \sum_k p_{0k}^{n-1} p_{kj} = q p_{0(j-1)}^{n-1} + p p_{0(j+1)}^{n-1} \text{for } j \in [1, b-1], \\ p_{00}^n &= 1, \text{ and } p_{0b}^n = 0. \end{aligned}$$

Assume here that $p \neq q$. Now it was shown above that $\lim_{n\to\infty} p_{0j}^n$ exists. Denote by P_j this limit. Then the above becomes much simpler if written as

$$P_j = qP_{j-1} + pP_{j+1} \text{ for } j \in [1, b-1], \tag{10.3}$$

$$P_0 = 1 \text{ and } P_b = 0. \tag{10.4}$$

It is only required to find a solution to the above difference equation with boundary conditions. To do this, look for a solution in the form $P_j = r^j$and use the difference equation with boundary conditions to find the correct values of r. Thus you need

$$r^j = qr^{j-1} + pr^{j+1}$$

and so to find r you need to have $pr^2 - r + q = 0$, and so the solutions for r are $r =$

$$\frac{1}{2p}\left(1+\sqrt{1-4pq}\right),\ \frac{1}{2p}\left(1-\sqrt{1-4pq}\right)$$

Now

$$\sqrt{1-4pq} = \sqrt{1-4p(1-p)} = \sqrt{1-4p+4p^2} = 1-2p.$$

Thus the two values of r simplify to

$$\frac{1}{2p}(1+1-2p) = \frac{q}{p},\quad \frac{1}{2p}(1-(1-2p)) = 1$$

Therefore, for any choice of $C_i, i = 1, 2,$

$$C_1 + C_2\left(\frac{q}{p}\right)^j$$

will solve the difference equation. Now choose C_1, C_2 to satisfy the boundary conditions 10.4. Thus you need to have

$$C_1 + C_2 = 1,\ C_1 + C_2\left(\frac{q}{p}\right)^b = 0$$

It follows that

$$C_2 = \frac{p^b}{p^b - q^b}, \quad C_1 = \frac{q^b}{q^b - p^b}$$

Thus $P_j =$

$$\frac{q^b}{q^b - p^b} + \frac{p^b}{p^b - q^b}\left(\frac{q}{p}\right)^j = \frac{q^b}{q^b - p^b} - \frac{p^{b-j}q^j}{q^b - p^b} = \frac{q^j\left(q^{b-j} - p^{b-j}\right)}{q^b - p^b}$$

To find the solution in the case of a fair game, one could take the $\lim_{p\to 1/2}$ of the above solution. Taking this limit, you get

$$P_j = \frac{b-j}{b}.$$

You could also verify directly in the case where $p = q = 1/2$ in 10.3 and 10.4 that $P_j = 1$ and $P_j = j$ are two solutions to the difference equation and proceeding as before.

10.4 Exercises

1. 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.

2. Show that if $\sum_i a_{ij} = 1$, then if $A = (a_{ij})$, then the sum of the entries of $A\mathbf{v}$ equals the sum of the entries of $\mathbf{v}$. Thus it does not matter whether $a_{ij} \geq 0$ for this to be so.

3. If A satisfies the conditions of the above problem, can it be concluded that $\lim_{n\to\infty} A^n$ exists?

4. Give an example of a non regular Markov matrix which has an eigenvalue equal to -1.

5. Show that when a Markov matrix is non defective, all of the above theory can be proved very easily. In particular, prove the theorem about the existence of $\lim_{n\to\infty} A^n$ if the eigenvalues are either 1 or have absolute value less than 1.

6. Find a formula for A^n where

$$A = \begin{pmatrix} \frac{5}{2} & -\frac{1}{2} & 0 & -1 \\ 5 & 0 & 0 & -4 \\ \frac{7}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{5}{2} \\ \frac{7}{2} & -\frac{1}{2} & 0 & -2 \end{pmatrix}$$

   Does $\lim_{n\to\infty} A^n$ exist? Note that all the rows sum to 1. **Hint:** This matrix is similar to a diagonal matrix. The eigenvalues are $1, -1, \frac{1}{2}, \frac{1}{2}$.

7. Find a formula for A^n where

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

Note that the rows sum to 1 in this matrix also. **Hint:** This matrix is not similar to a diagonal matrix but you can find the Jordan form and consider this in order to obtain a formula for this product. The eigenvalues are $1, -1, \frac{1}{2}, \frac{1}{2}$.

8. Find $\lim_{n\to\infty} A^n$ if it exists for the matrix
$$A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{3}{2} & 0 \\ \frac{3}{2} & \frac{3}{2} & \frac{3}{2} & 1 \end{pmatrix}$$
The eigenvalues are $\frac{1}{2}, 1, 1, 1$.

9. Give an example of a matrix A which has eigenvalues which are either equal to 1,-1, or have absolute value strictly less than 1 but which has the property that $\lim_{n\to\infty} A^n$ does not exist.

10. If A is an $n \times n$ matrix such that all the eigenvalues have absolute value less than 1, show $\lim_{n\to\infty} A^n = 0$.

11. Find an example of a 3×3 matrix A such that $\lim_{n\to\infty} A^n$ does not exist but $\lim_{r\to\infty} A^{5r}$ does exist.

12. If A is a Markov matrix and B is similar to A, does it follow that B is also a Markov matrix?

13. In Theorem 10.1.3 suppose everything is unchanged except that you assume either $\sum_j a_{ij} \leq 1$ or $\sum_i a_{ij} \leq 1$. Would the same conclusion be valid? What if you don't insist that each $a_{ij} \geq 0$? Would the conclusion hold in this case?

14. Let V be an n dimensional vector space and let $\mathbf{x} \in V$ and $\mathbf{x} \neq \mathbf{0}$. Consider $\beta_{\mathbf{x}} \equiv \mathbf{x}, A\mathbf{x}, \cdots, A^{m-1}\mathbf{x}$ where
$$A^m\mathbf{x} \in \text{span}\left(\mathbf{x}, A\mathbf{x}, \cdots, A^{m-1}\mathbf{x}\right)$$
and m is the smallest such that the above inclusion in the span takes place. Show that $\left\{\mathbf{x}, A\mathbf{x}, \cdots, A^{m-1}\mathbf{x}\right\}$ must be linearly independent. Next suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for V. Consider $\beta_{\mathbf{v}_i}$ as just discussed, having length m_i. Thus $A^{m_i}\mathbf{v}_i$ is a linearly combination of $\mathbf{v}_i$,$A\mathbf{v}_i, \cdots, A^{m-1}\mathbf{v}_i$ for m as small as possible. Let $p_{\mathbf{v}_i}(\lambda)$ be the monic polynomial which expresses this linear combination. Thus $p_{\mathbf{v}_i}(A)\mathbf{v}_i = 0$ and the degree of $p_{\mathbf{v}_i}(\lambda)$ is as small as possible for this to take place. Show that the minimal polynomial for A must be the monic polynomial which is the least common multiple of these polynomials $p_{\mathbf{v}_i}(\lambda)$.

15. If A is a complex Hermitian $n \times n$ matrix which has all eigenvalues nonnegative, show that there exists a complex Hermitian matrix B such that $BB = A$.

16. $\uparrow$Suppose A, B are $n \times n$ real Hermitian matrices and they both have all nonnegative eigenvalues. Show that $\det(A + B) \geq \det(A) + \det(B)$. **Hint:** Use the above problem and the Cauchy Binet theorem. Let $P^2 = A, Q^2 = B$ where P, Q are Hermitian and nonnegative. Then
$$A + B = \begin{pmatrix} P & Q \end{pmatrix} \begin{pmatrix} P \\ Q \end{pmatrix}.$$

17. Suppose $B = \begin{pmatrix} \alpha & \mathbf{c}^* \\ \mathbf{b} & A \end{pmatrix}$ is an $(n+1) \times (n+1)$ Hermitian nonnegative matrix where α is a scalar and A is $n \times n$. Show that α must be real, $\mathbf{c} = \mathbf{b}$, and $A = A^*, A$ is nonnegative, and that if $\alpha = 0$, then $\mathbf{b} = \mathbf{0}$. Otherwise, $\alpha > 0$.

18. ↑If A is an $n \times n$ complex Hermitian and nonnegative matrix, show that there exists an upper triangular matrix B such that $B^* B = A$. **Hint:** Prove this by induction. It is obviously true if $n = 1$. Now if you have an $(n+1) \times (n+1)$ Hermitian nonnegative matrix, then from the above problem, it is of the form $\begin{pmatrix} \alpha^2 & \alpha \mathbf{b}^* \\ \alpha \mathbf{b} & A \end{pmatrix}, \alpha$ real.

19. ↑ Suppose A is a nonnegative Hermitian matrix (all eigenvalues are nonnegative) which is partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11}, A_{22} are square matrices. Show that $\det(A) \leq \det(A_{11}) \det(A_{22})$. **Hint:** Use the above problem to factor A getting

$$A = \begin{pmatrix} B_{11}^* & 0^* \\ B_{12}^* & B_{22}^* \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}$$

Next argue that $A_{11} = B_{11}^* B_{11}, A_{22} = B_{12}^* B_{12} + B_{22}^* B_{22}$. Use the Cauchy Binet theorem to argue that $\det(A_{22}) = \det(B_{12}^* B_{12} + B_{22}^* B_{22}) \geq \det(B_{22}^* B_{22})$. Then explain why

$$\begin{aligned} \det(A) &= \det(B_{11}^*) \det(B_{22}^*) \det(B_{11}) \det(B_{22}) \\ &= \det(B_{11}^* B_{11}) \det(B_{22}^* B_{22}) \end{aligned}$$

20. ↑ Prove the inequality of Hadamard. If A is a Hermitian matrix which is nonnegative (all eigenvalues are nonnegative), then $\det(A) \leq \prod_i A_{ii}$.

Chapter 11

Inner Product Spaces

11.1 General Theory

It is assumed here that the field of scalars is either $\mathbb{R}$ or $\mathbb{C}$. The usual example of an inner product space is $\mathbb{C}^n$ or $\mathbb{R}^n$ as described earlier. However, there are many other inner product spaces and the topic is of such importance that it seems appropriate to discuss the general theory of these spaces.

Definition 11.1.1 *A vector space X is said to be a normed linear space if there exists a function, denoted by $|\cdot| : X \to [0, \infty)$ which satisfies the following axioms.*

1. $|x| \geq 0$ *for all* $x \in X$, *and* $|x| = 0$ *if and only if* $x = 0$.
2. $|ax| = |a| |x|$ *for all* $a \in \mathbb{F}$.
3. $|x + y| \leq |x| + |y|$.

This function $|\cdot|$ is called a norm.

The notation $||x||$ is also often used. Not all norms are created equal. There are many geometric properties which they may or may not possess. There is also a concept called an inner product which is discussed next. It turns out that the best norms come from an inner product.

Definition 11.1.2 *A mapping $(\cdot, \cdot) : V \times V \to \mathbb{F}$ is called an inner product if it satisfies the following axioms.*

1. $(x, y) = \overline{(y, x)}$.
2. $(x, x) \geq 0$ *for all* $x \in V$ *and equals zero if and only if* $x = 0$.
3. $(ax + by, z) = a(x, z) + b(y, z)$ *whenever* $a, b \in \mathbb{F}$.

Note that 2 and 3 imply $(x, ay + bz) = \overline{a}(x, y) + \overline{b}(x, z)$.
Then a norm is given by

$$(x, x)^{1/2} \equiv |x|.$$

It remains to verify this really is a norm.

Definition 11.1.3 *A normed linear space in which the norm comes from an inner product as just described is called an inner product space.*

Example 11.1.4 *Let $V = \mathbb{C}^n$ with the inner product given by $(\mathbf{x}, \mathbf{y}) \equiv \sum_{k=1}^{n} x_k \overline{y}_k$. This is an example of a complex inner product space already discussed.*

Example 11.1.5 *Let $V = \mathbb{R}^n$,, $(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} \equiv \sum_{j=1}^{n} x_j y_j$. This is an example of a real inner product space.*

Example 11.1.6 *Let V be any finite dimensional vector space and let $\{v_1, \cdots, v_n\}$ be a basis. Decree that*

$$(v_i, v_j) \equiv \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

and define the inner product by

$$(x, y) \equiv \sum_{i=1}^{n} x^i \overline{y^i}$$

where

$$x = \sum_{i=1}^{n} x^i v_i,\ y = \sum_{i=1}^{n} y^i v_i.$$

The above is well defined because $\{v_1, \cdots, v_n\}$ is a basis. Thus the components x_i associated with any given $x \in V$ are uniquely determined.

This example shows there is no loss of generality when studying finite dimensional vector spaces with field of scalars $\mathbb{R}$ or $\mathbb{C}$ in assuming the vector space is actually an inner product space. The following theorem was presented earlier with slightly different notation.

Theorem 11.1.7 *(Cauchy Schwarz) In any inner product space*

$$|(x, y)| \leq |x||y|.$$

where $|x| \equiv (x, x)^{1/2}$.

Proof: Let $\omega \in \mathbb{C}, |\omega| = 1$, and $\overline{\omega}(x, y) = |(x, y)| = \operatorname{Re}(x, y\omega)$. Let

$$F(t) = (x + ty\omega, x + t\omega y).$$

Then from the axioms of the inner product,

$$F(t) = |x|^2 + 2t \operatorname{Re}(x, \omega y) + t^2|y|^2 \geq 0.$$

This yields

$$|x|^2 + 2t|(x, y)| + t^2|y|^2 \geq 0.$$

If $|y| = 0$, then the inequality requires that $|(x, y)| = 0$ since otherwise, you could pick large negative t and contradict the inequality. If $|y| > 0$, it follows from the quadratic formula that

$$4|(x, y)|^2 - 4|x|^2|y|^2 \leq 0. \blacksquare$$

Earlier it was claimed that the inner product defines a norm. In this next proposition this claim is proved.

Proposition 11.1.8 *For an inner product space,* $|x| \equiv (x, x)^{1/2}$ *does specify a norm.*

Proof: All the axioms are obvious except the triangle inequality. To verify this,

$$\begin{aligned} |x + y|^2 &\equiv (x + y, x + y) \equiv |x|^2 + |y|^2 + 2 \operatorname{Re}(x, y) \\ &\leq |x|^2 + |y|^2 + 2 |(x, y)| \\ &\leq |x|^2 + |y|^2 + 2 |x| |y| = (|x| + |y|)^2. \blacksquare \end{aligned}$$

The best norms of all are those which come from an inner product because of the following identity which is known as the parallelogram identity.

Proposition 11.1.9 *If* $(V, (\cdot, \cdot))$ *is an inner product space then for* $|x| \equiv (x, x)^{1/2}$, *the following identity holds.*

$$|x + y|^2 + |x - y|^2 = 2 |x|^2 + 2 |y|^2 .$$

It turns out that the validity of this identity is equivalent to the existence of an inner product which determines the norm as described above. These sorts of considerations are topics for more advanced courses on functional analysis.

Definition 11.1.10 *A basis for an inner product space, $\{u_1, \cdots, u_n\}$ is an orthonormal basis if*

$$(u_k, u_j) = \delta_{kj} \equiv \begin{cases} 1 \text{ if } k = j \\ 0 \text{ if } k \neq j \end{cases}.$$

Note that if a list of vectors satisfies the above condition for being an orthonormal set, then the list of vectors is automatically linearly independent. To see this, suppose

$$\sum_{j=1}^{n} c^j u_j = 0$$

Then taking the inner product of both sides with u_k,

$$0 = \sum_{j=1}^{n} c^j (u_j, u_k) = \sum_{j=1}^{n} c^j \delta_{jk} = c^k.$$

11.2 The Gram Schmidt Process

Lemma 11.2.1 *Let X be an inner product space and let $\{x_1, \cdots, x_n\}$ be linearly independent. Then there exists an orthonormal basis for X, $\{u_1, \cdots, u_n\}$ which has the property that for each $k \leq n$, span$(x_1, \cdots, x_k) = \operatorname{span}(u_1, \cdots, u_k)$.*

Proof: Let $u_1 \equiv x_1 / |x_1|$. Thus for $k = 1$, $\operatorname{span}(u_1) = \operatorname{span}(x_1)$ and $\{u_1\}$ is an orthonormal set. Now suppose for some $k < n$, $u_1, \cdots, u_k$ have been chosen such that $(u_j, u_l) = \delta_{jl}$ and $\operatorname{span}(x_1, \cdots, x_k) = \operatorname{span}(u_1, \cdots, u_k)$. Then define

$$u_{k+1} \equiv \frac{x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j) u_j}{\left| x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j) u_j \right|}, \tag{11.1}$$

where the denominator is not equal to zero because the x_j form a basis and so

$$x_{k+1} \notin \operatorname{span}(x_1, \cdots, x_k) = \operatorname{span}(u_1, \cdots, u_k)$$

Thus by induction,

$$u_{k+1} \in \operatorname{span}(u_1, \cdots, u_k, x_{k+1}) = \operatorname{span}(x_1, \cdots, x_k, x_{k+1}).$$

Also, $x_{k+1} \in \operatorname{span}(u_1, \cdots, u_k, u_{k+1})$ which is seen easily by solving 11.1 for x_{k+1} and it follows

$$\operatorname{span}(x_1, \cdots, x_k, x_{k+1}) = \operatorname{span}(u_1, \cdots, u_k, u_{k+1}).$$

If $l \leq k$,

$$\begin{aligned}
(u_{k+1}, u_l) &= C\left((x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j)(u_j, u_l) \right) \\
&= C\left((x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j) \delta_{lj} \right) \\
&= C\left((x_{k+1}, u_l) - (x_{k+1}, u_l) \right) = 0.
\end{aligned}$$

The vectors, $\{u_j\}_{j=1}^n$, generated in this way are therefore an orthonormal basis because each vector has unit length. ■

The process by which these vectors were generated is called the Gram Schmidt process. The following corollary is obtained from the above process.

Corollary 11.2.2 *Let X be a finite dimensional inner product space of dimension n whose basis is $\{u_1, \cdots, u_k, x_{k+1}, \cdots, x_n\}$. Then if $\{u_1, \cdots, u_k\}$ is orthonormal, then the Gram Schmidt process applied to the given list of vectors in order leaves $\{u_1, \cdots, u_k\}$ unchanged.*

Lemma 11.2.3 *Suppose $\{u_j\}_{j=1}^n$ is an orthonormal basis for an inner product space X. Then for all $x \in X$,*

$$x = \sum_{j=1}^{n} (x, u_j) u_j.$$

Proof: Since $\{u_j\}_{j=1}^n$ is a basis, there exist unique scalars $\{\alpha_i\}$ such that

$$x = \sum_{j=1}^{n} \alpha_j u_j$$

It only remains to identify α_k. From the properties of the inner product,

$$(x, u_k) = \sum_{j=1}^{n} \alpha_j (u_j, u_k) = \sum_{j=1}^{n} \alpha_j \delta_{jk} = \alpha_k \blacksquare$$

The following theorem is of fundamental importance. First note that a subspace of an inner product space is also an inner product space because you can use the same inner product.

Theorem 11.2.4 *Let M be a finite dimensional subspace of X, an inner product space and let $\{e_i\}_{i=1}^m$ be an orthonormal basis for M. Then if $y \in X$ and $w \in M$,*

$$|y - w|^2 = \inf \left\{ |y - z|^2 : z \in M \right\} \tag{11.2}$$

if and only if

$$(y - w, z) = 0 \tag{11.3}$$

for all $z \in M$. Furthermore,

$$w = \sum_{i=1}^{m} (y, x_i) x_i \tag{11.4}$$

is the unique element of M which has this property. It is called the orthogonal projection.

Proof: First we show that if 11.3, then 11.2. Let $z \in M$ be arbitrary. Then

$$|y - z|^2 = |y - w + (w - z)|^2$$

$$= (y - w + (w - z), y - w + (w - z))$$

$$= |y - w|^2 + |z - w|^2 + 2 \operatorname{Re} (y - w, w - z)$$

The last term is given to be 0 and so

$$|y - z|^2 = |y - w|^2 + |z - w|^2$$

which verifies 11.2.

Next suppose 11.2. Is it true that 11.3 follows? Let $z \in M$ be arbitrary and let $|\theta| = 1, \bar{\theta}(x-w, w-z) = |(x-w, w-z)|$. Then let

$$\begin{aligned} p(t) &\equiv |x-w+t\theta(w-z)|^2 = |x-w|^2 + 2\operatorname{Re}(x-w, t\theta(w-z)) + t^2|w-z|^2 \\ &= |x-w|^2 + 2\operatorname{Re} t\bar{\theta}(x-w,(w-z)) + t^2|w-z|^2 \\ &= |x-w|^2 + 2t|(x-w,(w-z))| + t^2|w-z|^2 \end{aligned}$$

Then p has a minimum when $t = 0$ and so $p'(0) = 2|(x-w,(w-z))| = 0$ which shows 11.3. This proves the first part of the theorem since z is arbitrary.

It only remains to verify that w given in 11.4 satisfies 11.3 and is the only point of M which does so.

First, could there be two minimizers? Say w_1, w_2 both work. Then by the above characterization of minimizers,

$$\begin{aligned} (x-w_1, w_1-w_2) &= 0 \\ (x-w_2, w_1-w_2) &= 0 \end{aligned}$$

Subtracting gives $(w_1-w_2, w_1-w_2) = 0$. Hence the minimizer is unique.

Finally, it remains to show that the given formula works. Letting $\{e_1, \cdots, e_m\}$ be an orthonormal basis for M, such a thing existing by the Gramm Schmidt process,

$$\begin{aligned} \left(x - \sum_{i=1}^m (x,e_i)e_i, e_k\right) &= (x,e_k) - \sum_{i=1}^m (x,e_i)(e_i,e_k) \\ &= (x,e_k) - \sum_{i=1}^m (x,e_i)\delta_{ik} \\ &= (x,e_k) - (x,e_k) = 0 \end{aligned}$$

Since this inner product equals 0 for arbitrary e_k, it follows that

$$\left(x - \sum_{i=1}^m (x,e_i)e_i, z\right) = 0$$

for every $z \in M$ because each such z is a linear combination of the e_i. Hence $\sum_{i=1}^m (x,e_i)e_i$ is the unique minimizer. ■

Example 11.2.5 *Consider X equal to the continuous functions defined on $[-\pi,\pi]$ and let the inner product be given by*

$$\int_{-\pi}^{\pi} f(x)\overline{g(x)}dx$$

It is left to the reader to verify that this is an inner product. Letting e_k be the function $x \to \frac{1}{\sqrt{2\pi}}e^{ikx}$, define

$$M \equiv \operatorname{span}\left(\{e_k\}_{k=-n}^n\right).$$

Then you can verify that

$$(e_k, e_m) = \int_{-\pi}^{\pi}\left(\frac{1}{\sqrt{2\pi}}e^{-ikx}\right)\left(\frac{1}{\sqrt{2\pi}}\overline{e^{mix}}\right)dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(m-k)x} = \delta_{km}$$

then for a given function $f \in X$, the function from M which is closest to f in this inner product norm is

$$g = \sum_{k=-n}^{n} (f, e_k) e_k$$

In this case $(f, e_k) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x) e^{ikx} dx$. These are the Fourier coefficients. The above is the n^{th} partial sum of the Fourier series.

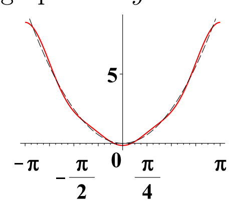
To show how this kind of thing approximates a given function, let $f(x) = x^2$. Let $M = \operatorname{span}\left(\left\{\frac{1}{\sqrt{2\pi}} e^{-ikx}\right\}_{k=-3}^{3}\right)$. Then, doing the computations, you find the closest point is of the form

$$\frac{1}{3}\sqrt{2}\pi^{\frac{5}{2}}\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{k=1}^{3}\left(\frac{(-1)^k 2}{k^2}\right)\sqrt{2}\sqrt{\pi}\frac{1}{\sqrt{2\pi}} e^{-ikx} + \sum_{k=1}^{3}\left(\frac{(-1)^k 2}{k^2}\right)\sqrt{2}\sqrt{\pi}\frac{1}{\sqrt{2\pi}} e^{ikx}$$

and now simplify to get

$$\frac{1}{3}\pi^2 + \sum_{k=1}^{3}(-1)^k\left(\frac{4}{k^2}\right)\cos kx$$

Then a graph of this along with the graph of $y = x^2$ is given below. In this graph, the dashed graph is of $y = x^2$ and the solid line is the graph of the above Fourier series approximation.



If we had taken the partial sum up to n much bigger, it would have been very hard to distinguish between the graph of the partial sum of the Fourier series and the graph of the function it is approximating. This is in contrast to approximation by Taylor series in which you only get approximation at a point of a function and its derivatives. These are very close near the point of interest but typically fail to approximate the function on the entire interval.

11.3 Riesz Representation Theorem

The next theorem is one of the most important results in the theory of inner product spaces. It is called the Riesz representation theorem.

Theorem 11.3.1 *Let $f \in \mathcal{L}(X, \mathbb{F})$ where X is an inner product space of dimension n. Then there exists a unique $z \in X$ such that for all $x \in X$,*

$$f(x) = (x, z).$$

Proof: First I will verify uniqueness. Suppose z_j works for $j = 1, 2$. Then for all $x \in X$,

$$0 = f(x) - f(x) = (x, z_1 - z_2)$$

and so $z_1 = z_2$.

It remains to verify existence. By Lemma 11.2.1, there exists an orthonormal basis, $\{u_j\}_{j=1}^{n}$. If there is such a z, then you would need $f(u_j) = (u_j, z)$ and so you would need $\overline{f(u_j)} = (z, u_j)$. Also you must have $z = \sum_i (z, u_j) u_j$. Therefore, define

$$z \equiv \sum_{j=1}^{n} \overline{f(u_j)} u_j.$$

Then using Lemma 11.2.3,

$$\begin{aligned}(x,z) &= \left(x,\sum_{j=1}^{n}\overline{f(u_j)}u_j\right)=\sum_{j=1}^{n}f(u_j)(x,u_j) \\ &= f\left(\sum_{j=1}^{n}(x,u_j)u_j\right)=f(x). \blacksquare\end{aligned}$$

Corollary 11.3.2 *Let $A\in\mathcal{L}(X,Y)$ where X and Y are two inner product spaces of finite dimension. Then there exists a unique $A^*\in\mathcal{L}(Y,X)$ such that*

$$(Ax,y)_Y=(x,A^*y)_X \tag{11.5}$$

for all $x\in X$ and $y\in Y$. The following formula holds

$$(\alpha A+\beta B)^*=\overline{\alpha}A^*+\overline{\beta}B^*$$

Proof: Let $f_y\in\mathcal{L}(X,\mathbb{F})$ be defined as

$$f_y(x)\equiv(Ax,y)_Y.$$

Then by the Riesz representation theorem, there exists a unique element of X, $A^*(y)$ such that

$$(Ax,y)_Y=(x,A^*(y))_X.$$

It only remains to verify that A^* is linear. Let a and b be scalars. Then for all $x\in X$,

$$(x,A^*(ay_1+by_2))_X\equiv(Ax,(ay_1+by_2))_Y$$

$$\equiv\overline{a}(Ax,y_1)+\overline{b}(Ax,y_2)\equiv$$

$$\overline{a}(x,A^*(y_1))+\overline{b}(x,A^*(y_2))=(x,aA^*(y_1)+bA^*(y_2)).$$

Since this holds for every x, it follows

$$A^*(ay_1+by_2)=aA^*(y_1)+bA^*(y_2)$$

which shows A^* is linear as claimed.

Consider the last assertion that * is conjugate linear.

$$\left(x,(\alpha A+\beta B)^*y\right)\equiv((\alpha A+\beta B)x,y)$$

$$\begin{aligned}&= \alpha(Ax,y)+\beta(Bx,y)=\alpha(x,A^*y)+\beta(x,B^*y)\\ &= (x,\overline{\alpha}A^*y)+\left(x,\overline{\beta}A^*y\right)=\left(x,\left(\overline{\alpha}A^*+\overline{\beta}A^*\right)y\right).\end{aligned}$$

Since x is arbitrary,

$$(\alpha A+\beta B)^*y=\left(\overline{\alpha}A^*+\overline{\beta}A^*\right)y$$

and since this is true for all y,

$$(\alpha A+\beta B)^*=\overline{\alpha}A^*+\overline{\beta}A^*. \blacksquare$$

Definition 11.3.3 *The linear map, A^* is called the adjoint of A. In the case when $A:X\to X$ and $A=A^*$, A is called a self adjoint map. Such a map is also called Hermitian.*

Theorem 11.3.4 *Let M be an $m \times n$ matrix. Then $M^* = \left(\overline{M}\right)^T$ in words, the transpose of the conjugate of M is equal to the adjoint.*

Proof: Using the definition of the inner product in $\mathbb{C}^n$,

$$(M\mathbf{x}, \mathbf{y}) = (\mathbf{x}, M^*\mathbf{y}) \equiv \sum_i x_i \overline{\sum_j (M^*)_{ij} y_j} = \sum_{i,j} \overline{(M^*)_{ij}}\, \overline{y_j} x_i.$$

Also

$$(M\mathbf{x}, \mathbf{y}) = \sum_j \sum_i M_{ji} \overline{y_j} x_i.$$

Since $\mathbf{x}, \mathbf{y}$ are arbitrary vectors, it follows that $M_{ji} = \overline{(M^*)_{ij}}$ and so, taking conjugates of both sides,

$$M^*_{ij} = \overline{M_{ji}} \quad \blacksquare$$

The next theorem is interesting. You have a p dimensional subspace of $\mathbb{F}^n$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Of course this might be "slanted". However, there is a linear transformation Q which preserves distances which maps this subspace to $\mathbb{F}^p$.

Theorem 11.3.5 *Suppose V is a subspace of $\mathbb{F}^n$ having dimension $p \leq n$. Then there exists a $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ such that*

$$QV \subseteq \operatorname{span}(\mathbf{e}_1, \cdots, \mathbf{e}_p)$$

and $|Q\mathbf{x}| = |\mathbf{x}|$ for all $\mathbf{x}$. Also

$$Q^*Q = QQ^* = I.$$

Proof: By Lemma 11.2.1 there exists an orthonormal basis for V, $\{\mathbf{v}_i\}_{i=1}^p$. By using the Gram Schmidt process this may be extended to an orthonormal basis of the whole space $\mathbb{F}^n$,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{v}_{p+1}, \cdots, \mathbf{v}_n\}.$$

Now define $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ by $Q(\mathbf{v}_i) \equiv \mathbf{e}_i$ and extend linearly. If $\sum_{i=1}^n x_i \mathbf{v}_i$ is an arbitrary element of $\mathbb{F}^n$,

$$\left| Q\left(\sum_{i=1}^n x_i \mathbf{v}_i \right) \right|^2 = \left| \sum_{i=1}^n x_i \mathbf{e}_i \right|^2 = \sum_{i=1}^n |x_i|^2 = \left| \sum_{i=1}^n x_i \mathbf{v}_i \right|^2.$$

It remains to verify that $Q^*Q = QQ^* = I$. To do so, let $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$. Then let ω be a complex number such that $|\omega| = 1$, $\omega(\mathbf{x}, Q^*Q\mathbf{y} - \mathbf{y}) = |(\mathbf{x}, Q^*Q\mathbf{y} - \mathbf{y})|$.

$$(Q(\omega\mathbf{x} + \mathbf{y}), Q(\omega\mathbf{x} + \mathbf{y})) = (\omega\mathbf{x} + \mathbf{y}, \omega\mathbf{x} + \mathbf{y}).$$

Thus

$$|Q\mathbf{x}|^2 + |Q\mathbf{y}|^2 + 2\operatorname{Re}\omega(Q\mathbf{x}, Q\mathbf{y}) = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\operatorname{Re}\omega(\mathbf{x}, \mathbf{y})$$

and since Q preserves norms, it follows that for all $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$,

$$\operatorname{Re}\omega(Q\mathbf{x}, Q\mathbf{y}) = \operatorname{Re}\omega(\mathbf{x}, Q^*Q\mathbf{y}) = \omega \operatorname{Re}(\mathbf{x}, \mathbf{y}).$$

Thus

$$0 = \operatorname{Re}\omega((\mathbf{x}, Q^*Q\mathbf{y}) - (\mathbf{x}, \mathbf{y})) = \operatorname{Re}\omega(\mathbf{x}, Q^*Q\mathbf{y} - \mathbf{y}) = |(\mathbf{x}, Q^*Q\mathbf{y} - \mathbf{y})|$$

$$\operatorname{Re}(\mathbf{x}, Q^*Q\mathbf{y} - \mathbf{y}) = 0 \tag{11.6}$$

for all $\mathbf{x}, \mathbf{y}$. Letting $\mathbf{x} = Q^*Q\mathbf{y} - \mathbf{y}$, it follows $Q^*Q\mathbf{y} = \mathbf{y}$. Similarly $QQ^* = I$. $\blacksquare$

Note that is is actually shown that $QV = \operatorname{span}(\mathbf{e}_1, \cdots, \mathbf{e}_p)$ and that in case $p = n$ one obtains that a linear transformation which maps an orthonormal basis to an orthonormal basis is unitary.

11.4 The Tensor Product of Two Vectors

Definition 11.4.1 *Let X and Y be inner product spaces and let $x \in X$ and $y \in Y$. Define the tensor product of these two vectors, $y \otimes x$, an element of $\mathcal{L}(X,Y)$ by*

$$y \otimes x(u) \equiv y(u,x)_X .$$

This is also called a rank one transformation because the image of this transformation is contained in the span of the vector, y.

The verification that this is a linear map is left to you. Be sure to verify this! The following lemma has some of the most important properties of this linear transformation.

Lemma 11.4.2 *Let X,Y,Z be inner product spaces. Then for α a scalar,*

$$(\alpha(y \otimes x))^* = \overline{\alpha} x \otimes y \tag{11.7}$$

$$(z \otimes y_1)(y_2 \otimes x) = (y_2, y_1) z \otimes x \tag{11.8}$$

Proof: Let $u \in X$ and $v \in Y$. Then

$$(\alpha(y \otimes x) u, v) = (\alpha(u,x) y, v) = \alpha(u,x)(y,v)$$

and

$$(u, \overline{\alpha} x \otimes y(v)) = (u, \overline{\alpha}(v,y) x) = \alpha(y,v)(u,x).$$

Therefore, this verifies 11.7.

To verify 11.8, let $u \in X$.

$$(z \otimes y_1)(y_2 \otimes x)(u) = (u,x)(z \otimes y_1)(y_2) = (u,x)(y_2,y_1) z$$

and

$$(y_2,y_1) z \otimes x(u) = (y_2,y_1)(u,x) z.$$

Since the two linear transformations on both sides of 11.8 give the same answer for every $u \in X$, it follows the two transformations are the same. ■

Definition 11.4.3 *Let X,Y be two vector spaces. Then define for $A,B \in \mathcal{L}(X,Y)$ and $\alpha \in \mathbb{F}$, new elements of $\mathcal{L}(X,Y)$ denoted by $A+B$ and αA as follows.*

$$(A+B)(x) \equiv Ax + Bx, \ (\alpha A) x \equiv \alpha(Ax).$$

Theorem 11.4.4 *Let X and Y be finite dimensional inner product spaces. Then $\mathcal{L}(X,Y)$ is a vector space with the above definition of what it means to multiply by a scalar and add. Let $\{v_1, \cdots, v_n\}$ be an orthonormal basis for X and $\{w_1, \cdots, w_m\}$ be an orthonormal basis for Y. Then a basis for $\mathcal{L}(X,Y)$ is*

$$\{w_j \otimes v_i : i = 1, \cdots, n, \ j = 1, \cdots, m\}.$$

Proof: It is obvious that $\mathcal{L}(X,Y)$ is a vector space. It remains to verify the given set is a basis. Consider the following:

$$\left(\left(A - \sum_{k,l}(Av_k, w_l) w_l \otimes v_k\right) v_p, w_r\right) = (Av_p, w_r) -$$

$$\sum_{k,l} (Av_k, w_l)(v_p, v_k)(w_l, w_r)$$

$$= (Av_p, w_r) - \sum_{k,l} (Av_k, w_l)\,\delta_{pk}\delta_{rl} = (Av_p, w_r) - (Av_p, w_r) = 0.$$

Letting $A - \sum_{k,l} (Av_k, w_l)\, w_l \otimes v_k = B$, this shows that $Bv_p = 0$ since w_r is an arbitrary element of the basis for Y. Since v_p is an arbitrary element of the basis for X, it follows $B = 0$ as hoped. This has shown $\{w_j \otimes v_i : i = 1, \cdots, n,\ j = 1, \cdots, m\}$ spans $\mathcal{L}(X, Y)$.

It only remains to verify the $w_j \otimes v_i$ are linearly independent. Suppose then that

$$\sum_{i,j} c_{ij} w_j \otimes v_i = 0$$

Then do both sides to v_s. By definition this gives

$$0 = \sum_{i,j} c_{ij} w_j (v_s, v_i) = \sum_{i,j} c_{ij} w_j \delta_{si} = \sum_j c_{sj} w_j$$

Now the vectors $\{w_1, \cdots, w_m\}$ are independent because it is an orthonormal set and so the above requires $c_{sj} = 0$ for each j. Since s was arbitrary, this shows the linear transformations, $\{w_j \otimes v_i\}$ form a linearly independent set. ■

Note this shows the dimension of $\mathcal{L}(X, Y) = nm$. The theorem is also of enormous importance because it shows you can always consider an arbitrary linear transformation as a sum of rank one transformations whose properties are easily understood. The following theorem is also of great interest.

Theorem 11.4.5 *Let $A = \sum_{i,j} c_{ij} w_i \otimes v_j \in \mathcal{L}(X, Y)$ where as before, the vectors, $\{w_i\}$ are an orthonormal basis for Y and the vectors, $\{v_j\}$ are an orthonormal basis for X. Then if the matrix of A has entries M_{ij}, it follows that $M_{ij} = c_{ij}$.*

Proof: Recall

$$Av_i \equiv \sum_k M_{ki} w_k$$

Also

$$\begin{aligned} Av_i &= \sum_{k,j} c_{kj} w_k \otimes v_j (v_i) = \sum_{k,j} c_{kj} w_k (v_i, v_j) \\ &= \sum_{k,j} c_{kj} w_k \delta_{ij} = \sum_k c_{ki} w_k \end{aligned}$$

Therefore,

$$\sum_k M_{ki} w_k = \sum_k c_{ki} w_k$$

and so $M_{ki} = c_{ki}$ for all k. This happens for each i. ■

11.5 Least Squares

A common problem in experimental work is to find a straight line which approximates as well as possible a collection of points in the plane $\{(x_i, y_i)\}_{i=1}^p$. The usual way of dealing with these problems is by the method of least squares and it turns out that all these sorts of approximation problems can be reduced to $A\mathbf{x} = \mathbf{b}$ where the problem is to find the best $\mathbf{x}$ for solving this equation even when there is no solution.

Lemma 11.5.1 *Let V and W be finite dimensional inner product spaces and let $A : V \to W$ be linear. For each $y \in W$ there exists $x \in V$ such that*

$$|Ax - y| \leq |Ax_1 - y|$$

for all $x_1 \in V$. Also, $x \in V$ is a solution to this minimization problem if and only if x is a solution to the equation, $A^*Ax = A^*y$.

Proof: By Theorem 11.2.4 on Page 276 there exists a point, Ax_0, in the finite dimensional subspace, $A(V)$, of W such that for all $x \in V, |Ax - y|^2 \geq |Ax_0 - y|^2$. Also, from this theorem, this happens if and only if $Ax_0 - y$ is perpendicular to every $Ax \in A(V)$. Therefore, the solution is characterized by $(Ax_0 - y, Ax) = 0$ for all $x \in V$ which is the same as saying $(A^*Ax_0 - A^*y, x) = 0$ for all $x \in V$. In other words the solution is obtained by solving $A^*Ax_0 = A^*y$ for x_0. ■

Consider the problem of finding the least squares regression line in statistics. Suppose you have given points in the plane, $\{(x_i, y_i)\}_{i=1}^n$ and you would like to find constants m and b such that the line $y = mx + b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find m, b such that you do the best you can to solve the system

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

which is of the form $\mathbf{y} = A\mathbf{x}$. In other words try to make $\left| A \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right|^2$ as small as possible. According to what was just shown, it is desired to solve the following for m and b.

$$A^*A \begin{pmatrix} m \\ b \end{pmatrix} = A^* \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Since $A^* = A^T$ in this case,

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

Solving this system of equations for m and b,

$$m = \frac{-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + \left(\sum_{i=1}^n x_i y_i\right) n}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}$$

and

$$b = \frac{-\left(\sum_{i=1}^n x_i\right) \sum_{i=1}^n x_i y_i + \left(\sum_{i=1}^n y_i\right) \sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}.$$

One could clearly do a least squares fit for curves of the form $y = ax^2 + bx + c$ in the same way. In this case you solve as well as possible for a, b, and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

using the same techniques.

11.6 Fredholm Alternative Again

The best context in which to study the Fredholm alternative is in inner product spaces. This is done here.

Definition 11.6.1 *Let S be a subset of an inner product space, X. Define*

$$S^{\perp} \equiv \{x \in X : (x,s) = 0 \text{ for all } s \in S\}.$$

The following theorem also follows from the above lemma. It is sometimes called the Fredholm alternative.

Theorem 11.6.2 *Let $A : V \to W$ where A is linear and V and W are inner product spaces. Then $A(V) = \ker(A^*)^{\perp}$.*

Proof: Let $y = Ax$ so $y \in A(V)$. Then if $A^* z = 0$,

$$(y,z) = (Ax,z) = (x, A^* z) = 0$$

showing that $y \in \ker(A^*)^{\perp}$. Thus $A(V) \subseteq \ker(A^*)^{\perp}$.

Now suppose $y \in \ker(A^*)^{\perp}$. Does there exists x such that $Ax = y$? Since this might not be immediately clear, take the least squares solution to the problem. Thus let x be a solution to $A^*Ax = A^*y$. It follows $A^*(y - Ax) = 0$ and so $y - Ax \in \ker(A^*)$ which implies from the assumption about y that $(y - Ax, y) = 0$. Also, since Ax is the closest point to y in $A(V)$, Theorem 11.2.4 on Page 276 implies that $(y - Ax, Ax_1) = 0$ for all $x_1 \in V$. In particular this is true for $x_1 = x$ and so $0 = (y - Ax, y) - \overbrace{(y - Ax, Ax)}^{=0} = |y - Ax|^2$, showing that $y = Ax$. Thus $A(V) \supseteq \ker(A^*)^{\perp}$. ■

Corollary 11.6.3 *Let $A, V,$ and W be as described above. If the only solution to $A^* y = 0$ is $y = 0$, then A is onto W.*

Proof: If the only solution to $A^* y = 0$ is $y = 0$, then $\ker(A^*) = \{0\}$ and so every vector from W is contained in $\ker(A^*)^{\perp}$ and by the above theorem, this shows $A(V) = W$. ■

11.7 Exercises

1. Find the best solution to the system

$$\begin{aligned} x + 2y &= 6 \\ 2x - y &= 5 \\ 3x + 2y &= 0 \end{aligned}$$

2. Find an orthonormal basis for $\mathbb{R}^3$, $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}$ given that $\mathbf{w}_1$ is a multiple of the vector $(1,1,2)$.

3. Suppose $A = A^T$ is a symmetric real $n \times n$ matrix which has all positive eigenvalues. Define

$$(\mathbf{x}, \mathbf{y}) \equiv (A\mathbf{x}, \mathbf{y}).$$

Show this is an inner product on $\mathbb{R}^n$. What does the Cauchy Schwarz inequality say in this case?

4. Let $||\mathbf{x}||_\infty \equiv \max\{|x_j| : j = 1, 2, \cdots, n\}$. Show this is a norm on $\mathbb{C}^n$. Here
$$\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T.$$
Show
$$||\mathbf{x}||_\infty \leq |\mathbf{x}| \equiv (\mathbf{x}, \mathbf{x})^{1/2}$$
where the above is the usual inner product on $\mathbb{C}^n$.

5. Let $||\mathbf{x}||_1 \equiv \sum_{j=1}^n |x_j|$ .Show this is a norm on $\mathbb{C}^n$. Here $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T$. Show
$$||\mathbf{x}||_1 \geq |\mathbf{x}| \equiv (\mathbf{x}, \mathbf{x})^{1/2}$$
where the above is the usual inner product on $\mathbb{C}^n$. Show there cannot exist an inner product such that this norm comes from the inner product as described above for inner product spaces.

6. Show that if $||\cdot||$ is any norm on any vector space, then $|||x|| - ||y||| \leq ||x - y||$.

7. Relax the assumptions in the axioms for the inner product. Change the axiom about $(x, x) \geq 0$ and equals 0 if and only if $x = 0$ to simply read $(x, x) \geq 0$. Show the Cauchy Schwarz inequality still holds in the following form. $|(x, y)| \leq (x, x)^{1/2} (y, y)^{1/2}$.

8. Let H be an inner product space and let $\{u_k\}_{k=1}^n$ be an orthonormal basis for H. Show
$$(x, y) = \sum_{k=1}^n (x, u_k) \overline{(y, u_k)}.$$

9. Let the vector space V consist of real polynomials of degree no larger than 3. Thus a typical vector is a polynomial of the form $a + bx + cx^2 + dx^3$. For $p, q \in V$ define the inner product, $(p, q) \equiv \int_0^1 p(x) q(x) dx$. Show this is indeed an inner product. Then state the Cauchy Schwarz inequality in terms of this inner product. Show $\{1, x, x^2, x^3\}$ is a basis for V. Finally, find an orthonormal basis for V. This is an example of some orthonormal polynomials.

10. Let P_n denote the polynomials of degree no larger than $n - 1$ which are defined on an interval $[a, b]$. Let $\{x_1, \cdots, x_n\}$ be n distinct points in $[a, b]$. Now define for $p, q \in P_n$,
$$(p, q) \equiv \sum_{j=1}^n p(x_j) \overline{q(x_j)}$$
Show this yields an inner product on P_n. **Hint:** Most of the axioms are obvious. The one which says $(p, p) = 0$ if and only if $p = 0$ is the only interesting one. To verify this one, note that a nonzero polynomial of degree no more than $n - 1$ has at most $n - 1$ zeros.

11. Let $C([0, 1])$ denote the vector space of continuous real valued functions defined on $[0, 1]$. Let the inner product be given as
$$(f, g) \equiv \int_0^1 f(x) g(x) dx$$
Show this is an inner product. Also let V be the subspace described in Problem 9. Using the result of this problem, find the vector in V which is closest to x^4.

12. A **regular Sturm Liouville problem** involves the differential equation, for an unknown function of x which is denoted here by y,

$$(p(x)y')' + (\lambda q(x) + r(x))y = 0,\ x \in [a,b]$$

and it is assumed that $p(t), q(t) > 0$ for any $t \in [a,b]$ and also there are 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

$$\int_a^b q(x)y(x)z(x)\,dx = 0. \tag{11.9}$$

and this defines an inner product. **Hint:** Do something like this:

$$\begin{aligned} (p(x)y')' z + (\lambda_1 q(x) + r(x))yz &= 0, \\ (p(x)z')' y + (\lambda_2 q(x) + r(x))zy &= 0. \end{aligned}$$

Now subtract and either use integration by parts or show

$$(p(x)y')' z - (p(x)z')' y = ((p(x)y')z - (p(x)z')y)'$$

and then integrate. Use the boundary conditions to show that $y'(a)z(a) - z'(a)y(a) = 0$ and $y'(b)z(b) - z'(b)y(b) = 0$. The formula, 11.9 is called an orthogonality relation. It turns out there are typically infinitely many eigenvalues and it is interesting to write given functions as an infinite series of these "eigenfunctions".

13. Consider the continuous functions defined on $[0,\pi]$, $C([0,\pi])$. Show $(f,g) \equiv \int_0^\pi fg\,dx$ is an inner product on this vector space. Show the functions $\left\{\sqrt{\frac{2}{\pi}}\sin(nx)\right\}_{n=1}^\infty$ are an orthonormal set. What does this mean about the dimension of the vector space $C([0,\pi])$? Now let $V_N = \text{span}\left(\sqrt{\frac{2}{\pi}}\sin(x), \cdots, \sqrt{\frac{2}{\pi}}\sin(Nx)\right)$. For $f \in C([0,\pi])$ find a formula for the vector in V_N which is closest to f with respect to the norm determined from the above inner product. This is called the N^{th} partial sum of the Fourier series of f. An important problem is to determine whether and in what way this Fourier series converges to the function f. The norm which comes from this inner product is sometimes called the mean square norm.

14. Consider the subspace $V \equiv \ker(A)$ where

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

Find an orthonormal basis for V. **Hint:** You might first find a basis and then use the Gram Schmidt procedure.

15. The Gram Schmidt process starts with a basis for a subspace $\{v_1, \cdots, v_n\}$ and produces an orthonormal basis for the same subspace $\{u_1, \cdots, u_n\}$ such that

$$\operatorname{span}(v_1, \cdots, v_k) = \operatorname{span}(u_1, \cdots, u_k)$$

for each k. Show that in the case of $\mathbb{R}^m$ the QR factorization does the same thing. More specifically, if

$$A = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$$

and if

$$A = QR \equiv \begin{pmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_n \end{pmatrix} R$$

then the vectors $\{\mathbf{q}_1, \cdots, \mathbf{q}_n\}$ is an orthonormal set of vectors and for each k,

$$\operatorname{span}(\mathbf{q}_1, \cdots, \mathbf{q}_k) = \operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k)$$

16. Verify the parallelogram identify for any inner product space,

$$|x + y|^2 + |x - y|^2 = 2|x|^2 + 2|y|^2 .$$

Why is it called the parallelogram identity?

17. Let H be an inner product space and let $K \subseteq H$ be a nonempty convex subset. This means that if $k_1, k_2 \in K$, then the line segment consisting of points of the form

$$tk_1 + (1 - t) k_2 \text{ for } t \in [0, 1]$$

is also contained in K. Suppose for each $x \in H$, there exists Px defined to be a point of K closest to x. Show that Px is unique so that P actually is a map. **Hint:** Suppose z_1 and z_2 both work as closest points. Consider the midpoint, $(z_1 + z_2)/2$ and use the parallelogram identity of Problem 16 in an auspicious manner.

18. In the situation of Problem 17 suppose K is a closed convex subset and that H is complete. This means every Cauchy sequence converges. Recall from calculus a sequence $\{k_n\}$ is a Cauchy sequence if for every $\varepsilon > 0$ there exists N_ε such that whenever $m, n > N_\varepsilon$, it follows $|k_m - k_n| < \varepsilon$. Let $\{k_n\}$ be a sequence of points of K such that

$$\lim_{n\to\infty} |x - k_n| = \inf \{|x - k| : k \in K\}$$

This is called a minimizing sequence. Show there exists a unique $k \in K$ such that $\lim_{n\to\infty} |k_n - k|$ and that $k = Px$. That is, there exists a well defined projection map onto the convex subset of H. **Hint:** Use the parallelogram identity in an auspicious manner to show $\{k_n\}$ is a Cauchy sequence which must therefore converge. Since K is closed it follows this will converge to something in K which is the desired vector.

19. Let H be an inner product space which is also complete and let P denote the projection map onto a convex closed subset, K. Show this projection map is characterized by the inequality

$$\operatorname{Re}(k - Px, x - Px) \leq 0$$

for all $k \in K$. That is, a point $z \in K$ equals Px if and only if the above variational inequality holds. This is what that inequality is called. This is because k is allowed to vary and the inequality continues to hold for all $k \in K$.

20. Using Problem 19 and Problems 17 - 18 show the projection map, P onto a closed convex subset is Lipschitz continuous with Lipschitz constant 1. That is

$$|Px - Py| \leq |x - y|$$

21. Give an example of two vectors in $\mathbb{R}^4$ or $\mathbb{R}^3$ $\mathbf{x}, \mathbf{y}$ and a subspace V such that $\mathbf{x} \cdot \mathbf{y} = 0$ but $P\mathbf{x} \cdot P\mathbf{y} \neq 0$ where P denotes the projection map which sends $\mathbf{x}$ to its closest point on V.

22. Suppose you are given the data, $(1, 2), (2, 4), (3, 8), (0, 0)$. Find the linear regression line using the formulas derived above. Then graph the given data along with your regression line.

23. Generalize the least squares procedure to the situation in which data is given and you desire to fit it with an expression of the form $y = af(x) + bg(x) + c$ where the problem would be to find a, b and c in order to minimize the error. Could this be generalized to higher dimensions? How about more functions?

24. Let $A \in \mathcal{L}(X, Y)$ where X and Y are finite dimensional vector spaces with the dimension of X equal to n. Define $\text{rank}(A) \equiv \dim(A(X))$ and $\text{nullity}(A) \equiv \dim(\ker(A))$. Show that $\text{nullity}(A) + \text{rank}(A) = \dim(X)$. **Hint:** Let $\{x_i\}_{i=1}^r$ be a basis for $\ker(A)$ and let $\{x_i\}_{i=1}^r \cup \{y_i\}_{i=1}^{n-r}$ be a basis for X. Then show that $\{Ay_i\}_{i=1}^{n-r}$ is linearly independent and spans AX.

25. Let A be an $m \times n$ matrix. Show the column rank of A equals the column rank of A^*A. Next verify column rank of A^*A is no larger than column rank of A^*. Next justify the following inequality to conclude the column rank of A equals the column rank of A^*.

$$\text{rank }(A) = \text{rank }(A^*A) \leq \text{rank }(A^*) \leq$$

$$= \text{rank }(AA^*) \leq \text{rank }(A).$$

Hint: Start with an orthonormal basis, $\{A\mathbf{x}_j\}_{j=1}^r$ of $A(\mathbb{F}^n)$ and verify $\{A^*A\mathbf{x}_j\}_{j=1}^r$ is a basis for $A^*A(\mathbb{F}^n)$.

26. Let A be a real $m \times n$ matrix and let $A = QR$ be the QR factorization with Q orthogonal and R upper triangular. Show that there exists a solution $\mathbf{x}$ to the equation

$$R^T R\mathbf{x} = R^T Q^T \mathbf{b}$$

and that this solution is also a least squares solution defined above such that $A^T A\mathbf{x} = A^T\mathbf{b}$.

11.8 The Determinant and Volume

The determinant is the essential algebraic tool which provides a way to give a unified treatment of the concept of p dimensional volume of a parallelepiped in $\mathbb{R}^M$. Here is the definition of what is meant by such a thing.

Definition 11.8.1 *Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^M$, $M \geq p$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ and it is defined as*

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^p s_j \mathbf{u}_j : s_j \in [0, 1] \right\} = UQ, \ Q = [0, 1]^p$$

where $U = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}$. *The volume of this parallelepiped is defined as*

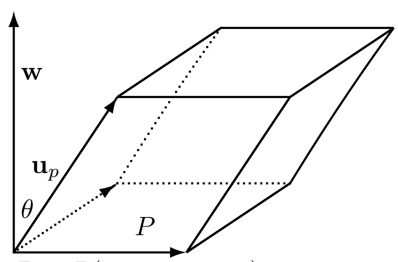
$$\textit{volume of } P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv v(P(\mathbf{u}_1, \cdots, \mathbf{u}_p)) \equiv (\det(G))^{1/2}.$$

where $G_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j$. *This* $G = U^T U$ *is called the metric tensor. If the vectors* $\mathbf{u}_i$ *are dependent, this definition will give the volume to be 0.*

First lets observe the last assertion is true. Say $\mathbf{u}_i = \sum_{j \neq i} \alpha_j \mathbf{u}_j$. Then the i^{th} row of G is a linear combination of the other rows using the scalars α_j and so from the properties of the determinant, the determinant of this matrix is indeed zero as it should be. Indeed, $\mathbf{u}_i \cdot \mathbf{u}_k = \sum_{j \neq i} \alpha_j \mathbf{u}_j \cdot \mathbf{u}_k$.

A parallelepiped is a sort of a squashed box. Here is a picture which shows the relationship between $P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})$ and $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$. In a sense, we can define the volume any way desired, but if it is to be reasonable, the following relationship must hold. The appropriate definition of the volume of $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ in terms of $P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})$ is $v(P(\mathbf{u}_1, \cdots, \mathbf{u}_p)) =$

$$|\mathbf{u}_p \cdot \mathbf{w}| \, v(P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})) \tag{11.10}$$



$P = P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})$

where $\mathbf{w}$ is any unit vector perpendicular to each of $\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}$. Note $|\mathbf{u}_p \cdot \mathbf{w}| = |\mathbf{u}_p| |\cos\theta|$ from the geometric meaning of the dot product. In the case where $p = 1$, the parallelepiped $P(\mathbf{v})$ consists of the single vector and the one dimensional volume should be $|\mathbf{v}| = (\mathbf{v}^T \mathbf{v})^{1/2} = (\mathbf{v} \cdot \mathbf{v})^{1/2}$. Now having made this definition, I will show that $\det(G)^{1/2}$ is the appropriate definition of $v(P(\mathbf{u}_1, \cdots, \mathbf{u}_p))$ for every p.

As just pointed out, this is the only reasonable definition of volume in the case of one vector. The next theorem shows that it is the only reasonable definition of volume of a parallelepiped in the case of p vectors because 11.10 holds.

Theorem 11.8.2 *If we desire 11.10 to hold for any* $\mathbf{w}$ *perpendicular to each* $\mathbf{u}_i$*, then we obtain the definition of 11.8.1 for* $v(P(\mathbf{u}_1, \cdots, \mathbf{u}_p))$ *in terms of determinants.*

Proof: So assume we want 11.10 to hold. Suppose the determinant formula holds for $P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1})$. It is necessary to show that if $\mathbf{w}$ is a unit vector perpendicular to each $\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}$ then $|\mathbf{u}_p \cdot \mathbf{w}| \, v(P(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}))$ reduces to $\det(G)^{1/2}$. By the Gram Schmidt procedure there is $(\mathbf{w}_1, \cdots, \mathbf{w}_p)$ an orthonormal basis for $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ such that $\operatorname{span}(\mathbf{w}_1, \cdots, \mathbf{w}_k) = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ for each $k \leq p$. We can pick $\mathbf{w}_p = \mathbf{w}$ the given unit vector perpendicular to each $\mathbf{u}_i$. First note that since $\{\mathbf{w}_k\}_{k=1}^p$ is an orthonormal basis for $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_p)$,

$$\mathbf{u}_j = \sum_{k=1}^p (\mathbf{u}_j \cdot \mathbf{w}_k) \mathbf{w}_k, \quad \mathbf{u}_j \cdot \mathbf{u}_i = \sum_{k=1}^p (\mathbf{u}_j \cdot \mathbf{w}_k)(\mathbf{u}_i \cdot \mathbf{w}_k)$$

Therefore, the ij^{th} entry of the $p \times p$ matrix $U^T U$ is just

$$\left(U^T U\right)_{ij} = \sum_{r=1}^p (\mathbf{u}_i \cdot \mathbf{w}_r)(\mathbf{w}_r \cdot \mathbf{u}_j)$$

which is the product of a $p \times p$ matrix M whose rj^{th} entry is $\mathbf{w}_r \cdot \mathbf{u}_j$ with its transpose. The vector $\mathbf{w}_p$ is a unit vector perpendicular to each $\mathbf{u}_j$ for $j \leq p - 1$ so $\mathbf{w}_p \cdot \mathbf{u}_j = 0$ if $j < p$.

Now consider the vector

$$\mathbf{N} \equiv \det \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_{p-1} & \mathbf{w}_p \\ \mathbf{u}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_1 \cdot \mathbf{w}_{p-1} & \overset{=0}{\mathbf{u}_1 \cdot \mathbf{w}_p} \\ \vdots & & \vdots & \vdots \\ \mathbf{u}_{p-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_{p-1} \cdot \mathbf{w}_{p-1} & \overset{=0}{\mathbf{u}_{p-1} \cdot \mathbf{w}_p} \end{pmatrix}$$

which results from formally expanding along the top row. Note you would get the same thing expanding along the last column because as just noted, the last column on the right is 0 except for the top entry, so every cofactor A_{1k} for the $1k^{th}$ position is $\pm$ a determinant which has a column of zeros. Thus $\mathbf{N}$ is a multiple of $\mathbf{w}_p$. Hence, for $j < p, \mathbf{N} \cdot \mathbf{u}_j = 0$. From what was just discussed and induction, $v\left(P\left(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}\right)\right) = \pm A_{1p} = \mathbf{N} \cdot \mathbf{w}_p$. Also $\mathbf{N} \cdot \mathbf{u}_p$ equals

$$\det \begin{pmatrix} \mathbf{u}_p \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_p \cdot \mathbf{w}_{p-1} & \mathbf{u}_p \cdot \mathbf{w}_p \\ \mathbf{u}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_1 \cdot \mathbf{w}_{p-1} & \overset{=0}{\mathbf{u}_1 \cdot \mathbf{w}_p} \\ \vdots & & \vdots & \vdots \\ \mathbf{u}_{p-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{u}_{p-1} \cdot \mathbf{w}_{p-1} & \overset{=0}{\mathbf{u}_{p-1} \cdot \mathbf{w}_p} \end{pmatrix} = \pm \det\left(M\right)$$

Thus from induction and expanding along the last column,

$$\begin{aligned} \left|\mathbf{u}_p \cdot \mathbf{w}_p\right| v\left(P\left(\mathbf{u}_1, \cdots, \mathbf{u}_{p-1}\right)\right) &= \left|\mathbf{N} \cdot \mathbf{u}_p\right| = \det\left(M^T M\right)^{1/2} \\ &= \det\left(U^T U\right)^{1/2} = \det\left(G\right)^{1/2}. \end{aligned}$$

Now $\mathbf{w}_p = \mathbf{w}$ the unit vector perpendicular to each $\mathbf{u}_j$ for $j \leq p-1$. Thus if 11.10, then the claimed determinant identity holds. ■

The theorem shows that the only reasonable definition of p dimensional volume of a parallelepiped is the one given in the above definition. Recall that these vectors are in $\mathbb{R}^M$. What is the role of $\mathbb{R}^M$? It is just to provide an inner product. That is its only function. If $p = M$, then $\det\left(U^T U\right) = \det\left(U^T\right)\det\left(U\right) = \det\left(U\right)^2$ and so $\det\left(G\right)^{1/2} = \left|\det\left(U\right)\right|$.

11.9 Finding an Orthogonal Basis

The Gram Schmidt process described above gives a way to generate an orthogonal set of vectors from a linearly independent set. Is there a convenient way to do this? Probably not. However, if you have access to a computer algebra system there might be a way which could help. In the following lemma, v_i will be a vector and it is assumed that $v_i, i = 1, ..., n$ are linearly independent.

Lemma 11.9.1 *Let $\{v_1, ..., v_n\}$ be linearly independent and consider the following formal derivative:*

$$\det \begin{pmatrix} (v_1, v_1) & (v_1, v_2) & \cdots & (v_1, v_{n-1}) & v_1 \\ (v_2, v_1) & (v_2, v_2) & \cdots & (v_2, v_{n-1}) & v_2 \\ \vdots & \vdots & & \vdots & \vdots \\ (v_{n-1}, v_1) & (v_{n-1}, v_2) & \cdots & (v_{n-1}, v_{n-1}) & v_{n-1} \\ (v_n, v_1) & (v_n, v_2) & \cdots & (v_n, v_{n-1}) & v_n \end{pmatrix}$$

Then the vector which results from expanding this determinant formally is perpendicular to each of $v_1, ..., v_{n-1}$.

Proof: It is of the form $\sum_{i=1}^{n} v_i C_i$ where C_i is a suitable $(n-1)\times(n-1)$ determinant. Thus the inner product of this with v_k for $k \leq n-1$ is the expansion of a determinant which has two equal columns. However, the inner product with v_n will be the Grammian of $\{v_1, ..., v_n\}$ which is not zero since these vectors v_i are independent. See Problem 11 on Page 293. ■

Example 11.9.2 *The vectors* $1, x, x^2, x^3$ *are linearly independent on* $[0,1]$*, the vector space being the continuous functions defined on* $[0,1]$*. You might show this. An inner product is given by* $\int_0^1 f(x) g(x)\,dx$*. Find an orthogonal basis for* $\operatorname{span}\left(1, x, x^2, x^3\right)$*.*

You could use the above lemma. $u_1(x) = 1$. Now I will assemble the formal determinants as given above.

$$\det\begin{pmatrix} 1 & 1 \\ \frac{1}{2} & x \end{pmatrix}, \det\begin{pmatrix} 1 & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{3} & x \\ \frac{1}{3} & \frac{1}{4} & x^2 \end{pmatrix}, \det\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & x \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & x^2 \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & x^3 \end{pmatrix}$$

Now the orthogonal basis is obtained from evaluating these determinants and adding 1 to the list. Thus an orthonormal basis is

$\left\{1, x - \frac{1}{2}, \frac{1}{12}x^2 - \frac{1}{12}x + \frac{1}{72}, \frac{1}{2160}x^3 - \frac{1}{1440}x^2 + \frac{1}{3600}x - \frac{1}{43\,200}\right\}$

Is this horrible? Yes it is. However, if you have a computer algebra system do it for you, it isn't so bad. For example, to get the last term, you just do

$$\begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix}\begin{pmatrix} 1 & x & x^2 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \\ x^3 & x^4 & x^5 \end{pmatrix}$$

Then you do the following.

$$\int_0^1 \begin{pmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \\ x^3 & x^4 & x^5 \end{pmatrix} dx = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \end{pmatrix}$$

You could get Matlab to do it for you. Then you add in the last column which consists of the original vectors. If you wanted an orthonormal basis, you could divide each vector by its magnitude. This was only painless because I let the computer do all the tedious busy work. However, I think it has independent interest because it gives a formula for a vector which will be orthogonal to a given set of linearly independent vectors.

11.10 Exercises

1. Here are three vectors in $\mathbb{R}^4$: $(1,2,0,3)^T, (2,1,-3,2)^T, (0,0,1,2)^T$. Find the three dimensional volume of the parallelepiped determined by these three vectors.

2. Here are two vectors in $\mathbb{R}^4$: $(1,2,0,3)^T, (2,1,-3,2)^T$. Find the volume of the parallelepiped determined by these two vectors.

3. Here are three vectors in $\mathbb{R}^2$: $(1,2)^T, (2,1)^T, (0,1)^T$. Find the three dimensional volume of the parallelepiped determined by these three vectors. Recall that from the above theorem, this should equal 0.

4. Find the equation of the plane through the three points $(1,2,3),(2,-3,1),(1,1,7)$.

5. Let T map a vector space V to itself. Explain why T is one to one if and only if T is onto. It is in the text, but do it again in your own words.

6. $\uparrow$Let all matrices be complex with complex field of scalars and let A be an $n\times n$ matrix and B a $m\times m$ matrix while X will be an $n\times m$ matrix. The problem is to consider solutions to Sylvester's equation. Solve the following equation for X
$$AX - XB = C$$
where C is an arbitrary $n\times m$ matrix. Show there exists a unique solution if and only if $\sigma(A)\cap\sigma(B)=\emptyset$. **Hint:** If $q(\lambda)$ is a polynomial, show first that if $AX-XB=0$, then $q(A)X - Xq(B)=0$. Next define the linear map T which maps the $n\times m$ matrices to the $n\times m$ matrices as follows.
$$TX \equiv AX - XB$$
Show that the only solution to $TX=0$ is $X=0$ so that T is one to one if and only if $\sigma(A)\cap\sigma(B)=\emptyset$. Do this by using the first part for $q(\lambda)$ the characteristic polynomial for B and then use the Cayley Hamilton theorem. Explain why $q(A)^{-1}$ exists if and only if the condition $\sigma(A)\cap\sigma(B)=\emptyset$.

7. Recall the Binet Cauchy theorem, Theorem 3.3.14. What is the geometric meaning of the Binet Cauchy theorem?

8. For W a subspace of V, W is said to have a complementary subspace [15] W' if $W\oplus W'=V$. Suppose that both W,W' are invariant with respect to $A\in\mathcal{L}(V,V)$. Show that for any polynomial $f(\lambda)$, if $f(A)x\in W$, then there exists $w\in W$ such that $f(A)x=f(A)w$. A subspace W is called A admissible if it is A invariant and the condition of this problem holds.

9. $\uparrow$ Return to Theorem 9.3.5 about the existence of a basis $\beta=\left\{\beta_{x_1},\cdots,\beta_{x_p}\right\}$ for V where $A\in\mathcal{L}(V,V)$. Adapt the statement and proof to show that if W is A admissible, then it has a complementary subspace which is also A invariant. **Hint:**

   The modified version of the theorem is: Suppose $A\in\mathcal{L}(V,V)$ and the minimal polynomial of A is $\phi(\lambda)^m$ where $\phi(\lambda)$ is a monic irreducible polynomial. Also suppose that W is an A admissible subspace. Then there exists a basis for V which is of the form $\beta=\left\{\beta_{x_1},\cdots,\beta_{x_p},v_1,\cdots,v_m\right\}$ where $\{v_1,\cdots,v_m\}$ is a basis of W. Thus span $\left(\beta_{x_1},\cdots,\beta_{x_p}\right)$ is the A invariant complementary subspace for W. You may want to use the fact that $\phi(A)(V)\cap W=\phi(A)(W)$ which follows easily because W is A admissible. Then use this fact to show that $\phi(A)(W)$ is also A admissible.

10. Let U,H be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let A be a linear map from U to H. Thus AU is a subspace of H. For $\mathbf{g}\in AU$, define $A^{-1}\mathbf{g}$ to be the unique element of $\{\mathbf{x}:A\mathbf{x}=\mathbf{g}\}$ which is closest to $\mathbf{0}$. Then define $(\mathbf{h},\mathbf{g})_{AU}\equiv\left(A^{-1}\mathbf{g},A^{-1}\mathbf{h}\right)_U$. Show that this is a well defined inner product. Let U,H be finite dimensional inner product spaces. (More generally, complete inner product spaces.) Let A be a linear map from U to H. Thus AU is a subspace of H. For $\mathbf{g}\in AU$, define $A^{-1}\mathbf{g}$ to be the unique element of $\{\mathbf{x}:A\mathbf{x}=\mathbf{g}\}$ which is closest to $\mathbf{0}$. Then define $(\mathbf{h},\mathbf{g})_{AU}\equiv\left(A^{-1}\mathbf{g},A^{-1}\mathbf{h}\right)_U$. Show that this is a well defined inner product and that if A is one to one, then $\|\mathbf{h}\|_{AU}=\left\|A^{-1}\mathbf{h}\right\|_U$ and $\|A\mathbf{x}\|_{AU}=\|\mathbf{x}\|_U$.

11. Suppose $\{v_1, ..., v_n\}$ is a linearly independent set of vectors in an inner product space. The Grammian determinant is

$$\det \begin{pmatrix} (v_1, v_1) & (v_1, v_2) & \cdots & (v_1, v_n) \\ (v_2, v_1) & (v_2, v_2) & \cdots & (v_2, v_n) \\ \vdots & \vdots & & \vdots \\ (v_n, v_1) & (v_n, v_2) & \cdots & (v_n, v_n) \end{pmatrix}$$

Show this is not zero.

Chapter 12

Self Adjoint Operators

12.1 Simultaneous Diagonalization

Recall the following definition of what it means for a matrix to be diagonalizable.

Definition 12.1.1 *Let A be an $n \times n$ matrix. It is said to be diagonalizable if there exists an invertible matrix S such that*

$$S^{-1}AS = D$$

where D is a diagonal matrix.

Also, here is a useful observation.

Observation 12.1.2 *If A is an $n \times n$ matrix and $AS = SD$ for D a diagonal matrix, then each column of S is an eigenvector or else it is the zero vector. This follows from observing that for $\mathbf{s}_k$ the k^{th} column of S and from the way we multiply matrices,*

$$A\mathbf{s}_k = \lambda_k \mathbf{s}_k$$

It is sometimes interesting to consider the problem of finding a single similarity transformation which will diagonalize all the matrices in some set.

Lemma 12.1.3 *Let A be an $n \times n$ matrix and let B be an $m \times m$ matrix. Denote by C the matrix*

$$C \equiv \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Then C is diagonalizable if and only if both A and B are diagonalizable.

Proof: Suppose $S_A^{-1}AS_A = D_A$ and $S_B^{-1}BS_B = D_B$ where D_A and D_B are diagonal matrices. You should use block multiplication to verify that $S \equiv \begin{pmatrix} S_A & 0 \\ 0 & S_B \end{pmatrix}$ is such that $S^{-1}CS = D_C$, a diagonal matrix.

Conversely, suppose C is diagonalized by $S = (\mathbf{s}_1, \cdots, \mathbf{s}_{n+m})$. Thus S has columns $\mathbf{s}_i$. For each of these columns, write in the form

$$\mathbf{s}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$$

where $\mathbf{x}_i \in \mathbb{F}^n$ and where $\mathbf{y}_i \in \mathbb{F}^m$. The result is

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where S_{11} is an $n \times n$ matrix and S_{22} is an $m \times m$ matrix. Then there is a diagonal matrix, D_1 being $n \times n$ and D_2 $m \times m$ such that

$$D = \operatorname{diag}(\lambda_1, \cdots, \lambda_{n+m}) = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

such that

$$\begin{aligned} & \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \\ = & \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \end{aligned}$$

Hence by block multiplication

$$AS_{11} = S_{11}D_1,\ BS_{22} = S_{22}D_2$$

$$BS_{21} = S_{21}D_1,\ AS_{12} = S_{12}D_2$$

It follows each of the $\mathbf{x}_i$ is an eigenvector of A or else is the zero vector and that each of the $\mathbf{y}_i$ is an eigenvector of B or is the zero vector. If there are n linearly independent $\mathbf{x}_i$, then A is diagonalizable by Theorem 8.3.12 on Page 8.3.12.

The row rank of the matrix $(\mathbf{x}_1, \cdots, \mathbf{x}_{n+m})$ must be n because if this is not so, the rank of S would be less than $n+m$ which would mean S^{-1} does not exist. Therefore, since the column rank equals the row rank, this matrix has column rank equal to n and this means there are n linearly independent eigenvectors of A implying that A is diagonalizable. Similar reasoning applies to B. ■

The following corollary follows from the same type of argument as the above.

Corollary 12.1.4 *Let A_k be an $n_k \times n_k$ matrix and let C denote the block diagonal*

$$\left(\sum_{k=1}^{r} n_k\right) \times \left(\sum_{k=1}^{r} n_k\right)$$

matrix given below.

$$C \equiv \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_r \end{pmatrix}.$$

Then C is diagonalizable if and only if each A_k is diagonalizable.

Definition 12.1.5 *A set, $\mathcal{F}$ of $n \times n$ matrices is said to be simultaneously diagonalizable if and only if there exists a single invertible matrix S such that for every $A \in \mathcal{F}$, $S^{-1}AS = D_A$ where D_A is a diagonal matrix. $\mathcal{F}$ is a commuting family of matrices if whenever $A, B \in \mathcal{F}$, $AB = BA$.*

Lemma 12.1.6 *If $\mathcal{F}$ is a set of $n \times n$ matrices which is simultaneously diagonalizable, then $\mathcal{F}$ is a commuting family of matrices.*

Proof: Let $A, B \in \mathcal{F}$ and let S be a matrix which has the property that $S^{-1}AS$ is a diagonal matrix for all $A \in \mathcal{F}$. Then $S^{-1}AS = D_A$ and $S^{-1}BS = D_B$ where D_A and D_B are diagonal matrices. Since diagonal matrices commute,

$$\begin{aligned} AB &= SD_AS^{-1}SD_BS^{-1} = SD_AD_BS^{-1} \\ &= SD_BD_AS^{-1} = SD_BS^{-1}SD_AS^{-1} = BA. \end{aligned}$$

Lemma 12.1.7 *Let D be a diagonal matrix of the form*

$$D \equiv \begin{pmatrix} \lambda_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{n_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r I_{n_r} \end{pmatrix}, \tag{12.1}$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix and $\lambda_i \neq \lambda_j$ for $i \neq j$ and suppose B is a matrix which commutes with D. Then B is a block diagonal matrix of the form

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_r \end{pmatrix} \tag{12.2}$$

where B_i is an $n_i \times n_i$ matrix.

Proof: Let $B = (B_{ij})$ where $B_{ii} = B_i$ a block matrix as above in 12.2.

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1r} \\ B_{21} & B_{22} & \ddots & B_{2r} \\ \vdots & \ddots & \ddots & \vdots \\ B_{r1} & B_{r2} & \cdots & B_{rr} \end{pmatrix}$$

Then by block multiplication, since B is given to commute with D,

$$\lambda_j B_{ij} = \lambda_i B_{ij}$$

Therefore, if $i \neq j, B_{ij} = 0$. ■

Lemma 12.1.8 *Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathcal{F}$ is diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable.*

Proof: First note that if every matrix in $\mathcal{F}$ has only one eigenvalue, there is nothing to prove. This is because for A such a matrix,

$$S^{-1}AS = \lambda I$$

and so

$$A = \lambda I$$

Thus all the matrices in $\mathcal{F}$ are diagonal matrices and you could pick any S to diagonalize them all. Therefore, without loss of generality, assume some matrix in $\mathcal{F}$ has more than one eigenvalue.

The significant part of the lemma is proved by induction on n. If $n = 1$, there is nothing to prove because all the 1×1 matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n-1$ where $n \geq 2$ and let $\mathcal{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathcal{F}$ which has more than one eigenvalue and let S be an invertible matrix such that $S^{-1}AS = D$ where D is of the form given in 12.1. By permuting the columns of S there is no loss of generality in assuming D has this form.

Now denote by $\widetilde{\mathcal{F}}$ the collection of matrices, $\left\{S^{-1}CS : C \in \mathcal{F}\right\}$. Note $\widetilde{\mathcal{F}}$ features the single matrix S.

It follows easily that $\widetilde{\mathcal{F}}$ is also a commuting family of diagonalizable matrices. By Lemma 12.1.7 every $B \in \widetilde{\mathcal{F}}$ is a block diagonal matrix of the form given in 12.2 because each of these commutes with D described above as $S^{-1}AS$ and so by block multiplication, the diagonal blocks B_i corresponding to different $B \in \widetilde{\mathcal{F}}$ commute.

By Corollary 12.1.4 each of these blocks is diagonalizable. This is because B is known to be so. Therefore, by induction, since all the blocks are no larger than $n-1 \times n-1$ thanks to the assumption that A has more than one eigenvalue, there exist invertible $n_i \times n_i$ matrices, T_i such that $T_i^{-1}B_iT_i$ is a diagonal matrix whenever B_i is one of the matrices making up the block diagonal of any $B \in \widetilde{\mathcal{F}}$. It follows that for T defined by

$$T \equiv \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix},$$

then $T^{-1}BT =$ a diagonal matrix for every $B \in \widetilde{\mathcal{F}}$ including D. Consider ST. It follows that for all $C \in \mathcal{F}$,

$$T^{-1} \overbrace{S^{-1}CS}^{\text{something in } \widetilde{\mathcal{F}}} T = (ST)^{-1} C (ST) = \text{ a diagonal matrix. } \blacksquare$$

Theorem 12.1.9 *Let $\mathcal{F}$ denote a family of matrices which are diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if $\mathcal{F}$ is a commuting family.*

Proof: If $\mathcal{F}$ is a commuting family, it follows from Lemma 12.1.8 that it is simultaneously diagonalizable. If it is simultaneously diagonalizable, then it follows from Lemma 12.1.6 that it is a commuting family. $\blacksquare$

12.2 Schur's Theorem

Recall that for a linear transformation, $L \in \mathcal{L}(V, V)$ for V a finite dimensional inner product space, it could be represented in the form

$$L = \sum_{ij} l_{ij} \mathbf{v}_i \otimes \mathbf{v}_j$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is an orthonormal basis. Of course different bases will yield different matrices, (l_{ij}). Schur's theorem gives the existence of a basis in an inner product space such that (l_{ij}) is particularly simple.

Definition 12.2.1 *Let $L \in \mathcal{L}(V, V)$ where V is a vector space. Then a subspace U of V is L invariant if $L(U) \subseteq U$.*

In what follows, $\mathbb{F}$ will be the field of scalars, usually $\mathbb{C}$ but maybe $\mathbb{R}$.

Theorem 12.2.2 *Let $L \in \mathcal{L}(H, H)$ for H a finite dimensional inner product space such that the restriction of $L^*$to every L invariant subspace has its eigenvalues in $\mathbb{F}$. Then there exist constants, c_{ij} for $i \leq j$ and an orthonormal basis, $\{\mathbf{w}_i\}_{i=1}^n$ such that*

$$L = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j$$

The constants, c_{ii} are the eigenvalues of L. Thus the matrix whose ij^{th} entry is c_{ij} is upper triangular.

Proof: If $\dim(H) = 1$, let $H = \text{span}(\mathbf{w})$ where $|\mathbf{w}| = 1$. Then $L\mathbf{w} = k\mathbf{w}$ for some k. Then

$$L = k\mathbf{w} \otimes \mathbf{w}$$

because by definition, $\mathbf{w} \otimes \mathbf{w}(\mathbf{w}) = \mathbf{w}$. Therefore, the theorem holds if H is 1 dimensional.

Now suppose the theorem holds for $n - 1 = \dim(H)$. Let $\mathbf{w}_n$ be an eigenvector for L^*. Dividing by its length, it can be assumed $|\mathbf{w}_n| = 1$. Say $L^*\mathbf{w}_n = \mu\mathbf{w}_n$. Using the Gram Schmidt process, there exists an orthonormal basis for H of the form $\{\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}, \mathbf{w}_n\}$. Then

$$(L\mathbf{v}_k, \mathbf{w}_n) = (\mathbf{v}_k, L^*\mathbf{w}_n) = (\mathbf{v}_k, \mu\mathbf{w}_n) = 0,$$

which shows

$$L : H_1 \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}) \to \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}).$$

Denote by L_1 the restriction of L to H_1. Since H_1 has dimension $n - 1$, the induction hypothesis yields an orthonormal basis, $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$ for H_1 such that

$$L_1 = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j. \tag{12.3}$$

Then $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ is an orthonormal basis for H because every vector in

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1})$$

has the property that its inner product with $\mathbf{w}_n$ is 0 so in particular, this is true for the vectors $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$. Now define c_{in} to be the scalars satisfying

$$L\mathbf{w}_n \equiv \sum_{i=1}^{n} c_{in} \mathbf{w}_i \tag{12.4}$$

and let

$$B \equiv \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j.$$

Then by 12.4,

$$B\mathbf{w}_n = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{nj} = \sum_{j=1}^{n} c_{in} \mathbf{w}_i = L\mathbf{w}_n.$$

If $1 \leq k \leq n - 1$,

$$B\mathbf{w}_k = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{kj} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i$$

while from 12.3,

$$L\mathbf{w}_k = L_1\mathbf{w}_k = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{jk} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i.$$

Since $L = B$ on the basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$, it follows $L = B$.

It remains to verify the constants, c_{kk} are the eigenvalues of L, solutions of the equation, $\det(\lambda I - L) = 0$. However, the definition of $\det(\lambda I - L)$ is the same as

$$\det(\lambda I - C)$$

where C is the upper triangular matrix which has c_{ij} for $i \leq j$ and zeros elsewhere. This equals 0 if and only if λ is one of the diagonal entries, one of the c_{kk}. ■

Now with the above Schur's theorem, the following diagonalization theorem comes very easily. Recall the following definition.

Definition 12.2.3 *Let $L \in \mathcal{L}(H, H)$ where H is a finite dimensional inner product space. Then L is Hermitian if $L^* = L$.*

Theorem 12.2.4 *Let $L \in \mathcal{L}(H, H)$ where H is an n dimensional inner product space. If L is Hermitian, then all of its eigenvalues λ_k are real and there exists an orthonormal basis of eigenvectors $\{\mathbf{w}_k\}$ such that*

$$L = \sum_k \lambda_k \mathbf{w}_k \otimes \mathbf{w}_k.$$

Proof: By Schur's theorem, Theorem 12.2.2, there exist $l_{ij} \in \mathbb{F}$ such that

$$L = \sum_{j=1}^{n} \sum_{i=1}^{j} l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j$$

Then by Lemma 11.4.2,

$$\begin{aligned} \sum_{j=1}^{n} \sum_{i=1}^{j} l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j &= L = L^* = \sum_{j=1}^{n} \sum_{i=1}^{j} \left(l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j \right)^* \\ &= \sum_{j=1}^{n} \sum_{i=1}^{j} \overline{l_{ij}} \mathbf{w}_j \otimes \mathbf{w}_i = \sum_{i=1}^{n} \sum_{j=1}^{i} \overline{l_{ji}} \mathbf{w}_i \otimes \mathbf{w}_j \end{aligned}$$

By independence, if $i = j$, $l_{ii} = \overline{l_{ii}}$ and so these are all real. If $i < j$, it follows from independence again that $l_{ij} = 0$ because the coefficients corresponding to $i < j$ are all 0 on the right side. Similarly if $i > j$, it follows $l_{ij} = 0$. Letting $\lambda_k = l_{kk}$, this shows

$$L = \sum_k \lambda_k \mathbf{w}_k \otimes \mathbf{w}_k$$

That each of these $\mathbf{w}_k$ is an eigenvector corresponding to λ_k is obvious from the definition of the tensor product. ■

12.3 Spectral Theory of Self Adjoint Operators

The following theorem is about the eigenvectors and eigenvalues of a self adjoint operator. Such operators are also called Hermitian as in the case of matrices. The proof given generalizes to the situation of a compact self adjoint operator on a Hilbert space and leads to many very useful results. It is also a very elementary proof because it does not use the fundamental theorem of algebra and it contains a way, very important in applications, of finding the eigenvalues. This proof depends more directly on the methods of analysis than the preceding material. Recall the following notation.

Definition 12.3.1 *Let X be an inner product space and let $S \subseteq X$. Then*

$$S^{\perp} \equiv \{x \in X : (x, s) = 0 \text{ for all } s \in S\}.$$

Note that even if S is not a subspace, $S^{\perp}$ is.

Theorem 12.3.2 *Let $A \in \mathcal{L}(X, X)$ be self adjoint (Hermitian) where X is a finite dimensional inner product space of dimension n. Thus $A = A^*$. Then there exists an orthonormal basis of eigenvectors, $\{v_j\}_{j=1}^n$.*

Proof: Consider (Ax, x). This quantity is always a real number because

$$\overline{(Ax, x)} = (x, Ax) = (x, A^* x) = (Ax, x)$$

thanks to the assumption that A is self adjoint. Now define

$$\lambda_1 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_1 \equiv X\}.$$

Claim: λ_1 is finite and there exists $v_1 \in X$ with $|v_1| = 1$ such that $(Av_1, v_1) = \lambda_1$.

Proof of claim: Let $\{u_j\}_{j=1}^n$ be an orthonormal basis for X and for $x \in X$, let $(x_1, \cdots, x_n)$ be defined as the components of the vector x. Thus,

$$x = \sum_{j=1}^n x_j u_j.$$

Since this is an orthonormal basis, it follows from the axioms of the inner product that

$$|x|^2 = \sum_{j=1}^n |x_j|^2.$$

Thus

$$(Ax, x) = \left(\sum_{k=1}^n x_k A u_k, \sum_{j=1} x_j u_j \right) = \sum_{k,j} x_k \overline{x_j} (Au_k, u_j),$$

a real valued continuous function of $(x_1, \cdots, x_n)$ which is defined on the compact set

$$K \equiv \{(x_1, \cdots, x_n) \in \mathbb{F}^n : \sum_{j=1}^n |x_j|^2 = 1\}.$$

Therefore, it achieves its minimum from the extreme value theorem. Then define

$$v_1 \equiv \sum_{j=1}^n x_j u_j$$

where $(x_1, \cdots, x_n)$ is the point of K at which the above function achieves its minimum. This proves the claim.

I claim that λ_1 is an eigenvalue and v_1 is an eigenvector. Letting $w \in X_1 \equiv X$, the function of the real variable, t, given by

$$f(t) \equiv \frac{(A(v_1 + tw), v_1 + tw)}{|v_1 + tw|^2} = \frac{(Av_1, v_1) + 2t \operatorname{Re}(Av_1, w) + t^2 (Aw, w)}{|v_1|^2 + 2t \operatorname{Re}(v_1, w) + t^2 |w|^2}$$

achieves its minimum when $t = 0$. Therefore, the derivative of this function evaluated at $t = 0$ must equal zero. Using the quotient rule, this implies, since $|v_1| = 1$ that

$$2 \operatorname{Re}(Av_1, w) |v_1|^2 - 2 \operatorname{Re}(v_1, w)(Av_1, v_1) = 2 (\operatorname{Re}(Av_1, w) - \operatorname{Re}(v_1, w) \lambda_1) = 0.$$

Thus $\operatorname{Re}(Av_1 - \lambda_1 v_1, w) = 0$ for all $w \in X$. This implies $Av_1 = \lambda_1 v_1$. To see this, let $w \in X$ be arbitrary and let θ be a complex number with $|\theta| = 1$ and

$$|(Av_1 - \lambda_1 v_1, w)| = \theta (Av_1 - \lambda_1 v_1, w).$$

Then

$$|(Av_1 - \lambda_1 v_1, w)| = \operatorname{Re}\left(Av_1 - \lambda_1 v_1, \overline{\theta} w\right) = 0.$$

Since this holds for all w, $Av_1 = \lambda_1 v_1$.

Continuing with the proof of the theorem, let $X_2 \equiv \{v_1\}^\perp$. This is a closed subspace of X and $A : X_2 \to X_2$ because for $x \in X_2$,

$$(Ax, v_1) = (x, Av_1) = \lambda_1 (x, v_1) = 0.$$

Let

$$\lambda_2 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_2\}$$

As before, there exists $v_2 \in X_2$ such that $Av_2 = \lambda_2 v_2$, $\lambda_1 \leq \lambda_2$. Now let $X_3 \equiv \{v_1, v_2\}^\perp$ and continue in this way. As long as $k < n$, it will be the case that $\{v_1, \cdots, v_k\}^\perp \neq \{0\}$. This is because for $k < n$ these vectors cannot be a spanning set and so there exists some $w \notin \operatorname{span}(v_1, \cdots, v_k)$. Then letting z be the closest point to w from $\operatorname{span}(v_1, \cdots, v_k)$, it follows that $w - z \in \{v_1, \cdots, v_k\}^\perp$. Thus there is an decreasing sequence of eigenvalues $\{\lambda_k\}_{k=1}^n$ and a corresponding sequence of eigenvectors, $\{v_1, \cdots, v_n\}$ with this being an orthonormal set. ■

Contained in the proof of this theorem is the following important corollary.

Corollary 12.3.3 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional inner product space. Then all the eigenvalues are real and for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exists an orthonormal set of vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \inf \{(Ax, x) : |x| = 1, x \in X_k\}$$

where

$$X_k \equiv \{u_1, \cdots, u_{k-1}\}^\perp, X_1 \equiv X.$$

Corollary 12.3.4 *Let $A \in \mathcal{L}(X, X)$ be self adjoint (Hermitian) where X is a finite dimensional inner product space. Then the largest eigenvalue of A is given by*

$$\max \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\} \tag{12.5}$$

and the minimum eigenvalue of A is given by

$$\min \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\}. \tag{12.6}$$

Proof: The proof of this is just like the proof of Theorem 12.3.2. Simply replace inf with sup and obtain a decreasing list of eigenvalues. This establishes 12.5. The claim 12.6 follows from Theorem 12.3.2. ■

Another important observation is found in the following corollary.

Corollary 12.3.5 *Let $A \in \mathcal{L}(X, X)$ where A is self adjoint. Then $A = \sum_i \lambda_i v_i \otimes v_i$ where $Av_i = \lambda_i v_i$ and $\{v_i\}_{i=1}^n$ is an orthonormal basis.*

Proof : If v_k is one of the orthonormal basis vectors, $Av_k = \lambda_k v_k$. Also,

$$\sum_i \lambda_i v_i \otimes v_i (v_k) = \sum_i \lambda_i v_i (v_k, v_i) = \sum_i \lambda_i \delta_{ik} v_i = \lambda_k v_k.$$

Since the two linear transformations agree on a basis, it follows they must coincide. ■

By Theorem 11.4.5 this says the matrix of A with respect to this basis $\{v_i\}_{i=1}^n$ is the diagonal matrix having the eigenvalues $\lambda_1, \cdots, \lambda_n$ down the main diagonal.

The result of Courant and Fischer which follows resembles Corollary 12.3.3 but is more useful because it does not depend on a knowledge of the eigenvectors.

Theorem 12.3.6 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional inner product space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1, \cdots, w_{k-1}} \left\{ \min \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \tag{12.7}$$

where if $k = 1$, $\{w_1, \cdots, w_{k-1}\}^{\perp} \equiv X$.

Proof: From Theorem 12.3.2, there exist eigenvalues and eigenvectors with $\{u_1, \cdots, u_n\}$ orthonormal and $\lambda_i \leq \lambda_{i+1}$.

$$(Ax, x) = \sum_{j=1}^n (Ax, u_j) \overline{(x, u_j)} = \sum_{j=1}^n \lambda_j (x, u_j)(u_j, x) = \sum_{j=1}^n \lambda_j |(x, u_j)|^2$$

Recall that $(z, w) = \sum_j (z, u_j) \overline{(w, u_i)}$. Then let $Y = \{w_1, \cdots, w_{k-1}\}^{\perp}$

$$\inf \{(Ax, x) : |x| = 1, x \in Y\} = \inf \left\{ \sum_{j=1}^n \lambda_j |(x, u_j)|^2 : |x| = 1, x \in Y \right\}$$

$$\leq \inf \left\{ \sum_{j=1}^k \lambda_j |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\}. \tag{12.8}$$

The reason this is so is that the infimum is taken over a smaller set. Therefore, the infimum gets larger. Now 12.8 is no larger than

$$\inf \left\{ \lambda_k \sum_{j=1}^n |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\} \leq \lambda_k$$

because since $\{u_1, \cdots, u_n\}$ is an orthonormal basis, $|x|^2 = \sum_{j=1}^n |(x, u_j)|^2$. It follows, since $\{w_1, \cdots, w_{k-1}\}$ is arbitrary,

$$\sup_{w_1, \cdots, w_{k-1}} \left\{ \inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \leq \lambda_k. \tag{12.9}$$

Then from Corollary 12.3.3,

$$\lambda_k = \inf \left\{ (Ax, x) : |x| = 1, x \in \{u_1, \cdots, u_{k-1}\}^{\perp} \right\} \leq$$

$$\sup_{w_1,\cdots,w_{k-1}} \left\{ \inf \left\{ (Ax,x) : |x| = 1, x \in \{w_1,\cdots,w_{k-1}\}^{\perp} \right\} \right\} \leq \lambda_k$$

Hence these are all equal and this proves the theorem. ■

The following corollary is immediate.

Corollary 12.3.7 *Let $A \in \mathcal{L}(X,X)$ be self adjoint where X is a finite dimensional inner product space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1,\cdots,u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1,\cdots,w_{k-1}} \left\{ \min \left\{ \frac{(Ax,x)}{|x|^2} : x \neq 0, x \in \{w_1,\cdots,w_{k-1}\}^{\perp} \right\} \right\} \tag{12.10}$$

where if $k = 1, \{w_1,\cdots,w_{k-1}\}^{\perp} \equiv X$.

Here is a version of this for which the roles of max and min are reversed.

Corollary 12.3.8 *Let $A \in \mathcal{L}(X,X)$ be self adjoint where X is a finite dimensional inner product space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1,\cdots,u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \min_{w_1,\cdots,w_{n-k}} \left\{ \max \left\{ \frac{(Ax,x)}{|x|^2} : x \neq 0, x \in \{w_1,\cdots,w_{n-k}\}^{\perp} \right\} \right\} \tag{12.11}$$

where if $k = n, \{w_1,\cdots,w_{n-k}\}^{\perp} \equiv X$.

12.4 Positive and Negative Linear Transformations

The notion of a positive definite or negative definite linear transformation is very important in many applications. In particular it is used in versions of the second derivative test for functions of many variables. Here the main interest is the case of a linear transformation which is an $n \times n$ matrix but the theorem is stated and proved using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

Definition 12.4.1 *A self adjoint $A \in \mathcal{L}(X,X)$, is positive definite if whenever $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x},\mathbf{x}) > 0$ and A is negative definite if for all $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x},\mathbf{x}) < 0$. A is positive semidefinite or just nonnegative for short if for all $\mathbf{x}$, $(A\mathbf{x},\mathbf{x}) \geq 0$. A is negative semidefinite or nonpositive for short if for all $\mathbf{x}$, $(A\mathbf{x},\mathbf{x}) \leq 0$.*

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

Lemma 12.4.2 *Let X be a finite dimensional inner product space. A self adjoint $A \in \mathcal{L}(X,X)$ is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.*

Proof: Suppose first that A is positive definite and let λ be an eigenvalue. Then for $\mathbf{x}$ an eigenvector corresponding to λ, $\lambda(\mathbf{x},\mathbf{x}) = (\lambda\mathbf{x},\mathbf{x}) = (A\mathbf{x},\mathbf{x}) > 0$. Therefore, $\lambda > 0$ as claimed.

Now suppose all the eigenvalues of A are positive. From Theorem 12.3.2 and Corollary 12.3.5, $A = \sum_{i=1}^n \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i$ where the λ_i are the positive eigenvalues and $\{\mathbf{u}_i\}$ are an orthonormal set of eigenvectors. Therefore, letting $\mathbf{x} \neq \mathbf{0}$,

$$\begin{aligned}(A\mathbf{x},\mathbf{x}) &= \left(\left(\sum_{i=1}^n \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i\right)\mathbf{x},\mathbf{x}\right) = \left(\sum_{i=1}^n \lambda_i \mathbf{u}_i(\mathbf{x},\mathbf{u}_i),\mathbf{x}\right) \\ &= \left(\sum_{i=1}^n \lambda_i (\mathbf{x},\mathbf{u}_i)(\mathbf{u}_i,\mathbf{x})\right) = \sum_{i=1}^n \lambda_i \left|(\mathbf{u}_i,\mathbf{x})\right|^2 > 0\end{aligned}$$

because, since $\{\mathbf{u}_i\}$ is an orthonormal basis, $|\mathbf{x}|^2 = \sum_{i=1}^n \left|(\mathbf{u}_i,\mathbf{x})\right|^2$.

To establish the claim about negative definite, it suffices to note that A is negative definite if and only if $-A$ is positive definite and the eigenvalues of A are (-1) times the eigenvalues of $-A$. The claims about positive semidefinite and negative semidefinite are obtained similarly. ■

The next theorem is about a way to recognize whether a self adjoint $n \times n$ complex matrix A is positive or negative definite without having to find the eigenvalues. In order to state this theorem, here is some notation.

Definition 12.4.3 *Let A be an $n \times n$ matrix. Denote by A_k the $k \times k$ matrix obtained by deleting the $k+1,\cdots,n$ columns and the $k+1,\cdots,n$ rows from A. Thus $A_n = A$ and A_k is the $k \times k$ submatrix of A which occupies the upper left corner of A. The determinants of these submatrices are called the principle minors.*

The following theorem is proved in [8]. For the sake of simplicity, we state this for real matrices since this is also where the main interest lies.

Theorem 12.4.4 *Let A be a self adjoint $n \times n$ matrix. Then A is positive definite if and only if $\det(A_k) > 0$ for every $k = 1,\cdots,n$.*

Proof: This theorem is proved by induction on n. It is clearly true if $n = 1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\det(A) > 0$, it follows that all the eigenvalues are nonzero. Are they all positive? Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\det(A)$. Pick two, λ_1 and λ_2 and let $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ where $\mathbf{u}_i \neq \mathbf{0}$ for $i = 1,2$ and $(\mathbf{u}_1,\mathbf{u}_2) = 0$. Now if $\mathbf{y} \equiv \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$ is an element of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$, then since these are eigenvalues and $(\mathbf{u}_1,\mathbf{u}_2)_{\mathbb{R}^n} = 0$, a short computation shows

$$(A(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2), \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = |\alpha_1|^2 \lambda_1 |\mathbf{u}_1|^2 + |\alpha_2|^2 \lambda_2 |\mathbf{u}_2|^2 < 0.$$

Now letting $\mathbf{x} \in \mathbb{R}^{n-1}$, $\mathbf{x} \neq \mathbf{0}$, the induction hypothesis implies

$$\left(\mathbf{x}^T, 0\right) A \begin{pmatrix} \mathbf{x} \\ 0 \end{pmatrix} = \mathbf{x}^T A_{n-1}\mathbf{x} = (A_{n-1}\mathbf{x},\mathbf{x}) > 0.$$

The dimension of $\{\mathbf{z} \in \mathbb{R}^n : z_n = 0\}$ is $n-1$ and the dimension of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2) = 2$ and so there must be some nonzero $\mathbf{x} \in \mathbb{R}^n$ which is in both of these subspaces of $\mathbb{R}^n$. However, the first computation would require that $(A\mathbf{x},\mathbf{x}) < 0$ while the second would require that $(A\mathbf{x},\mathbf{x}) > 0$. This contradiction shows that all the eigenvalues must be positive. This proves the if part of the theorem.

To show the converse, note that, as above, $(A\mathbf{x},\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. Suppose that A is positive definite. Then this is equivalent to having

$$\mathbf{x}^T A\mathbf{x} \geq \delta \|\mathbf{x}\|^2$$

Note that for $\mathbf{x} \in \mathbb{R}^k$,

$$\begin{pmatrix} \mathbf{x}^T & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} = \mathbf{x}^T A_k \mathbf{x} \geq \delta \|\mathbf{x}\|^2$$

From Lemma 12.4.2, this implies that all the eigenvalues of A_k are positive. Hence from Lemma 12.4.2, it follows that $\det(A_k) > 0$, being the product of its eigenvalues. ■

Corollary 12.4.5 *Let A be a self adjoint $n \times n$ matrix. Then A is negative definite if and only if $\det(A_k)(-1)^k > 0$ for every $k = 1, \cdots, n$.*

Proof: This is immediate from the above theorem by noting that, as in the proof of Lemma 12.4.2, A is negative definite if and only if $-A$ is positive definite. Therefore, $\det(-A_k) > 0$ for all $k = 1, \cdots, n$, is equivalent to having A negative definite. However, $\det(-A_k) = (-1)^k \det(A_k)$. ■

12.5 The Square Root

With the above theory, it is possible to take fractional powers of certain elements of $\mathcal{L}(X, X)$ where X is a finite dimensional inner product space. I will give two treatments of this, the first pertaining to the square root only and the second more generally pertaining to the k^{th} root of a self adjoint nonnegative matrix.

Theorem 12.5.1 *Let $A \in \mathcal{L}(X, X)$ be self adjoint and nonnegative. Then there exists a unique self adjoint nonnegative $B \in \mathcal{L}(X, X)$ such that $B^2 = A$ and B commutes with every element of $\mathcal{L}(X, X)$ which commutes with A.*

Proof: By Theorem 12.3.2, there exists an orthonormal basis of eigenvectors of A, say $\{v_i\}_{i=1}^n$ such that $Av_i = \lambda_i v_i$. Therefore, by Theorem 12.2.4, $A = \sum_i \lambda_i v_i \otimes v_i$ where each $\lambda_i \geq 0$.

Now by Lemma 12.4.2, each $\lambda_i \geq 0$. Therefore, it makes sense to define

$$B \equiv \sum_i \lambda_i^{1/2} v_i \otimes v_i.$$

It is easy to verify that

$$(v_i \otimes v_i)(v_j \otimes v_j) = \begin{cases} 0 \text{ if } i \neq j \\ v_i \otimes v_i \text{ if } i = j \end{cases}.$$

Therefore, a short computation verifies that $B^2 = \sum_i \lambda_i v_i \otimes v_i = A$. If C commutes with A, then for some c_{ij},

$$C = \sum_{ij} c_{ij} v_i \otimes v_j$$

and so since they commute,

$$\sum_{i,j,k} c_{ij} v_i \otimes v_j \lambda_k v_k \otimes v_k = \sum_{i,j,k} c_{ij} \lambda_k \delta_{jk} v_i \otimes v_k = \sum_{i,k} c_{ik} \lambda_k v_i \otimes v_k$$

$$\begin{aligned} &= \sum_{i,j,k} c_{ij}\lambda_k v_k \otimes v_k v_i \otimes v_j = \sum_{i,j,k} c_{ij}\lambda_k \delta_{ki} v_k \otimes v_j = \sum_{j,k} c_{kj}\lambda_k v_k \otimes v_j \\ &= \sum_{k,i} c_{ik}\lambda_i v_i \otimes v_k \end{aligned}$$

Then by independence,

$$c_{ik}\lambda_i = c_{ik}\lambda_k$$

Therefore, $c_{ik}\lambda_i^{1/2} = c_{ik}\lambda_k^{1/2}$ which amounts to saying that B also commutes with C. It is clear that this operator is self adjoint. This proves existence.

Suppose B_1 is another square root which is self adjoint, nonnegative and commutes with every linear transformation which commutes with A. Since both B, B_1 are nonnegative,

$$\begin{aligned} (B(B-B_1)x, (B-B_1)x) &\geq 0, \\ (B_1(B-B_1)x, (B-B_1)x) &\geq 0 \end{aligned} \tag{12.12}$$

Now, adding these together, and using the fact that the two commute,

$$\left(\left(B^2 - B_1^2\right)x, (B-B_1)x\right) = ((A-A)x, (B-B_1)x) = 0.$$

It follows that both inner products in 12.12 equal 0. Next use the existence part of this to take the square root of B and B_1 which is denoted by $\sqrt{B}, \sqrt{B_1}$ respectively. Then

$$\begin{aligned} 0 &= \left(\sqrt{B}(B-B_1)x, \sqrt{B}(B-B_1)x\right) \\ 0 &= \left(\sqrt{B_1}(B-B_1)x, \sqrt{B_1}(B-B_1)x\right) \end{aligned}$$

which implies $\sqrt{B}(B-B_1)x = \sqrt{B_1}(B-B_1)x = 0$. Thus also,

$$B(B-B_1)x = B_1(B-B_1)x = 0$$

Hence

$$0 = (B(B-B_1)x - B_1(B-B_1)x, x) = ((B-B_1)x, (B-B_1)x)$$

and so, since x is arbitrary, $B_1 = B$. ■

12.6 Fractional Powers

The main result is the following theorem.

Theorem 12.6.1 *Let A be a self adjoint and nonnegative $n \times n$ matrix (all eigenvalues are nonnegative) and let k be a positive integer. Then there exists a unique self adjoint nonnegative matrix B such that $B^k = A$.*

Proof: By Theorem 12.3.2 or Corollary 6.4.12, there exists an orthonormal basis of eigenvectors of A, say $\{v_i\}_{i=1}^n$ such that $Av_i = \lambda_i v_i$ with each λ_i real. In particular, there exists a unitary matrix U such that

$$U^*AU = D, \quad A = UDU^*$$

where D has nonnegative diagonal entries. Define B in the obvious way.

$$B \equiv UD^{1/k}U^*$$

Then it is clear that B is self adjoint and nonnegative. Also it is clear that $B^k = A$. What of uniqueness? Let $p(t)$ be a polynomial whose graph contains the ordered pairs $\left(\lambda_i, \lambda_i^{1/k}\right)$ where the λ_i are the diagonal entries of D, the eigenvalues of A. Then

$$p(A) = UP(D)U^* = UD^{1/k}U^* \equiv B$$

Suppose then that $C^k = A$ and C is also self adjoint and nonnegative.

$$CB = Cp(A) = Cp\left(C^k\right) = p\left(C^k\right)C = p(A)C = BC$$

and so $\{B, C\}$ is a commuting family of non defective matrices. By Theorem 12.1.9 this family of matrices is simultaneously diagonalizable. Hence there exists a single S such that

$$S^{-1}BS = D_B, \quad S^{-1}CS = D_C$$

Where D_C, D_B denote diagonal matrices. Hence, raising to the power k, it follows that

$$A = B^k = SD_B^k S^{-1}, \quad A = C^k = SD_C^k S^{-1}$$

Hence

$$SD_B^k S^{-1} = SD_C^k S^{-1}$$

and so $D_B^k = D_C^k$. Since the entries of the two diagonal matrices are nonnegative, this implies $D_B = D_C$ and so $S^{-1}BS = S^{-1}CS$ which shows $B = C$. ■

A similar result holds for a general finite dimensional inner product space. See Problem 22 in the exercises.

12.7 Square Roots and Polar Decompositions

An application of Theorem 12.3.2, is the following fundamental result, important in geometric measure theory and continuum mechanics. It is sometimes called the right polar decomposition. The notation used is that which is seen in continuum mechanics, see for example Gurtin [12]. Don't confuse the U in this theorem with a unitary transformation. It is not so. When the following theorem is applied in continuum mechanics, F is normally the deformation gradient, the derivative of a nonlinear map from some subset of three dimensional space to three dimensional space. In this context, U is called the right Cauchy Green strain tensor. It is a measure of how a body is stretched independent of rigid motions. First, here is a simple lemma.

Lemma 12.7.1 *Suppose $R \in \mathcal{L}(X, Y)$ where X, Y are inner product spaces and R preserves distances. Then $R^*R = I$.*

Proof: Since R preserves distances, $|R\mathbf{u}| = |\mathbf{u}|$ for every $\mathbf{u}$. Let $\mathbf{u}, \mathbf{v}$ be arbitrary vectors in X and let $\theta \in \mathbb{C}$, $|\theta| = 1$, and $\theta(R^*R\mathbf{u} - \mathbf{u}, \mathbf{v}) = |(R^*R\mathbf{u} - \mathbf{u}, \mathbf{v})|$. Therefore from the axioms of the inner product,

$$\begin{aligned} |\mathbf{u}|^2 + |\mathbf{v}|^2 + 2\operatorname{Re}\theta(\mathbf{u}, \mathbf{v}) &= |\theta\mathbf{u}|^2 + |\mathbf{v}|^2 + \theta(\mathbf{u}, \mathbf{v}) + \bar{\theta}(\mathbf{v}, \mathbf{u}) \\ &= |\theta\mathbf{u} + \mathbf{v}|^2 = (R(\theta\mathbf{u} + \mathbf{v}), R(\theta\mathbf{u} + \mathbf{v})) \end{aligned}$$

$$= (R\theta\mathbf{u}, R\theta\mathbf{u}) + (R\mathbf{v}, R\mathbf{v}) + (R\theta\mathbf{u}, R\mathbf{v}) + (R\mathbf{v}, R\theta\mathbf{u})$$

$$\begin{aligned} &= |\theta\mathbf{u}|^2 + |\mathbf{v}|^2 + \theta(R^*R\mathbf{u}, \mathbf{v}) + \bar{\theta}(\mathbf{v}, R^*R\mathbf{u}) \\ &= |\mathbf{u}|^2 + |\mathbf{v}|^2 + 2\operatorname{Re}\theta(R^*R\mathbf{u}, \mathbf{v}) \end{aligned}$$

and so for all $\mathbf{u}, \mathbf{v}$,

$$2 \operatorname{Re} \theta \left(R^* R\mathbf{u} - \mathbf{u}, \mathbf{v}\right) = 2 \left|\left(R^* R\mathbf{u} - \mathbf{u}, \mathbf{v}\right)\right| = 0$$

Now let $\mathbf{v} = R^* R\mathbf{u} - \mathbf{u}$. It follows that $R^* R\mathbf{u} - \mathbf{u} = \mathbf{0}$. ■

The decomposition in the following is called the right polar decomposition.

Theorem 12.7.2 *Let X be a inner product space of dimension n and let Y be a inner product space of dimension $m \geq n$ and let $F \in \mathcal{L}(X, Y)$. Then there exists $R \in \mathcal{L}(X, Y)$ and $U \in \mathcal{L}(X, X)$ such that*

$$F = RU,\ U = U^*, (U \text{ is Hermitian}),$$

all eigenvalues of U are non negative,

$$U^2 = F^* F, R^* R = I,$$

and $|R\mathbf{x}| = |\mathbf{x}|$.

Proof: $(F^* F)^* = F^* F$ and so by Theorem 12.3.2, there is an orthonormal basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$F^* F\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ F^* F = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i.$$

It is also clear that $\lambda_i \geq 0$ because

$$\lambda_i \left(\mathbf{v}_i, \mathbf{v}_i\right) = \left(F^* F\mathbf{v}_i, \mathbf{v}_i\right) = \left(F\mathbf{v}_i, F\mathbf{v}_i\right) \geq 0.$$

Let

$$U \equiv \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{v}_i \otimes \mathbf{v}_i.$$

Then $U^2 = F^* F$, $U = U^*$, and the eigenvalues of U, $\left\{\lambda_i^{1/2}\right\}_{i=1}^{n}$ are all non negative.

Let $\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r\}$ be an orthonormal basis for $U(X)$. By the Gram Schmidt procedure there exists an extension to an orthonormal basis for X,

$$\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r, \mathbf{y}_{r+1}, \cdots, \mathbf{y}_n\}.$$

Next note that $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ is also an orthonormal set of vectors in Y because

$$\left(F\mathbf{x}_k, F\mathbf{x}_j\right) = \left(F^* F\mathbf{x}_k, \mathbf{x}_j\right) = \left(U^2\mathbf{x}_k, \mathbf{x}_j\right) = \left(U\mathbf{x}_k, U\mathbf{x}_j\right) = \delta_{jk}.$$

By the Gram Schmidt procedure, there exists an extension of $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ to an orthonormal basis for Y,

$$\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r, \mathbf{z}_{r+1}, \cdots, \mathbf{z}_m\}.$$

Since $m \geq n$, there are at least as many $\mathbf{z}_k$ as there are $\mathbf{y}_k$. Now for $\mathbf{x} \in X$, since

$$\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r, \mathbf{y}_{r+1}, \cdots, \mathbf{y}_n\}$$

is an orthonormal basis for X, 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

$$R\mathbf{x} \equiv \sum_{k=1}^{r} c_k F \mathbf{x}_k + \sum_{k=r+1}^{n} d_k \mathbf{z}_k \tag{12.13}$$

Thus

$$|R\mathbf{x}|^2 = \sum_{k=1}^{r} |c_k|^2 + \sum_{k=r+1}^{n} |d_k|^2 = |\mathbf{x}|^2 .$$

Therefore, by Lemma 12.7.1 $R^*R = I$.

Then also there exist unique scalars b_k such that for a given $\mathbf{x} \in X$,

$$U\mathbf{x} = \sum_{k=1}^{r} b_k U \mathbf{x}_k \tag{12.14}$$

and so from 12.13,

$$RU\mathbf{x} = \sum_{k=1}^{r} b_k F \mathbf{x}_k = F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right)$$

Is $F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$?

$$\left(F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x}), F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x})\right)$$

$$\begin{aligned}
&= \left((F^*F)\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U \mathbf{x}_k - U\mathbf{x}, \sum_{k=1}^{r} b_k U \mathbf{x}_k - U\mathbf{x}\right) = 0
\end{aligned}$$

Because from 12.14, $U\mathbf{x} = \sum_{k=1}^{r} b_k U \mathbf{x}_k$. Therefore, $RU\mathbf{x} = F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$. ■

The following corollary follows as a simple consequence of this theorem. It is called the left polar decomposition.

Corollary 12.7.3 *Let $F \in \mathcal{L}(X,Y)$ and suppose $n \geq m$ where X is a inner product space of dimension n and Y is a inner product space of dimension m. Then there exists a Hermitian $U \in \mathcal{L}(X,X)$, and an element of $\mathcal{L}(X,Y)$, R, such that*

$$F = UR, \ RR^* = I.$$

Proof: Recall that $L^{**} = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 12.7.2 to $F^* \in \mathcal{L}(Y,X)$. Thus, $F^* = R^*U$ where R^* and U satisfy the conditions of that theorem. Then $F = UR$ and $RR^* = R^{**}R^* = I$. ■

The following existence theorem for the polar decomposition of an element of $\mathcal{L}(X,X)$ is a corollary.

Corollary 12.7.4 *Let $F \in \mathcal{L}(X,X)$. Then there exists a Hermitian $W \in \mathcal{L}(X,X)$, and a unitary matrix Q such that $F = WQ$, and there exists a Hermitian $U \in \mathcal{L}(X,X)$ and a unitary R, such that $F = RU$.*

This corollary has a fascinating relation to the question whether a given linear transformation is normal. Recall that an $n \times n$ matrix A, is normal if $AA^* = A^*A$. Retain the same definition for an element of $\mathcal{L}(X,X)$.

Theorem 12.7.5 *Let $F \in \mathcal{L}(X,X)$. Then F is normal if and only if in Corollary 12.7.4 $RU = UR$ and $QW = WQ$.*

Proof: I will prove the statement about $RU = UR$ and leave the other part as an exercise. First suppose that $RU = UR$ and show F is normal. To begin with,

$$UR^* = (RU)^* = (UR)^* = R^*U.$$

Therefore,

$$\begin{aligned} F^*F &= UR^*RU = U^2 \\ FF^* &= RUUR^* = URR^*U = U^2 \end{aligned}$$

which shows F is normal.

Now suppose F is normal. Is $RU = UR$? Since F is normal,

$$FF^* = RUUR^* = RU^2R^*$$

and

$$F^*F = UR^*RU = U^2.$$

Therefore, $RU^2R^* = U^2$, and both are nonnegative and self adjoint. Therefore, the square roots of both sides must be equal by the uniqueness part of the theorem on fractional powers. It follows that the square root of the first, RUR^* must equal the square root of the second, U. Therefore, $RUR^* = U$ and so $RU = UR$. This proves the theorem in one case. The other case in which W and Q commute is left as an exercise. ∎

12.8 An Application to Statistics

A random vector is a function $\mathbf{X} : \Omega \to \mathbb{R}^p$ where Ω is a probability space. This means that there exists a σ algebra of measurable sets $\mathcal{F}$ and a probability measure $P : \mathcal{F} \to [0,1]$. In practice, people often don't worry too much about the underlying probability space and instead pay more attention to the distribution measure of the random variable. For E a suitable subset of $\mathbb{R}^p$, this measure gives the probability that $\mathbf{X}$ has values in E. There are often excellent reasons for believing that a random vector is normally distributed. This means that the probability that $\mathbf{X}$ has values in a set E is given by

$$\int_E \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^* \Sigma^{-1} (\mathbf{x}-\mathbf{m})\right) d\mathbf{x}$$

The expression in the integral is called the normal probability density function. There are two parameters, $\mathbf{m}$ and Σ where $\mathbf{m}$ is called the mean and Σ is called the covariance matrix. It is a symmetric matrix which has all real eigenvalues which are all positive. While it may be reasonable to assume this is the distribution, in general, you won't know $\mathbf{m}$ and Σ and in order to use this formula to predict anything, you would need to know these quantities. I am following a nice discussion given in Wikipedia which makes use of the existence of square roots.

What people do to estimate these is to take n independent observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and try to predict what $\mathbf{m}$ and Σ should be based on these observations. One criterion used for making this determination is the method of maximum likelihood. In this method, you seek to choose the two parameters in such a way as to maximize the likelihood which is given as

$$\prod_{i=1}^{n} \frac{1}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m})\right).$$

For convenience the term $(2\pi)^{p/2}$ was ignored. Maximizing the above is equivalent to maximizing the ln of the above. So taking ln,

$$\frac{n}{2} \ln\left(\det\left(\Sigma^{-1}\right)\right) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m})$$

Note that the above is a function of the entries of $\mathbf{m}$. Take the partial derivative with respect to m_l. Since the matrix Σ^{-1} is symmetric this implies

$$\sum_{i=1}^{n} \sum_{r} (x_{ir} - m_r) \Sigma_{rl}^{-1} = 0 \text{ each } l.$$

Written in terms of vectors,

$$\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} = \mathbf{0}$$

and so, multiplying by Σ on the right and then taking adjoints, this yields

$$\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) = \mathbf{0},\ n\mathbf{m} = \sum_{i=1}^{n} \mathbf{x}_i,\ \ \mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \equiv \bar{\mathbf{x}}.$$

Now that $\mathbf{m}$ is determined, it remains to find the best estimate for Σ. $(\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m})$ is a scalar, so since $\operatorname{trace}(AB) = \operatorname{trace}(BA)$,

$$\begin{aligned} (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m}) &= \operatorname{trace}\left((\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1} (\mathbf{x}_i - \mathbf{m})\right) \\ &= \operatorname{trace}\left((\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1}\right) \end{aligned}$$

Therefore, the thing to maximize is

$$\begin{aligned} & n \ln\left(\det\left(\Sigma^{-1}\right)\right) - \sum_{i=1}^{n} \operatorname{trace}\left((\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1}\right) \\ = \ & n \ln\left(\det\left(\Sigma^{-1}\right)\right) - \operatorname{trace}\left(\overbrace{\left(\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^*\right)}^{S} \Sigma^{-1}\right) \end{aligned}$$

We assume that S has rank p. Thus it is a self adjoint matrix which has all positive eigenvalues. Therefore, from the property of the trace, the thing to maximize is

$$n \ln\left(\det\left(\Sigma^{-1}\right)\right) - \operatorname{trace}\left(S^{1/2} \Sigma^{-1} S^{1/2}\right)$$

Now let $B = S^{1/2}\Sigma^{-1}S^{1/2}$. Then B is positive and self adjoint also and so there exists U unitary such that $B = U^*DU$ where D is the diagonal matrix having the positive scalars $\lambda_1, \cdots, \lambda_p$ down the main diagonal. Solving for Σ^{-1} in terms of B, this yields $S^{-1/2}BS^{-1/2} = \Sigma^{-1}$ and so

$$\ln\left(\det\left(\Sigma^{-1}\right)\right) = \ln\left(\det\left(S^{-1/2}\right)\det(B)\det\left(S^{-1/2}\right)\right) = \ln\left(\det\left(S^{-1}\right)\right) + \ln\left(\det(B)\right)$$

which yields

$$C(S) + n\ln(\det(B)) - \operatorname{trace}(B)$$

as the thing to maximize. Of course this yields

$$\begin{aligned} & C(S) + n\ln\left(\prod_{i=1}^p \lambda_i\right) - \sum_{i=1}^p \lambda_i \\ = \quad & C(S) + n\sum_{i=1}^p \ln(\lambda_i) - \sum_{i=1}^p \lambda_i \end{aligned}$$

as the quantity to be maximized. To do this, take $\partial/\partial\lambda_k$ and set equal to 0. This yields $\lambda_k = n$. Therefore, from the above, $B = U^*nIU = nI$. Also from the above,

$$B^{-1} = \frac{1}{n}I = S^{-1/2}\Sigma S^{-1/2}$$

and so

$$\Sigma = \frac{1}{n}S = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^*$$

This has shown that the maximum likelihood estimates are

$$\mathbf{m} = \bar{\mathbf{x}} \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i, \ \Sigma = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^*.$$

12.9 The Singular Value Decomposition

In this section, A will be an $m \times n$ matrix. To begin with, here is a simple lemma.

Lemma 12.9.1 *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 12.9.2 *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. In all that follows, I will write the following partitioned matrix

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ denotes an $r \times r$ diagonal matrix of the form

$$\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

and the bottom row of zero matrices in the partitioned matrix, as well as the right columns of zero matrices are each of the right size so that the resulting matrix is $m \times n$. Either could vanish completely. However, I will write it in the above form. It is easy to make the necessary adjustments in the other two cases.

Theorem 12.9.3 *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, arranged in order of decreasing size.

Proof: By the above lemma and Theorem 12.3.2 there exists an orthonormal basis, $\{\mathbf{v}_i\}_{i=1}^n$ for $\mathbb{F}^n$ 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 $i > k$. Let the eigenvalues σ_i^2 be arranged in decreasing order. It is desired to have

$$AV = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

and so if $U = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{pmatrix}$, one needs to have for $j \leq k$, $\sigma_j \mathbf{u}_j = A\mathbf{v}_j$. Thus let

$$\mathbf{u}_j \equiv \sigma_j^{-1} A\mathbf{v}_j, \ j \leq k$$

Then for $i, j \leq k$,

$$\begin{aligned} (\mathbf{u}_i, \mathbf{u}_j) &= \sigma_j^{-1}\sigma_i^{-1} (A\mathbf{v}_i, A\mathbf{v}_j) = \sigma_j^{-1}\sigma_i^{-1} (A^* A\mathbf{v}_i, \mathbf{v}_j) \\ &= \sigma_j^{-1}\sigma_i^{-1}\sigma_i^2 (\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij} \end{aligned}$$

Now extend to an orthonormal basis of $\mathbb{F}^m$, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}, \cdots, \mathbf{u}_m\}$. If $i > k$,

$$(A\mathbf{v}_i, A\mathbf{v}_i) = (A^* A\mathbf{v}_i, \mathbf{v}_i) = 0 (\mathbf{v}_i, \mathbf{v}_i) = 0$$

so $A\mathbf{v}_i = \mathbf{0}$. Then for σ given as above in the statement of the theorem, it follows that

$$AV = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}, \ U^* AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \ \blacksquare$$

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 12.9.4 *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, they are each one to one and onto and so it follows that

$$\text{rank}(A) = \text{rank}(U^*AV) = \text{rank}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \text{number of singular values.}$$

Also since U, V are unitary,

$$\text{rank}(A^*) = \text{rank}(V^*A^*U) = \text{rank}\left((U^*AV)^*\right)$$

$$= \text{rank}\left(\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}^*\right) = \text{number of singular values. } \blacksquare$$

12.10 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. Here is its definition.

Definition 12.10.1 *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 $\sum_{ij} |a_{ij}|^2$ so essentially, it treats the matrix as a vector in $\mathbb{F}^{m\times n}$.

Lemma 12.10.2 *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, $U^*AV = \Sigma$. Then

$$||\Sigma||_F^2 = ||A||_F^2 \tag{12.15}$$

and the following hold for the Frobenius norm. If U, V are unitary and of the right size,

$$||UA||_F = ||A||_F\,,\ ||UAV||_F = ||A||_F\,. \tag{12.16}$$

Proof: From the definition and letting U, V be unitary and of the right size,

$$||UA||_F^2 \equiv \text{trace}(UAA^*U^*) = \text{trace}(U^*UAA^*) = \text{trace}(AA^*) = ||A||_F^2$$

Also,

$$||AV||_F^2 \equiv \text{trace}(AVV^*A^*) = \text{trace}(AA^*) = ||A||_F^2\,.$$

It follows

$$||\Sigma||_F^2 = ||U^*AV||_F^2 = ||AV||_F^2 = ||A||_F^2\,. \ \blacksquare$$

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

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix},\ U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},$$

$$||A - A'||_F^2 = \left|\left| U \begin{pmatrix} \sigma - \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^* \right|\right|_F^2 = \sum_{k=l+1}^{r} \sigma_k^2$$

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. Suppose $\tilde{B}$ approximates $A = \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix}$ as well as possible out of all matrices $\tilde{B}$ having rank no more than $l < r$ the size of the matrix $\sigma_{r\times r}$. Suppose the rank of $\tilde{B}$ is l. Then obviously no column $\mathbf{x}_j$ of $\tilde{B}$ in a basis for the column space can have $j > r$ since if so, the approximation of A could be improved by simply making this column into a zero column. Therefore there are $\begin{pmatrix} r \\ l \end{pmatrix}$ choices for columns for a basis for the column space of $\tilde{B}$. Suppose you pick the first l for instance. Thus the first column of $\tilde{B}$ should be $\sigma_1\mathbf{e}_1$ to make the approximation up to the first column as good as possible. Now consider approximating as well as possible up to the first two columns. Clearly the second column should be $\sigma_2\mathbf{e}_2$ and in this way, the approximation up to the first two columns is exact. Continue this way till the l^{th} column. Then since $\tilde{B}$ has rank l, all other columns should be zero columns since you cannot have a nonzero entry in any diagonal position and keep the rank of $\tilde{B}$ only l. Then since it is desired to get the best approximation of A you wouldn't want any off diagonal nonzero terms either. The square of the error in doing this, picking the first l columns as a basis would be $\sum_{j=l+1}^{r} \sigma_j^2$. On the other hand, if you picked other columns than the first l in the basis for the column space of $\tilde{B}$, you would have a larger error because you would include sums involving the larger singular values. Thus letting σ' denote the $l \times l$ upper left corner of σ, $\tilde{B}$ should be of the form $\begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}$. For example,

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

is best approximated by the rank 2 matrix

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

Now suppose A is an $m \times n$ matrix. Let U, V be unitary and of the right size such that

$$U^* A V = \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix}$$

Then suppose B approximates A as well as possible in the Frobenius norm. Then you would want

$$\|A - B\| = \|U^* A V - U^* B V\| = \left\| \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix} - U^* B V \right\|$$

to be as small as possible. Therefore, from the above discussion, you should have

$$U^* B V = \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}, B = U \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^*$$

whereas

$$A = U \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix} V^*$$

12.11 Least Squares and Singular Value Decomposition

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 11.5.1 shows that there is a solution to this problem which can be found by solving the system $A^* A \mathbf{x} = A^* \mathbf{y}$. Each $\mathbf{x}$ which solves this system solves the minimization problem as was shown in the lemma just mentioned. Now consider this equation for the solutions of the minimization problem in terms of the singular value decomposition.

$$\overbrace{V \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*}^{A^*} \overbrace{U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*}^{A} \mathbf{x} = \overbrace{V \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*}^{A^*} \mathbf{y}.$$

Therefore, this yields the following upon using block multiplication and multiplying on the left by V^*.

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^* \mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^* \mathbf{y}. \tag{12.17}$$

One solution to this equation which is very easy to spot is

$$\mathbf{x} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \mathbf{y}. \tag{12.18}$$

12.12 The Moore Penrose Inverse

The particular solution of the least squares problem given in 12.18 is important enough that it motivates the following definition.

Definition 12.12.1 *Let A be an $m \times n$ matrix. Then the Moore Penrose inverse of A, denoted by A^+ is defined as*

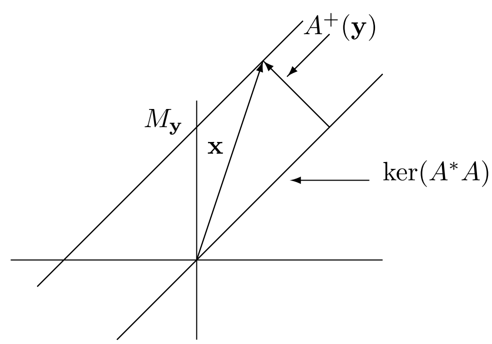
$$A^+ \equiv V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*.$$

Here

$$U^* A V = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

as above.

Thus $A^+\mathbf{y}$ is a solution to the minimization problem to find $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{y}|$. In fact, one can say more about this. In the following picture $M_{\mathbf{y}}$ denotes the set of least squares solutions $\mathbf{x}$ such that $A^*A\mathbf{x} = A^*\mathbf{y}$.



Then $A^+(\mathbf{y})$ is as given in the picture.

Proposition 12.12.2 *$A^+\mathbf{y}$ is the solution to the problem of minimizing $|A\mathbf{x} - \mathbf{y}|$ for all $\mathbf{x}$ which has smallest norm. Thus*

$$\left|AA^+\mathbf{y} - \mathbf{y}\right| \leq |A\mathbf{x} - \mathbf{y}| \text{ for all } \mathbf{x}$$

and if $\mathbf{x}_1$ satisfies $|A\mathbf{x}_1 - \mathbf{y}| \leq |A\mathbf{x} - \mathbf{y}|$ for all $\mathbf{x}$, then $|A^+\mathbf{y}| \leq |\mathbf{x}_1|$.

Proof: Consider $\mathbf{x}$ satisfying 12.17, equivalently $A^*A\mathbf{x} = A^*\mathbf{y}$,

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^*\mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

which has smallest norm. This is equivalent to making $|V^*\mathbf{x}|$ as small as possible because V^* is unitary and so it preserves norms. For $\mathbf{z}$ a vector, denote by $(\mathbf{z})_k$ the vector in $\mathbb{F}^k$ which consists of the first k entries of $\mathbf{z}$. Then if $\mathbf{x}$ is a solution to 12.17

$$\begin{pmatrix} \sigma^2 (V^*\mathbf{x})_k \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sigma (U^*\mathbf{y})_k \\ \mathbf{0} \end{pmatrix}$$

and so $(V^*\mathbf{x})_k = \sigma^{-1}(U^*\mathbf{y})_k$. Thus the first k entries of $V^*\mathbf{x}$ are determined. In order to make $|V^*\mathbf{x}|$ as small as possible, the remaining $n-k$ entries should equal zero. Therefore,

$$V^*\mathbf{x} = \begin{pmatrix} (V^*\mathbf{x})_k \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma^{-1}(U^*\mathbf{y})_k \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

and so

$$\mathbf{x} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y} \equiv A^+\mathbf{y} \blacksquare$$

Lemma 12.12.3 *The matrix A^+ satisfies the following conditions.*

$$AA^+A = A,\ A^+AA^+ = A^+,\ A^+A \text{ and } AA^+ \text{ are Hermitian.} \tag{12.19}$$

Proof: This is routine. Recall

$$A = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*$$

and

$$A^+ = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$$

so you just plug in and verify it works. $\blacksquare$

A much more interesting observation is that A^+ is characterized as being the unique matrix which satisfies 12.19. This is the content of the following Theorem. The conditions are sometimes called the Penrose conditions.

Theorem 12.12.4 *Let A be an $m \times n$ matrix. Then a matrix A_0, is the Moore Penrose inverse of A if and only if A_0 satisfies*

$$AA_0A = A,\ A_0AA_0 = A_0,\ A_0A \text{ and } AA_0 \text{ are Hermitian.} \tag{12.20}$$

Proof: From the above lemma, the Moore Penrose inverse satisfies 12.20. Suppose then that A_0 satisfies 12.20. It is necessary to verify that $A_0 = A^+$. Recall that from the singular value decomposition, there exist unitary matrices, U and V such that

$$U^*AV = \Sigma \equiv \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},\ A = U\Sigma V^*.$$

Recall that

$$A^+ = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$$

Let

$$A_0 = V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^* \tag{12.21}$$

where P is $r \times r$, the same size as the diagonal matrix composed of the singular values on the main diagonal.

Next use the first equation of 12.20 to write

$$\overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} = \overbrace{U\Sigma V^*}^{A}.$$

Then multiplying both sides on the left by U^* and on the right by V,

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma P \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \tag{12.22}$$

Therefore, $P = \sigma^{-1}$. From the requirement that AA_0 is Hermitian,

$$\overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*$$

must be Hermitian. Therefore, it is necessary that

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} \sigma P & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix}$$

is Hermitian. Then

$$\begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ Q^* \sigma & 0 \end{pmatrix}$$

and so $Q = 0$.

Next,

$$\overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} = V \begin{pmatrix} P\sigma & 0 \\ R\sigma & 0 \end{pmatrix} V^* = V \begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix} V^*$$

is Hermitian. Therefore, also

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}$$

is Hermitian. Thus $R = 0$ because

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}^* = \begin{pmatrix} I & \sigma^* R^* \\ 0 & 0 \end{pmatrix}$$

which requires $R\sigma = 0$. Now multiply on right by σ^{-1} to find that $R = 0$.

Use 12.21 and the second equation of 12.20 to write

$$\overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} = \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0}.$$

which implies

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}.$$

This yields from the above in which is was shown that R, Q are both 0

$$\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{12.23}$$

$$= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}. \tag{12.24}$$

Therefore, $S = 0$ also and so

$$V^*A_0U \equiv \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

which says

$$A_0 = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \equiv A^+. \blacksquare$$

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in 12.20. It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 3.

12.13 Exercises

1. Show $(A^*)^* = A$ and $(AB)^* = B^*A^*$.

2. Prove Corollary 12.3.8.

3. Show that if A is an $n \times n$ matrix which has an inverse then $A^+ = A^{-1}$.

4. Using the singular value decomposition, show that for any square matrix A, it follows that A^*A is unitarily similar to AA^*.

5. 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, $\text{trace}(M) \equiv \sum_{i=1}^n M_{ii}$. The resulting norm, $||\cdot||_F$ is called the Frobenius norm and it can be used to measure the distance between two matrices.

6. 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.

7. 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.

8. 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 7 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 6.1.7 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 $\lim_{k\to\infty} p_k(A_k) = p(A)$.

9. Prove that Theorem 12.4.4 and Corollary 12.4.5 can be strengthened so that the condition on the A_k is necessary as well as sufficient. **Hint:** Consider vectors of the form $\begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{x} \in \mathbb{F}^k$.

10. Show directly that if A is an $n\times n$ matrix and $A=A^*$ (A is Hermitian) then all the eigenvalues are real and eigenvectors can be assumed to be real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).

11. Let $\mathbf{v}_1,\cdots,\mathbf{v}_n$ be an orthonormal basis for $\mathbb{F}^n$. Let Q be a matrix whose i^{th} column is $\mathbf{v}_i$. Show
$$Q^*Q=QQ^*=I.$$

12. Show that an $n\times n$ matrix Q is unitary if and only if it preserves distances. This means $|Q\mathbf{v}|=|\mathbf{v}|$. This was done in the text but you should try to do it for yourself.

13. Suppose $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ and $\{\mathbf{w}_1,\cdots,\mathbf{w}_n\}$ are two orthonormal bases for $\mathbb{F}^n$ and suppose Q is an $n\times n$ matrix satisfying $Q\mathbf{v}_i=\mathbf{w}_i$. Then show Q is unitary. If $|\mathbf{v}|=1$, show there is a unitary transformation which maps $\mathbf{v}$ to $\mathbf{e}_1$.

14. Finish the proof of Theorem 12.7.5.

15. Let A be a Hermitian matrix so $A=A^*$ and suppose all eigenvalues of A are larger than δ^2. Show
$$(A\mathbf{v},\mathbf{v})\geq\delta^2|\mathbf{v}|^2$$
Where here, the inner product is $(\mathbf{v},\mathbf{u})\equiv\sum_{j=1}^n v_j\overline{u_j}$.

16. The discrete Fourier transform maps $\mathbb{C}^n\to\mathbb{C}^n$ as follows.
$$F(\mathbf{x})=\mathbf{z}\text{ where }z_k=\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}e^{-i\frac{2\pi}{n}jk}x_j.$$
Show that F^{-1} exists and is given by the formula
$$F^{-1}(\mathbf{z})=\mathbf{x}\text{ where }x_j=\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}e^{i\frac{2\pi}{n}jk}z_k$$
Here is one way to approach this problem. Note $\mathbf{z}=U\mathbf{x}$ where
$$U=\frac{1}{\sqrt{n}}\begin{pmatrix} e^{-i\frac{2\pi}{n}0\cdot 0} & e^{-i\frac{2\pi}{n}1\cdot 0} & e^{-i\frac{2\pi}{n}2\cdot 0} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 0}\\ e^{-i\frac{2\pi}{n}0\cdot 1} & e^{-i\frac{2\pi}{n}1\cdot 1} & e^{-i\frac{2\pi}{n}2\cdot 1} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 1}\\ e^{-i\frac{2\pi}{n}0\cdot 2} & e^{-i\frac{2\pi}{n}1\cdot 2} & e^{-i\frac{2\pi}{n}2\cdot 2} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 2}\\ \vdots & \vdots & \vdots & & \vdots\\ e^{-i\frac{2\pi}{n}0\cdot(n-1)} & e^{-i\frac{2\pi}{n}1\cdot(n-1)} & e^{-i\frac{2\pi}{n}2\cdot(n-1)} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot(n-1)}\end{pmatrix}$$
Now argue U is unitary and use this to establish the result. To show this verify each row has length 1 and the inner product of two different rows gives 0. Now $U_{kj}=e^{-i\frac{2\pi}{n}jk}$ and so $(U^*)_{kj}=e^{i\frac{2\pi}{n}jk}$.

17. Let f be a periodic function having period 2π. The Fourier series of f is an expression of the form
$$\sum_{k=-\infty}^{\infty}c_ke^{ikx}\equiv\lim_{n\to\infty}\sum_{k=-n}^{n}c_ke^{ikx}$$
and the idea is to find c_k such that the above sequence converges in some way to f. If
$$f(x)=\sum_{k=-\infty}^{\infty}c_ke^{ikx}$$

and you formally multiply both sides by e^{-imx} and then integrate from 0 to 2π, interchanging the integral with the sum without any concern for whether this makes sense, show it is reasonable from this to expect

$$c_m = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-imx} dx.$$

Now suppose you only know $f(x)$ at equally spaced points $2\pi j/n$ for $j = 0, 1, \cdots, n$. Consider the Riemann sum for this integral obtained from using the left endpoint of the subintervals determined from the partition $\left\{ \frac{2\pi}{n} j \right\}_{j=0}^{n}$. How does this compare with the discrete Fourier transform? What happens as $n \to \infty$ to this approximation?

18. Suppose A is a real 3×3 orthogonal matrix (Recall this means $AA^T = A^T A = I$.) having determinant 1. Show it must have an eigenvalue equal to 1. Note this shows there exists a vector $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{x}$. **Hint:** Show first or recall that any orthogonal matrix must preserve lengths. That is, $|A\mathbf{x}| = |\mathbf{x}|$.

19. Let A be a complex $m \times n$ matrix. Using the description of the Moore Penrose inverse in terms of the singular value decomposition, show that

$$\lim_{\delta \to 0+} (A^* A + \delta I)^{-1} A^* = A^+$$

where the convergence happens in the Frobenius norm. Also verify, using the singular value decomposition, that the inverse exists in the above formula. Observe that this shows that the Moore Penrose inverse is unique.

20. Show that $A^+ = (A^* A)^+ A^*$. **Hint:** You might use the description of A^+ in terms of the singular value decomposition.

21. In Theorem 12.6.1. Show that every matrix which commutes with A also commutes with $A^{1/k}$ the unique nonnegative self adjoint k^{th} root.

22. Let X be a finite dimensional inner product space and let $\beta = \{u_1, \cdots, u_n\}$ be an orthonormal basis for X. Let $A \in \mathcal{L}(X, X)$ be self adjoint and nonnegative and let M be its matrix with respect to the given orthonormal basis. Show that M is nonnegative, self adjoint also. Use this to show that A has a unique nonnegative self adjoint k^{th} root.

23. Let A be a complex $m \times n$ matrix having singular value decomposition $U^* AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$ as explained above, where σ is $k \times k$. Show that

$$\ker(A) = \text{span}(V\mathbf{e}_{k+1}, \cdots, V\mathbf{e}_n),$$

the last $n - k$ columns of V.

24. The principal submatrices of an $n \times n$ matrix A are A_k where A_k consists those entries which are in the first k rows and first k columns of A. Suppose A is a real symmetric matrix and that $\mathbf{x} \to \langle A\mathbf{x}, \mathbf{x} \rangle$ is positive definite. This means that if $\mathbf{x} \neq \mathbf{0}$, then $\langle A\mathbf{x}, \mathbf{x} \rangle > 0$. Show that each of the principal submatrices are positive definite. **Hint:** Consider $\begin{pmatrix} \mathbf{x}^T & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{x}$ consists of k entries.

25. $\uparrow$Show that if A is a symmetric positive definite $n \times n$ real matrix, then A has an LU factorization with the property that each entry on the main diagonal in U is positive. **Hint:** This is pretty clear if A is 1×1. Assume true for $(n-1) \times (n-1)$. Then

$$A = \begin{pmatrix} \hat{A} & \mathbf{a} \\ \mathbf{a}^T & a_{nn} \end{pmatrix}$$

Then as above, $\hat{A}$ is positive definite. Thus it has an LU factorization with all positive entries on the diagonal of U. Notice that, using block multiplication,

$$A = \begin{pmatrix} LU & \mathbf{a} \\ \mathbf{a}^T & a_{nn} \end{pmatrix} = \begin{pmatrix} L & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} U & L^{-1}\mathbf{a} \\ \mathbf{a}^T & a_{nn} \end{pmatrix}$$

Now consider that matrix on the right. Argue that it is of the form $\tilde{L}\tilde{U}$ where $\tilde{U}$ has all positive diagonal entries except possibly for the one in the n^{th} row and n^{th} column. Now explain why $\det(A) > 0$ and argue that in fact all diagonal entries of $\tilde{U}$ are positive.

26. $\uparrow$Let A be a real symmetric $n \times n$ matrix and $A = LU$ where L has all ones down the diagonal and U has all positive entries down the main diagonal. Show that $A = LDH$ where L is lower triangular and H is upper triangular, each having all ones down the diagonal and D a diagonal matrix having all positive entries down the main diagonal. In fact, these are the diagonal entries of U.

27. $\uparrow$Show that if L, L_1 are lower triangular with ones down the main diagonal and H, H_1 are upper triangular with all ones down the main diagonal and D, D_1 are diagonal matrices having all positive diagonal entries, and if $LDH = L_1 D_1 H_1$, then $L = L_1, H = H_1, D = D_1$. **Hint:** Explain why $D_1^{-1} L_1^{-1} L D = H_1 H^{-1}$. Then explain why the right side is upper triangular and the left side is lower triangular. Conclude these are both diagonal matrices. However, there are all ones down the diagonal in the expression on the right. Hence $H = H_1$. Do something similar to conclude that $L = L_1$ and then that $D = D_1$.

28. $\uparrow$Show that if A is a symmetric real matrix such that $\mathbf{x} \to \langle A\mathbf{x}, \mathbf{x}\rangle$ is positive definite, then there exists a lower triangular matrix L having all positive entries down the diagonal such that $A = LL^T$. **Hint:** From the above, $A = LDH$ where L, H are respectively lower and upper triangular having all ones down the diagonal and D is a diagonal matrix having all positive entries. Then argue from the above problem and symmetry of A that $H = L^T$. Now modify L by making it equal to $LD^{1/2}$. This is called the Cholesky factorization.

29. Given $F \in \mathcal{L}(X, Y)$ where X, Y are inner product spaces and $\dim(X) = n \leq m = \dim(Y)$, there exists R, U such that U is nonnegative and Hermitian and $R^*R = I$ such that $F = RU$. Show that U is actually unique and that R is determined on $U(X)$.

Chapter 13

Norms

In this chapter, X and Y are finite dimensional vector spaces which have a norm. The following is a definition.

Definition 13.0.1 *A linear space X is a normed linear space if there is a norm defined on X, $||\cdot||$ satisfying*

$$||\mathbf{x}|| \geq 0, \ \ ||\mathbf{x}|| = 0 \text{ if and only if } \mathbf{x} = 0,$$

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

$$||c\mathbf{x}|| = |c|\,||\mathbf{x}||$$

whenever c is a scalar. A set, $U \subseteq X$, a normed linear space is open if for every $p \in U$, there exists $\delta > 0$ such that

$$B(p, \delta) \equiv \{x : ||x - p|| < \delta\} \subseteq U.$$

Thus, a set is open if every point of the set is an interior point. Also, $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$ *means* $\lim_{n\to\infty} \|\mathbf{x}_n - \mathbf{x}\| = 0$. *This is written sometimes as $\mathbf{x}_n \to \mathbf{x}$.*

Note first that

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

so

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|\,.$$

Similarly

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

and so

$$|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\|\,. \tag{13.1}$$

To begin with recall the Cauchy Schwarz inequality which is stated here for convenience in terms of the inner product space, $\mathbb{C}^n$.

Theorem 13.0.2 *The following inequality holds for a_i and $b_i \in \mathbb{C}$.*

$$\left|\sum_{i=1}^{n} a_i \overline{b}_i\right| \leq \left(\sum_{i=1}^{n} |a_i|^2\right)^{1/2} \left(\sum_{i=1}^{n} |b_i|^2\right)^{1/2}. \tag{13.2}$$

Let X be a finite dimensional normed linear space with norm $||\cdot||$ where the field of scalars is denoted by $\mathbb{F}$ and is understood to be either $\mathbb{R}$ or $\mathbb{C}$. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X. If $\mathbf{x} \in X$, denote by x_i the i^{th} component of $\mathbf{x}$ with respect to this basis. Thus

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{v}_i.$$

Definition 13.0.3 *For $\mathbf{x} \in X$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ a basis, define a new norm by*

$$|\mathbf{x}| \equiv \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}.$$

where

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{v}_i.$$

Similarly, for $\mathbf{y} \in Y$ *with basis* $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$*, and* y_i *its components with respect to this basis,*

$$|\mathbf{y}| \equiv \left(\sum_{i=1}^{m} |y_i|^2\right)^{1/2}$$

For $A \in \mathcal{L}(X, Y)$*, the space of linear mappings from* X *to* Y*,*

$$||A|| \equiv \sup\{|A\mathbf{x}| : |\mathbf{x}| \leq 1\}. \tag{13.3}$$

The first thing to show is that the two norms, $||\cdot||$ and $|\cdot|$, are equivalent. This means the conclusion of the following theorem holds.

Theorem 13.0.4 *Let* $(X, ||\cdot||)$ *be a finite dimensional normed linear space and let* $|\cdot|$ *be described above relative to a given basis,* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$*. Then* $|\cdot|$ *is a norm and there exist constants* $\delta, \Delta > 0$ *independent of* $\mathbf{x}$ *such that*

$$\delta\, ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta\, ||\mathbf{x}||\,. \tag{13.4}$$

Proof: All of the above properties of a norm are obvious except the second, the triangle inequality. To establish this inequality, use the Cauchy Schwarz inequality to write

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

and this proves the second property above.

It remains to show the equivalence of the two norms. By the Cauchy Schwarz inequality again,

$$\begin{aligned}
||\mathbf{x}|| &\equiv \left|\left|\sum_{i=1}^{n} x_i \mathbf{v}_i\right|\right| \leq \sum_{i=1}^{n} |x_i|\, ||\mathbf{v}_i|| \leq |\mathbf{x}| \left(\sum_{i=1}^{n} ||\mathbf{v}_i||^2\right)^{1/2} \\
&\equiv \delta^{-1}\,|\mathbf{x}|\,.
\end{aligned}$$

This proves the first half of the inequality.

Suppose the second half of the inequality is not valid. Then there exists a sequence $\mathbf{x}^k \in X$ such that

$$\left|\mathbf{x}^k\right| > k\left|\left|\mathbf{x}^k\right|\right|,\ k = 1, 2, \cdots.$$

Then define

$$\mathbf{y}^k \equiv \frac{\mathbf{x}^k}{|\mathbf{x}^k|}.$$

It follows

$$\left|\mathbf{y}^k\right| = 1, \quad \left|\mathbf{y}^k\right| > k\left|\left|\mathbf{y}^k\right|\right|. \tag{13.5}$$

Letting y_i^k be the components of $\mathbf{y}^k$ with respect to the given basis, it follows the vector

$$\left(y_1^k, \cdots, y_n^k\right)$$

is a unit vector in $\mathbb{F}^n$. By the Heine Borel theorem, there exists a subsequence, still denoted by k such that

$$\left(y_1^k, \cdots, y_n^k\right) \to \left(y_1, \cdots, y_n\right).$$

It follows from 13.5 and this that for

$$\mathbf{y} = \sum_{i=1}^n y_i \mathbf{v}_i,$$

$$0 = \lim_{k\to\infty} \left|\left|\mathbf{y}^k\right|\right| = \lim_{k\to\infty} \left|\left|\sum_{i=1}^n y_i^k \mathbf{v}_i\right|\right| = \left|\left|\sum_{i=1}^n y_i \mathbf{v}_i\right|\right|$$

but not all the y_i equal zero. The last equation follows easily from 13.1 and

$$\left|\left|\left|\sum_{i=1}^n y_i^k \mathbf{v}_i\right|\right| - \left|\left|\sum_{i=1}^n y_i \mathbf{v}_i\right|\right|\right| \leq \left|\left|\sum_{i=1}^n \left(y_i^k - y_i\right) \mathbf{v}_i\right|\right| \leq \sum_{i=1}^n \left|y_i^k - y_i\right| \|\mathbf{v}_i\|$$

This contradicts the assumption that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis and proves the second half of the inequality. ■

Definition 13.0.5 *Let $(X, ||\cdot||)$ be a normed linear space and let $\{x_n\}_{n=1}^\infty$ be a sequence of vectors. Then this is called a Cauchy sequence if for all $\varepsilon > 0$ there exists N such that if $m, n \geq N$, then*

$$||x_n - x_m|| < \varepsilon.$$

This is written more briefly as

$$\lim_{m,n\to\infty} ||x_n - x_m|| = 0.$$

Definition 13.0.6 *A normed linear space, $(X, ||\cdot||)$ is called a Banach space if it is complete. This means that, whenever, $\{\mathbf{x}_n\}$ is a Cauchy sequence there exists a unique $\mathbf{x} \in X$ such that $\lim_{n\to\infty} ||\mathbf{x} - \mathbf{x}_n|| = 0$.*

Corollary 13.0.7 *If $(X, ||\cdot||)$ is a finite dimensional normed linear space with the field of scalars $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, then $(X, ||\cdot||)$ is a Banach space.*

Proof: Let $\{\mathbf{x}^k\}$ be a Cauchy sequence. Then letting the components of $\mathbf{x}^k$ with respect to the given basis be

$$x_1^k, \cdots, x_n^k,$$

it follows from Theorem 13.0.4, that

$$\left(x_1^k, \cdots, x_n^k\right)$$

is a Cauchy sequence in $\mathbb{F}^n$ and so

$$\left(x_1^k, \cdots, x_n^k\right) \to \left(x_1, \cdots, x_n\right) \in \mathbb{F}^n.$$

Thus, letting $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{v}_i$, it follows from the equivalence of the two norms shown above that

$$\lim_{k\to\infty} \left|\mathbf{x}^k - \mathbf{x}\right| = \lim_{k\to\infty} \left|\left|\mathbf{x}^k - \mathbf{x}\right|\right| = 0. \blacksquare$$

Corollary 13.0.8 *Suppose X is a finite dimensional linear space with the field of scalars either $\mathbb{C}$ or $\mathbb{R}$ and $||\cdot||$ and $|||\cdot|||$ are two norms on X. Then there exist positive constants, δ and Δ, independent of $\mathbf{x} \in X$ such that*

$$\delta\,|||\mathbf{x}||| \leq ||\mathbf{x}|| \leq \Delta\,|||\mathbf{x}|||\,.$$

Thus any two norms are equivalent.

This is very important because it shows that all questions of convergence can be considered relative to any norm with the same outcome.

Proof: Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X and let $|\cdot|$ be the norm taken with respect to this basis which was described earlier. Then by Theorem 13.0.4, there are positive constants $\delta_1, \Delta_1, \delta_2, \Delta_2$, all independent of $\mathbf{x} \in X$ such that

$$\delta_2\,|||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_2\,|||\mathbf{x}|||\,,\ \ \delta_1\,||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta_1\,||\mathbf{x}||\,.$$

Then

$$\delta_2\,|||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_1\,||\mathbf{x}|| \leq \frac{\Delta_1}{\delta_1}\,|\mathbf{x}| \leq \frac{\Delta_1\Delta_2}{\delta_1}\,|||\mathbf{x}|||$$

and so

$$\frac{\delta_2}{\Delta_1}\,|||\mathbf{x}||| \leq ||\mathbf{x}|| \leq \frac{\Delta_2}{\delta_1}\,|||\mathbf{x}||| \ \ \blacksquare$$

Definition 13.0.9 *Let X and Y be normed linear spaces with norms $||\cdot||_X$ and $||\cdot||_Y$ respectively. Then $\mathcal{L}(X,Y)$ denotes the space of linear transformations, called bounded linear transformations, mapping X to Y which have the property that*

$$||A|| \equiv \sup\{||Ax||_Y : ||x||_X \leq 1\} < \infty.$$

Then $||A||$ is referred to as the operator norm of the bounded linear transformation A.

It is an easy exercise to verify that $||\cdot||$ is a norm on $\mathcal{L}(X,Y)$ and it is always the case that

$$||Ax||_Y \leq ||A||\,||x||_X\,.$$

Furthermore, you should verify that you can replace ≤ 1 with $= 1$ in the definition. Thus

$$||A|| \equiv \sup\{||Ax||_Y : ||x||_X = 1\}\,.$$

Theorem 13.0.10 *Let X and Y be finite dimensional normed linear spaces of dimension n and m respectively and denote by $||\cdot||$ the norm on either X or Y. Then if A is any linear function mapping X to Y, then $A \in \mathcal{L}(X,Y)$ and $(\mathcal{L}(X,Y), ||\cdot||)$ is a complete normed linear space of dimension nm with*

$$||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}||\,.$$

Also if $A \in \mathcal{L}(X,Y)$ and $B \in \mathcal{L}(Y,Z)$ where X, Y, Z are normed linear spaces,

$$\|BA\| \leq \|B\|\,\|A\|$$

Proof: It is necessary to show the norm defined on linear transformations really is a norm. Again the first and third properties listed above for norms are obvious. It remains to show the second and verify $||A|| < \infty$. Letting $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis and $|\cdot|$ defined with respect to this basis as above, there exist constants $\delta, \Delta > 0$ such that

$$\delta ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta ||\mathbf{x}|| .$$

Then,

$$||A + B|| \equiv \sup\{||(A + B)(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\}$$
$$\leq \sup\{||A\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} + \sup\{||B\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \equiv ||A|| + ||B|| .$$

Next consider the claim that $||A|| < \infty$. This follows from

$$||A(\mathbf{x})|| = \left|\left|A\left(\sum_{i=1}^n x_i \mathbf{v}_i\right)\right|\right| \leq \sum_{i=1}^n |x_i| \, ||A(\mathbf{v}_i)||$$

$$\leq |\mathbf{x}| \left(\sum_{i=1}^n ||A(\mathbf{v}_i)||^2\right)^{1/2} \leq \Delta ||\mathbf{x}|| \left(\sum_{i=1}^n ||A(\mathbf{v}_i)||^2\right)^{1/2} < \infty.$$

Thus $||A|| \leq \Delta \left(\sum_{i=1}^n ||A(\mathbf{v}_i)||^2\right)^{1/2}$.

Next consider the assertion about the dimension of $\mathcal{L}(X, Y)$. It follows from Theorem 8.2.3. By Corollary 13.0.7 $(\mathcal{L}(X, Y), ||\cdot||)$ is complete. If $\mathbf{x} \neq \mathbf{0}$,

$$||A\mathbf{x}|| \frac{1}{||\mathbf{x}||} = \left|\left|A \frac{\mathbf{x}}{||\mathbf{x}||}\right|\right| \leq ||A||$$

Consider the last claim.

$$\|BA\| \equiv \sup_{\|x\| \leq 1} \|B(A(x))\| \leq \|B\| \sup_{\|x\| \leq 1} \|Ax\| = \|B\| \, \|A\| \quad \blacksquare$$

Note by Corollary 13.0.8 you can define a norm any way desired on any finite dimensional linear space which has the field of scalars $\mathbb{R}$ or $\mathbb{C}$ and any other way of defining a norm on this space yields an equivalent norm. Thus, it doesn't much matter as far as notions of convergence are concerned which norm is used for a finite dimensional space. In particular in the space of $m \times n$ matrices, you can use the operator norm defined above, or some other way of giving this space a norm. A popular choice for a norm is the Frobenius norm discussed earlier but reviewed here.

Definition 13.0.11 *Make the space of $m \times n$ matrices into a inner product space by defining*

$$(A, B) \equiv \operatorname{trace}(AB^*) .$$

Another way of describing a norm for an $n \times n$ matrix is as follows.

Definition 13.0.12 *Let A be an $m \times n$ matrix. Define the spectral norm of A, written as $||A||_2$ to be*

$$\max\left\{\lambda^{1/2} : \lambda \text{ is an eigenvalue of } A^*A\right\}.$$

That is, the largest singular value of A. (Note the eigenvalues of A^*A are all positive because if $A^*A\mathbf{x} = \lambda\mathbf{x}$, then

$$\lambda |\mathbf{x}|^2 = \lambda(\mathbf{x}, \mathbf{x}) = (A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \geq 0.)$$

Actually, this is nothing new. It turns out that $||\cdot||_2$ is nothing more than the operator norm for A taken with respect to the usual Euclidean norm,

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

Proposition 13.0.13 *The following holds.*

$$||A||_2 = \sup\{|A\mathbf{x}| : |\mathbf{x}| = 1\} \equiv ||A||.$$

Proof: Note that A^*A is Hermitian and so by Corollary 12.3.4,

$$\begin{aligned} ||A||_2 &= \max\left\{(A^*A\mathbf{x},\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} = \max\left\{(A\mathbf{x},A\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &= \max\{|A\mathbf{x}| : |\mathbf{x}| = 1\} = ||A||. \blacksquare \end{aligned}$$

Here is another proof of this proposition. Recall there are unitary matrices of the right size U, V such that $A = U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*$ where the matrix on the inside is as described in the section on the singular value decomposition. Then since unitary matrices preserve norms,

$$\begin{aligned} ||A|| &= \sup_{|\mathbf{x}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*\mathbf{x}\right| = \sup_{|V^*\mathbf{x}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*\mathbf{x}\right| \\ &= \sup_{|\mathbf{y}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\mathbf{y}\right| = \sup_{|\mathbf{y}|\leq 1}\left|\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\mathbf{y}\right| = \sigma_1 \equiv ||A||_2 \end{aligned}$$

This completes the alternate proof.

From now on, $||A||_2$ will mean either the operator norm of A taken with respect to the usual Euclidean norm or the largest singular value of A, whichever is most convenient.

An interesting application of the notion of equivalent norms on $\mathbb{R}^n$ is the process of giving a norm on a finite Cartesian product of normed linear spaces.

Definition 13.0.14 *Let X_i, $i = 1, \cdots, n$ be normed linear spaces with norms, $||\cdot||_i$. For*

$$\mathbf{x} \equiv (x_1, \cdots, x_n) \in \prod_{i=1}^{n} X_i$$

define $\theta : \prod_{i=1}^{n} X_i \to \mathbb{R}^n$ by

$$\theta(\mathbf{x}) \equiv (||x_1||_1, \cdots, ||x_n||_n)$$

Then if $||\cdot||$ is any norm on $\mathbb{R}^n$, define a norm on $\prod_{i=1}^{n} X_i$, also denoted by $||\cdot||$ by

$$||\mathbf{x}|| \equiv ||\theta\mathbf{x}||.$$

The following theorem follows immediately from Corollary 13.0.8.

Theorem 13.0.15 *Let X_i and $||\cdot||_i$ be given in the above definition and consider the norms on $\prod_{i=1}^{n} X_i$ described there in terms of norms on $\mathbb{R}^n$. Then any two of these norms on $\prod_{i=1}^{n} X_i$ obtained in this way are equivalent.*

For example, define

$$||\mathbf{x}||_1 \equiv \sum_{i=1}^{n} |x_i|,$$

$$||\mathbf{x}||_\infty \equiv \max\{|x_i|, i = 1, \cdots, n\},$$

or

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

and all three are equivalent norms on $\prod_{i=1}^{n} X_i$.

13.1 The p Norms

In addition to $||\cdot||_1$ and $||\cdot||_\infty$ mentioned above, it is common to consider the so called p norms for $\mathbf{x} \in \mathbb{C}^n$.

Definition 13.1.1 *Let $\mathbf{x} \in \mathbb{C}^n$. Then define for $p \geq 1$,*

$$||\mathbf{x}||_p \equiv \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

The following inequality is called Holder's inequality.

Proposition 13.1.2 *For $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,*

$$\sum_{i=1}^{n} |x_i| |y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}$$

The proof will depend on the following lemma.

Lemma 13.1.3 *If $a, b \geq 0$ and p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

Proof of the Proposition: If $\mathbf{x}$ or $\mathbf{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 13.1.3,

$$\sum_{i=1}^{n} \frac{|x_i|}{A} \frac{|y_i|}{B} \leq \sum_{i=1}^{n} \left[\frac{1}{p}\left(\frac{|x_i|}{A}\right)^p + \frac{1}{p'}\left(\frac{|y_i|}{B}\right)^{p'}\right]$$

$$= \frac{1}{p}\frac{1}{A^p}\sum_{i=1}^{n} |x_i|^p + \frac{1}{p'}\frac{1}{B^p}\sum_{i=1}^{n} |y_i|^{p'} = \frac{1}{p} + \frac{1}{p'} = 1$$

and so

$$\sum_{i=1}^{n} |x_i| |y_i| \leq AB = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}. \blacksquare$$

Theorem 13.1.4 *The p norms do indeed satisfy the axioms of a norm.*

Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $\frac{p}{p'} = p - 1$. Then using the Holder inequality,

$$\begin{aligned}
||\mathbf{x}+\mathbf{y}||^p &= \sum_{i=1}^n |x_i + y_i|^p \\
&\leq \sum_{i=1}^n |x_i + y_i|^{p-1} |x_i| + \sum_{i=1}^n |x_i + y_i|^{p-1} |y_i| \\
&= \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}} |x_i| + \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}} |y_i| \\
&\leq \left(\sum_{i=1}^n |x_i + y_i|^p \right)^{1/p'} \left[\left(\sum_{i=1}^n |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p \right)^{1/p} \right] \\
&= ||\mathbf{x}+\mathbf{y}||^{p/p'} \left(||\mathbf{x}||_p + ||\mathbf{y}||_p \right)
\end{aligned}$$

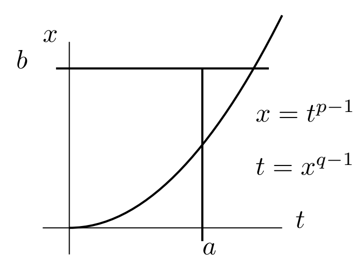
so dividing by $||\mathbf{x}+\mathbf{y}||^{p/p'}$, it follows

$$||\mathbf{x}+\mathbf{y}||^p \, ||\mathbf{x}+\mathbf{y}||^{-p/p'} = ||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$$

$\left(p - \frac{p}{p'} = p \left(1 - \frac{1}{p'} \right) = p \frac{1}{p} = 1. \right)$. ■

It only remains to prove Lemma 13.1.3.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1} dt + \int_0^b x^{q-1} dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$.

Alternate proof of the lemma: For $a, b \geq 0$, let b be fixed and

$$f(a) \equiv \frac{1}{p} a^p + \frac{1}{q} b^q - ab, \ t > 0$$

If $b = 0$, it is clear that $f(a) \geq 0$ for all a. Then assume $b > 0$. It is clear since $p > 1$ that $\lim_{a \to \infty} f(a) = \infty$.

$$f'(a) = a^{p-1} - b$$

This is negative for small a and then eventually is positive. Consider the minimum value of f which must occur at $a > 0$ thanks to the observation that the function is initially strictly decreasing. At this point,

$$0 = f'(a) = a^{p-1} - b = a^{(p/q)} - b$$

and so $a^p = b^q$ at the point where this function has a minimum. Thus at this value of a,

$$f(a) = \frac{1}{p}a^p + \frac{1}{q}a^p - aa^{p-1} = a^p - a^p = 0$$

Hence $f(a) \geq 0$ for all $a \geq 0$ and this proves the inequality. Equality occurs when $a^p = b^q$. ■

Now $||A||_p$ may be considered as the operator norm of A taken with respect to $||\cdot||_p$. In the case when $p = 2$, this is just the spectral norm. There is an easy estimate for $||A||_p$ in terms of the entries of A.

Theorem 13.1.5 *The following holds.*

$$||A||_p \leq \left(\sum_k \left(\sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q}$$

Proof: Let $||\mathbf{x}||_p \leq 1$ and let $A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$ where the $\mathbf{a}_k$ are the columns of A. Then

$$A\mathbf{x} = \left(\sum_k x_k \mathbf{a}_k \right)$$

and so by Holder's inequality,

$$||A\mathbf{x}||_p \equiv \left|\left| \sum_k x_k \mathbf{a}_k \right|\right|_p \leq \sum_k |x_k| \, ||\mathbf{a}_k||_p \leq$$

$$\leq \left(\sum_k |x_k|^p \right)^{1/p} \left(\sum_k ||\mathbf{a}_k||_p^q \right)^{1/q} \leq \left(\sum_k \left(\sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q} \blacksquare$$

13.2 The Condition Number

Let $A \in \mathcal{L}(X, X)$ be a linear transformation where X is a finite dimensional vector space and consider the problem $Ax = b$ where it is assumed there is a unique solution to this problem. How does the solution change if A is changed a little bit and if b is changed a little bit? This is clearly an interesting question because you often do not know A and b exactly. If a small change in these quantities results in a large change in the solution, x, then it seems clear this would be undesirable. In what follows $||\cdot||$ when applied to a linear transformation will always refer to the operator norm. Recall the following property of the operator norm in Theorem 13.0.10.

Lemma 13.2.1 *Let $A, B \in \mathcal{L}(X, X)$ where X is a normed vector space as above. Then for $||\cdot||$ denoting the operator norm,*

$$\|AB\| \leq \|A\| \, \|B\| .$$

Lemma 13.2.2 *Let $A, B \in \mathcal{L}(X, X), A^{-1} \in \mathcal{L}(X, X),$ and suppose $\|B\| < 1/\left\|A^{-1}\right\|$. Then $(A + B)^{-1}, \left(I + A^{-1}B\right)^{-1}$ exists and*

$$\left\| \left(I + A^{-1}B\right)^{-1} \right\| \leq \left(1 - \left\|A^{-1}B\right\|\right)^{-1} \tag{13.6}$$

$$\left\|(A+B)^{-1}\right\| \leq \left\|A^{-1}\right\| \left|\frac{1}{1-\left\|A^{-1}B\right\|}\right|. \tag{13.7}$$

The above formula makes sense because $\left|\left|A^{-1}B\right|\right| < 1$.

Proof: By Lemma 13.0.10,

$$\left\|A^{-1}B\right\| \leq \left\|A^{-1}\right\| \|B\| < \left\|A^{-1}\right\| \frac{1}{\left\|A^{-1}\right\|} = 1 \tag{13.8}$$

Then from the triangle inequality,

$$\begin{aligned} \left\|\left(I+A^{-1}B\right)x\right\| &\geq \|x\| - \left\|A^{-1}Bx\right\| \\ &\geq \|x\| - \left\|A^{-1}B\right\| \|x\| = \left(1-\left\|A^{-1}B\right\|\right)\|x\| \end{aligned}$$

It follows that $I + A^{-1}B$ is one to one because from 13.8, $1 - \left|\left|A^{-1}B\right|\right| > 0$. Thus if $\left(I+A^{-1}B\right)x = 0$, then $x = 0$. Thus $I + A^{-1}B$ is also onto, taking a basis to a basis. Then a generic $y \in X$ is of the form $y = \left(I+A^{-1}B\right)x$ and the above shows that

$$\left\|\left(I+A^{-1}B\right)^{-1} y\right\| \leq \left(1-\left|\left|A^{-1}B\right|\right|\right)^{-1} \|y\|$$

which verifies 13.6. Thus $(A+B) = A\left(I+A^{-1}B\right)$ is one to one and this with Lemma 13.0.10 implies 13.7. ■

Proposition 13.2.3 *Suppose A is invertible, $b \neq 0$, $Ax = b$, and $(A+B)x_1 = b_1$ where $||B|| < 1/\left|\left|A^{-1}\right|\right|$. Then*

$$\frac{\|x_1 - x\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\| \|A\|}{1-\left\|A^{-1}B\right\|} \left(\frac{\|b_1 - b\|}{\|b\|} + \frac{\|B\|}{\|A\|}\right)$$

Proof: This follows from the above lemma.

$$\begin{aligned} \frac{\|x_1 - x\|}{\|x\|} &= \frac{\left\|\left(I+A^{-1}B\right)^{-1} A^{-1}b_1 - A^{-1}b\right\|}{\left\|A^{-1}b\right\|} \\ &\leq \frac{1}{1-\left\|A^{-1}B\right\|} \frac{\left\|A^{-1}b_1 - \left(I+A^{-1}B\right)A^{-1}b\right\|}{\left\|A^{-1}b\right\|} \end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{1-\left\|A^{-1}B\right\|} \frac{\left\|A^{-1}(b_1 - b)\right\| + \left\|A^{-1}BA^{-1}b\right\|}{\left\|A^{-1}b\right\|} \\ &\leq \frac{\left\|A^{-1}\right\|}{1-\left\|A^{-1}B\right\|} \left(\frac{\|b_1 - b\|}{\left\|A^{-1}b\right\|} + \|B\|\right) \end{aligned}$$

because $A^{-1}b/\left\|A^{-1}b\right\|$ is a unit vector. Now multiply and divide by $\|A\|$. Then

$$\begin{aligned} &\leq \frac{\left\|A^{-1}\right\| \|A\|}{1-\left\|A^{-1}B\right\|} \left(\frac{\|b_1 - b\|}{\|A\| \left\|A^{-1}b\right\|} + \frac{\|B\|}{\|A\|}\right) \\ &\leq \frac{\left\|A^{-1}\right\| \|A\|}{1-\left\|A^{-1}B\right\|} \left(\frac{\|b_1 - b\|}{\|b\|} + \frac{\|B\|}{\|A\|}\right). \quad ■ \end{aligned}$$

This shows that the number, $\left\|A^{-1}\right\| \|A\|$, controls how sensitive the relative change in the solution of $Ax = b$ is to small changes in A and b. This number is called the condition number. It is bad when this number is large because a small relative change in b, for example could yield a large relative change in x.

Recall that for A an $n \times n$ matrix, $\|A\|_2 = \sigma_1$ where σ_1 is the largest singular value. The largest singular value of A^{-1} is therefore, $1/\sigma_n$ where σ_n is the smallest singular value of A. Therefore, the condition number reduces to σ_1/σ_n, the ratio of the largest to the smallest singular value of A provided the norm is the usual Euclidean norm.

13.3 The Spectral Radius

Even though it is in general impractical to compute the Jordan form, its existence is all that is needed in order to prove an important theorem about something which is relatively easy to compute. This is the spectral radius of a matrix.

Definition 13.3.1 *Define $\sigma(A)$ to be the eigenvalues of A. Also,*

$$\rho(A) \equiv \max(|\lambda| : \lambda \in \sigma(A))$$

The number, $\rho(A)$ is known as the spectral radius of A.

Recall the following symbols and their meaning.

$$\limsup_{n\to\infty} a_n, \ \liminf_{n\to\infty} a_n$$

They are respectively the largest and smallest limit points of the sequence $\{a_n\}$ where $\pm\infty$ is allowed in the case where the sequence is unbounded. They are also defined as

$$\begin{aligned} \limsup_{n\to\infty} a_n &\equiv \lim_{n\to\infty} (\sup\{a_k : k \geq n\}), \\ \liminf_{n\to\infty} a_n &\equiv \lim_{n\to\infty} (\inf\{a_k : k \geq n\}). \end{aligned}$$

Thus, the limit of the sequence exists if and only if these are both equal to the same real number. Also note that the

Lemma 13.3.2 *Let J be a $p \times p$ Jordan matrix*

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{pmatrix}$$

where each J_k is of the form

$$J_k = \lambda_k I + N_k$$

in which N_k is a nilpotent matrix having zeros down the main diagonal and ones down the super diagonal. Then

$$\lim_{n\to\infty} ||J^n||^{1/n} = \rho$$

where $\rho = \max\{|\lambda_k|, k = 1, \ldots, n\}$. Here the norm is the operator norm.

Proof: Consider one of the blocks, $|\lambda_k| < \rho$. Here J_k is $p \times p$.

$$\frac{1}{\rho^n} J_k^n = \frac{1}{\rho^n} \sum_{i=0}^{p} \begin{pmatrix} n \\ i \end{pmatrix} N_k^i \lambda_k^{n-i}$$

Then

$$\left\| \frac{1}{\rho^n} J_k^n \right\| \leq \sum_{i=0}^{p} \binom{n}{i} \left\| N_k^i \right\| \frac{\left| \lambda_k^{n-i} \right|}{\rho^{n-i}} \frac{1}{\rho^i} \tag{13.9}$$

Now there are p numbers $\left\| N_k^i \right\|$ so you could pick the largest, C. Also

$$\frac{\left| \lambda_k^{n-i} \right|}{\rho^{n-i}} \leq \frac{\left| \lambda_k^{n-p} \right|}{\rho^{n-p}}$$

so 13.9 is dominated by

$$\leq Cn^p \frac{\left| \lambda_k^{n-p} \right|}{\rho^{n-p}} \sum_{i=0}^{p} \frac{1}{\rho^i} \equiv \hat{C} \frac{\left| \lambda_k^{n-p} \right|}{\rho^{n-p}}$$

The ratio or root test shows that this converges to 0 as $n \to \infty$.

What happens when $|\lambda_k| = \rho$?

$$\frac{1}{\rho^n} J_k^n = \omega^n I + \sum_{i=1}^{p} \binom{n}{i} N_k^i \omega^{n-i} \frac{1}{\rho^i}$$

where $|\omega| = 1$.

$$\frac{1}{\rho^n} \left\| J_k^n \right\| \leq 1 + n^p C$$

where $C = \max \left\{ \left\| N_k^i \right\|, i = 1, \cdots, p, k = 1..., s \right\} \sum_{i=1}^{p} \frac{1}{\rho^i}$. Thus

$$\frac{1}{\rho^n} \left\| J^n \right\| \leq \frac{1}{\rho^n} \sum_{k=1}^{s} \left\| J_k^n \right\| \leq s \left(1 + n^p C \right) = sn^p C \left(\frac{1}{n^p C} + 1 \right)$$

and so

$$\frac{1}{\rho} \limsup_{n \to \infty} \left\| J^n \right\|^{1/n} \leq \limsup_{n \to \infty} s^{1/n} \left(n^p C \right)^{1/n} \left(\frac{1}{n^p C} + 1 \right)^{1/n} = 1$$

$$\limsup_{n \to \infty} \left\| J^n \right\|^{1/n} \leq \rho$$

Next let $\mathbf{x}$ be an eigenvector for $\lambda, |\lambda| = \rho$ and let $\|\mathbf{x}\| = 1$. Then

$$\rho^n = \rho^n \|\mathbf{x}\| = \|J^n \mathbf{x}\| \leq \|J^n\|$$

and so

$$\rho \leq \|J^n\|^{1/n}$$

Hence

$$\rho \geq \limsup_{n \to \infty} \|J^n\|^{1/n} \geq \liminf_{n \to \infty} \|J^n\|^{1/n} \geq \rho \blacksquare$$

The following theorem is due to Gelfand around 1941.

Theorem 13.3.3 *(Gelfand) Let A be a complex $p \times p$ matrix. Then if ρ is the absolute value of its largest eigenvalue,*

$$\lim_{n \to \infty} ||A^n||^{1/n} = \rho.$$

Here $||\cdot||$ is any norm on $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$.

Proof: First assume $||\cdot||$ is the operator norm with respect to the usual Euclidean metric on $\mathbb{C}^n$. Then letting J denote the Jordan form of $A, S^{-1}AS = J$, it follows from Lemma 13.3.2

$$\begin{aligned}\limsup_{n\to\infty} ||A^n||^{1/n} &= \limsup_{n\to\infty} \left|\left|SJ^nS^{-1}\right|\right|^{1/n} \leq \limsup_{n\to\infty} \left(\|S\| \left\|S^{-1}\right\| \|J^n\|\right)^{1/n} \\ &\leq \limsup_{n\to\infty} \left(||S|| \left|\left|S^{-1}\right|\right| ||J^n||\right)^{1/n} = \rho\end{aligned}$$

Letting λ be the largest eigenvalue of $A, |\lambda| = \rho$, and $A\mathbf{x} = \lambda\mathbf{x}$ where $\|\mathbf{x}\| = 1$,

$$\|A^n\| \geq \|A^n\mathbf{x}\| = \rho^n$$

and so

$$\liminf_{n\to\infty} \|A^n\|^{1/n} \geq \rho \geq \limsup_{n\to\infty} \|A^n\|^{1/n}$$

If follows that $\liminf_{n\to\infty} ||A^n||^{1/n} = \limsup_{n\to\infty} ||A^n||^{1/n} = \lim_{n\to\infty} ||A^n||^{1/n} = \rho$.

Now by equivalence of norms, if $|||\cdot|||$ is any other norm for the set of complex $p \times p$ matrices, there exist constants δ, Δ such that

$$\delta ||A^n|| \leq |||A^n||| \leq \Delta ||A^n||$$

Then

$$\delta^{1/n} \|A^n\|^{1/n} \leq |||A^n|||^{1/n} \leq \Delta^{1/n} \|A^n\|^{1/n}$$

The limits exist and equal ρ for the ends of the above inequality. Hence, by the squeezing theorem, $\rho = \lim_{n\to\infty} |||A^n|||^{1/n}$. ■

Example 13.3.4 *Consider* $\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}$. *Estimate the absolute value of the largest eigenvalue.*

A laborious computation reveals the eigenvalues are 5, and 10. Therefore, the right answer in this case is 10. Consider $\left|\left|A^7\right|\right|^{1/7}$ where the norm is obtained by taking the maximum of all the absolute values of the entries. Thus

$$\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}^7 = \begin{pmatrix} 8015\,625 & -1984\,375 & 3968\,750 \\ -3968\,750 & 6031\,250 & 7937\,500 \\ 1984\,375 & 1984\,375 & 6031\,250 \end{pmatrix}$$

and taking the seventh root of the largest entry gives

$$\rho(A) \approxeq 8015\,625^{1/7} = 9.688\,951\,236\,71.$$

Of course the interest lies primarily in matrices for which the exact roots to the characteristic equation are not known and in the theoretical significance.

13.4 Series and Sequences of Linear Operators

Before beginning this discussion, it is necessary to define what is meant by convergence in $\mathcal{L}(X, Y)$.

Definition 13.4.1 *Let* $\{A_k\}_{k=1}^{\infty}$ *be a sequence in* $\mathcal{L}(X,Y)$ *where* X,Y *are finite dimensional normed linear spaces. Then* $\lim_{n\to\infty} A_k = A$ *if for every* $\varepsilon > 0$ *there exists* N *such that if* $n > N$, *then*

$$||A - A_n|| < \varepsilon.$$

Here the norm refers to any of the norms defined on $\mathcal{L}(X,Y)$. *By Corollary 13.0.8 and Theorem 8.2.3 it doesn't matter which one is used. Define the symbol for an infinite sum in the usual way. Thus*

$$\sum_{k=1}^{\infty} A_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} A_k$$

Lemma 13.4.2 *Suppose* $\{A_k\}_{k=1}^{\infty}$ *is a sequence in* $\mathcal{L}(X,Y)$ *where* X,Y *are finite dimensional normed linear spaces. Then if*

$$\sum_{k=1}^{\infty} ||A_k|| < \infty,$$

It follows that

$$\sum_{k=1}^{\infty} A_k \tag{13.10}$$

exists (converges). In words, absolute convergence implies convergence. Also,

$$\left\|\sum_{k=1}^{\infty} A_k\right\| \leq \sum_{k=1}^{\infty} \|A_k\|$$

Proof: For $p \leq m \leq n$,

$$\left\|\sum_{k=1}^{n} A_k - \sum_{k=1}^{m} A_k\right\| \leq \sum_{k=p}^{\infty} ||A_k||$$

and so for p large enough, this term on the right in the above inequality is less than ε. Since ε is arbitrary, this shows the partial sums of 13.10 are a Cauchy sequence. Therefore by Corollary 13.0.7 it follows that these partial sums converge. As to the last claim,

$$\left\|\sum_{k=1}^{n} A_k\right\| \leq \sum_{k=1}^{n} \|A_k\| \leq \sum_{k=1}^{\infty} \|A_k\|$$

Therefore, passing to the limit,

$$\left\|\sum_{k=1}^{\infty} A_k\right\| \leq \sum_{k=1}^{\infty} \|A_k\| . \blacksquare$$

Why is this last step justified? (Recall the triangle inequality $|\|A\| - \|B\|| \leq \|A - B\|$.)

Now here is a useful result for differential equations.

Theorem 13.4.3 *Let* X *be a finite dimensional inner product space and let* $A \in \mathcal{L}(X,X)$. *Define*

$$\Phi(t) \equiv \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}$$

Then the series converges for each $t \in \mathbb{R}$. Also

$$\Phi'(t) \equiv \lim_{h \to 0} \frac{\Phi(t+h) - \Phi(t)}{h} = \sum_{k=1}^{\infty} \frac{t^{k-1} A^k}{(k-1)!} = A \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = A\Phi(t)$$

Also $A\Phi(t) = \Phi(t) A$ and for all $t, \Phi(t)\Phi(-t) = I$ so $\Phi(t)^{-1} = \Phi(-t)$, $\Phi(0) = I$. (It is understood that $A^0 = I$ in the above formula.)

Proof: First consider the claim about convergence.

$$\sum_{k=0}^{\infty} \left\| \frac{t^k A^k}{k!} \right\| \leq \sum_{k=0}^{\infty} \frac{|t|^k \|A\|^k}{k!} = e^{|t|\|A\|} < \infty$$

so it converges by Lemma 13.4.2.

$$\begin{aligned} \frac{\Phi(t+h) - \Phi(t)}{h} &= \frac{1}{h} \sum_{k=0}^{\infty} \frac{\left((t+h)^k - t^k\right) A^k}{k!} \\ &= \frac{1}{h} \sum_{k=0}^{\infty} \frac{\left(k (t + \theta_k h)^{k-1} h\right) A^k}{k!} = \sum_{k=1}^{\infty} \frac{(t + \theta_k h)^{k-1} A^k}{(k-1)!} \end{aligned}$$

this by the mean value theorem. Note that the series converges thanks to Lemma 13.4.2. Here $\theta_k \in (0,1)$. Thus

$$\left\| \frac{\Phi(t+h) - \Phi(t)}{h} - \sum_{k=1}^{\infty} \frac{t^{k-1} A^k}{(k-1)!} \right\| = \left\| \sum_{k=1}^{\infty} \frac{\left((t + \theta_k h)^{k-1} - t^{k-1}\right) A^k}{(k-1)!} \right\|$$

$$= \left\| \sum_{k=1}^{\infty} \frac{\left((k-1)(t + \tau_k \theta_k h)^{k-2} \theta_k h\right) A^k}{(k-1)!} \right\| = |h| \left\| \sum_{k=2}^{\infty} \frac{\left((t + \tau_k \theta_k h)^{k-2} \theta_k\right) A^k}{(k-2)!} \right\|$$

$$\leq |h| \sum_{k=2}^{\infty} \frac{(|t| + |h|)^{k-2} \|A\|^{k-2}}{(k-2)!} \|A\|^2 = |h| e^{(|t|+|h|)\|A\|} \|A\|^2$$

so letting $|h| < 1$, this is no larger than $|h| e^{(|t|+1)\|A\|} \|A\|^2$. Hence the desired limit is valid. It is obvious that $A\Phi(t) = \Phi(t) A$. Also the formula shows that

$$\Phi'(t) = A\Phi(t) = \Phi(t) A, \ \Phi(0) = I.$$

Now consider the claim about $\Phi(-t)$. The above computation shows that $\Phi'(-t) = A\Phi(-t)$ and so $\frac{d}{dt}(\Phi(-t)) = -\Phi'(-t) = -A\Phi(-t)$. Now let x, y be two vectors in X. Consider

$$(\Phi(-t)\Phi(t) x, y)_X$$

Then this equals (x, y) when $t = 0$. Take its derivative.

$$\begin{aligned} & ((-\Phi'(-t)\Phi(t) + \Phi(-t)\Phi'(t)) x, y)_X \\ = \ & ((-A\Phi(-t)\Phi(t) + \Phi(-t) A\Phi(t)) x, y)_X \\ = \ & (0, y)_X = 0 \end{aligned}$$

Hence this scalar valued function equals a constant and so the constant must be $(x, y)_X$. Hence for all $x, y, (\Phi(-t)\Phi(t)x - x, y)_X = 0$ for all x, y and this is so in particular for $y = \Phi(-t)\Phi(t)x - x$ which shows that $\Phi(-t)\Phi(t) = I$. ■

As a special case, suppose $\lambda \in \mathbb{C}$ and consider

$$\sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!}$$

where $t \in \mathbb{R}$. In this case, $A_k = \frac{t^k \lambda^k}{k!}$ and you can think of it as being in $\mathcal{L}(\mathbb{C}, \mathbb{C})$. Then the following corollary is of great interest.

Corollary 13.4.4 *Let*

$$f(t) \equiv \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} \equiv 1 + \sum_{k=1}^{\infty} \frac{t^k \lambda^k}{k!}$$

Then this function is a well defined complex valued function and furthermore, it satisfies the initial value problem,

$$y' = \lambda y,\ y(0) = 1$$

Furthermore, if $\lambda = a + ib$,

$$|f|(t) = e^{at}.$$

Proof: The first part is a special case of the above theorem. Note that for $f(t) = u(t) + iv(t)$, both u, v are differentiable. This is because

$$u = \frac{f + \overline{f}}{2},\ v = \frac{f - \overline{f}}{2i}.$$

Then from the differential equation,

$$(a + ib)(u + iv) = u' + iv'$$

and equating real and imaginary parts,

$$u' = au - bv,\ v' = av + bu.$$

Then a short computation shows

$$\left(u^2 + v^2\right)' = 2uu' + 2vv' = 2u(au - bv) + 2v(av + bu) = 2a\left(u^2 + v^2\right)$$

$$\left(u^2 + v^2\right)(0) = |f|^2(0) = 1$$

Now in general, if

$$y' = cy,\ y(0) = 1,$$

with c real it follows $y(t) = e^{ct}$. To see this,

$$y' - cy = 0$$

and so, multiplying both sides by e^{-ct} you get

$$\frac{d}{dt}\left(ye^{-ct}\right) = 0$$

and so ye^{-ct} equals a constant which must be 1 because of the initial condition $y(0) = 1$. Thus

$$\left(u^2 + v^2\right)(t) = e^{2at}$$

and taking square roots yields the desired conclusion. ■

Definition 13.4.5 *The function in Corollary 13.4.4 given by that power series is denoted as*

$$\exp(\lambda t) \text{ or } e^{\lambda t}.$$

The next lemma is normally discussed in advanced calculus courses but is proved here for the convenience of the reader. It is known as the root test.

Definition 13.4.6 *For $\{a_n\}$ any sequence of real numbers*

$$\lim\sup_{n\to\infty} a_n \equiv \lim_{n\to\infty} (\sup\{a_k : k \geq n\})$$

Similarly

$$\lim\inf_{n\to\infty} a_n \equiv \lim_{n\to\infty} (\inf\{a_k : k \geq n\})$$

In case A_n is an increasing (decreasing) sequence which is unbounded above (below) then it is understood that $\lim_{n\to\infty} A_n = \infty$ $(-\infty)$ respectively. Thus either of $\lim\sup$ or $\lim\inf$ can equal $+\infty$ or $-\infty$. However, the important thing about these is that unlike the limit, these always exist.

It is convenient to think of these as the largest point which is the limit of some subsequence of $\{a_n\}$ and the smallest point which is the limit of some subsequence of $\{a_n\}$ respectively. Thus $\lim_{n\to\infty} a_n$ exists and equals some point of $[-\infty, \infty]$ if and only if the two are equal.

Lemma 13.4.7 *Let $\{a_p\}$ be a sequence of nonnegative terms and let*

$$r = \lim\sup_{p\to\infty} a_p^{1/p}.$$

Then if $r < 1$, it follows the series, $\sum_{k=1}^{\infty} a_k$ converges and if $r > 1$, then a_p fails to converge to 0 so the series diverges. If A is an $n \times n$ matrix and

$$r = \lim\sup_{p\to\infty} ||A^p||^{1/p}, \tag{13.11}$$

then if $r > 1$, then $\sum_{k=0}^{\infty} A^k$ fails to converge and if $r < 1$ then the series converges. Note that the series converges when the spectral radius is less than one and diverges if the spectral radius is larger than one. In fact, $\lim\sup_{p\to\infty} ||A^p||^{1/p} = \lim_{p\to\infty} ||A^p||^{1/p}$ from Theorem 13.3.3.

Proof: Suppose $r < 1$. Then there exists N such that if $p > N$,

$$a_p^{1/p} < R$$

where $r < R < 1$. Therefore, for all such p, $a_p < R^p$ and so by comparison with the geometric series, $\sum R^p$, it follows $\sum_{p=1}^{\infty} a_p$ converges.

Next suppose $r > 1$. Then letting $1 < R < r$, it follows there are infinitely many values of p at which

$$R < a_p^{1/p}$$

which implies $R^p < a_p$, showing that a_p cannot converge to 0 and so the series cannot converge either.

To see the last claim, if $r > 1$, then $||A^p||$ fails to converge to 0 and so $\left\{\sum_{k=0}^{m} A^k\right\}_{m=0}^{\infty}$ is not a Cauchy sequence. Hence $\sum_{k=0}^{\infty} A^k \equiv \lim_{m\to\infty} \sum_{k=0}^{m} A^k$ cannot exist. If $r < 1$, then for all n large enough, $\|A^n\|^{1/n} \leq r < 1$ for some r so $\|A^n\| \leq r^n$. Hence $\sum_n \|A^n\|$ converges and so by Lemma 13.4.2, it follows that $\sum_{k=1}^{\infty} A^k$ also converges. ■

Now denote by $\sigma(A)^p$ the collection of all numbers of the form λ^p where $\lambda \in \sigma(A)$.

Lemma 13.4.8 $\sigma(A^p) = \sigma(A)^p \equiv \{\lambda^p : \lambda \in \sigma(A)\}$.

Proof: In dealing with $\sigma(A^p)$, it suffices to deal with $\sigma(J^p)$ where J is the Jordan form of A because J^p and A^p are similar. Thus if $\lambda \in \sigma(A^p)$, then $\lambda \in \sigma(J^p)$ and so $\lambda = \alpha$ where α is one of the entries on the main diagonal of J^p. These entries are of the form λ^p where $\lambda \in \sigma(A)$. Thus $\lambda \in \sigma(A)^p$ and this shows $\sigma(A^p) \subseteq \sigma(A)^p$.

Now take $\alpha \in \sigma(A)$ and consider α^p.

$$\alpha^p I - A^p = \left(\alpha^{p-1} I + \cdots + \alpha A^{p-2} + A^{p-1}\right)(\alpha I - A)$$

and so $\alpha^p I - A^p$ fails to be one to one which shows that $\alpha^p \in \sigma(A^p)$ which shows that $\sigma(A)^p \subseteq \sigma(A^p)$. ■

13.5 Iterative Methods for Linear Systems

Consider the problem of solving the equation

$$A\mathbf{x} = \mathbf{b} \tag{13.12}$$

where A is an $n \times n$ matrix. In many applications, the matrix A is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. To solve them, it is common to use an iterative technique. I am following the treatment given to this subject by Nobel and Daniel [21].

Definition 13.5.1 *The Jacobi iterative technique, also called the method of simultaneous corrections is defined as follows. Let $\mathbf{x}^1$ be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors, $\mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \cdots$ and hopefully this sequence of vectors will converge to the solution to 13.12. The vectors in this list are called iterates and they are obtained according to the following procedure. Letting $A = (a_{ij})$,*

$$a_{ii} x_i^{r+1} = -\sum_{j \neq i} a_{ij} x_j^r + b_i. \tag{13.13}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & * & \cdots & * \\ * & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{13.14}$$

The matrix on the left in 13.14 is obtained by retaining the main diagonal of A and setting every other entry equal to zero. The matrix on the right in 13.14 is obtained from A by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 13.5.2 *Use the Jacobi method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Of course this is solved most easily using row reductions. The Jacobi method is useful when the matrix is very large. This example is just to illustrate how the method works. First lets solve it using row operations. The exact solution from row reduction is $\begin{pmatrix} \frac{6}{29} & \frac{11}{29} & \frac{8}{29} & \frac{25}{29} \end{pmatrix}$, which in terms of decimals is approximately equal to

$$\begin{pmatrix} 0.207 & 0.379 & 0.276 & 0.862 \end{pmatrix}^T.$$

In terms of the matrices, the Jacobi iteration is of the form

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1^{r} \\ x_2^{r} \\ x_3^{r} \\ x_4^{r} \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left, [1]this iteration reduces to

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1^{r} \\ x_2^{r} \\ x_3^{r} \\ x_4^{r} \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}. \tag{13.15}$$

Now iterate this starting with $\mathbf{x}^1 \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}^T$.

Thus

$$\mathbf{x}^2 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}$$

Then

$$\mathbf{x}^3 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}}^{\mathbf{x}_2} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}$$

Continuing this way one finally gets

$$\mathbf{x}^6 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}}^{\mathbf{x}5} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .216 \\ .386 \\ .295 \\ .871 \end{pmatrix}.$$

[1]You certainly would not compute the invese in solving a large system. This is just to show you how the method works for this simple example. You would use the first description in terms of indices.

You can keep going like this. Recall the solution is approximately equal to

$$\begin{pmatrix} 0.206 & 0.379 & 0.275 & 0.862 \end{pmatrix}^T$$

so you see that with no care at all and only 6 iterations, an approximate solution has been obtained which is not too far off from the actual solution.

Definition 13.5.3 *The Gauss Seidel method, also called the method of successive corrections is given as follows. For $A = (a_{ij})$, the iterates for the problem $A\mathbf{x} = \mathbf{b}$ are obtained according to the formula*

$$\sum_{j=1}^{i} a_{ij} x_j^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^r + b_i. \tag{13.16}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{13.17}$$

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 13.16 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 13.5.4 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

In terms of matrices, this procedure is

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left[2] this yields

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}$$

As before, I will be totally unoriginal in the choice of $\mathbf{x}^1$. Let it equal the zero vector. Therefore, $\mathbf{x}^2 = \begin{pmatrix} \frac{1}{3} & \frac{5}{12} & \frac{13}{30} & \frac{47}{60} \end{pmatrix}^T$. Now

$$\mathbf{x}^3 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}}^{\mathbf{x}^2} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix}.$$

Continuing this way,

$$\mathbf{x}^4 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .219 \\ .368\,75 \\ .283\,3 \\ .858\,35 \end{pmatrix}$$

and so

$$\mathbf{x}^5 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} .219 \\ .368\,75 \\ .283\,3 \\ .858\,35 \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .210\,42 \\ .376\,57 \\ .277\,7 \\ .861\,15 \end{pmatrix}.$$

This is fairly close to the answer. You could continue doing these iterates and it appears they converge to the solution. Now consider the following example.

Example 13.5.5 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 1 & 4 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

The exact solution is given by doing row operations on the augmented matrix. When this is done the solution is seen to be $\begin{pmatrix} 6.0 & -1.25 & 1.0 & 0.5 \end{pmatrix}$.The Gauss Seidel iterations are of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

[2] As in the case of the Jacobi iteration, the computer would not do this. It would use the iteration procedure in terms of the entries of the matrix directly. Otherwise all benefit to using this method is lost.

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

$$\mathbf{x}^{r+1} = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \mathbf{x}^r + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix}.$$

This time, I will pick an initial vector close to the answer. Let $\mathbf{x}^1 = \begin{pmatrix} 6 & -1 & 1 & \frac{1}{2} \end{pmatrix}^T$. This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

$$\mathbf{x}^2 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix}$$

It appears that it moved the initial guess far from the solution even though you started with one which was initially close to the solution. This is discouraging. However, you can't expect the method to work well after only one iteration. Unfortunately, if you do multiple iterations, the iterates never seem to get close to the actual solution. Why is the process which worked so well in the other examples not working here? A better question might be: Why does either process ever work at all?

Both iterative procedures for solving

$$A\mathbf{x} = \mathbf{b} \tag{13.18}$$

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

$$\mathbf{x}^{r+1} = -B^{-1}C\mathbf{x}^r + B^{-1}\mathbf{b}. \tag{13.19}$$

This is what you would never have the computer do but this is what will allow the statement of a theorem which gives the condition for convergence of these and all other similar methods. Recall the definition of the spectral radius of $M, \rho(M)$, in Definition 13.3.1 on Page 335.

Theorem 13.5.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 13.19 converge to the unique solution of 13.18.*

I will prove this theorem in the next section. The proof depends on analysis which should not be surprising because it involves a statement about convergence of sequences.

What is an easy to verify sufficient condition which will imply the above holds? It is easy to give one in the case of the Jacobi method. Suppose the matrix A is diagonally dominant.

That is $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$. Then B would be the diagonal matrix consisting of the entries a_{ii}. You need to find the size of λ where

$$B^{-1}C\mathbf{x} = \lambda\mathbf{x}$$

Thus you need

$$(\lambda B - C)\,\mathbf{x} = \mathbf{0}$$

Now if $|\lambda| \geq 1$, then the matrix $\lambda B - C$ is diagonally dominant and so this matrix will be invertible so λ is not an eigenvalue. Hence the only eigenvalues have absolute value less than 1.

You might try a similar argument in the case of the Gauss Seidel method.

13.6 Theory of Convergence

Definition 13.6.1 *A normed vector space, E with norm $||\cdot||$ is called a Banach space if it is also complete. This means that every Cauchy sequence converges. Recall that a sequence $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence if for every $\varepsilon > 0$ there exists N such that whenever $m, n > N$,*

$$||x_n - x_m|| < \varepsilon.$$

Thus whenever $\{x_n\}$ is a Cauchy sequence, there exists x such that

$$\lim_{n\to\infty} ||x - x_n|| = 0.$$

Example 13.6.2 *Let E be a Banach space and let Ω be a nonempty subset of a normed linear space F. Let $B(\Omega; E)$ denote those functions f for which*

$$||f|| \equiv \sup\{||f(x)||_E : x \in \Omega\} < \infty$$

Denote by $BC(\Omega; E)$ the set of functions in $B(\Omega; E)$ which are also continuous.

Lemma 13.6.3 *The above $\|\cdot\|$ is a norm on $B(\Omega; E)$. The subspace $BC(\Omega; E)$ with the given norm is a Banach space.*

Proof: It is obvious $||\cdot||$ is a norm. It only remains to verify $BC(\Omega; E)$ is complete. Let $\{f_n\}$ be a Cauchy sequence. Since $\|f_n - f_m\| \to 0$ as $m, n \to \infty$, it follows that $\{f_n(x)\}$ is a Cauchy sequence in E for each x. Let $f(x) \equiv \lim_{n\to\infty} f_n(x)$. Then for any $x \in \Omega$.

$$||f_n(x) - f_m(x)||_E \leq ||f_n - f_m|| < \varepsilon$$

whenever m, n are large enough, say as large as N. For $n \geq N$, let $m \to \infty$. Then passing to the limit, it follows that for all x,

$$||f_n(x) - f(x)||_E \leq \varepsilon$$

and so for all x,

$$\|f(x)\|_E \leq \varepsilon + \|f_n(x)\|_E \leq \varepsilon + \|f_n\|.$$

It follows that $\|f\| \leq \|f_n\| + \varepsilon$ and $\|f - f_n\| \leq \varepsilon$.

It remains to verify that f is continuous.

$$\begin{aligned} \|f(x) - f(y)\|_E &\leq \|f(x) - f_n(x)\|_E + \|f_n(x) - f_n(y)\|_E + \|f_n(y) - f(y)\|_E \\ &\leq 2\|f - f_n\| + \|f_n(x) - f_n(y)\|_E < \frac{2\varepsilon}{3} + \|f_n(x) - f_n(y)\|_E \end{aligned}$$

for all n large enough. Now pick such an n. By continuity, the last term is less than $\frac{\varepsilon}{3}$ if $\|x - y\|$ is small enough. Hence f is continuous as well. ■

The most familiar example of a Banach space is $\mathbb{F}^n$. The following lemma is of great importance so it is stated in general.

Lemma 13.6.4 *Suppose $T : E \to E$ where E is a Banach space with norm $|\cdot|$. Also suppose*

$$|T\mathbf{x} - T\mathbf{y}| \leq r\,|\mathbf{x} - \mathbf{y}| \tag{13.20}$$

for some $r \in (0,1)$. Then there exists a unique fixed point, $\mathbf{x} \in E$ such that

$$T\mathbf{x} = \mathbf{x}. \tag{13.21}$$

Letting $\mathbf{x}^1 \in E$, this fixed point $\mathbf{x}$, is the limit of the sequence of iterates,

$$\mathbf{x}^1, T\mathbf{x}^1, T^2\mathbf{x}^1, \cdots. \tag{13.22}$$

In addition to this, there is a nice estimate which tells how close $\mathbf{x}^1$ is to $\mathbf{x}$ in terms of things which can be computed.

$$\left|\mathbf{x}^1 - \mathbf{x}\right| \leq \frac{1}{1-r}\left|\mathbf{x}^1 - T\mathbf{x}^1\right|. \tag{13.23}$$

Proof: This follows easily when it is shown that the above sequence, $\left\{T^k\mathbf{x}^1\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Note that

$$\left|T^2\mathbf{x}^1 - T\mathbf{x}^1\right| \leq r\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.$$

Suppose

$$\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \leq r^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right|. \tag{13.24}$$

Then

$$\begin{aligned}\left|T^{k+1}\mathbf{x}^1 - T^k\mathbf{x}^1\right| &\leq r\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \\ &\leq rr^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right| = r^k\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.\end{aligned}$$

By induction, this shows that for all $k \geq 2$, 13.24 is valid. Now let $k > l \geq N$.

$$\begin{aligned}\left|T^k\mathbf{x}^1 - T^l\mathbf{x}^1\right| &= \left|\sum_{j=l}^{k-1}\left(T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right)\right| \leq \sum_{j=l}^{k-1}\left|T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right| \\ &\leq \sum_{j=N}^{k-1} r^j\left|T\mathbf{x}^1 - \mathbf{x}^1\right| \leq \left|T\mathbf{x}^1 - \mathbf{x}^1\right|\frac{r^N}{1-r}\end{aligned}$$

which converges to 0 as $N \to \infty$. Therefore, this is a Cauchy sequence so it must converge to $\mathbf{x} \in E$. Then

$$\mathbf{x} = \lim_{k\to\infty} T^k\mathbf{x}^1 = \lim_{k\to\infty} T^{k+1}\mathbf{x}^1 = T\lim_{k\to\infty} T^k\mathbf{x}^1 = T\mathbf{x}.$$

This shows the existence of the fixed point. To show it is unique, suppose there were another one, $\mathbf{y}$. Then

$$|\mathbf{x} - \mathbf{y}| = |T\mathbf{x} - T\mathbf{y}| \leq r\,|\mathbf{x} - \mathbf{y}|$$

and so $\mathbf{x} = \mathbf{y}$.

It remains to verify the estimate.

$$\begin{aligned}\left|\mathbf{x}^1 - \mathbf{x}\right| &\leq \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + \left|T\mathbf{x}^1 - \mathbf{x}\right| = \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + \left|T\mathbf{x}^1 - T\mathbf{x}\right| \\ &\leq \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + r\left|\mathbf{x}^1 - \mathbf{x}\right|\end{aligned}$$

and solving the inequality for $\left|\mathbf{x}^1 - \mathbf{x}\right|$ gives the estimate desired. ■

The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

Corollary 13.6.5 *Suppose $T : E \to E$, for some constant C*

$$|T\mathbf{x} - T\mathbf{y}| \leq C\,|\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in E$, and for some $N \in \mathbb{N}$,

$$\left|T^N\mathbf{x} - T^N\mathbf{y}\right| \leq r\,|\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in E$ where $r \in (0,1)$. Then there exists a unique fixed point for T and it is still the limit of the sequence, $\left\{T^k\mathbf{x}^1\right\}$ for any choice of $\mathbf{x}^1$.

Proof: From Lemma 13.6.4 there exists a unique fixed point for T^N denoted here as $\mathbf{x}$. Therefore, $T^N\mathbf{x} = \mathbf{x}$. Now doing T to both sides,

$$T^N T\mathbf{x} = T\mathbf{x}.$$

By uniqueness, $T\mathbf{x} = \mathbf{x}$ because the above equation shows $T\mathbf{x}$ is a fixed point of T^N and there is only one fixed point of T^N. In fact, there is only one fixed point of T because a fixed point of T is automatically a fixed point of T^N.

It remains to show $T^k\mathbf{x}^1 \to \mathbf{x}$, the unique fixed point of T^N. If this does not happen, there exists $\varepsilon > 0$ and a subsequence, still denoted by T^k such that

$$\left|T^k\mathbf{x}^1 - \mathbf{x}\right| \geq \varepsilon$$

Now $k = j_k N + r_k$ where $r_k \in \{0, \cdots, N-1\}$ and j_k is a positive integer such that $\lim_{k\to\infty} j_k = \infty$. Then there exists a single $r \in \{0, \cdots, N-1\}$ such that for infinitely many $k, r_k = r$. Taking a further subsequence, still denoted by T^k it follows

$$\left|T^{j_k N + r}\mathbf{x}^1 - \mathbf{x}\right| \geq \varepsilon \tag{13.25}$$

However,

$$T^{j_k N + r}\mathbf{x}^1 = T^r T^{j_k N}\mathbf{x}^1 \to T^r\mathbf{x} = \mathbf{x}$$

and this contradicts 13.25. ■

Theorem 13.6.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 13.19 converge to the unique solution of 13.18.*

Proof: Consider the iterates in 13.19. Let $T\mathbf{x} = B^{-1}C\mathbf{x} + B^{-1}\mathbf{b}$. Then

$$\left|T^k\mathbf{x} - T^k\mathbf{y}\right| = \left|\left(B^{-1}C\right)^k\mathbf{x} - \left(B^{-1}C\right)^k\mathbf{y}\right| \leq \left|\left|\left(B^{-1}C\right)^k\right|\right|\,|\mathbf{x} - \mathbf{y}|.$$

Here $||\cdot||$ refers to any of the operator norms. It doesn't matter which one you pick because they are all equivalent. I am writing the proof to indicate the operator norm taken with respect to the usual norm on E. Since $\rho\left(B^{-1}C\right) < 1$, it follows from Gelfand's theorem, Theorem 13.3.3 on Page 336, there exists N such that if $k \geq N$, then for some $r^{1/k} < 1$,

$$\left|\left|\left(B^{-1}C\right)^k\right|\right|^{1/k} < r^{1/k} < 1.$$

Consequently,

$$\left|T^N\mathbf{x} - T^N\mathbf{y}\right| \leq r\,|\mathbf{x} - \mathbf{y}|.$$

Also $|T\mathbf{x} - T\mathbf{y}| \leq \left|\left|B^{-1}C\right|\right|\,|\mathbf{x} - \mathbf{y}|$ and so Corollary 13.6.5 applies and gives the conclusion of this theorem. ■

13.7 Exercises

1. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

2. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

3. Solve the system
$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

4. If you are considering a system of the form $A\mathbf{x} = \mathbf{b}$ and A^{-1} does not exist, will either the Gauss Seidel or Jacobi methods work? Explain. What does this indicate about finding eigenvectors for a given eigenvalue?

5. For $||\mathbf{x}||_\infty \equiv \max\{|x_j| : j = 1, 2, \cdots, n\}$, the parallelogram identity does not hold. Explain.

6. A norm $||\cdot||$ is said to be strictly convex if whenever $||x|| = ||y||, x \neq y$, it follows
$$\left|\left|\frac{x+y}{2}\right|\right| < ||x|| = ||y|| .$$
Show the norm $|\cdot|$ which comes from an inner product is strictly convex.

7. A norm $||\cdot||$ is said to be uniformly convex if whenever $||x_n||, ||y_n||$ are equal to 1 for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} ||x_n + y_n|| = 2$, it follows $\lim_{n\to\infty} ||x_n - y_n|| = 0$. Show the norm $|\cdot|$ coming from an inner product is always uniformly convex. Also show that uniform convexity implies strict convexity which is defined in Problem 6.

8. Suppose $A : \mathbb{C}^n \to \mathbb{C}^n$ is a one to one and onto matrix. Define
$$||\mathbf{x}|| \equiv |A\mathbf{x}| .$$
Show this is a norm.

9. If X is a finite dimensional normed vector space and $A, B \in \mathcal{L}(X, X)$ such that $||B|| < ||A||$, can it be concluded that $\left|\left|A^{-1}B\right|\right| < 1$?

10. Let X be a vector space with a norm $||\cdot||$ and let $V = \text{span}(v_1, \cdots, v_m)$ be a finite dimensional subspace of X such that $\{v_1, \cdots, v_m\}$ is a basis for V. Show V is a closed subspace of X. This means that if $w_n \to w$ and each $w_n \in V$, then so is w. Next show that if $w \notin V$,
$$\text{dist}(w, V) \equiv \inf\{||w - v|| : v \in V\} > 0$$
is a continuous function of w and
$$|\text{dist}(w, V) - \text{dist}(w_1, V)| \leq \|w_1 - w\|$$
Next show that if $w \notin V$, there exists z such that $||z|| = 1$ and $\text{dist}(z, V) > 1/2$. For those who know some advanced calculus, show that if X is an infinite dimensional vector space having norm $||\cdot||$, then the closed unit ball in X cannot be compact. Thus closed and bounded is never compact in an infinite dimensional normed vector space.

11. Suppose $\rho(A) < 1$ for $A \in \mathcal{L}(V, V)$ where V is a p dimensional vector space having a norm $||\cdot||$. You can use $\mathbb{R}^p$ or $\mathbb{C}^p$ if you like. Show there exists a new norm $|||\cdot|||$ such that with respect to this new norm, $|||A||| < 1$ where $|||A|||$ denotes the operator norm of A taken with respect to this new norm on V,
$$|||A||| \equiv \sup\{|||A\mathbf{x}||| : |||\mathbf{x}||| \leq 1\}$$
Hint: You know from Gelfand's theorem that
$$||A^n||^{1/n} < r < 1$$
provided n is large enough, this operator norm taken with respect to $||\cdot||$. Show there exists $0 < \lambda < 1$ such that
$$\rho\left(\frac{A}{\lambda}\right) < 1.$$
You can do this by arguing the eigenvalues of A/λ are the scalars μ/λ where $\mu \in \sigma(A)$. Now let $\mathbb{Z}_+$ denote the nonnegative integers.
$$|||\mathbf{x}||| \equiv \sup_{n \in \mathbb{Z}_+} \left|\left|\frac{A^n}{\lambda^n}\mathbf{x}\right|\right|$$
First show this is actually a norm. Next explain why
$$|||A\mathbf{x}||| \equiv \lambda \sup_{n \in \mathbb{Z}_+} \left|\left|\frac{A^{n+1}}{\lambda^{n+1}}\mathbf{x}\right|\right| \leq \lambda\, |||\mathbf{x}|||\,.$$

12. Establish a similar result to Problem 11 without using Gelfand's theorem. Use an argument which depends directly on the Jordan form or a modification of it.

13. Using Problem 11 give an easier proof of Theorem 13.6.6 without having to use Corollary 13.6.5. It would suffice to use a different norm of this problem and the contraction mapping principle of Lemma 13.6.4.

14. A matrix A is diagonally dominant if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$. Show that the Gauss Seidel method converges if A is diagonally dominant.

15. Suppose $f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n$ converges if $|\lambda| < R$. Show that if $\rho(A) < R$ where A is an $n \times n$ matrix, then
$$f(A) \equiv \sum_{n=0}^{\infty} a_n A^n$$
converges in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$. **Hint:** Use Gelfand's theorem and the root test.

16. Referring to Corollary 13.4.4, for $\lambda = a + ib$ show

$$\exp(\lambda t) = e^{at}(\cos(bt) + i\sin(bt)).$$

Hint: Let $y(t) = \exp(\lambda t)$ and let $z(t) = e^{-at}y(t)$. Show

$$z'' + b^2 z = 0,\ z(0) = 1, z'(0) = ib.$$

Now letting $z = u + iv$ where u, v are real valued, show

$$\begin{aligned} u'' + b^2 u &= 0,\ u(0) = 1, u'(0) = 0 \\ v'' + b^2 v &= 0,\ v(0) = 0, v'(0) = b. \end{aligned}$$

Next show $u(t) = \cos(bt)$ and $v(t) = \sin(bt)$ work in the above and that there is at most one solution to

$$w'' + b^2 w = 0\ w(0) = \alpha, w'(0) = \beta.$$

Thus $z(t) = \cos(bt) + i\sin(bt)$ and so $y(t) = e^{at}(\cos(bt) + i\sin(bt))$. To show there is at most one solution to the above problem, suppose you have two, w_1, w_2. Subtract them. Let $f = w_1 - w_2$. Thus

$$f'' + b^2 f = 0$$

and f is real valued. Multiply both sides by f' and conclude

$$\frac{d}{dt}\left(\frac{(f')^2}{2} + b^2\frac{f^2}{2}\right) = 0$$

Thus the expression in parenthesis is constant. Explain why this constant must equal 0.

17. Let $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. Show the following power series converges in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$.

$$\Psi(t) \equiv \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}$$

This was done in the chapter. Go over it and be sure you understand it. This is how you can define $\exp(tA)$. Next show that $\Psi'(t) = A\Psi(t), \Psi(0) = I$. Next let $\Phi(t) = \sum_{k=0}^{\infty} \frac{t^k(-A)^k}{k!}$. Show each $\Phi(t), \Psi(t)$ each commute with A. Next show that $\Phi(t)\Psi(t) = I$ for all t. Finally, solve the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f},\ \mathbf{x}(0) = \mathbf{x}_0$$

in terms of Φ and Ψ. This yields most of the substance of a typical differential equations course.

18. In Problem 17 $\Psi(t)$ is defined by the given series. Denote by $\exp(t\sigma(A))$ the numbers $\exp(t\lambda)$ where $\lambda \in \sigma(A)$. Show $\exp(t\sigma(A)) = \sigma(\Psi(t))$. This is like Lemma 13.4.8. Letting J be the Jordan canonical form for A, explain why

$$\Psi(t) \equiv \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = S\sum_{k=0}^{\infty} \frac{t^k J^k}{k!} S^{-1}$$

and you note that in J^k, the diagonal entries are of the form λ^k for λ an eigenvalue of A. Also $J = D + N$ where N is nilpotent and commutes with D. Argue then that

$$\sum_{k=0}^{\infty} \frac{t^k J^k}{k!}$$

is an upper triangular matrix which has on the diagonal the expressions $e^{\lambda t}$ where $\lambda \in \sigma(A)$. Thus conclude

$$\sigma(\Psi(t)) \subseteq \exp(t\sigma(A))$$

Next take $e^{t\lambda} \in \exp(t\sigma(A))$ and argue it must be in $\sigma(\Psi(t))$. You can do this as follows:

$$\begin{aligned}
\Psi(t) - e^{t\lambda} I &= \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} - \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} I = \sum_{k=0}^{\infty} \frac{t^k}{k!} \left(A^k - \lambda^k I \right) \\
&= \left(\sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j \right) (A - \lambda I)
\end{aligned}$$

Now you need to argue

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j$$

converges to something in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. To do this, use the ratio test and Lemma 13.4.2 after first using the triangle inequality. Since $\lambda \in \sigma(A)$, $\Psi(t) - e^{t\lambda} I$ is not one to one and so this establishes the other inclusion. You fill in the details. This theorem is a special case of theorems which go by the name "spectral mapping theorem".

19. Suppose $\Psi(t) \in \mathcal{L}(V, W)$ where V, W are finite dimensional inner product spaces and $t \to \Psi(t)$ is continuous for $t \in [a, b]$: For every $\varepsilon > 0$ there there exists $\delta > 0$ such that if $|s - t| < \delta$ then $||\Psi(t) - \Psi(s)|| < \varepsilon$. Show $t \to (\Psi(t) v, w)$ is continuous. Here it is the inner product in W. Also define what it means for $t \to \Psi(t) v$ to be continuous and show this is continuous. Do it all for differentiable in place of continuous. Next show $t \to ||\Psi(t)||$ is continuous.

20. If $z(t) \in W$, a finite dimensional inner product space, what does it mean for $t \to z(t)$ to be continuous or differentiable? If z is continuous, define

$$\int_a^b z(t)\, dt \in W$$

as follows.

$$\left(w, \int_a^b z(t)\, dt \right) \equiv \int_a^b (w, z(t))\, dt.$$

Show that this definition is well defined and furthermore the triangle inequality,

$$\left| \int_a^b z(t)\, dt \right| \leq \int_a^b |z(t)|\, dt,$$

and fundamental theorem of calculus,

$$\frac{d}{dt} \left(\int_a^t z(s)\, ds \right) = z(t)$$

hold along with any other interesting properties of integrals which are true.

21. For V, W two inner product spaces, define
$$\int_a^b \Psi(t)\, dt \in \mathcal{L}(V,W)$$
as follows.
$$\left(w, \int_a^b \Psi(t)\, dt\,(v)\right) \equiv \int_a^b (w, \Psi(t)\, v)\, dt.$$
Show this is well defined and does indeed give $\int_a^b \Psi(t)\, dt \in \mathcal{L}(V,W)$. Also show the triangle inequality
$$\left|\left|\int_a^b \Psi(t)\, dt\right|\right| \leq \int_a^b ||\Psi(t)||\, dt$$
where $||\cdot||$ is the operator norm and verify the fundamental theorem of calculus holds.
$$\left(\int_a^t \Psi(s)\, ds\right)' = \Psi(t).$$
Also verify the usual properties of integrals continue to hold such as the fact the integral is linear and
$$\int_a^b \Psi(t)\, dt + \int_b^c \Psi(t)\, dt = \int_a^c \Psi(t)\, dt$$
and similar things. **Hint:** On showing the triangle inequality, it will help if you use the fact that
$$|w|_W = \sup_{|v|\leq 1} |(w,v)|.$$
You should show this also.

22. Prove Gronwall's inequality. Suppose $u(t) \geq 0$ and for all $t \in [0,T]$,
$$u(t) \leq u_0 + \int_0^t K u(s)\, ds.$$
where K is some nonnegative constant. Then
$$u(t) \leq u_0 e^{Kt}.$$
Hint: $w(t) = \int_0^t u(s)\, ds$. Then using the fundamental theorem of calculus, $w(t)$ satisfies the following.
$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0,\ w(0) = 0.$$
Now use the usual techniques you saw in an introductory differential equations class. Multiply both sides of the above inequality by e^{-Kt} and note the resulting left side is now a total derivative. Integrate both sides from 0 to t and see what you have got.

23. With Gronwall's inequality and the integral defined in Problem 21 with its properties listed there, prove there is at most one solution to the initial value problem
$$\mathbf{y}' = A\mathbf{y},\ \mathbf{y}(0) = \mathbf{y}_0.$$

Hint: If there are two solutions, subtract them and call the result $\mathbf{z}$. Then

$$\mathbf{z}' = A\mathbf{z},\ \mathbf{z}(0) = \mathbf{0}.$$

It follows

$$\mathbf{z}(t) = \mathbf{0} + \int_0^t A\mathbf{z}(s)\,ds$$

and so

$$||\mathbf{z}(t)|| \leq \int_0^t \|A\|\,||\mathbf{z}(s)||\,ds$$

Now consider Gronwall's inequality of Problem 22.

24. Suppose A is a matrix which has the property that whenever $\mu \in \sigma(A)$, $\operatorname{Re}\mu < 0$. Consider the initial value problem

$$\mathbf{y}' = A\mathbf{y}, \mathbf{y}(0) = \mathbf{y}_0.$$

The existence and uniqueness of a solution to this equation has been established above in preceding problems, Problem 17 to 23. Show that in this case where the real parts of the eigenvalues are all negative, the solution to the initial value problem satisfies

$$\lim_{t\to\infty} \mathbf{y}(t) = \mathbf{0}.$$

Hint: A nice way to approach this problem is to show you can reduce it to the consideration of the initial value problem

$$\mathbf{z}' = J_\varepsilon \mathbf{z},\ \mathbf{z}(0) = \mathbf{z}_0$$

where J_ε is the modified Jordan canonical form where instead of ones down the main diagonal, there are ε down the main diagonal (Problem 19). Then

$$\mathbf{z}' = D\mathbf{z} + N_\varepsilon \mathbf{z}$$

where D is the diagonal matrix obtained from the eigenvalues of A and N_ε is a nilpotent matrix commuting with D which is very small provided ε is chosen very small. Now let $\Psi(t)$ be the solution of

$$\Psi' = -D\Psi,\ \Psi(0) = I$$

described earlier as

$$\sum_{k=0}^{\infty} \frac{(-1)^k t^k D^k}{k!}.$$

Thus $\Psi(t)$ commutes with D and N_ε. Tell why. Next argue

$$(\Psi(t)\,\mathbf{z})' = \Psi(t)\,N_\varepsilon \mathbf{z}(t)$$

and integrate from 0 to t. Then

$$\Psi(t)\,\mathbf{z}(t) - \mathbf{z}_0 = \int_0^t \Psi(s)\,N_\varepsilon \mathbf{z}(s)\,ds.$$

It follows

$$||\Psi(t)\,\mathbf{z}(t)|| \leq ||z_0|| + \int_0^t ||N_\varepsilon||\,||\Psi(s)\,\mathbf{z}(s)||\,ds.$$

It follows from Gronwall's inequality

$$||\Psi(t)\mathbf{z}(t)|| \leq ||z_0|| \, e^{||N_\varepsilon||t}$$

Now look closely at the form of $\Psi(t)$ to get an estimate which is interesting. Explain why

$$\Psi(t) = \begin{pmatrix} e^{\mu_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\mu_n t} \end{pmatrix}$$

and now observe that if ε is chosen small enough, $||N_\varepsilon||$ is so small that each component of $\mathbf{z}(t)$ converges to 0.

25. Using Problem 24 show that if A is a matrix having the real parts of all eigenvalues less than 0 then if

$$\Psi'(t) = A\Psi(t), \; \Psi(0) = I$$

it follows

$$\lim_{t\to\infty} \Psi(t) = 0.$$

Hint: Consider the columns of $\Psi(t)$?

26. Let $\Psi(t)$ be a fundamental matrix satisfying

$$\Psi'(t) = A\Psi(t), \; \Psi(0) = I.$$

Show $\Psi(t)^n = \Psi(nt)$. **Hint:** Subtract and show the difference satisfies

$$\Phi' = A\Phi, \; \Phi(0) = 0.$$

Use uniqueness.

27. If the real parts of the eigenvalues of A are all negative, show that for every positive t,

$$\lim_{n\to\infty} \Psi(nt) = 0.$$

Hint: Pick $\operatorname{Re}(\sigma(A)) < -\lambda < 0$ and use Problem 18 about the spectrum of $\Psi(t)$ and Gelfand's theorem for the spectral radius along with Problem 26 to argue that $\left|\left|\Psi(nt)/e^{-\lambda nt}\right|\right| < 1$ for all n large enough.

28. Let H be a Hermitian matrix. $(H = H^*)$. Show that $e^{iH} \equiv \sum_{n=0}^{\infty} \frac{(iH)^n}{n!}$ is unitary.

29. Show the converse of the above exercise. If V is unitary, then $V = e^{iH}$ for some H Hermitian.

30. If U is unitary and does not have -1 as an eigenvalue so that $(I+U)^{-1}$ exists, show that

$$H = i(I-U)(I+U)^{-1}$$

is Hermitian. Then, verify that

$$U = (I+iH)(I-iH)^{-1}.$$

31. Suppose that $A \in \mathcal{L}(V,V)$ where V is a normed linear space. Also suppose that $\|A\| < 1$ where this refers to the operator norm on A. Verify that

$$(I-A)^{-1} = \sum_{i=0}^{\infty} A^i$$

This is called the Neumann series. Suppose now that you only know the algebraic condition $\rho(A) < 1$. Is it still the case that the Neumann series converges to $(I-A)^{-1}$?

Chapter 14

Numerical Methods, Eigenvalues

14.1 The Power Method for Eigenvalues

This chapter discusses numerical methods for finding eigenvalues. However, to do this correctly, you must include numerical analysis considerations which are distinct from linear algebra. The purpose of this chapter is to give an introduction to some numerical methods without leaving the context of linear algebra. In addition, some examples are given which make use of computer algebra systems. For a more thorough discussion, you should see books on numerical methods in linear algebra like some listed in the references.

I will use $\approxeq$ to signify "approximately equal".

Let A be a complex $p \times p$ matrix and suppose that it has distinct eigenvalues

$$\{\lambda_1, \cdots, \lambda_m\}$$

and that $|\lambda_1| > |\lambda_k|$ for all k. Also let the Jordan form of A be

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}$$

with J_1 an $m_1 \times m_1$ matrix.

$$J_k = \lambda_k I_k + N_k$$

where $N_k^{r_k} \neq 0$ but $N_k^{r_k+1} = 0$. Also let

$$P^{-1}AP = J,\ A = PJP^{-1}.$$

Now fix $\mathbf{x} \in \mathbb{F}^p$. Take $A\mathbf{x}$ and let s_1 be the entry of the vector $A\mathbf{x}$ which has largest absolute value. Thus $A\mathbf{x}/s_1$ is a vector $\mathbf{y}_1$ which has a component of 1 and every other entry of this vector has magnitude no larger than 1. If the scalars $\{s_1, \cdots, s_{n-1}\}$ and vectors $\{\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}\}$ have been obtained, let $\mathbf{y}_n \equiv A\mathbf{y}_{n-1}/s_n$ where s_n is the entry of $A\mathbf{y}_{n-1}$ which has largest absolute value. Thus

$$\mathbf{y}_n = \frac{AA\mathbf{y}_{n-2}}{s_n s_{n-1}} \cdots = \frac{A^n\mathbf{x}}{s_n s_{n-1} \cdots s_1} \tag{14.1}$$

$$\begin{aligned} &= \frac{1}{s_n s_{n-1} \cdots s_1} P \begin{pmatrix} J_1^n & & \\ & \ddots & \\ & & J_m^n \end{pmatrix} P^{-1}\mathbf{x} \\ &= \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} P \begin{pmatrix} \lambda_1^{-n} J_1^n & & \\ & \ddots & \\ & & \lambda_1^{-n} J_m^n \end{pmatrix} P^{-1}\mathbf{x} \end{aligned} \tag{14.2}$$

Consider one of the blocks in the Jordan form. First consider the k^{th} of these blocks, $k > 1$. It equals

$$\lambda_1^{-n} J_k^n = \sum_{i=0}^{r_k} \binom{n}{i} \lambda_1^{-n} \lambda_k^{n-i} N_k^i$$

which clearly converges to 0 as $n \to \infty$ since $|\lambda_1| > |\lambda_k|$. An application of the ratio test or root test for each term in the sum will show this. When $k = 1$, this block is

$$\lambda_1^{-n} J_1^n = \lambda_1^{-n} J_k^n = \sum_{i=0}^{r_1} \binom{n}{i} \lambda_1^{-n} \lambda_1^{n-i} N_1^i = \binom{n}{r_1} \left[\lambda_1^{-r_1} N_1^{r_1} + e_n \right]$$

where $\lim_{n\to\infty} e_n = 0$ because it is a sum of bounded matrices which are multiplied by $\binom{n}{i} / \binom{n}{r_1}$. This quotient converges to 0 as $n \to \infty$ because $i < r_1$. It follows that 14.2 is of the form

$$\mathbf{y}_n = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1} + e_n & 0 \\ 0 & E_n \end{pmatrix} P^{-1}\mathbf{x} \equiv \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} \mathbf{w}_n$$

where $E_n \to 0, e_n \to 0$. Let $\left(P^{-1}\mathbf{x}\right)_{m_1}$ denote the first m_1 entries of the vector $P^{-1}\mathbf{x}$. Unless a very unlucky choice for $\mathbf{x}$ was picked, it will follow that $\left(P^{-1}\mathbf{x}\right)_{m_1} \notin \ker\left(N_1^{r_1}\right)$. Then for large n, $\mathbf{y}_n$ is close to the vector

$$\frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1} & 0 \\ 0 & 0 \end{pmatrix} P^{-1}\mathbf{x} \equiv \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} \mathbf{w} \equiv \mathbf{z} \neq \mathbf{0}$$

However, this is an eigenvector because

$$(A - \lambda_1 I)\,\mathbf{w} = \overbrace{P\,(J - \lambda_1 I)\,P^{-1}}^{A-\lambda_1 I} P \begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1} & 0 \\ 0 & 0 \end{pmatrix} P^{-1}\mathbf{x} =$$

$$P \begin{pmatrix} N_1 & & \\ & \ddots & \\ & & J_m - \lambda_1 I \end{pmatrix} P^{-1} P \begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1} & & \\ & \ddots & \\ & & 0 \end{pmatrix} P^{-1}\mathbf{x}$$

$$= \quad P \begin{pmatrix} N_1 \lambda_1^{-r_1} N_1^{r_1} & 0 \\ 0 & 0 \end{pmatrix} P^{-1}\mathbf{x} = \mathbf{0}$$

Recall $N_1^{r_1+1} = 0$. Now you could recover an approximation to the eigenvalue as follows.

$$\frac{(A\mathbf{y}_n, \mathbf{y}_n)}{(\mathbf{y}_n, \mathbf{y}_n)} \approxeq \frac{(A\mathbf{z}, \mathbf{z})}{(\mathbf{z}, \mathbf{z})} = \lambda_1$$

Here $\approxeq$ means "approximately equal". However, there is a more convenient way to identify the eigenvalue in terms of the scaling factors s_k.

$$\left\| \frac{\lambda_1^n}{s_n \cdots s_1} \binom{n}{r_1} (\mathbf{w}_n - \mathbf{w}) \right\|_\infty \approxeq 0$$

Pick the largest nonzero entry of $\mathbf{w}, w_l$. Then for large n, it is also likely the case that the largest entry of $\mathbf{w}_n$ will be in the l^{th} position because $\mathbf{w}_m$ is close to $\mathbf{w}$. From the construction,

$$\frac{\lambda_1^n}{s_n \cdots s_1} \binom{n}{r_1} w_{nl} = 1 \approxeq \frac{\lambda_1^n}{s_n \cdots s_1} \binom{n}{r_1} w_l$$

In other words, for large n

$$\frac{\lambda_1^n}{s_n \cdots s_1} \binom{n}{r_1} \approxeq 1/w_l$$

Therefore, for large n,

$$\frac{\lambda_1^n}{s_n \cdots s_1}\binom{n}{r_1} \approxeq \frac{\lambda_1^{n+1}}{s_{n+1}s_n \cdots s_1}\binom{n+1}{r_1}$$

and so

$$\binom{n}{r_1} \Big/ \binom{n+1}{r_1} \approxeq \frac{\lambda_1}{s_{n+1}}$$

But $\lim_{n\to\infty} \binom{n}{r_1}/\binom{n+1}{r_1} = 1$ and so, for large n it must be the case that $\lambda_1 \approxeq s_{n+1}$.

This has proved the following theorem which justifies the power method.

Theorem 14.1.1 *Let A be a complex $p \times p$ matrix such that the eigenvalues are*

$$\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$$

with $|\lambda_1| > |\lambda_j|$ for all $j \neq 1$. Then for $\mathbf{x}$ a given vector, let

$$\mathbf{y}_1 = \frac{A\mathbf{x}}{s_1}$$

where s_1 is an entry of $A\mathbf{x}$ which has the largest absolute value. If the scalars $\{s_1, \cdots, s_{n-1}\}$ and vectors $\{\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}\}$ have been obtained, let

$$\mathbf{y}_n \equiv \frac{A\mathbf{y}_{n-1}}{s_n}$$

where s_n is the entry of $A\mathbf{y}_{n-1}$ which has largest absolute value. Then it is probably the case that $\{s_n\}$ will converge to λ_1 and $\{\mathbf{y}_n\}$ will converge to an eigenvector associated with λ_1. If it doesn't, you picked an incredibly inauspicious initial vector $\mathbf{x}$.

In summary, here is the procedure.

Finding the largest eigenvalue with its eigenvector.

1. Start with a vector, $\mathbf{u}_1$ which you hope is not unlucky.
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. If things don't work well, try another $\mathbf{u}_1$. You were miraculously unlucky in your choice.

Example 14.1.2 *Find the largest eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$.

You can begin with $\mathbf{u}_1 = (1, \cdots, 1)^T$ and apply the above procedure. However, you can accelerate the process if you begin with $A^n \mathbf{u}_1$ and then divide by the largest entry to get the first approximate eigenvector. Thus

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}^{20} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2.5558 \times 10^{21} \\ -1.2779 \times 10^{21} \\ -3.6562 \times 10^{15} \end{pmatrix}$$

Divide by the largest entry to obtain a good aproximation.

$$\begin{pmatrix} 2.5558 \times 10^{21} \\ -1.2779 \times 10^{21} \\ -3.6562 \times 10^{15} \end{pmatrix} \frac{1}{2.5558 \times 10^{21}} = \begin{pmatrix} 1.0 \\ -0.5 \\ -1.4306 \times 10^{-6} \end{pmatrix}$$

Now begin with this one.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.5 \\ -1.4306 \times 10^{-6} \end{pmatrix} = \begin{pmatrix} 12.000 \\ -6.0000 \\ 4.2918 \times 10^{-6} \end{pmatrix}$$

Divide by 12 to get the next iterate.

$$\begin{pmatrix} 12.000 \\ -6.0000 \\ 4.2918 \times 10^{-6} \end{pmatrix} \frac{1}{12} = \begin{pmatrix} 1.0 \\ -0.5 \\ 3.5765 \times 10^{-7} \end{pmatrix}$$

Another iteration will reveal that the scaling factor is still 12. Thus this is an approximate eigenvalue. In fact, it is **the** largest eigenvalue and the corresponding eigenvector is $\begin{pmatrix} 1.0 & -0.5 & 0 \end{pmatrix}$. The process has worked very well.

14.1.1 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. One can find complex eigenvalues using this method. The situation is this: You have a number α which is close to λ, some eigenvalue of an $n \times n$ matrix A. You don't know λ but you know that α is closer to λ than to any other eigenvalue. Your problem is to find both λ and an eigenvector which goes with λ. Another way to look at this is to start with α and seek the eigenvalue λ, which is closest to α along with an eigenvector associated with λ. If α is an eigenvalue of A, then you have what you want. Therefore, I will always assume α is not an eigenvalue of A and so $(A - \alpha I)^{-1}$ exists. The method is based on the following lemma.

Lemma 14.1.3 *Let $\{\lambda_k\}_{k=1}^n$ be the eigenvalues of A. If $\mathbf{x}_k$ is an eigenvector of A for the eigenvalue λ_k, then $\mathbf{x}_k$ is an eigenvector for $(A - \alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_k - \alpha}$. Conversely, if*

$$(A - \alpha I)^{-1} \mathbf{y} = \frac{1}{\lambda - \alpha} \mathbf{y} \tag{14.3}$$

and $\mathbf{y} \neq \mathbf{0}$, then $A\mathbf{y} = \lambda \mathbf{y}$.

Proof: Let λ_k and $\mathbf{x}_k$ be as described in the statement of the lemma. Then

$$(A - \alpha I)\,\mathbf{x}_k = (\lambda_k - \alpha)\,\mathbf{x}_k$$

and so

$$\frac{1}{\lambda_k - \alpha}\mathbf{x}_k = (A - \alpha I)^{-1}\,\mathbf{x}_k.$$

Suppose 14.3. Then $\mathbf{y} = \frac{1}{\lambda - \alpha}\left[A\mathbf{y} - \alpha\mathbf{y}\right]$. Solving for $A\mathbf{y}$ leads to $A\mathbf{y} = \lambda\mathbf{y}$. ■

Now assume α is closer to λ than to any other eigenvalue. Then the magnitude of $\frac{1}{\lambda-\alpha}$ is greater than the magnitude of all the other eigenvalues of $(A - \alpha I)^{-1}$. Therefore, the power method applied to $(A - \alpha I)^{-1}$ will yield $\frac{1}{\lambda-\alpha}$. You end up with $s_{n+1} \approxeq \frac{1}{\lambda-\alpha}$ and solve for λ.

14.1.2 The Explicit Description of the Method

Here is how you use this method to find the eigenvalue closest to α and the corresponding eigenvector.

1. Find $(A - \alpha I)^{-1}$.

2. Pick $\mathbf{u}_1$. If you are not phenomenally unlucky, the iterations will converge.

3. If $\mathbf{u}_k$ has been obtained,

$$\mathbf{u}_{k+1} = \frac{(A - \alpha I)^{-1}\,\mathbf{u}_k}{s_{k+1}}$$

   where s_{k+1} is the entry of $(A - \alpha I)^{-1}\,\mathbf{u}_k$ which has largest absolute value.

4. When the scaling factors, s_k are not changing much and the $\mathbf{u}_k$ are not changing much, find the approximation to the eigenvalue by solving

$$s_{k+1} = \frac{1}{\lambda - \alpha}$$

   for λ. The eigenvector is approximated by $\mathbf{u}_{k+1}$.

5. Check your work by multiplying by the original matrix to see how well what you have found works.

Thus this amounts to the power method for the matrix $(A - \alpha I)^{-1}$ but you are free to pick α.

Example 14.1.4 *Find the eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$ *which is closest to* -7. *Also find an eigenvector which goes with this eigenvalue.*

In this case the eigenvalues are $-6, 0,$ and 12 so the correct answer is -6 for the eigenvalue. Then from the above procedure, I will start with an initial vector, $\mathbf{u}_1 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T$. Then I must solve the following equation.

$$\left(\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} + 7\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Simplifying the matrix on the left, I must solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and then divide by the entry which has largest absolute value to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ .\,184 \\ -.\,76 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .\,184 \\ -.\,76 \end{pmatrix}$$

and divide by the largest entry, $1.\,051\,5$ to get

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ .0\,266 \\ -.\,970\,61 \end{pmatrix}$$

Solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .0\,266 \\ -.\,970\,61 \end{pmatrix}$$

and divide by the largest entry, $1.\,01$ to get

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ 3.\,845\,4 \times 10^{-3} \\ -.\,996\,04 \end{pmatrix}.$$

These scaling factors are pretty close after these few iterations. Therefore, the predicted eigenvalue is obtained by solving the following for λ.

$$\frac{1}{\lambda + 7} = 1.01$$

which gives $\lambda = -6.\,01$. You see this is pretty close. In this case the eigenvalue closest to -7 was -6.

How would you know what to start with for an initial guess? You might apply Gerschgorin's theorem. However, sometimes you can begin with a better estimate.

Example 14.1.5 *Consider the symmetric matrix* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}$. *Find the middle eigenvalue and an eigenvector which goes with it.*

Since A is symmetric, it follows it has three real eigenvalues which are solutions to

$$\begin{aligned} p(\lambda) &= \det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right) \\ &= \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0 \end{aligned}$$

If you use your graphing calculator to graph this polynomial, you find there is an eigenvalue somewhere between $-.9$ and $-.8$ and that this is the middle eigenvalue. Of course you could zoom in and find it very accurately without much trouble but what about the eigenvector which goes with it? If you try to solve

$$\left((-.8)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

there will be only the zero solution because the matrix on the left will be invertible and the same will be true if you replace $-.8$ with a better approximation like $-.86$ or $-.855$. This is because all these are only approximations to the eigenvalue and so the matrix in the above is nonsingular for all of these. Therefore, you will only get the zero solution and

Eigenvectors are never equal to zero!

However, there exists such an eigenvector and you can find it using the shifted inverse power method. Pick $\alpha = -.855$. Then you solve

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} + .855\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

or in other words,

$$\begin{pmatrix} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and after finding the solution, divide by the largest entry -67.944, to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ -.58921 \\ -.23044 \end{pmatrix}$$

After a couple more iterations, you obtain

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ -.58777 \\ -.22714 \end{pmatrix} \tag{14.4}$$

Then doing it again, the scaling factor is -513.42 and the next iterate is

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ -.58778 \\ -.22714 \end{pmatrix}$$

Clearly the $\mathbf{u}_k$ are not changing much. This suggests an approximate eigenvector for this eigenvalue which is close to $-.855$ is the above $\mathbf{u}_3$ and an eigenvalue is obtained by solving

$$\frac{1}{\lambda + .855} = -513.42,$$

which yields $\lambda = -0.85695$ Lets check this.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.58778 \\ -.22714 \end{pmatrix} = \begin{pmatrix} -0.85698 \\ 0.50366 \\ 0.1946 \end{pmatrix}.$$

$$-0.85695 \begin{pmatrix} 1.0 \\ -.58777 \\ -.22714 \end{pmatrix} = \begin{pmatrix} -0.85695 \\ 0.50369 \\ 0.19465 \end{pmatrix}$$

Thus the vector of 14.4 is very close to the desired eigenvector, just as $-.8569$ is very close to the desired eigenvalue. For practical purposes, I have found both the eigenvector and the eigenvalue.

Example 14.1.6 *Find the eigenvalues and eigenvectors of the matrix* $A = \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

This is only a 3×3 matrix and so it is not hard to estimate the eigenvalues. Just get the characteristic equation, graph it using a calculator and zoom in to find the eigenvalues. If you do this, you find there is an eigenvalue near -1.2, one near $-.4$, and one near 5.5. (The characteristic equation is $2 + 8\lambda + 4\lambda^2 - \lambda^3 = 0$.) Of course I have no idea what the eigenvectors are.

Lets first try to find the eigenvector and a better approximation for the eigenvalue near -1.2. In this case, let $\alpha = -1.2$. Then

$$(A - \alpha I)^{-1} = \begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}.$$

As before, it helps to get things started if you raise to a power and then go from the approximate eigenvector obtained.

$$\begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}^7 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2.2956 \times 10^{11} \\ 1.1291 \times 10^{11} \\ 2.0865 \times 10^{11} \end{pmatrix}$$

Then the next iterate will be

$$\begin{pmatrix} -2.2956 \times 10^{11} \\ 1.1291 \times 10^{11} \\ 2.0865 \times 10^{11} \end{pmatrix} \frac{1}{-2.2956 \times 10^{11}} = \begin{pmatrix} 1.0 \\ -0.49185 \\ -0.90891 \end{pmatrix}$$

Next iterate:

$$\begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49185 \\ -0.90891 \end{pmatrix} = \begin{pmatrix} -54.115 \\ 26.615 \\ 49.184 \end{pmatrix}$$

Divide by largest entry

$$\begin{pmatrix} -54.115 \\ 26.615 \\ 49.184 \end{pmatrix} \frac{1}{-54.115} = \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix}$$

You can see the vector didn't change much and so the next scaling factor will not be much different than this one. Hence you need to solve for λ

$$\frac{1}{\lambda + 1.2} = -54.115$$

Then $\lambda = -1.2185$ is an approximate eigenvalue and

$$\begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix}$$

is an approximate eigenvector. How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix} = \begin{pmatrix} -1.2185 \\ 0.5993 \\ 1.1075 \end{pmatrix}$$

$$(-1.2185) \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix} = \begin{pmatrix} -1.2185 \\ 0.59928 \\ 1.1075 \end{pmatrix}$$

You can see that for practical purposes, this has found the eigenvalue closest to -1.2185 and the corresponding eigenvector.

The other eigenvectors and eigenvalues can be found similarly. In the case of $-.4$, you could let $\alpha = -.4$ and then

$$(A - \alpha I)^{-1} = \begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -.40322581 & 11.370968 & -7.2580645 \\ .40322581 & 3.6290323 & -2.7419355 \end{pmatrix}.$$

Following the procedure of the power method, you find that after about 5 iterations, the scaling factor is 9.7573139, they are not changing much, and

$$\mathbf{u}_5 = \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix}.$$

Thus the approximate eigenvalue is

$$\frac{1}{\lambda + .4} = 9.7573139$$

which shows $\lambda = -.29751278$ is an approximation to the eigenvalue near .4. How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix} = \begin{pmatrix} .23236104 \\ -.29751272 \\ -.07873752 \end{pmatrix}.$$

$$-.29751278\begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix} = \begin{pmatrix} .23242436 \\ -.29751278 \\ -7.8822108\times 10^{-2} \end{pmatrix}.$$

It works pretty well. For practical purposes, the eigenvalue and eigenvector have now been found. If you want better accuracy, you could just continue iterating. One can find the eigenvector corresponding to the eigenvalue nearest 5.5 the same way.

14.1.3 Complex Eigenvalues

What about complex eigenvalues? If your matrix is real, you won't see these by graphing the characteristic equation on your calculator. Will the shifted inverse power method find these eigenvalues and their associated eigenvectors? The answer is yes. However, for a real matrix, you must pick α to be complex. This is because the eigenvalues occur in conjugate pairs so if you don't pick it complex, it will be the same distance between any conjugate pair of complex numbers and so nothing in the above argument for convergence implies you will get convergence to a complex number. Also, the process of iteration will yield only real vectors and scalars.

Example 14.1.7 *Find the complex eigenvalues and corresponding eigenvectors for the matrix*

$$\begin{pmatrix} 5 & -8 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Here the characteristic equation is $\lambda^3 - 5\lambda^2 + 8\lambda - 6 = 0$. One solution is $\lambda = 3$. The other two are $1+i$ and $1-i$. I will apply the process to $\alpha = i$ to find the eigenvalue closest to i.

$$(A-\alpha I)^{-1} = \begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix}$$

Then let $\mathbf{u}_1 = (1,1,1)^T$ for lack of any insight into anything better.

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix}^{20}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} -0.40000+0.8i \\ 0.20000+0.6i \\ 0.40000+0.2i \end{pmatrix}$$

Now divide by the largest entry to get the next iterate. This yields for an approximate eigenvector approximately

$$\begin{pmatrix} -0.40000+0.8i \\ 0.20000+0.6i \\ 0.40000+0.2i \end{pmatrix}\frac{1}{-0.40000+0.8i} = \begin{pmatrix} 1.0 \\ 0.5-0.5i \\ -0.5i \end{pmatrix}$$

Now leaving off extremely small terms,

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.5-0.5i \\ -0.5i \end{pmatrix} =$$

$$\begin{pmatrix} 1.0 \\ 0.5-0.5i \\ -0.5i \end{pmatrix}$$

so it appears that an eigenvector is the above and an eigenvalue can be obtained by solving

$$\frac{1}{\lambda - i} = 1, \text{ so } \lambda = 1 + i$$

The method has successfully found the complex eigenvalue closest to i as well as the eigenvector. Note that I used essentially 20 iterations of the method.

This illustrates an interesting topic which leads to many related topics. If you have a polynomial, $x^4 + ax^3 + bx^2 + cx + d$, you can consider it as the characteristic polynomial of a certain matrix, called a **companion matrix**. In this case,

$$\begin{pmatrix} -a & -b & -c & -d \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The above example was just a companion matrix for $\lambda^3 - 5\lambda^2 + 8\lambda - 6$. You can see the pattern which will enable you to obtain a companion matrix for any polynomial of the form $\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n$. This illustrates that one way to find the complex zeros of a polynomial is to use the shifted inverse power method on a companion matrix for the polynomial. Doubtless there are better ways but this does illustrate how impressive this procedure is. Do you have a better way?

Note that the shifted inverse power method is a way you can begin with something close but not equal to an eigenvalue and end up with something close to an eigenvector.

14.1.4 Rayleigh Quotients and Estimates for Eigenvalues

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. Recall a matrix A is Hermitian if $A = A^*$ where A^* means to take the transpose of the conjugate of A. In the case of a real matrix, Hermitian reduces to symmetric. Recall also that for $\mathbf{x} \in \mathbb{F}^n$,

$$|\mathbf{x}|^2 = \mathbf{x}^*\mathbf{x} = \sum_{j=1}^{n} |x_j|^2 .$$

Recall the following corollary found on Page 166 which is stated here for convenience.

Corollary 14.1.8 *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,

$$\begin{aligned}\frac{\mathbf{x}^* A\mathbf{x}}{|\mathbf{x}|^2} &= \frac{\left(\sum_{i=1}^n \overline{a}_i \mathbf{x}_i^*\right)\left(\sum_{j=1}^n a_j \lambda_j \mathbf{x}_j\right)}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \mathbf{x}_i^* \mathbf{x}_j}{\sum_{i=1}^n |a_i|^2} \\ &= \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \delta_{ij}}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{i=1}^n |a_i|^2 \lambda_i}{\sum_{i=1}^n |a_i|^2} \in [\lambda_1, \lambda_n] .\end{aligned}$$

In other words, the Rayleigh quotient is always between the largest and the smallest eigenvalues of A. When $\mathbf{x} = \mathbf{x}_n$, the Rayleigh quotient equals the largest eigenvalue and when $\mathbf{x} = \mathbf{x}_1$ the Rayleigh quotient equals the smallest eigenvalue. Suppose you calculate a Rayleigh quotient. How close is it to some eigenvalue?

Theorem 14.1.9 *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

$$|\lambda_q - q| \leq \frac{|A\mathbf{x} - q\mathbf{x}|}{|\mathbf{x}|}. \tag{14.5}$$

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 14.5. ■

Example 14.1.10 *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

$$\left|\lambda_q-\frac{19}{3}\right|\leq\frac{\left|\begin{pmatrix}1&2&3\\2&2&1\\3&1&4\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}-\frac{19}{3}\begin{pmatrix}1\\1\\1\end{pmatrix}\right|}{\sqrt{3}}=\frac{1}{\sqrt{3}}\begin{pmatrix}-\frac{1}{3}\\-\frac{4}{3}\\\frac{5}{3}\end{pmatrix}$$

$$=\frac{\sqrt{\frac{1}{9}+\left(\frac{4}{3}\right)^2+\left(\frac{5}{3}\right)^2}}{\sqrt{3}}=1.\,247\,2$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the shifted inverse power method is all about.

Solve

$$\left(\begin{pmatrix}1&2&3\\2&2&1\\3&1&4\end{pmatrix}-\frac{19}{3}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right)\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

In other words solve

$$\begin{pmatrix}-\frac{16}{3}&2&3\\2&-\frac{13}{3}&1\\3&1&-\frac{7}{3}\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

and divide by the entry which is largest, $3.\,870\,7$, to get

$$\mathbf{u}_2=\begin{pmatrix}.\,699\,25\\.\,493\,89\\1.0\end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .699\,25 \\ .493\,89 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, $2.\,997\,9$ to get

$$\mathbf{u}_3 = \begin{pmatrix} .714\,73 \\ .522\,63 \\ 1.\,0 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .714\,73 \\ .522\,63 \\ 1.\,0 \end{pmatrix}$$

and divide by the largest entry, $3.\,045\,4$, to get

$$\mathbf{u}_4 = \begin{pmatrix} .713\,7 \\ .520\,56 \\ 1.0 \end{pmatrix}$$

Solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .713\,7 \\ .520\,56 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, $3.\,042\,1$ to get

$$\mathbf{u}_5 = \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix}$$

You can see these scaling factors are not changing much. The predicted eigenvalue is then about

$$\frac{1}{3.\,042\,1} + \frac{19}{3} = 6.\,662\,1.$$

How close is this?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.\,755\,2 \\ 3.\,469 \\ 6.\,662\,1 \end{pmatrix}$$

while

$$6.\,662\,1 \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.\,755\,3 \\ 3.\,469\,2 \\ 6.\,662\,1 \end{pmatrix}.$$

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

14.2 The QR Algorithm

14.2.1 Basic Properties and Definition

Recall the theorem about the QR factorization in Theorem 5.7.5. It says that given an $n \times n$ real matrix A, there exists a real orthogonal matrix Q and an upper triangular matrix R such that $A = QR$ and that this factorization can be accomplished by a systematic procedure. One such procedure was given in proving this theorem.

Theorem 14.2.1 *Let A be an $m \times n$ complex matrix. Then there exists a unitary Q and R, where R is all zero below the main diagonal ($R_{ij} = 0$if $i > j$) such that $A = QR$.*

Proof: This is obvious if $m = 1$.

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} = (1) \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$$

Suppose true for $m - 1$ and let

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}, \ A \text{ is } m \times n$$

There exists Q_1 a unitary matrix such that $Q_1 (\mathbf{a}_1 / |\mathbf{a}_1|) = \mathbf{e}_1$ in case $\mathbf{a}_1 \neq \mathbf{0}$. Thus $Q_1 \mathbf{a}_1 = |\mathbf{a}_1| \mathbf{e}_1$. If $\mathbf{a}_1 = \mathbf{0}$, let $Q_1 = I$. Thus

$$Q_1 A = \begin{pmatrix} a & \mathbf{b} \\ \mathbf{0} & A_1 \end{pmatrix}$$

where A_1 is $(m-1) \times (n-1)$. If $n = 1$, this obtains

$$Q_1 A = \begin{pmatrix} a \\ \mathbf{0} \end{pmatrix}, \ A = Q_1^* \begin{pmatrix} a \\ \mathbf{0} \end{pmatrix}, \text{ let } Q = Q_1^*.$$

That which is desired is obtained. So assume $n > 1$. By induction, there exists Q_2' an $(m-1) \times (n-1)$ unitary matrix such that $Q_2' A_1 = R'$, $R_{ij}' = 0$ if $i > j$. Then

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2' \end{pmatrix} Q_1 A = \begin{pmatrix} a & \mathbf{b} \\ \mathbf{0} & R' \end{pmatrix} = R$$

Since the product of unitary matrices is unitary, there exists Q unitary such that $Q^* A = R$ and so $A = QR$. ■ ►►

The QR algorithm is described in the following definition.

Definition 14.2.2 *The QR algorithm is the following. In the description of this algorithm, Q is unitary and R is upper triangular having nonnegative entries on the main diagonal. Starting with A an $n \times n$ matrix, form*

$$A_0 \equiv A = Q_1 R_1 \tag{14.6}$$

Then

$$A_1 \equiv R_1 Q_1. \tag{14.7}$$

In general given

$$A_k = R_k Q_k, \tag{14.8}$$

obtain A_{k+1} by

$$A_k = Q_{k+1} R_{k+1}, \ A_{k+1} = R_{k+1} Q_{k+1} \tag{14.9}$$

This algorithm was proposed by Francis in 1961. The sequence $\{A_k\}$ is the desired sequence of iterates. Now with the above definition of the algorithm, here are its properties. The next lemma shows each of the A_k is unitarily similar to A and the amazing thing about this algorithm is that often it becomes increasingly easy to find the eigenvalues of the A_k.

Lemma 14.2.3 *Let A be an $n \times n$ matrix and let the Q_k and R_k be as described in the algorithm. Then each A_k is unitarily similar to A and denoting by $Q^{(k)}$ the product $Q_1 Q_2 \cdots Q_k$ and $R^{(k)}$ the product $R_k R_{k-1} \cdots R_1$, it follows that*

$$A^k = Q^{(k)} R^{(k)}$$

(The matrix on the left is A raised to the k^{th} power.)

$$A = Q^{(k)} A_k Q^{(k)*},\ A_k = Q^{(k)*} A Q^{(k)}.$$

Proof: From the algorithm, $R_{k+1} = A_{k+1} Q_{k+1}^*$ and so

$$A_k = Q_{k+1} R_{k+1} = Q_{k+1} A_{k+1} Q_{k+1}^*$$

Now iterating this, it follows

$$A_{k-1} = Q_k A_k Q_k^* = Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^*$$

$$A_{k-2} = Q_{k-1} A_{k-1} Q_{k-1}^* = Q_{k-1} Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* Q_{k-1}^*$$

etc. Thus, after $k-2$ more iterations,

$$A = Q^{(k+1)} A_{k+1} Q^{(k+1)*}$$

The product of unitary matrices is unitary and so this proves the first claim of the lemma.

Now consider the part about A^k. From the algorithm, this is clearly true for $k = 1$. $(A^1 = QR)$ Suppose then that

$$A^k = Q_1 Q_2 \cdots Q_k R_k R_{k-1} \cdots R_1$$

What was just shown indicated

$$A = Q_1 Q_2 \cdots Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*$$

and now from the algorithm, $A_{k+1} = R_{k+1} Q_{k+1}$ and so

$$A = Q_1 Q_2 \cdots Q_{k+1} R_{k+1} Q_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*$$

Then

$$A^{k+1} = AA^k =$$

$$\overbrace{Q_1 Q_2 \cdots Q_{k+1} R_{k+1} Q_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*}^{A} Q_1 \cdots Q_k R_k R_{k-1} \cdots R_1$$

$$= Q_1 Q_2 \cdots Q_{k+1} R_{k+1} R_k R_{k-1} \cdots R_1 \equiv Q^{(k+1)} R^{(k+1)} \blacksquare$$

Here is another very interesting lemma.

Lemma 14.2.4 *Suppose $Q^{(k)}, Q$ are unitary and R_k is upper triangular such that the diagonal entries on R_k are all positive and*

$$Q = \lim_{k \to \infty} Q^{(k)} R_k$$

Then

$$\lim_{k \to \infty} Q^{(k)} = Q,\ \lim_{k \to \infty} R_k = I.$$

Also the QR factorization of A is unique whenever A^{-1} exists.

Proof: Let

$$Q = (\mathbf{q}_1, \cdots, \mathbf{q}_n), \; Q^{(k)} = \left(\mathbf{q}_1^k, \cdots, \mathbf{q}_n^k\right)$$

where the $\mathbf{q}$ are the columns. Also denote by r_{ij}^k the ij^{th} entry of R_k. Thus

$$Q^{(k)} R_k = \left(\mathbf{q}_1^k, \cdots, \mathbf{q}_n^k\right) \begin{pmatrix} r_{11}^k & & * \\ & \ddots & \\ 0 & & r_{nn}^k \end{pmatrix}$$

It follows

$$r_{11}^k \mathbf{q}_1^k \to \mathbf{q}_1$$

and so

$$r_{11}^k = \left| r_{11}^k \mathbf{q}_1^k \right| \to 1$$

Therefore,

$$\mathbf{q}_1^k \to \mathbf{q}_1.$$

Next consider the second column.

$$r_{12}^k \mathbf{q}_1^k + r_{22}^k \mathbf{q}_2^k \to \mathbf{q}_2$$

Taking the inner product of both sides with $\mathbf{q}_1^k$ it follows

$$\lim_{k\to\infty} r_{12}^k = \lim_{k\to\infty} \left(\mathbf{q}_2 \cdot \mathbf{q}_1^k\right) = (\mathbf{q}_2 \cdot \mathbf{q}_1) = 0.$$

Therefore,

$$\lim_{k\to\infty} r_{22}^k \mathbf{q}_2^k = \mathbf{q}_2$$

and since $r_{22}^k > 0$, it follows as in the first part that $r_{22}^k \to 1$. Hence

$$\lim_{k\to\infty} \mathbf{q}_2^k = \mathbf{q}_2.$$

Continuing this way, it follows

$$\lim_{k\to\infty} r_{ij}^k = 0$$

for all $i \neq j$ and

$$\lim_{k\to\infty} r_{jj}^k = 1, \; \lim_{k\to\infty} \mathbf{q}_j^k = \mathbf{q}_j.$$

Thus $R_k \to I$ and $Q^{(k)} \to Q$. This proves the first part of the lemma.

The second part follows immediately. If $QR = Q'R' = A$ where A^{-1} exists, then

$$Q^* Q' = R\left(R'\right)^{-1}$$

and I need to show both sides of the above are equal to I. The left side of the above is unitary and the right side is upper triangular having positive entries on the diagonal. This is because the inverse of such an upper triangular matrix having positive entries on the main diagonal is still upper triangular having positive entries on the main diagonal and the product of two such upper triangular matrices gives another of the same form having positive entries on the main diagonal. Suppose then that $Q = R$ where Q is unitary and R is upper triangular having positive entries on the main diagonal. Let $Q_k = Q$ and $R_k = R$. It follows

$$IR_k \to R = Q$$

and so from the first part, $R_k \to I$ but $R_k = R$ and so $R = I$. Thus applying this to $Q^* Q' = R(R')^{-1}$ yields both sides equal I. ■

A case of all this is of great interest. Suppose A has a largest eigenvalue λ which is real. Then A^n is of the form $\left(A^{n-1}\mathbf{a}_1, \cdots, A^{n-1}\mathbf{a}_n\right)$ and so likely each of these columns will be pointing roughly in the direction of an eigenvector of A which corresponds to this eigenvalue. Then when you do the QR factorization of this, it follows from the fact that R is upper triangular, that the first column of Q will be a multiple of $A^{n-1}\mathbf{a}_1$ and so will end up being roughly parallel to the eigenvector desired. Also this will require the entries below the top in the first column of $A_n = Q^T A Q$ will all be small because they will be of the form $\mathbf{q}_i^T A \mathbf{q}_1 \approxeq \lambda \mathbf{q}_i^T \mathbf{q}_1 = 0$. Therefore, A_n will be of the form

$$\begin{pmatrix} \lambda' & \mathbf{a} \\ \mathbf{e} & B \end{pmatrix}$$

where $\mathbf{e}$ is small. It follows that λ' will be close to λ and $\mathbf{q}_1$ will be close to an eigenvector for λ. Then if you like, you could do the same thing with the matrix B to obtain approximations for the other eigenvalues. Finally, you could use the shifted inverse power method to get more exact solutions.

14.2.2 The Case of Real Eigenvalues

With these lemmas, it is possible to prove that for the QR algorithm and certain conditions, the sequence A_k converges pointwise to an upper triangular matrix having the eigenvalues of A down the diagonal. I will assume all the matrices are real here.

This convergence won't always happen. Consider for example the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. You can verify quickly that the algorithm will return this matrix for each k. The problem here is that, although the matrix has the two eigenvalues $-1, 1$, they have the same absolute value. The QR algorithm works in somewhat the same way as the power method, exploiting differences in the size of the eigenvalues.

If A has all real eigenvalues and you are interested in finding these eigenvalues along with the corresponding eigenvectors, you could always consider $A + \lambda I$ instead where λ is sufficiently large and positive that $A + \lambda I$ has all positive eigenvalues. (Recall Gerschgorin's theorem.) Then if μ is an eigenvalue of $A + \lambda I$ with

$$(A + \lambda I)\mathbf{x} = \mu \mathbf{x}$$

then

$$A\mathbf{x} = (\mu - \lambda)\mathbf{x}$$

so to find the eigenvalues of A you just subtract λ from the eigenvalues of $A + \lambda I$. Thus there is no loss of generality in assuming at the outset that the eigenvalues of A are all positive. Here is the theorem. It involves a technical condition which will often hold. The proof presented here follows [27] and is a special case of that presented in this reference.

Before giving the proof, note that the product of upper triangular matrices is upper triangular. If they both have positive entries on the main diagonal so will the product. Furthermore, the inverse of an upper triangular matrix is upper triangular. I will use these simple facts without much comment whenever convenient.

Theorem 14.2.5 *Let A be a real matrix having eigenvalues*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$$

and let

$$A = SDS^{-1} \tag{14.10}$$

where

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

and suppose S^{-1} has an LU factorization. Then the matrices A_k in the QR algorithm described above converge to an upper triangular matrix T' having the eigenvalues of A, $\lambda_1, \cdots, \lambda_n$ descending on the main diagonal. The matrices $Q^{(k)}$ converge to Q', an orthogonal matrix which equals Q except for possibly having some columns multiplied by -1 for Q the unitary part of the QR factorization of S,

$$S = QR,$$

and

$$\lim_{k\to\infty} A_k = T' = Q'^T AQ'$$

Proof: From Lemma 14.2.3

$$A^k = Q^{(k)}R^{(k)} = SD^kS^{-1} \tag{14.11}$$

Let $S = QR$ where this is just a QR factorization which is known to exist and let $S^{-1} = LU$ which is assumed to exist. Thus

$$Q^{(k)}R^{(k)} = QRD^kLU \tag{14.12}$$

and so

$$Q^{(k)}R^{(k)} = QRD^kLU = QRD^kLD^{-k}D^kU$$

That matrix in the middle, D^kLD^{-k} satisfies

$$\left(D^kLD^{-k}\right)_{ij} = \lambda_i^k L_{ij}\lambda_j^{-k} \text{ for } j \leq i, \ 0 \text{ if } j > i.$$

Thus for $j < i$ the expression converges to 0 because $\lambda_j > \lambda_i$ when this happens. When $i = j$ it reduces to 1. Thus the matrix in the middle is of the form $I + E_k$ where $E_k \to 0$. Then it follows

$$A^k = Q^{(k)}R^{(k)} = QR\left(I + E_k\right)D^kU$$
$$= Q\left(I + RE_kR^{-1}\right)RD^kU \equiv Q\left(I + F_k\right)RD^kU$$

where $F_k \to 0$. Then let $I + F_k = Q_kR_k$ where this is another QR factorization. Then it reduces to

$$Q^{(k)}R^{(k)} = QQ_kR_kRD^kU$$

This looks really interesting because by Lemma 14.2.4 $Q_k \to I$ and $R_k \to I$ because $Q_kR_k = (I + F_k) \to I$. So it follows QQ_k is an orthogonal matrix converging to Q while

$$R_kRD^kU\left(R^{(k)}\right)^{-1}$$

is upper triangular, being the product of upper triangular matrices. Unfortunately, it is not known that the diagonal entries of this matrix are nonnegative because of the U. Let Λ be just like the identity matrix but having some of the ones replaced with -1 in such a way that ΛU is an upper triangular matrix having positive diagonal entries. Note $\Lambda^2 = I$ and also Λ commutes with a diagonal matrix. Thus

$$Q^{(k)} R^{(k)} = QQ_k R_k R D^k \Lambda^2 U = QQ_k R_k R \Lambda D^k (\Lambda U)$$

At this point, one does some inspired massaging to write the above in the form

$$\begin{aligned} & QQ_k \left(\Lambda D^k\right) \left[\left(\Lambda D^k\right)^{-1} R_k R \Lambda D^k\right] (\Lambda U) \\ = \ & Q \left(Q_k \Lambda\right) D^k \left[\left(\Lambda D^k\right)^{-1} R_k R \Lambda D^k\right] (\Lambda U) \\ = \ & Q \left(Q_k \Lambda\right) \overbrace{D^k \left[\left(\Lambda D^k\right)^{-1} R_k R \Lambda D^k\right] (\Lambda U)}^{\equiv G_k} \end{aligned}$$

Now I claim the middle matrix in $[\cdot]$ is upper triangular and has all positive entries on the diagonal. This is because it is an upper triangular matrix which is similar to the upper triangular matrix $R_k R$ and so it has the same eigenvalues (diagonal entries) as $R_k R$. Thus the matrix $G_k \equiv D^k \left[\left(\Lambda D^k\right)^{-1} R_k R \Lambda D^k\right] (\Lambda U)$ is upper triangular and has all positive entries on the diagonal. Multiply on the right by G_k^{-1} to get

$$Q^{(k)} R^{(k)} G_k^{-1} = QQ_k \Lambda \to Q'$$

where Q' is essentially equal to Q but might have some of the columns multiplied by -1. This is because $Q_k \to I$ and so $Q_k \Lambda \to \Lambda$. Now by Lemma 14.2.4, it follows

$$Q^{(k)} \to Q', \ R^{(k)} G_k^{-1} \to I.$$

It remains to verify A_k converges to an upper triangular matrix. Recall that from 14.11 and the definition below this ($S = QR$)

$$A = SDS^{-1} = (QR) D (QR)^{-1} = QRDR^{-1} Q^T = QTQ^T$$

Where T is an upper triangular matrix. This is because it is the product of upper triangular matrices R, D, R^{-1}. Thus $Q^T A Q = T$. If you replace Q with Q' in the above, it still results in an upper triangular matrix T' having the same diagonal entries as T. This is because

$$T = Q^T A Q = \left(Q' \Lambda\right)^T A \left(Q' \Lambda\right) = \Lambda Q'^T A Q' \Lambda$$

and considering the ii^{th} entry yields

$$\left(Q^T A Q\right)_{ii} \equiv \sum_{j,k} \Lambda_{ij} \left(Q'^T A Q'\right)_{jk} \Lambda_{ki} = \Lambda_{ii} \Lambda_{ii} \left(Q'^T A Q'\right)_{ii} = \left(Q'^T A Q'\right)_{ii}$$

Recall from Lemma 14.2.3, $A_k = Q^{(k)T} A Q^{(k)}$. Thus taking a limit and using the first part,

$$A_k = Q^{(k)T} A Q^{(k)} \to Q'^T A Q' = T'. \blacksquare$$

An easy case is for A symmetric. Recall Corollary 6.4.13. By this corollary, there exists an orthogonal (real unitary) matrix Q such that

$$Q^T A Q = D$$

where D is diagonal having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right.

Corollary 14.2.6 *Let A be a real symmetric $n \times n$ matrix having eigenvalues*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$$

and let Q be defined by

$$QDQ^T = A, \ D = Q^T AQ, \tag{14.13}$$

where Q is orthogonal and D is a diagonal matrix having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right. Let Q^T have an LU factorization. Then in the QR algorithm, the matrices $Q^{(k)}$ converge to Q' where Q' is the same as Q except having some columns multiplied by (-1). Thus the columns of Q' are eigenvectors of A. The matrices A_k converge to D.

Proof: This follows from Theorem 14.2.5. Here $S = Q, S^{-1} = Q^T$. Thus

$$Q = S = QR$$

and $R = I$. By Theorem 14.2.5 and Lemma 14.2.3,

$$A_k = Q^{(k)T} AQ^{(k)} \to Q'^T AQ' = Q^T AQ = D.$$

because formula 14.13 is unaffected by replacing Q with Q'. ■

When using the QR algorithm, it is not necessary to check technical condition about S^{-1} having an LU factorization. The algorithm delivers a sequence of matrices which are similar to the original one. If that sequence converges to an upper triangular matrix, then the algorithm worked. Furthermore, the technical condition is sufficient but not necessary. The algorithm will work even without the technical condition.

Example 14.2.7 *Find the eigenvalues and eigenvectors of the matrix*

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

It is a symmetric matrix but other than that, I just pulled it out of the air. By Lemma 14.2.3 it follows $A_k = Q^{(k)T} AQ^{(k)}$. And so to get to the answer quickly I could have the computer raise A to a power and then take the QR factorization of what results to get the k^{th} iteration using the above formula. Lets pick $k = 10$.

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{10} = \begin{pmatrix} 4.2273 \times 10^7 & 2.5959 \times 10^7 & 1.8611 \times 10^7 \\ 2.5959 \times 10^7 & 1.6072 \times 10^7 & 1.1506 \times 10^7 \\ 1.8611 \times 10^7 & 1.1506 \times 10^7 & 8.2396 \times 10^6 \end{pmatrix}$$

Now take QR factorization of this. The computer will do that also.

This yields

$$\begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix} \cdot$$
$$\begin{pmatrix} 5.2983 \times 10^7 & 3.2627 \times 10^7 & 2.338 \times 10^7 \\ 0 & 1.2172 \times 10^5 & 71946. \\ 0 & 0 & 277.03 \end{pmatrix}$$

Next it follows

$$A_{10} = \begin{pmatrix} .79785 & -.59912 & -6.6943\times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix}^T \cdot$$

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} .79785 & -.59912 & -6.6943\times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix}$$

and this equals

$$\begin{pmatrix} 6.0571 & 3.698\times 10^{-3} & 3.4346\times 10^{-5} \\ 3.698\times 10^{-3} & 3.2008 & -4.0643\times 10^{-4} \\ 3.4346\times 10^{-5} & -4.0643\times 10^{-4} & -.2579 \end{pmatrix}$$

By Gerschgorin's theorem, the eigenvalues are pretty close to the diagonal entries of the above matrix. Note I didn't use the theorem, just Lemma 14.2.3 and Gerschgorin's theorem to verify the eigenvalues are close to the above numbers. The eigenvectors are close to

$$\begin{pmatrix} .79785 \\ .48995 \\ .35126 \end{pmatrix}, \begin{pmatrix} -.59912 \\ .70912 \\ .37176 \end{pmatrix}, \begin{pmatrix} -6.6943\times 10^{-2} \\ -.50706 \\ .85931 \end{pmatrix}$$

Lets check one of these.

$$\left(\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix} - 6.0571 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} .79785 \\ .48995 \\ .35126 \end{pmatrix}$$

$$= \begin{pmatrix} -2.1972\times 10^{-3} \\ 2.5439\times 10^{-3} \\ 1.3931\times 10^{-3} \end{pmatrix} \approxeq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Now lets see how well the smallest approximate eigenvalue and eigenvector works.

$$\left(\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix} - (-.2579) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} -6.6943\times 10^{-2} \\ -.50706 \\ .85931 \end{pmatrix}$$

$$= \begin{pmatrix} 2.704\times 10^{-4} \\ -2.7377\times 10^{-4} \\ -1.3695\times 10^{-4} \end{pmatrix} \approxeq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

For practical purposes, this has found the eigenvalues and eigenvectors.

14.2.3 The QR Algorithm in the General Case

In the case where A has distinct positive eigenvalues it was shown above that under reasonable conditions related to a certain matrix having an LU factorization the QR algorithm produces a sequence of matrices $\{A_k\}$ which converges to an upper triangular matrix. What if A is just an $n \times n$ matrix having possibly complex eigenvalues but A is nondefective? What happens with the QR algorithm in this case? The short answer to this question is that the A_k of the algorithm **typically cannot converge**. However, this does not mean the algorithm is not useful in finding eigenvalues. It turns out the sequence of matrices $\{A_k\}$ have the appearance of a block upper triangular matrix for large k in the sense that the entries below the blocks on the main diagonal are small. Then looking at these blocks gives a way to approximate the eigenvalues. An important example of the concept of a block triangular matrix is the real Schur form for a matrix discussed in Theorem 6.4.7 but the concept as described here allows for any size block centered on the diagonal.

First it is important to note a simple fact about unitary diagonal matrices. In what follows Λ will denote a unitary matrix which is also a diagonal matrix. These matrices are just the identity matrix with some of the ones replaced with a number of the form $e^{i\theta}$ for some θ. The important property of multiplication of any matrix by Λ on either side is that it leaves all the zero entries the same and also preserves the absolute values of the other entries. Thus a block triangular matrix multiplied by Λ on either side is still block triangular. If the matrix is close to being block triangular this property of being close to a block triangular matrix is also preserved by multiplying on either side by Λ. Other patterns depending only on the size of the absolute value occurring in the matrix are also preserved by multiplying on either side by Λ. In other words, in looking for a pattern in a matrix, multiplication by Λ is irrelevant.

Now let A be an $n \times n$ matrix having real or complex entries. By Lemma 14.2.3 and the assumption that A is nondefective, there exists an invertible S,

$$A^k = Q^{(k)} R^{(k)} = S D^k S^{-1} \tag{14.14}$$

where

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

and by rearranging the columns of S, D can be made such that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| .$$

Assume S^{-1} has an LU factorization. Then

$$A^k = S D^k L U = S D^k L D^{-k} D^k U.$$

Consider the matrix in the middle, $D^k L D^{-k}$. The ij^{th} entry is of the form

$$\left(D^k L D^{-k} \right)_{ij} = \begin{cases} \lambda_i^k L_{ij} \lambda_j^{-k} \text{ if } j < i \\ 1 \text{ if } i = j \\ 0 \text{ if } j > i \end{cases}$$

and these all converge to 0 whenever $|\lambda_i| < |\lambda_j|$. Thus

$$D^k L D^{-k} = (L_k + E_k)$$

where L_k is a lower triangular matrix which has all ones down the diagonal and some subdiagonal terms of the form

$$\lambda_i^k L_{ij} \lambda_j^{-k} \tag{14.15}$$

for which $|\lambda_i| = |\lambda_j|$ while $E_k \to 0$. (Note the entries of L_k are all bounded independent of k but some may fail to converge.) Then

$$Q^{(k)}R^{(k)} = S\left(L_k + E_k\right)D^k U$$

Let

$$SL_k = Q_k R_k \tag{14.16}$$

where this is the QR factorization of SL_k. Then

$$\begin{aligned} Q^{(k)}R^{(k)} &= \left(Q_k R_k + SE_k\right)D^k U \\ &= Q_k\left(I + Q_k^* SE_k R_k^{-1}\right)R_k D^k U \\ &= Q_k\left(I + F_k\right)R_k D^k U \end{aligned}$$

where $F_k \to 0$. Let $I + F_k = Q_k' R_k'$. Then $Q^{(k)}R^{(k)} = Q_k Q_k' R_k' R_k D^k U$. By Lemma 14.2.4

$$Q_k' \to I \text{ and } R_k' \to I. \tag{14.17}$$

Now let Λ_k be a diagonal unitary matrix which has the property that $\Lambda_k^* D^k U$ is an upper triangular matrix which has all the diagonal entries positive. Then

$$Q^{(k)}R^{(k)} = Q_k Q_k' \Lambda_k \left(\Lambda_k^* R_k' R_k \Lambda_k\right)\Lambda_k^* D^k U$$

That matrix in the middle has all positive diagonal entries because it is itself an upper triangular matrix, being the product of such, and is similar to the matrix $R_k' R_k$ which is upper triangular with positive diagonal entries. By Lemma 14.2.4 again, this time using the uniqueness assertion,

$$Q^{(k)} = Q_k Q_k' \Lambda_k, \; R^{(k)} = \left(\Lambda_k^* R_k' R_k \Lambda_k\right)\Lambda_k^* D^k U$$

Note the term $Q_k Q_k' \Lambda_k$ must be real because the algorithm gives all $Q^{(k)}$ as real matrices. By 14.17 it follows that for k large enough $Q^{(k)} \approxeq Q_k \Lambda_k$ where $\approxeq$ means the two matrices are close. Recall $A_k = Q^{(k)T} A Q^{(k)}$ and so for large k,

$$A_k \approxeq \left(Q_k \Lambda_k\right)^* A \left(Q_k \Lambda_k\right) = \Lambda_k^* Q_k^* A Q_k \Lambda_k$$

As noted above, the form of $\Lambda_k^* Q_k^* A Q_k \Lambda_k$ in terms of which entries are large and small is not affected by the presence of Λ_k and Λ_k^*. Thus, in considering what form this is in, it suffices to consider $Q_k^* A Q_k$.

This could get pretty complicated but I will consider the case where

$$\text{if } \left|\lambda_i\right| = \left|\lambda_{i+1}\right|, \text{ then } \left|\lambda_{i+2}\right| < \left|\lambda_{i+1}\right|. \tag{14.18}$$

This is typical of the situation where the eigenvalues are all distinct and the matrix A is real so the eigenvalues occur as conjugate pairs. Then in this case, L_k above is lower triangular with some nonzero terms on the diagonal right below the main diagonal but zeros everywhere else. Thus maybe $\left(L_k\right)_{s+1,s} \neq 0$ Recall 14.16 which implies

$$Q_k = SL_k R_k^{-1} \tag{14.19}$$

where R_k^{-1} is upper triangular. Also recall from the definition of S in 14.14, it follows that $S^{-1}AS = D$. Thus the columns of S are eigenvectors of A, the i^{th} being an eigenvector for λ_i. Now from the form of L_k, it follows $L_k R_k^{-1}$ is a block upper triangular matrix denoted by T_B and so $Q_k = ST_B$. It follows from the above construction in 14.15 and the given assumption on the sizes of the eigenvalues, there are finitely many 2×2 blocks centered on the main diagonal along with possibly some diagonal entries. Therefore, for large k the matrix $A_k = Q^{(k)T} A Q^{(k)}$ is approximately of the same form as that of

$$Q_k^* A Q_k = T_B^{-1} S^{-1} A S T_B = T_B^{-1} D T_B$$

which is a block upper triangular matrix. As explained above, multiplication by the various diagonal unitary matrices does not affect this form. Therefore, for large k, A_k is approximately a block upper triangular matrix.

How would this change if the above assumption on the size of the eigenvalues were relaxed but the matrix was still nondefective with appropriate matrices having an LU factorization as above? It would mean the blocks on the diagonal would be larger. This immediately makes the problem more cumbersome to deal with. However, in the case that the eigenvalues of A are distinct, the above situation really is typical of what occurs and in any case can be quickly reduced to this case.

To see this, suppose condition 14.18 is violated and $\lambda_j, \cdots, \lambda_{j+p}$ are complex eigenvalues having nonzero imaginary parts such that each has the same absolute value but they are all distinct. Then let $\mu > 0$ and consider the matrix $A + \mu I$. Thus the corresponding eigenvalues of $A + \mu I$ are $\lambda_j + \mu, \cdots, \lambda_{j+p} + \mu$. A short computation shows $|\lambda_j + \mu|, \cdots, |\lambda_{j+p} + \mu|$ are all distinct and so the above situation of 14.18 is obtained. Of course, if there are repeated eigenvalues, it may not be possible to reduce to the case above and you would end up with large blocks on the main diagonal which could be difficult to deal with.

So how do you identify the eigenvalues? You know A_k and behold that it is close to a block upper triangular matrix T_B'. You know A_k is also similar to A. Therefore, T_B' has eigenvalues which are close to the eigenvalues of A_k and hence those of A provided k is sufficiently large. See Theorem 6.9.2 which depends on complex analysis or the exercise on Page 184 which gives another way to see this. Thus you find the eigenvalues of this block triangular matrix T_B' and assert that these are good approximations of the eigenvalues of A_k and hence to those of A. How do you find the eigenvalues of a block triangular matrix? This is easy from Lemma 6.4.6. Say

$$T_B' = \begin{pmatrix} B_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & B_m \end{pmatrix}$$

Then forming $\lambda I - T_B'$ and taking the determinant, it follows from Lemma 6.4.6 this equals

$$\prod_{j=1}^{m} \det \left(\lambda I_j - B_j \right)$$

and so all you have to do is take the union of the eigenvalues for each B_j. In the case emphasized here this is very easy because these blocks are just 2×2 matrices.

How do you identify approximate eigenvectors from this? First try to find the approximate eigenvectors for A_k. Pick an approximate eigenvalue λ, an exact eigenvalue for T_B'. Then find $\mathbf{v}$ solving $T_B' \mathbf{v} = \lambda \mathbf{v}$. It follows since T_B' is close to A_k that $A_k \mathbf{v} \approxeq \lambda \mathbf{v}$ and so

$$Q^{(k)} A Q^{(k)T} \mathbf{v} = A_k \mathbf{v} \approxeq \lambda \mathbf{v}$$

Hence

$$A Q^{(k)T} \mathbf{v} \approxeq \lambda Q^{(k)T} \mathbf{v}$$

and so $Q^{(k)T} \mathbf{v}$ is an approximation to the eigenvector which goes with the eigenvalue of A which is close to λ.

Example 14.2.8 *Here is a matrix.*

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

It happens that the eigenvalues of this matrix are $1, 1+i, 1-i$. *Lets apply the* QR *algorithm as if the eigenvalues were not known.*

Applying the QR algorithm to this matrix yields the following sequence of matrices.

$$A_1 = \begin{pmatrix} 1.2353 & 1.9412 & 4.3657 \\ -.39215 & 1.5425 & 5.3886 \times 10^{-2} \\ -.16169 & -.18864 & .22222 \end{pmatrix}$$

$$\vdots$$

$$A_{12} = \begin{pmatrix} 9.1772 \times 10^{-2} & .63089 & -2.0398 \\ -2.8556 & 1.9082 & -3.1043 \\ 1.0786 \times 10^{-2} & 3.4614 \times 10^{-4} & 1.0 \end{pmatrix}$$

At this point the bottom two terms on the left part of the bottom row are both very small so it appears the real eigenvalue is near 1.0. The complex eigenvalues are obtained from solving

$$\det\left(\lambda\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 9.1772 \times 10^{-2} & .63089 \\ -2.8556 & 1.9082 \end{pmatrix}\right) = 0$$

This yields

$$\lambda = 1.0 - .98828i,\ 1.0 + .98828i$$

Example 14.2.9 *The equation* $x^4+x^3+4x^2+x-2=0$ *has exactly two real solutions. You can see this by graphing it. However, the rational root theorem from algebra shows neither of these solutions are rational. Also, graphing it does not yield any information about the complex solutions. Lets use the* QR *algorithm to approximate all the solutions, real and complex.*

A matrix whose characteristic polynomial is the given polynomial is

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

Using the QR algorithm yields the following sequence of iterates for A_k

$$A_1 = \begin{pmatrix} .99999 & -2.5927 & -1.7588 & -1.2978 \\ 2.1213 & -1.7778 & -1.6042 & -.99415 \\ 0 & .34246 & -.32749 & -.91799 \\ 0 & 0 & -.44659 & .10526 \end{pmatrix}$$

$$\vdots$$

$$A_9 = \begin{pmatrix} -.834\,12 & -4.168\,2 & -1.939 & -.778\,3 \\ 1.05 & .145\,14 & .217\,1 & 2.547\,4 \times 10^{-2} \\ 0 & 4.026\,4 \times 10^{-4} & -.850\,29 & -.616\,08 \\ 0 & 0 & -1.826\,3 \times 10^{-2} & .539\,39 \end{pmatrix}$$

Now this is similar to A and the eigenvalues are close to the eigenvalues obtained from the two blocks on the diagonal,

$$\begin{pmatrix} -.834\,12 & -4.168\,2 \\ 1.05 & .145\,14 \end{pmatrix}, \begin{pmatrix} -.850\,29 & -.616\,08 \\ -1.826\,3 \times 10^{-2} & .539\,39 \end{pmatrix}$$

since $4.026\,4 \times 10^{-4}$ is small. After routine computations involving the quadratic formula, these are seen to be

$$-.858\,34,\ .547\,44,\ -.344\,49 - 2.033\,9i,\ -.344\,49 + 2.033\,9i$$

When these are plugged in to the polynomial equation, you see that each is close to being a solution of the equation.

It seems like most of the attention to the QR algorithm has to do with finding ways to get it to "converge" faster. Great and marvelous are the clever tricks which have been proposed to do this but my intent is to present the basic ideas, not to go in to the numerous refinements of this algorithm. However, there is one thing which is usually done. It involves reducing to the case of an upper Hessenberg matrix which is one which is zero below the main sub diagonal. Every matrix is unitarily similar to one of these.

Let A be an invertible $n \times n$ matrix. Let Q_1' be a unitary matrix

$$Q_1' \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} \sqrt{\sum_{j=2}^{n} |a_{j1}|^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv \begin{pmatrix} a \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The vector Q_1' is multiplying is just the bottom $n-1$ entries of the first column of A. Then let Q_1 be

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1' \end{pmatrix}$$

It follows

$$Q_1 A Q_1^* = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1' \end{pmatrix} A Q_1^* = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a & & & \\ \vdots & & A_1' & \\ 0 & & & \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1'^* \end{pmatrix}$$

$$= \begin{pmatrix} * & * & \cdots & * \\ a & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

Now let Q_2' be the $n-2 \times n-2$ matrix which does to the first column of A_1 the same sort of thing that the $n-1 \times n-1$ matrix Q_1' did to the first column of A. Let

$$Q_2 \equiv \begin{pmatrix} I & 0 \\ 0 & Q_2' \end{pmatrix}$$

where I is the 2×2 identity. Then applying block multiplication,

$$Q_2 Q_1 A Q_1^* Q_2^* = \begin{pmatrix} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$$

where A_2 is now an $n-2 \times n-2$ matrix. Continuing this way you eventually get a unitary matrix Q which is a product of those discussed above such that

$$QAQ^T = \begin{pmatrix} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & * & & \vdots \\ \vdots & \vdots & \ddots & \ddots & * \\ 0 & 0 & & * & * \end{pmatrix}$$

This matrix equals zero below the subdiagonal. It is called an upper Hessenberg matrix.

It happens that in the QR algorithm, if A_k is upper Hessenberg, so is A_{k+1}. To see this, note that 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 as shown before,

$$A_{k+1} = R_k A_k R_k^{-1}$$

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.

Since upper Hessenberg matrices stay that way in the algorithm and it is closer to being upper triangular, it is reasonable to suppose the QR algorithm will yield good results more quickly for this upper Hessenberg matrix than for the original matrix. This would be especially true if the matrix is good sized. The other important thing to observe is that, starting with an upper Hessenberg matrix, the algorithm will restrict the size of the blocks which occur to being 2×2 blocks which are easy to deal with. These blocks allow you to identify the complex roots.

14.3 Exercises

In these exercises which call for a computation, don't waste time on them unless you use a computer or calculator which can raise matrices to powers and take QR factorizations.

1. In Example 14.1.10 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

2. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   In this case the exact eigenvalues are $\pm\sqrt{3}, 6$. Compare with the exact answers.

3. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   The exact eigenvalues are $2, 4 + \sqrt{15}, 4 - \sqrt{15}$. Compare your numerical results with the exact values. Is it much fun to compute the exact eigenvectors?

4. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

5. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

6. Consider the matrix $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix}$ and the vector $(1, 1, 1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

7. Consider the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix}$ and the vector $(1, 1, 1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

8. Consider the matrix $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$ and the vector $(1, 1, 1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

9. Using Gerschgorin's theorem, find upper and lower bounds for the eigenvalues of $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$.

10. Tell how to find a matrix whose characteristic polynomial is a given monic polynomial. This is called a companion matrix. Find the roots of the polynomial $x^3 + 7x^2 + 3x + 7$.

11. Find the roots to $x^4 + 3x^3 + 4x^2 + x + 1$. It has two complex roots.

12. Suppose A is a real symmetric matrix and the technique of reducing to an upper Hessenberg matrix is followed. Show the resulting upper Hessenberg matrix is actually equal to 0 on the top as well as the bottom.

Appendix A

Matrix Calculator on the Web

A.1 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/ To get to it, you can use the address or google matrix calculator.

When you go to this site, 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 start the next row. After entering the matrix, you select what you want it to do. You will see that it also solves systems of equations.

Appendix B

Positive Matrices

Earlier theorems about Markov matrices were presented. These were matrices in which all the entries were nonnegative and either the columns or the rows added to 1. It turns out that many of the theorems presented can be generalized to positive matrices. When this is done, the resulting theory is mainly due to Perron and Frobenius. I will give an introduction to this theory here following Karlin and Taylor [19].

Definition B.0.1 *For A a matrix or vector, the notation, $A >> 0$ will mean every entry of A is positive. By $A > 0$ is meant that every entry is nonnegative and at least one is positive. By $A \geq 0$ is meant that every entry is nonnegative. Thus the matrix or vector consisting only of zeros is ≥ 0. An expression like $A >> B$ will mean $A - B >> 0$ with similar modifications for $>$ and $\geq$.*

For the sake of this section only, define the following for $\mathbf{x} = (x_1, \cdots, x_n)^T$, a vector.

$$|\mathbf{x}| \equiv (|x_1|, \cdots, |x_n|)^T .$$

Thus $|\mathbf{x}|$ is the vector which results by replacing each entry of $\mathbf{x}$ with its absolute value[1]*. Also define for $\mathbf{x} \in \mathbb{C}^n$,*

$$||\mathbf{x}||_1 \equiv \sum_k |x_k| .$$

Lemma B.0.2 *Let $A >> 0$ and let $\mathbf{x} > \mathbf{0}$. Then $A\mathbf{x} >> \mathbf{0}$.*

Proof: $(A\mathbf{x})_i = \sum_j A_{ij} x_j > 0$ because all the $A_{ij} > 0$ and at least one $x_j > 0$.

Lemma B.0.3 *Let $A >> 0$. Define*

$$S \equiv \{\lambda : A\mathbf{x} > \lambda\mathbf{x} \text{ for some } \mathbf{x} >> \mathbf{0}\} ,$$

and let

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\} .$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\} .$$

Then

$$\sup (S) = \sup (S_1) .$$

Proof: Let $\lambda \in S$. Then there exists $\mathbf{x} >> \mathbf{0}$ such that $A\mathbf{x} > \lambda\mathbf{x}$. Consider $\mathbf{y} \equiv \mathbf{x}/ ||\mathbf{x}||_1$. Then $||\mathbf{y}||_1 = 1$ and $A\mathbf{y} > \lambda\mathbf{y}$. Therefore, $\lambda \in S_1$ and so $S \subseteq S_1$. Therefore, $\sup (S) \leq \sup (S_1)$.

Now let $\lambda \in S_1$. Then there exists $\mathbf{x} \geq \mathbf{0}$ such that $||\mathbf{x}||_1 = 1$ so $\mathbf{x} > \mathbf{0}$ and $A\mathbf{x} > \lambda\mathbf{x}$. Letting $\mathbf{y} \equiv A\mathbf{x}$, it follows from Lemma B.0.2 that $A\mathbf{y} >> \lambda\mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. Thus $\lambda \in S$ and so $S_1 \subseteq S$ which shows that $\sup (S_1) \leq \sup (S)$. ■

This lemma is significant because the set, $\{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\} \equiv K$ is a compact set in $\mathbb{R}^n$. Define

$$\lambda_0 \equiv \sup (S) = \sup (S_1) . \tag{2.1}$$

The following theorem is due to Perron.

[1] This notation is just about the most abominable thing imaginable because it is the same notation but entirely different meaning than the norm. However, it saves space in the presentation of this theory of positive matrices and avoids the use of new symbols. Please forget about it when you leave this section.

Theorem B.0.4 *Let $A >> 0$ be an $n \times n$ matrix and let λ_0 be given in 2.1. Then*

1. *$\lambda_0 > 0$ and there exists $\mathbf{x}_0 >> \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$ so λ_0 is an eigenvalue for A.*
2. *If $A\mathbf{x} = \mu \mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq \lambda_0$. Then $|\mu| < \lambda_0$.*
3. *The eigenspace for λ_0 has dimension 1.*

Proof: To see $\lambda_0 > 0$, consider the vector, $\mathbf{e} \equiv (1, \cdots, 1)^T$. Then

$$(A\mathbf{e})_i = \sum_j A_{ij} > 0$$

and so λ_0 is at least as large as

$$\min_i \sum_j A_{ij}.$$

Let $\{\lambda_k\}$ be an increasing sequence of numbers from S_1 converging to λ_0. Letting $\mathbf{x}_k$ be the vector from K which occurs in the definition of S_1, these vectors are in a compact set. Therefore, there exists a subsequence, still denoted by $\mathbf{x}_k$ such that $\mathbf{x}_k \to \mathbf{x}_0 \in K$ and $\lambda_k \to \lambda_0$. Then passing to the limit,

$$A\mathbf{x}_0 \geq \lambda_0 \mathbf{x}_0, \ \mathbf{x}_0 > \mathbf{0}.$$

If $A\mathbf{x}_0 > \lambda_0 \mathbf{x}_0$, then letting $\mathbf{y} \equiv A\mathbf{x}_0$, it follows from Lemma B.0.2 that $A\mathbf{y} >> \lambda_0 \mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. But this contradicts the definition of λ_0 as the supremum of the elements of S because since $A\mathbf{y} >> \lambda_0 \mathbf{y}$, it follows $A\mathbf{y} >> (\lambda_0 + \varepsilon)\mathbf{y}$ for ε a small positive number. Therefore, $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$. It remains to verify that $\mathbf{x}_0 >> \mathbf{0}$. But this follows immediately from

$$0 < \sum_j A_{ij} x_{0j} = (A\mathbf{x}_0)_i = \lambda_0 x_{0i}.$$

This proves 1.

Next suppose $A\mathbf{x} = \mu \mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Then $|A\mathbf{x}| = |\mu| |\mathbf{x}|$. But this implies $A|\mathbf{x}| \geq |\mu| |\mathbf{x}|$. (See the above abominable definition of $|\mathbf{x}|$.)

Case 1: $|\mathbf{x}| \neq \mathbf{x}$ and $|\mathbf{x}| \neq -\mathbf{x}$.

In this case, $A|\mathbf{x}| > |A\mathbf{x}| = |\mu| |\mathbf{x}|$ and letting $\mathbf{y} = A|\mathbf{x}|$, it follows $\mathbf{y} >> \mathbf{0}$ and $A\mathbf{y} >> |\mu| \mathbf{y}$ which shows $A\mathbf{y} >> (|\mu| + \varepsilon)\mathbf{y}$ for sufficiently small positive ε and verifies $|\mu| < \lambda_0$.

Case 2: $|\mathbf{x}| = \mathbf{x}$ or $|\mathbf{x}| = -\mathbf{x}$

In this case, the entries of $\mathbf{x}$ are all real and have the same sign. Therefore, $A|\mathbf{x}| = |A\mathbf{x}| = |\mu| |\mathbf{x}|$. Now let $\mathbf{y} \equiv |\mathbf{x}| / ||\mathbf{x}||_1$. Then $A\mathbf{y} = |\mu| \mathbf{y}$ and so $|\mu| \in S_1$ showing that $|\mu| \leq \lambda_0$. But also, the fact the entries of $\mathbf{x}$ all have the same sign shows $\mu = |\mu|$ and so $\mu \in S_1$. Since $\mu \neq \lambda_0$, it must be that $\mu = |\mu| < \lambda_0$. This proves 2.

It remains to verify 3. Suppose then that $A\mathbf{y} = \lambda_0 \mathbf{y}$ and for all scalars $\alpha, \alpha \mathbf{x}_0 \neq \mathbf{y}$. Then

$$A \operatorname{Re} \mathbf{y} = \lambda_0 \operatorname{Re} \mathbf{y}, A \operatorname{Im} \mathbf{y} = \lambda_0 \operatorname{Im} \mathbf{y}.$$

If $\operatorname{Re} \mathbf{y} = \alpha_1 \mathbf{x}_0$ and $\operatorname{Im} \mathbf{y} = \alpha_2 \mathbf{x}_0$ for real numbers, α_i,then $\mathbf{y} = (\alpha_1 + i\alpha_2)\mathbf{x}_0$ and it is assumed this does not happen. Therefore, either

$$t \operatorname{Re} \mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}$$

or

$$t \operatorname{Im} \mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}.$$

Assume the first holds. Then varying $t \in \mathbb{R}$, there exists a value of t such that $\mathbf{x}_0 + t \operatorname{Re} \mathbf{y} > \mathbf{0}$ but it is not the case that $\mathbf{x}_0 + t \operatorname{Re} \mathbf{y} >> 0$. Then $A(\mathbf{x}_0 + t \operatorname{Re} \mathbf{y}) >> 0$ by Lemma B.0.2. But this implies $\lambda_0 (\mathbf{x}_0 + t \operatorname{Re} \mathbf{y}) >> 0$ which is a contradiction. Hence there exist real numbers, α_1 and α_2 such that $\operatorname{Re} \mathbf{y} = \alpha_1 \mathbf{x}_0$ and $\operatorname{Im} \mathbf{y} = \alpha_2 \mathbf{x}_0$ showing that $\mathbf{y} = (\alpha_1 + i\alpha_2) \mathbf{x}_0$. This proves 3.

It is possible to obtain a simple corollary to the above theorem.

Corollary B.0.5 *If $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$, then all the conclusions of the above theorem hold.*

Proof: There exists $\mu_0 > 0$ such that $A^m \mathbf{y}_0 = \mu_0 \mathbf{y}_0$ for $\mathbf{y}_0 >> 0$ by Theorem B.0.4 and

$$\mu_0 = \sup \{\mu : A^m \mathbf{x} \geq \mu \mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Let $\lambda_0^m = \mu_0$. Then

$$(A - \lambda_0 I) \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right) \mathbf{y}_0 = (A^m - \lambda_0^m I) \mathbf{y}_0 = \mathbf{0}$$

and so letting $\mathbf{x}_0 \equiv \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right) \mathbf{y}_0$, it follows $\mathbf{x}_0 >> 0$ and $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$.

Suppose now that $A\mathbf{x} = \mu \mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Suppose $|\mu| \geq \lambda_0$. Multiplying both sides by A, it follows $A^m \mathbf{x} = \mu^m \mathbf{x}$ and $|\mu^m| = |\mu|^m \geq \lambda_0^m = \mu_0$ and so from Theorem B.0.4, since $|\mu^m| \geq \mu_0$, and μ^m is an eigenvalue of A^m, it follows that $\mu^m = \mu_0$. But by Theorem B.0.4 again, this implies $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c and hence $A\mathbf{y}_0 = \mu \mathbf{y}_0$. Since $\mathbf{y}_0 >> \mathbf{0}$, it follows $\mu \geq 0$ and so $\mu = \lambda_0$, a contradiction. Therefore, $|\mu| < \lambda_0$.

Finally, if $A\mathbf{x} = \lambda_0 \mathbf{x}$, then $A^m \mathbf{x} = \lambda_0^m \mathbf{x}$ and so $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c. Consequently,

$$\begin{aligned} \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right) \mathbf{x} &= c \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right) \mathbf{y}_0 \\ &= c\mathbf{x}_0. \end{aligned}$$

Hence

$$m \lambda_0^{m-1} \mathbf{x} = c \mathbf{x}_0$$

which shows the dimension of the eigenspace for λ_0 is one. ■

The following corollary is an extremely interesting convergence result involving the powers of positive matrices.

Corollary B.0.6 *Let $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$. Then for λ_0 given in 2.1, there exists a rank one matrix P such that $\lim_{m \to \infty} \left|\left| \left(\frac{A}{\lambda_0}\right)^m - P \right|\right| = 0$.*

Proof: Considering A^T, and the fact that A and A^T have the same eigenvalues, Corollary B.0.5 implies the existence of a vector, $\mathbf{v} >> \mathbf{0}$ such that

$$A^T \mathbf{v} = \lambda_0 \mathbf{v}.$$

Also let $\mathbf{x}_0$ denote the vector such that $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$ with $\mathbf{x}_0 >> \mathbf{0}$. First note that $\mathbf{x}_0^T \mathbf{v} > 0$ because both these vectors have all entries positive. Therefore, $\mathbf{v}$ may be scaled such that

$$\mathbf{v}^T \mathbf{x}_0 = \mathbf{x}_0^T \mathbf{v} = 1. \tag{2.2}$$

Define

$$P \equiv \mathbf{x}_0 \mathbf{v}^T.$$

Thanks to 2.2,

$$\frac{A}{\lambda_0}P = \mathbf{x}_0\mathbf{v}^T = P,\; P\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T = P, \tag{2.3}$$

and

$$P^2 = \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0\mathbf{v}^T = \mathbf{v}^T\mathbf{x}_0 = P. \tag{2.4}$$

Therefore,

$$\begin{aligned}\left(\frac{A}{\lambda_0} - P\right)^2 &= \left(\frac{A}{\lambda_0}\right)^2 - 2\left(\frac{A}{\lambda_0}\right)P + P^2 \\ &= \left(\frac{A}{\lambda_0}\right)^2 - P.\end{aligned}$$

Continuing this way, using 2.3 repeatedly, it follows

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m = \left(\frac{A}{\lambda_0}\right)^m - P. \tag{2.5}$$

The eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are of interest because it is powers of this matrix which determine the convergence of $\left(\frac{A}{\lambda_0}\right)^m$ to P. Therefore, let μ be a nonzero eigenvalue of this matrix. Thus

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x} = \mu\mathbf{x} \tag{2.6}$$

for $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq 0$. Applying P to both sides and using the second formula of 2.3 yields

$$\mathbf{0} = (P - P)\,\mathbf{x} = \left(P\left(\frac{A}{\lambda_0}\right) - P^2\right)\mathbf{x} = \mu P\mathbf{x}.$$

But since $P\mathbf{x} = \mathbf{0}$, it follows from 2.6 that

$$A\mathbf{x} = \lambda_0\mu\mathbf{x}$$

which implies $\lambda_0\mu$ is an eigenvalue of A. Therefore, by Corollary B.0.5 it follows that either $\lambda_0\mu = \lambda_0$ in which case $\mu = 1$, or $\lambda_0\,|\mu| < \lambda_0$ which implies $|\mu| < 1$. But if $\mu = 1$, then $\mathbf{x}$ is a multiple of $\mathbf{x}_0$ and 2.6 would yield

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x}_0 = \mathbf{x}_0$$

which says $\mathbf{x}_0 - \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0 = \mathbf{x}_0$ and so by 2.2, $\mathbf{x}_0 = \mathbf{0}$ contrary to the property that $\mathbf{x}_0 >> \mathbf{0}$. Therefore, $|\mu| < 1$ and so this has shown that the absolute values of all eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are less than 1. By Gelfand's theorem, Theorem 13.3.3, it follows

$$\left\|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right\|^{1/m} < r < 1$$

whenever m is large enough. Now by 2.5 this yields

$$\left\|\left(\frac{A}{\lambda_0}\right)^m - P\right\| = \left\|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right\| \leq r^m$$

whenever m is large enough. It follows

$$\lim_{m\to\infty}\left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = 0$$

as claimed.

What about the case when $A > 0$ but maybe it is not the case that $A >> 0$? As before,

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\}.$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and

$$\lambda_0 \equiv \sup(S_1) \tag{2.7}$$

Theorem B.0.7 *Let $A > 0$ and let λ_0 be defined in 2.7. Then there exists $\mathbf{x}_0 > \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$.*

Proof: Let E consist of the matrix which has a one in every entry. Then from Theorem B.0.4 it follows there exists $\mathbf{x}_\delta >> \mathbf{0}$, $||\mathbf{x}_\delta||_1 = 1$, such that $(A + \delta E)\mathbf{x}_\delta = \lambda_{0\delta}\mathbf{x}_\delta$ where

$$\lambda_{0\delta} \equiv \sup\{\lambda : (A + \delta E)\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Now if $\alpha < \delta$

$$\{\lambda : (A + \alpha E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\} \subseteq$$
$$\{\lambda : (A + \delta E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and so $\lambda_{0\delta} \geq \lambda_{0\alpha}$ because $\lambda_{0\delta}$ is the sup of the second set and $\lambda_{0\alpha}$ is the sup of the first. It follows the limit, $\lambda_1 \equiv \lim_{\delta\to 0+}\lambda_{0\delta}$ exists. Taking a subsequence and using the compactness of K, there exists a subsequence, still denoted by δ such that as $\delta \to 0$, $\mathbf{x}_\delta \to \mathbf{x} \in K$. Therefore,

$$A\mathbf{x} = \lambda_1\mathbf{x}$$

and so, in particular, $A\mathbf{x} \geq \lambda_1\mathbf{x}$ and so $\lambda_1 \leq \lambda_0$. But also, if $\lambda \leq \lambda_0$,

$$\lambda\mathbf{x} \leq A\mathbf{x} < (A + \delta E)\mathbf{x}$$

showing that $\lambda_{0\delta} \geq \lambda$ for all such λ. But then $\lambda_{0\delta} \geq \lambda_0$ also. Hence $\lambda_1 \geq \lambda_0$, showing these two numbers are the same. Hence $A\mathbf{x} = \lambda_0\mathbf{x}$. ■

If $A^m >> 0$ for some m and $A > 0$, it follows that the dimension of the eigenspace for λ_0 is one and that the absolute value of every other eigenvalue of A is less than λ_0. If it is only assumed that $A > 0$, not necessarily $>> 0$, this is no longer true. However, there is something which is very interesting which can be said. First here is an interesting lemma.

Lemma B.0.8 *Let M be a matrix of the form*

$$M = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix}$$

or

$$M = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$

where A is an $r \times r$ matrix and C is an $(n-r) \times (n-r)$ matrix. Then $\det(M) = \det(A)\det(B)$ and $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Proof: To verify the claim about the determinants, note

$$\begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}$$

Therefore,

$$\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}.$$

But it is clear from the method of Laplace expansion that

$$\det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} = \det A$$

and from the multilinear properties of the determinant and row operations that

$$\det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det C.$$

The case where M is upper block triangular is similar.

This immediately implies $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Theorem B.0.9 *Let $A > 0$ and let λ_0 be given in 2.7. If λ is an eigenvalue for A such that $|\lambda| = \lambda_0$, then λ/λ_0 is a root of unity. Thus $(\lambda/\lambda_0)^m = 1$ for some $m \in \mathbb{N}$.*

Proof: Applying Theorem B.0.7 to A^T, there exists $\mathbf{v} > \mathbf{0}$ such that $A^T\mathbf{v} = \lambda_0 \mathbf{v}$. In the first part of the argument it is assumed $\mathbf{v} >> \mathbf{0}$. Now suppose $A\mathbf{x} = \lambda\mathbf{x}, \mathbf{x} \neq \mathbf{0}$ and that $|\lambda| = \lambda_0$. Then

$$A|\mathbf{x}| \geq |\lambda| |\mathbf{x}| = \lambda_0 |\mathbf{x}|$$

and it follows that if $A|\mathbf{x}| > |\lambda| |\mathbf{x}|$, then since $\mathbf{v} >> \mathbf{0}$,

$$\lambda_0 (\mathbf{v}, |\mathbf{x}|) < (\mathbf{v}, A|\mathbf{x}|) = \left(A^T\mathbf{v}, |\mathbf{x}|\right) = \lambda_0 (\mathbf{v}, |\mathbf{x}|),$$

a contradiction. Therefore,

$$A|\mathbf{x}| = \lambda_0 |\mathbf{x}|. \tag{2.8}$$

It follows that

$$\left|\sum_j A_{ij} x_j\right| = \lambda_0 |\mathbf{x}_i| = \sum_j A_{ij} |x_j|$$

and so the complex numbers,

$$A_{ij}x_j, \ A_{ik}x_k$$

must have the same argument for every k, j because equality holds in the triangle inequality. Therefore, there exists a complex number, μ_i such that

$$A_{ij}x_j = \mu_i A_{ij} |x_j| \tag{2.9}$$

and so, letting $r \in \mathbb{N}$,

$$A_{ij}x_j\mu_j^r = \mu_i A_{ij} |x_j| \mu_j^r.$$

Summing on j yields

$$\sum_j A_{ij}x_j\mu_j^r = \mu_i \sum_j A_{ij} |x_j| \mu_j^r. \tag{2.10}$$

Also, summing 2.9 on j and using that λ is an eigenvalue for $\mathbf{x}$, it follows from 2.8 that

$$\lambda x_i = \sum_j A_{ij} x_j = \mu_i \sum_j A_{ij} |x_j| = \mu_i \lambda_0 |x_i| . \tag{2.11}$$

From 2.10 and 2.11,

$$\begin{aligned}
\sum_j A_{ij} x_j \mu_j^r &= \mu_i \sum_j A_{ij} |x_j| \mu_j^r \\
&= \mu_i \sum_j A_{ij} \overbrace{\mu_j |x_j|}^{\text{see } 2.11} \mu_j^{r-1} \\
&= \mu_i \sum_j A_{ij} \left(\frac{\lambda}{\lambda_0} \right) x_j \mu_j^{r-1} \\
&= \mu_i \left(\frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} x_j \mu_j^{r-1}
\end{aligned}$$

Now from 2.10 with r replaced by $r-1$, this equals

$$\begin{aligned}
\mu_i^2 \left(\frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} |x_j| \mu_j^{r-1} &= \mu_i^2 \left(\frac{\lambda}{\lambda_0} \right) \sum_j A_{ij} \mu_j |x_j| \mu_j^{r-2} \\
&= \mu_i^2 \left(\frac{\lambda}{\lambda_0} \right)^2 \sum_j A_{ij} x_j \mu_j^{r-2}.
\end{aligned}$$

Continuing this way,

$$\sum_j A_{ij} x_j \mu_j^r = \mu_i^k \left(\frac{\lambda}{\lambda_0} \right)^k \sum_j A_{ij} x_j \mu_j^{r-k}$$

and eventually, this shows

$$\begin{aligned}
\sum_j A_{ij} x_j \mu_j^r &= \mu_i^r \left(\frac{\lambda}{\lambda_0} \right)^r \sum_j A_{ij} x_j \\
&= \left(\frac{\lambda}{\lambda_0} \right)^r \lambda \left(x_i \mu_i^r \right)
\end{aligned}$$

and this says $\left(\frac{\lambda}{\lambda_0} \right)^{r+1}$ is an eigenvalue for $\left(\frac{A}{\lambda_0} \right)$ with the eigenvector being

$$\left(x_1 \mu_1^r, \cdots, x_n \mu_n^r \right)^T .$$

Now recall that $r \in \mathbb{N}$ was arbitrary and so this has shown that $\left(\frac{\lambda}{\lambda_0} \right)^2, \left(\frac{\lambda}{\lambda_0} \right)^3, \left(\frac{\lambda}{\lambda_0} \right)^4, \cdots$ are each eigenvalues of $\left(\frac{A}{\lambda_0} \right)$ which has only finitely many and hence this sequence must repeat. Therefore, $\left(\frac{\lambda}{\lambda_0} \right)$ is a root of unity as claimed. This proves the theorem in the case that $\mathbf{v} >> \mathbf{0}$.

Now it is necessary to consider the case where $\mathbf{v} > \mathbf{0}$ but it is not the case that $\mathbf{v} >> \mathbf{0}$.

Then in this case, there exists a permutation matrix P such that

$$P\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix} \equiv \mathbf{v}_1$$

Then

$$\lambda_0 \mathbf{v} = A^T \mathbf{v} = A^T P \mathbf{v}_1.$$

Therefore,

$$\lambda_0 \mathbf{v}_1 = P A^T P \mathbf{v}_1 = G \mathbf{v}_1$$

Now $P^2 = I$ because it is a permutation matrix. Therefore, the matrix $G \equiv PA^T P$ and A are similar. Consequently, they have the same eigenvalues and it suffices from now on to consider the matrix G rather than A. Then

$$\lambda_0 \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix}$$

where M_1 is $r \times r$ and M_4 is $(n-r) \times (n-r)$. It follows from block multiplication and the assumption that A and hence G are > 0 that

$$G = \begin{pmatrix} A' & B \\ 0 & C \end{pmatrix}.$$

Now let λ be an eigenvalue of G such that $|\lambda| = \lambda_0$. Then from Lemma B.0.8, either $\lambda \in \sigma(A')$ or $\lambda \in \sigma(C)$. Suppose without loss of generality that $\lambda \in \sigma(A')$. Since $A' > 0$ it has a largest positive eigenvalue λ_0' which is obtained from 2.7. Thus $\lambda_0' \leq \lambda_0$ but λ being an eigenvalue of A', has its absolute value bounded by λ_0' and so $\lambda_0 = |\lambda| \leq \lambda_0' \leq \lambda_0$ showing that $\lambda_0 \in \sigma(A')$. Now if there exists $\mathbf{v} >> \mathbf{0}$ such that $A'^T \mathbf{v} = \lambda_0 \mathbf{v}$, then the first part of this proof applies to the matrix A and so (λ/λ_0) is a root of unity. If such a vector, $\mathbf{v}$ does not exist, then let A' play the role of A in the above argument and reduce to the consideration of

$$G' \equiv \begin{pmatrix} A'' & B' \\ 0 & C' \end{pmatrix}$$

where G' is similar to A' and $\lambda, \lambda_0 \in \sigma(A'')$. Stop if $A''^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Otherwise, decompose A'' similar to the above and add another prime. Continuing this way you must eventually obtain the situation where $\left(A'^{\cdots\prime}\right)^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Indeed, this happens no later than when $A'^{\cdots\prime}$ is a 1×1 matrix. ■

Appendix C

Functions of Matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \tag{3.1}$$

for all $|\lambda| < R$. There is a formula for $f(A) \equiv \sum_{n=0}^{\infty} a_n A^n$ which makes sense whenever $\rho(A) < R$. Thus you can speak of $\sin(A)$ or e^A for A an $n \times n$ matrix. To begin with, define

$$f_P(\lambda) \equiv \sum_{n=0}^{P} a_n \lambda^n$$

so for $k < P$

$$\begin{aligned} f_P^{(k)}(\lambda) &= \sum_{n=k}^{P} a_n n \cdots (n-k+1) \lambda^{n-k} \\ &= \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}. \end{aligned} \tag{3.2}$$

Thus

$$\frac{f_P^{(k)}(\lambda)}{k!} = \sum_{n=k}^{P} a_n \binom{n}{k} \lambda^{n-k} \tag{3.3}$$

To begin with consider $f(J_m(\lambda))$ where $J_m(\lambda)$ is an $m \times m$ Jordan block. Thus $J_m(\lambda) = D + N$ where $N^m = 0$ and N commutes with D. Therefore, letting $P > m$

$$\begin{aligned} \sum_{n=0}^{P} a_n J_m(\lambda)^n &= \sum_{n=0}^{P} a_n \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{P} \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{m-1} N^k \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k}. \end{aligned} \tag{3.4}$$

From 3.3 this equals

$$\sum_{k=0}^{m-1} N^k \operatorname{diag}\left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right) \tag{3.5}$$

where for $k = 0, \cdots, m-1$, define $\operatorname{diag}_k(a_1, \cdots, a_{m-k})$ the $m \times m$ matrix which equals zero everywhere except on the k^{th} super diagonal where this diagonal is filled with the numbers, $\{a_1, \cdots, a_{m-k}\}$ from the upper left to the lower right. With no subscript, it is just the diagonal matrices having the indicated entries. Thus in 4×4 matrices, $\operatorname{diag}_2(1,2)$ would be the matrix

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

Then from 3.5 and 3.2,

$$\sum_{n=0}^{P} a_n J_m(\lambda)^n = \sum_{k=0}^{m-1} \operatorname{diag}_k \left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right).$$

Therefore, $\sum_{n=0}^{P} a_n J_m(\lambda)^n =$

$$\begin{pmatrix} f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \frac{f_P^{(2)}(\lambda)}{2!} & \cdots & \frac{f_P^{(m-1)}(\lambda)}{(m-1)!} \\ & f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \ddots & \vdots \\ & & f_P(\lambda) & \ddots & \frac{f_P^{(2)}(\lambda)}{2!} \\ & & & \ddots & \frac{f_P'(\lambda)}{1!} \\ 0 & & & & f_P(\lambda) \end{pmatrix} \tag{3.6}$$

Now let A be an $n \times n$ matrix with $\rho(A) < R$ where R is given above. Then the Jordan form of A is of the form

$$J = \begin{pmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_r \end{pmatrix} \tag{3.7}$$

where $J_k = J_{m_k}(\lambda_k)$ is an $m_k \times m_k$ Jordan block and $A = S^{-1}JS$. Then, letting $P > m_k$ for all k,

$$\sum_{n=0}^{P} a_n A^n = S^{-1} \sum_{n=0}^{P} a_n J^n S,$$

and because of block multiplication of matrices,

$$\sum_{n=0}^{P} a_n J^n = \begin{pmatrix} \sum_{n=0}^{P} a_n J_1^n & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \sum_{n=0}^{P} a_n J_r^n \end{pmatrix}$$

and from 3.6 $\sum_{n=0}^{P} a_n J_k^n$ converges as $P \to \infty$ to the $m_k \times m_k$ matrix

$$\begin{pmatrix} f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \frac{f^{(2)}(\lambda_k)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \ddots & \vdots \\ 0 & 0 & f(\lambda_k) & \ddots & \frac{f^{(2)}(\lambda_k)}{2!} \\ \vdots & & \ddots & \ddots & \frac{f'(\lambda_k)}{1!} \\ 0 & 0 & \cdots & 0 & f(\lambda_k) \end{pmatrix} \tag{3.8}$$

There is no convergence problem because $|\lambda| < R$ for all $\lambda \in \sigma(A)$. This has proved the following theorem.

Theorem C.0.1 *Let f be given by 3.1 and suppose $\rho(A) < R$ where R is the radius of convergence of the power series in 3.1. Then the series,*

$$\sum_{k=0}^{\infty} a_n A^n \tag{3.9}$$

converges in the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ with respect to any of the norms on this space and furthermore,

$$\sum_{k=0}^{\infty} a_n A^n = S^{-1} \begin{pmatrix} \sum_{n=0}^{\infty} a_n J_1^n & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \sum_{n=0}^{\infty} a_n J_r^n \end{pmatrix} S$$

where $\sum_{n=0}^{\infty} a_n J_k^n$ is an $m_k \times m_k$ matrix of the form given in 3.8 where $A = S^{-1}JS$ and the Jordan form of A, J is given by 3.7. Therefore, you can define $f(A)$ by the series in 3.9.

Here is a simple example.

Example C.0.2 *Find* $\sin(A)$ *where* $A = \begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4 \end{pmatrix}$.

In this case, the Jordan canonical form of the matrix is not too hard to find.

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

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

Then from the above theorem $\sin(J)$ is given by

$$\sin \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix}.$$

Therefore, $\sin(A) =$

$$\begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix} \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = M$$

where the columns of M are as follows from left to right,

$$\begin{pmatrix} \sin 4 \\ \frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 \\ 0 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 \end{pmatrix}, \begin{pmatrix} \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 + \frac{3}{2}\sin 2 - 2\cos 2 \\ -\cos 2 \\ -\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}, \begin{pmatrix} -\cos 2 \\ \sin 2 \\ \sin 2 - \cos 2 \\ \cos 2 - \sin 2 \end{pmatrix}$$
$$\begin{pmatrix} \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 - 2\cos 2 \\ -\cos 2 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}.$$

Perhaps this isn't the first thing you would think of. Of course the ability to get this nice closed form description of $\sin(A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

The following corollary is known as the spectral mapping theorem.

Corollary C.0.3 *Let A be an $n \times n$ matrix and let $\rho(A) < R$ where for $|\lambda| < R$,*

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n.$$

Then $f(A)$ is also an $n \times n$ matrix and furthermore, $\sigma(f(A)) = f(\sigma(A))$. Thus the eigenvalues of $f(A)$ are exactly the numbers $f(\lambda)$ where λ is an eigenvalue of A. Furthermore, the algebraic multiplicity of $f(\lambda)$ coincides with the algebraic multiplicity of λ.

All of these things can be generalized to linear transformations defined on infinite dimensional spaces and when this is done the main tool is the Dunford integral along with the methods of complex analysis. It is good to see it done for finite dimensional situations first because it gives an idea of what is possible. Actually, some of the most interesting functions in applications do not come in the above form as a power series expanded about 0. One example of this situation has already been encountered in the proof of the right polar decomposition with the square root of an Hermitian transformation which had all nonnegative eigenvalues. Another example is that of taking the positive part of an Hermitian matrix. This is important in some physical models where something may depend on the positive part of the strain which is a symmetric real matrix. Obviously there is no way to consider this as a power series expanded about 0 because the function $f(r) = r^+$ is not even differentiable at 0. Therefore, a totally different approach must be considered. First the notion of a positive part is defined.

Definition C.0.4 *Let A be an Hermitian matrix. Thus it suffices to consider A as an element of $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ according to the usual notion of matrix multiplication. Then there exists an orthonormal basis of eigenvectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ such that*

$$A = \sum_{j=1}^{n} \lambda_j \mathbf{u}_j \otimes \mathbf{u}_j,$$

for λ_j the eigenvalues of A, all real. Define

$$A^+ \equiv \sum_{j=1}^{n} \lambda_j^+ \mathbf{u}_j \otimes \mathbf{u}_j$$

where $\lambda^+ \equiv \frac{|\lambda|+\lambda}{2}$.

This gives us a nice definition of what is meant but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix A. The following addresses this question.

Theorem C.0.5 *If A, B be Hermitian matrices, then for $|\cdot|$ the Frobenius norm,*

$$\left|A^{+} - B^{+}\right| \leq |A - B|\,.$$

Proof: Let $A = \sum_i \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i$ and let $B = \sum_j \mu_j \mathbf{w}_j \otimes \mathbf{w}_j$ where $\{\mathbf{v}_i\}$ and $\{\mathbf{w}_j\}$ are orthonormal bases of eigenvectors.

$$\left|A^{+} - B^{+}\right|^2 = \text{trace}\left(\sum_i \lambda_i^{+} \mathbf{v}_i \otimes \mathbf{v}_i - \sum_j \mu_j^{+} \mathbf{w}_j \otimes \mathbf{w}_j\right)^2 =$$

$$\text{trace}\left[\sum_i \left(\lambda_i^{+}\right)^2 \mathbf{v}_i \otimes \mathbf{v}_i + \sum_j \left(\mu_j^{+}\right)^2 \mathbf{w}_j \otimes \mathbf{w}_j\right.$$

$$\left.-\sum_{i,j} \lambda_i^{+} \mu_j^{+} \left(\mathbf{w}_j, \mathbf{v}_i\right) \mathbf{v}_i \otimes \mathbf{w}_j - \sum_{i,j} \lambda_i^{+} \mu_j^{+} \left(\mathbf{v}_i, \mathbf{w}_j\right) \mathbf{w}_j \otimes \mathbf{v}_i\right]$$

Since the trace of $\mathbf{v}_i \otimes \mathbf{w}_j$ is $(\mathbf{v}_i, \mathbf{w}_j)$, a fact which follows from $(\mathbf{v}_i, \mathbf{w}_j)$ being the only possibly nonzero eigenvalue,

$$= \sum_i \left(\lambda_i^{+}\right)^2 + \sum_j \left(\mu_j^{+}\right)^2 - 2\sum_{i,j} \lambda_i^{+} \mu_j^{+} \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2. \tag{3.10}$$

Since these are orthonormal bases,

$$\sum_i \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2 = 1 = \sum_j \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2$$

and so 3.10 equals

$$= \sum_i \sum_j \left(\left(\lambda_i^{+}\right)^2 + \left(\mu_j^{+}\right)^2 - 2\lambda_i^{+} \mu_j^{+}\right) \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2.$$

Similarly,

$$|A - B|^2 = \sum_i \sum_j \left(\left(\lambda_i\right)^2 + \left(\mu_j\right)^2 - 2\lambda_i \mu_j\right) \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2.$$

Now it is easy to check that $\left(\lambda_i\right)^2 + \left(\mu_j\right)^2 - 2\lambda_i \mu_j \geq \left(\lambda_i^{+}\right)^2 + \left(\mu_j^{+}\right)^2 - 2\lambda_i^{+} \mu_j^{+}$. ■

Bibliography

[1] **Apostol T.,** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Artin M.,** *Algebra,* Pearson *2011.*

[3] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[4] **Baker, A.** *Transcendental Number Theory*, Cambridge University Press 1975.

[5] **Chahal J.S.,** *Historical Perspective of Mathematics 2000 B.C. - 2000 A.D. Kendrick Press, Inc. (2007)*

[6] **Coddington and Levinson**, *Theory of Ordinary Differential Equations* McGraw Hill 1955.

[7] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[8] **Edwards C.H.,** *Advanced Calculus of several Variables,* Dover 1994.

[9] **Evans L.C.** *Partial Differential Equations,* Berkeley Mathematics Lecture Notes. 1993.

[10] **Friedberg S. Insel A. and Spence L.**, *Linear Algebra*, Prentice Hall, 2003.

[11] **Golub, G. and Van Loan, C.**,*Matrix Computations*, Johns Hopkins University Press, 1996.

[12] **Gurtin M.,** *An introduction to continuum mechanics,* Academic press 1981.

[13] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[14] **Herstein I. N.**, *Topics In Algebra*, Xerox, 1964.

[15] **Hofman K. and Kunze R.,** *Linear Algebra,* Prentice Hall, 1971.

[16] **Householder A.** *The theory of matrices in numerical analysis* , Dover, 1975.

[17] **Horn R. and Johnson C.,** *matrix Analysis,* Cambridge University Press, 1985.

[18] **Jacobsen N.** *Basic Algebra* Freeman 1974.

[19] **Karlin S. and Taylor H.,** *A First Course in Stochastic Processes,* Academic Press, 1975.

[20] **Marcus M., and Minc H.,** *A Survey Of Matrix Theory and Matrix Inequalities,* Allyn and Bacon, INc. Boston, 1964

[21] **Nobel B. and Daniel J.,** *Applied Linear Algebra, Prentice Hall, 1977.*

[22] **E. J. Putzer,** American Mathematical Monthly, Vol. 73 (1966), pp. 2-7.

[23] **Rudin W.,** *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[24] **Rudin W.,** *Functional Analysis, McGraw Hill, 1991.*

[25] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[26] **Strang Gilbert**, *Linear Algebra and its Applications,* Harcourt Brace Jovanovich 1980.

[27] **Wilkinson, J.H.,** *The Algebraic Eigenvalue Problem, Clarendon Press Oxford 1965.*

Index