Math 570 notes

Kenneth Kuttler

February 18, 2003

Contents

Determinants

Here we give a discussion of the most important properties of determinants. There are more elegant ways to proceed and the reader is encouraged to consult a more advanced algebra book to read these. Another very good source is Apostol [1]. The goal here is to present all of the major theorems on determinants with a minimum of abstract algebra as quickly as possible. In this section and elsewhere $\mathbb{F}$ will denote the field of scalars, usually $\mathbb{R}$ or $\mathbb{C}$. To begin with we make a simple definition.

Definition 1.1 *Let $(k_1, \cdots, k_n)$ be an ordered list of n integers. We define*

$$\pi(k_1, \cdots, k_n) \equiv \prod \{(k_s - k_r) : r < s\}.$$

In words, we consider all terms of the form $(k_s - k_r)$ where k_s comes after k_r in the ordered list and then multiply all these together. We also make the following definition.

$$sgn(k_1, \cdots, k_n) \equiv \begin{cases} 1 \text{ if } \pi(k_1, \cdots, k_n) > 0 \\ -1 \text{ if } \pi(k_1, \cdots, k_n) < 0 \\ 0 \text{ if } \pi(k_1, \cdots, k_n) = 0 \end{cases}$$

This is called the sign of the permutation $\binom{1 \cdots n}{k_1 \cdots k_n}$ in the case when there are no repeats in the ordered list, $(k_1, \cdots, k_n)$ and $\{k_1, \cdots, k_n\} = \{1, \cdots, n\}$.

Lemma 1.2 *Let $(k_1, \cdots k_i \cdots, k_j, \cdots, k_n)$ be a list of n integers. Then*

$$\pi(k_1, \cdots, k_i \cdots, k_j, \cdots, k_n) = -\pi(k_1, \cdots, k_j \cdots, k_i, \cdots, k_n)$$

and

$$sgn(k_1, \cdots, k_i \cdots, k_j, \cdots, k_n) = -sgn(k_1, \cdots, k_j \cdots, k_i, \cdots, k_n)$$

In words, if we switch two entries the sign changes.

Proof: The two lists are

$$(k_1, \cdots, k_i, \cdots, k_j, \cdots, k_n) \tag{1.1}$$

and

$$(k_1, \cdots, k_j, \cdots, k_i, \cdots, k_n). \tag{1.2}$$

Suppose there are $r-1$ numbers between k_i and k_j in the first list and consequently $r-1$ numbers between k_j and k_i in the second list. In computing $\pi(k_1, \cdots, k_i, \cdots, k_j, \cdots, k_n)$ we have $r-1$ terms of the form $(k_j - k_p)$ where the k_p are those numbers occurring in the list between k_i and k_j. Corresponding to these

terms we have $r-1$ terms of the form $(k_p - k_j)$ in the computation of $\pi(k_1, \cdots, k_j, \cdots, k_i, \cdots, k_n)$. These differences produce a $(-1)^{r-1}$ in going from $\pi(k_1, \cdots, k_i, \cdots, k_j, \cdots, k_n)$ to $\pi(k_1, \cdots, k_j, \cdots, k_i, \cdots, k_n)$. We also have the $r-1$ terms $(k_p - k_i)$ in computing $\pi(k_1, \cdots, k_i \cdots, k_j, \cdots, k_n)$ and the $r-1$ terms, $(k_i - k_p)$ in computing $\pi(k_1, \cdots, k_j, \cdots, k_i, \cdots, k_n)$, producing another $(-1)^{r-1}$. Thus, in considering the differences in π, we see these terms just considered do not change the sign. However, we have $(k_j - k_i)$ in the first product and $(k_i - k_j)$ in the second and all other factors in the computation of π match up in the two computations so it follows $\pi(k_1, \cdots, k_i, \cdots, k_j, \cdots, k_n) = -\pi(k_1, \cdots, k_j, \cdots, k_i, \cdots, k_n)$ as claimed.

Corollary 1.3 *Suppose $(k_1, \cdots, k_n)$ is obtained by making p switches in the ordered list, $(1, \cdots, n)$. Then*

$$(-1)^p = sgn(k_1, \cdots, k_n). \tag{1.3}$$

Proof: We observe that $\operatorname{sgn}(1, \cdots, n) = 1$ and according to Lemma 1.2, each time we switch two entries we multiply by (-1). Therefore, making the p switches, we obtain $(-1)^p = (-1)^p \operatorname{sgn}(1, \cdots, n) = \operatorname{sgn}(k_1, \cdots, k_n)$ as claimed.

We now are ready to define the determinant of an $n \times n$ matrix.

Definition 1.4 *Let $(a_{ij}) = A$ denote an $n \times n$ matrix. We define*

$$\det(A) \equiv \sum_{(k_1, \cdots, k_n)} sgn(k_1, \cdots, k_n)\, a_{1k_1} \cdots a_{nk_n}$$

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, we know $sgn(k_1, \cdots, k_n) = 0$.

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

$$\det(A(r_1, \cdots, r_n)) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n)\, a_{r_1 k_1} \cdots a_{r_n k_n} \tag{1.4}$$

and

$$A(1, \cdots, n) = A.$$

Proposition 1.5 *Let $(r_1, \cdots, r_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. Then*

$$\begin{aligned} sgn(r_1, \cdots, r_n) \det(A) &= \sum_{(k_1, \cdots, k_n)} sgn(k_1, \cdots, k_n)\, a_{r_1 k_1} \cdots a_{r_n k_n} && (1.5) \\ &= \det(A(r_1, \cdots, r_n)). && (1.6) \end{aligned}$$

In words, if we take the determinant of the matrix obtained by letting the p^{th} row be the r_p row of A, then the determinant of this modified matrix equals the expression on the left in 1.5.

Proof: Let $(1, \cdots, n) = (1, \cdots, r, \cdots s, \cdots, n)$ so $r < s$.

$$\det(A(1, \cdots, r, \cdots s, \cdots, n)) = \tag{1.7}$$

$$\sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n)\, a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} -\operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= -\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{1.8}$$

Consequently,

$$\det(A(1,\cdots,s,\cdots,r,\cdots,n)) = -\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = -\det(A)$$

Now letting $A(1,\cdots,s,\cdots,r,\cdots,n)$ play the role of A, and continuing in this way, we eventually arrive at the conclusion

$$\det(A(r_1,\cdots,r_n)) = (-1)^p \det(A)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Corollary 1.3 this implies

$$\det(A(r_1,\cdots,r_n)) = \operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1,\cdots,r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 1.7 -1.8 shows that $A(r_1,\cdots,r_n) = 0$ and we also know that $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ so the formula holds in this case also.

Definition 1.6 *If $A = (a_{ij})$ is an $n\times n$ matrix, then we define the transpose, $A^T = \left(a_{ij}^T\right)$ by $a_{ij}^T \equiv a_{ji}$. Thus to form the transpose we let the i^{th} column become the i^{th} row.*

Corollary 1.7 *We have the following formula for* $\det(A)$.

$$\det(A) = \frac{1}{n!}\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} sgn(r_1,\cdots,r_n)\, sgn(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}. \tag{1.9}$$

Also $\det\left(A^T\right) = \det(A)$.

Proof: From Proposition 1.5, if the r_i are distinct,

$$\det(A) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, we know $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ and so there is no contribution to the sum.) we obtain

$$n!\det(A) = \sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary.

Corollary 1.8 *If we switch two rows or two columns in an $n\times n$ matrix, A, 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* $\det(A) = 0$.

Proof: By Proposition 1.5 when we switch two rows the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 1.7 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, then

$$\det(A) = \det\left(A^T\right) = -\det\left(A_1^T\right) = -\det(A_1).$$

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 $\det(A) = 0$.

Definition 1.9 *If A and B are $n \times n$ matrices, $A = (a_{ij})$ and $B = (b_{ij})$, we form the product, $AB = (c_{ij})$ by defining*

$$c_{ij} \equiv \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{1.10}$$

This is just the usual rule for matrix multiplication.More generally if $A = (a_{ij})$ is an $m \times n$ matrix and $B = (b_{ij})$ is an $n \times k$ matrix, then AB is an $m \times k$ matrix such that $AB = (c_{ij})$ where 1.10 holds. We call two such matrices conformable.

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

Theorem 1.10 *Let A and B be $n \times n$ matrices. Then* $\det(AB) = \det(A)\det(B)$.

Proof: We will denote by c_{ij} the ij^{th} entry of AB. Thus by Proposition 1.5,

$$\begin{aligned}
\det(AB) &= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1}\cdots c_{nk_n} \\
&= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n) \left(\sum_{r_1} a_{1r_1} b_{r_1k_1}\right)\cdots\left(\sum_{r_n} a_{nr_n} b_{r_nk_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1\cdots r_n)\, a_{1r_1}\cdots a_{nr_n} \det(B) = \det(A)\det(B).
\end{aligned}$$

This proves the theorem.

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column.

Definition 1.11 *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then we define a new matrix, $cof(A)$ by $cof(A) = (c_{ij})$ where to obtain c_{ij} we 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 and then multiply this number by $(-1)^{i+j}$. The determinant of the $n-1 \times n-1$ matrix just described is called the ij^{th} minor of A. To make the formulas easier to remember, we shall write $cof(A)_{ij}$ for the ij^{th} entry of the cofactor matrix.*

The main result is the following monumentally important theorem. It states that you can expand an $n \times n$ matrix along any row or column. This is often taken as a definition in elementary courses but how anyone in their right mind could believe without a proof that you always get the same answer by expanding along any row or column is totally beyond my powers of comprehension.

Theorem 1.12 *Let A be an $n \times n$ matrix. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij}\, cof(A)_{ij} = \sum_{i=1}^{n} a_{ij}\, cof(A)_{ij}\,. \tag{1.11}$$

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: We will prove this by using the definition and then doing a computation and verifying that we have what we want.

$$\det(A) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_r}\cdots a_{nk_n}$$

$$= \sum_{k_r=1}^{n} \left(\sum_{(k_1,\cdots,k_r,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{(r-1)k_{(r-1)}} a_{(r+1)k_{(r+1)}} a_{nk_n} \right) a_{rk_r}$$

$$= \sum_{j=1}^{n} (-1)^{r-1} \cdot$$

$$\left(\sum_{(k_1,\cdots,j,\cdots,k_n)} \operatorname{sgn}(j,k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n)\, a_{1k_1}\cdots a_{(r-1)k_{(r-1)}} a_{(r+1)k_{(r+1)}} a_{nk_n} \right) a_{rj}. \tag{1.12}$$

We need to consider for fixed j the term

$$\sum_{(k_1,\cdots,j,\cdots,k_n)} \operatorname{sgn}(j,k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n)\, a_{1k_1}\cdots a_{(r-1)k_{(r-1)}} a_{(r+1)k_{(r+1)}} a_{nk_n}. \tag{1.13}$$

We may assume all the indices in $(k_1,\cdots,j,\cdots,k_n)$ are distinct. We define $(l_1,\cdots,l_{n-1})$ as follows. If $k_\alpha < j$, then $l_\alpha \equiv k_\alpha$. If $k_\alpha > j$, then $l_\alpha \equiv k_\alpha - 1$. Thus every choice of the ordered list, $(k_1,\cdots,j,\cdots,k_n)$, corresponds to an ordered list, $(l_1,\cdots,l_{n-1})$ of indices from $\{1,\cdots,n-1\}$. Now define

$$b_{\alpha l_\alpha} \equiv \begin{cases} a_{\alpha k_\alpha} \text{ if } \alpha < r, \\ a_{(\alpha+1)k_\alpha} \text{ if } n-1 \geq \alpha > r \end{cases}$$

where here k_α corresponds to l_α as just described. Thus $(b_{\alpha\beta})$ is the $n-1 \times n-1$ matrix which results from deleting the r^{th} row and the j^{th} column. In computing

$$\pi(j,k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n),$$

we note there are exactly $j-1$ of the k_i which are less than j. Therefore,

$$\operatorname{sgn}(k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n)(-1)^{j-1} = \operatorname{sgn}(j,k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n).$$

But it also follows from the definition that

$$\operatorname{sgn}(k_1,\cdots,k_{r-1},k_{r+1}\cdots,k_n) = \operatorname{sgn}(l_1\cdots,l_{n-1})$$

and so the term in 1.13 equals

$$(-1)^{j-1} \sum_{(l_1,\cdots,l_{n-1})} \operatorname{sgn}(l_1,\cdots,l_{n-1})\, b_{1l_1}\cdots b_{(n-1)l_{(n-1)}}$$

Using this in 1.12 we see

$$\begin{aligned}
\det(A) &= \sum_{j=1}^{n} (-1)^{r-1}(-1)^{j-1} \left(\sum_{(l_1,\cdots,l_{n-1})} \operatorname{sgn}(l_1,\cdots,l_{n-1})\, b_{1l_1}\cdots b_{(n-1)l_{(n-1)}} \right) a_{rj} \\
&= \sum_{j=1}^{n} (-1)^{r+j} \left(\sum_{(l_1,\cdots,l_{n-1})} \operatorname{sgn}(l_1,\cdots,l_{n-1})\, b_{1l_1}\cdots b_{(n-1)l_{(n-1)}} \right) a_{rj} \\
&= \sum_{j=1}^{n} a_{rj}\operatorname{cof}(A)_{rj}
\end{aligned}$$

as claimed. Now to get the second half of 1.11, we can apply the first part to A^T and write for $A^T = \left(a_{ij}^T\right)$

$$\begin{aligned}
\det(A) &= \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} \\
&= \sum_{j=1}^{n} a_{ji}\operatorname{cof}(A)_{ji} = \sum_{i=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}\,.
\end{aligned}$$

This proves the theorem. We leave it as an exercise to show that $\operatorname{cof}\left(A^T\right)_{ij} = \operatorname{cof}(A)_{ji}$.

Note that this gives us an easy way to write a formula for the inverse of an $n \times n$ matrix.

Definition 1.13 *We say an $n \times n$ matrix, A has an inverse, A^{-1} if and only if $AA^{-1} = A^{-1}A = I$ where $I = (\delta_{ij})$ for*

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

Theorem 1.14 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} C_{ji}$$

for C_{ij} the ij^{th} cofactor of A.

Proof: By Theorem 1.12 and letting $(a_{ir}) = A$, if we assume $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} C_{ir} \det(A)^{-1} = \det(A)\det(A)^{-1} = 1.$$

Now we consider

$$\sum_{i=1}^{n} a_{ir} C_{ik} \det(A)^{-1}$$

when $k \neq r$. We replace the k^{th} column with the r^{th} column to obtain a matrix, B_k whose determinant equals zero by Corollary 1.8. However, expanding this matrix along the k^{th} column yields

$$0 = \det(B_k)\det(A)^{-1} = \sum_{i=1}^{n} a_{ir} C_{ik} \det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir}C_{ik}\det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 1.12, we can also write using similar reasoning,

$$\sum_{j=1}^{n} a_{rj}C_{kj}\det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then A^{-1} exists and if $A^{-1} = \left(a_{ij}^{-1}\right)$,

$$a_{ij}^{-1} = C_{ji}\det(A)^{-1}.$$

Now suppose A^{-1} exists. Then by Theorem 1.10,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A)\det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. This proves the theorem.

This theorem says that to find the inverse, we can 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. The term, adjoint, should be reserved for something much more interesting which will be discussed later. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

By Problem 1 matrix multiplication is associative. Therefore, in case we are solving a system of equations,

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

for $\mathbf{x}$, it follows that if A^{-1} exists, we can write

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, we just presented a formula for A^{-1}. Using this formula, we see

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1}y_j = \sum_{j=1}^{n} \frac{1}{\det(A)}\mathrm{cof}(A)_{ji}\,y_j.$$

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 we have replaced the i^{th} column of A with the column vector, $(y_1 \cdots, y_n)^T$, taken its determinant and divided by $\det(A)$. This formula is known as Cramer's rule.

Definition 1.15 *We say 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, we give the following corollary of Theorem 1.12.

Corollary 1.16 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

1.1 The characteristic polynomial

Definition 1.17 *Let A be an $n \times n$ matrix. The characteristic polynomial is defined as*

$$p_A(t) \equiv \det(tI - A).$$

A principal submatrix of A is one lying in the same set of k rows and columns and a principal minor is the determinant of a principal submatrix. There are $\binom{n}{k}$ principal minors of A. How do we get a typical principal submatrix? We pick k rows, say $r_1, \cdots, r_k$ and consider the $k \times k$ matrix which results from using exactly those entries of these k rows which are also in one of the $r_1, \cdots, r_k$ columns. We denote by $E_k(A)$ the sum of the principal $k \times k$ minors of A.

We write a formula for the characteristic polynomial in terms of the $E_k(A)$.

$$p_A(t) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1, \cdots, k_n)(t\delta_{1k_1} - a_{1k_1}) \cdots (t\delta_{nk_n} - a_{nk_n})$$

Consider the terms which are multiplied by t^r. A typical such term would be

$$t^r (-1)^{n-r} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1, \cdots, k_n) \delta_{m_1 k_{m_1}} \cdots \delta_{m_r k_{m_r}} a_{s_1 k_{s_1}} \cdots a_{s_{(n-r)} k_{s_{(n-r)}}} \tag{1.14}$$

where $\{m_1, \cdots, m_r, s_1, \cdots, s_{n-r}\} = \{1, \cdots, n\}$. From the definition of determinant, the sum in the above expression is the determinant of a matrix like

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

where the starred rows are simply the original rows of the matrix, A. Using the row operation which involves replacing a row with a multiple of another row added to itself, we can use the ones to zero out everything above them and below them, obtaining a modified matrix which has the same determinant (See Problem 6). In the given example this would result in a matrix of the form

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

and so the sum in 1.14 is just the principal minor corresponding to the subset $\{m_1, \cdots, m_r\}$ of $\{1, \cdots, n\}$. For each of the $\binom{n}{r}$ such choices, there is such a term equal to the principal minor determined in this way and so the sum of these equals the coefficient of the t^r term. Therefore, the coefficient of t^r equals $(-1)^{n-r} E_{n-r}(A)$. It follows

$$\begin{aligned} p_A(t) &= \sum_{r=0}^{n} t^r (-1)^{n-r} E_{n-r}(A) \\ &= (-1)^n E_n(A) + (-1)^{n-1} t E_{n-1}(A) + \cdots + (-1) t^{n-1} E_1(A) + t^n. \end{aligned}$$

Definition 1.18 *The solutions to $p_A(t) = 0$ are called the eigenvalues of A.*

We know also that

$$p_A(t) = \prod_{k=1}^{n} (t - \lambda_k)$$

where λ_k are the roots of the equation, $p_A(t) = 0$. (Note these might be complex numbers.) Therefore, expanding the above polynomial,

$$E_k(A) = S_k(\lambda_1, \cdots, \lambda_n)$$

where $S_k(\lambda_1, \cdots, \lambda_n)$, called the k^{th} elementary symmetric function of the numbers $\lambda_1, \cdots, \lambda_n$, is defined as the sum of all possible products of k elements of $\{\lambda_1, \cdots, \lambda_n\}$. Therefore,

$$p_A(t) = t^n - S_1(\lambda_1, \cdots, \lambda_n) t^{n-1} + S_2(\lambda_1, \cdots, \lambda_n) t^{n-2} + \cdots \pm S_n(\lambda_1, \cdots, \lambda_n).$$

A remarkable and profound theorem is the Cayley Hamilton theorem which states that every matrix satisfies its characteristic equation. We give a simple proof of this theorem using the following lemma.

Lemma 1.19 *Suppose for all $|\lambda|$ large enough, we have*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = 0,$$

where the A_i are $n \times n$ matrices. Then each $A_i = 0$.

Proof: Multiply by λ^{-m} to obtain

$$A_0\lambda^{-m} + A_1\lambda^{-m+1} + \cdots + A_{m-1}\lambda^{-1} + A_m = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_m = 0$. With this, multiply by λ to obtain

$$A_0\lambda^{-m+1} + A_1\lambda^{-m+2} + \cdots + A_{m-1} = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_{m-1} = 0$. Continue multiplying by λ and letting $\lambda \to \infty$ to obtain that all the $A_i = 0$. This proves the lemma.

With the lemma, we have the following simple corollary.

Corollary 1.20 *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.

Proof: Subtract and use the result of the lemma.

With this preparation, we can now give an easy proof of the Cayley Hamilton theorem.

Theorem 1.21 *Let A be an $n \times n$ matrix and let $p(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $p(A) = 0$.*

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then λ cannot be in the finite list of eigenvalues of A and so for such λ, $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 1.14 we may write

$$C(\lambda) = p(\lambda)(\lambda I - A)^{-1}.$$

Note that each entry in $C(\lambda)$ is a polynomial in λ having degree no more than $n-1$. Therefore, collecting the terms, we may write

$$C(\lambda) = C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}$$

for C_j some $n \times n$ matrix. It follows that for all $|\lambda|$ large enough,

$$(A - \lambda I)\left(C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}\right) = p(\lambda) I$$

and so we are in the situation of Corollary 1.20. It follows the matrix coefficients corresponding to equal powers of λ are equal on both sides of this equation. Therefore, we may replace λ with A and the two will be equal. Thus

$$0 = (A - A)\left(C_0 + C_1 A + \cdots + C_{n-1}A^{n-1}\right) = p(A) I = p(A).$$

This proves the Cayley Hamilton theorem.

We will return to this important theorem later.

1.2 Exercises

1. Show that matrix multiplication is associative. That is, $(AB)C = A(BC)$.
2. Show the inverse of a matrix, if it exists, is unique. Thus if $AB = BA = I$, then $B = A^{-1}$.
3. In the proof of Theorem 1.14 we asserted that $\det(I) = 1$. Here $I = (\delta_{ij})$. Prove this assertion. Also prove Corollary 1.16.
4. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be vectors in $\mathbb{F}^n$ and let $M(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ denote the matrix whose i^{th} column equals $\mathbf{v}_i$. Define

   $$d(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \det(M(\mathbf{v}_1, \cdots, \mathbf{v}_n)).$$

   Prove that d is linear in each variable, (multilinear), that

   $$d(\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_n) = -d(\mathbf{v}_1, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_n), \tag{1.15}$$

   and

   $$d(\mathbf{e}_1, \cdots, \mathbf{e}_n) = 1 \tag{1.16}$$

   where here $\mathbf{e}_j$ is the vector in $\mathbb{F}^n$ which has a zero in every position except the j^{th} position in which it has a one.
5. Suppose $f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \to \mathbb{F}$ satisfies 1.15 and 1.16 and is linear in each variable. Show that $f = d$.
6. Show that if we replace a row (column) of an $n \times n$ matrix A with itself added to some multiple of another row (column) then the new matrix has the same determinant as the original one.
7. If $A = (a_{ij})$, show $\det(A) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{k_1 1} \cdots a_{k_n n}$.
8. Use the result of Problem 6 to evaluate by hand the determinant

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

9. Find the inverse if it exists of the matrix,

$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

10. Let $Ly = y^{(n)} + a_{n-1}(x)\, y^{(n-1)} + \cdots + a_1(x)\, y' + a_0(x)\, y$ where the a_i are given continuous functions defined on a closed interval, (a,b) and y is some function which has n derivatives so it makes sense to write Ly. Suppose we have $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The Wronskian of these functions, y_i is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify that $W' + a_{n-1}(x)\, W = 0$. Give an explicit solution of this linear differential equation, Abel's formula, and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a,b) or never.

11. We say 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. Thus $E_k(A) = E_k(B)$ and for this reason, the $E_k(A)$ are called invariants of the matrix. Show that $E_n(A) = \det(A)$ and $E_1(A) = tr(A) \equiv$ sum of the entries down the main diagonal. **Hint:** We know λ_i is a solution of $p_A(t) = 0$ if and only if $\det(\lambda_i I - A) = 0$ if and only if $(\lambda_i I - A)^{-1}$ does not exist. Now verify that the set of λ for which $(\lambda I - A)^{-1}$ does not exist is the same as the set of λ for which $(\lambda I - B)^{-1}$ does not exist. Thus the λ_i are the same in each case and now you can use the description of the characteristic polynomial in terms of the symmetric functions of the λ_i. You could also simply verify that $\det(\lambda I - A) = \det(\lambda I - B)$ without too much trouble.

Vector Spaces

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.)

$$\alpha(\mathbf{v}+\mathbf{w})=\alpha\mathbf{v}+\alpha\mathbf{w}, \tag{2.1}$$

$$(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}, \tag{2.2}$$

$$\alpha(\beta\mathbf{v})=\alpha\beta(\mathbf{v}), \tag{2.3}$$

$$1\mathbf{v}=\mathbf{v}. \tag{2.4}$$

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.

Definition 2.1 *If $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}\subseteq V$, a vector space, then*

$$span(\mathbf{v}_1,\cdots,\mathbf{v}_n)\equiv\{\sum_{i=1}^{n}\alpha_i\mathbf{v}_i:\alpha_i\in\mathbb{F}\}.$$

A subset, $W\subseteq V$ is said to be a subspace if it is also a vector space with the same field of scalars. Thus $W\subseteq V$ is a subspace if $ax+by\in W$ whenever $a,b\in\mathbb{F}$ and $x,y\in W$. The span of a set of vectors as just described is an example of a subspace.

Definition 2.2 *If* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \subseteq V$, *we say 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 we say $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is a basis for* V *if*

$$span\,(\mathbf{v}_1, \cdots, \mathbf{v}_n) = V$$

and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is linearly independent. We say the set of vectors is linearly dependent if it is not linearly independent.*

Theorem 2.3 *If*

$$span\,(\mathbf{u}_1, \cdots, \mathbf{u}_r) \subseteq span\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$$

and $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ *are linearly independent, then* $r \leq s$.

Proof: Let $V \equiv \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$ and suppose $r > s$. Let $A_l \equiv \{\mathbf{u}_1, \cdots, \mathbf{u}_l\}$, $A_0 = \emptyset$, and let B_{s-l} denote a subset of the vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_s\}$ which contains $s-l$ vectors and has the property that $\text{span}\,(A_l, B_{s-l}) = V$. Note that the assumption of the theorem says $\text{span}\,(A_0, B_s) = V$.

Now we describe an exchange operation for $\text{span}\,(A_l, B_{s-l}) = V$. Since $r > s$, we know $l < r$. Letting

$$B_{s-l} \equiv \{\mathbf{z}_1, \cdots, \mathbf{z}_{s-l}\} \subseteq \{\mathbf{v}_1, \cdots, \mathbf{v}_s\},$$

it follows that there exist constants, c_i and d_i such that

$$\mathbf{u}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{u}_i + \sum_{i=1}^{s-l} d_i \mathbf{z}_i,$$

and not all the d_i can equal zero. If they were all equal to zero, it would follow that the set, $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ would be dependent since one of the vectors in it would be a linear combination of the others.

Let $d_k \neq 0$. Then we can solve for $\mathbf{z}_k$ in terms of the vectors of A_l and the remaining vectors of B_{s-l} as follows.

$$\mathbf{z}_k = \frac{1}{d_k}\mathbf{u}_{l+1} - \sum_{i=1}^{l} \frac{c_i}{d_k}\mathbf{u}_i - \sum_{i \neq k} \frac{d_i}{d_k}\mathbf{z}_i.$$

Now this implies that $V = \text{span}\,(A_{l+1}, B_{s-l-1})$, where $B_{s-l-1} = B_{s-l} \setminus \{\mathbf{z}_k\}$, a set obtained by deleting $\mathbf{z}_k$ from B_{k-l}. You see, we exchanged a vector in B_{s-l} with one from $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ and kept the span the same. Starting with $V = \text{span}\,(A_0, B_s)$, we do the exchange operation until we obtain $V = \text{span}\,(A_{s-1}, \mathbf{z})$ where $\mathbf{z} \in \{\mathbf{v}_1, \cdots, \mathbf{v}_s\}$. Then one more application of the exchange operation yields $V = \text{span}\,(A_s)$. But this implies $\mathbf{u}_r \in \text{span}\,(A_s) = \text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_s)$, contradicting the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$. It follows that $r \leq s$ as claimed.

Corollary 2.4 *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 2.3, $m \leq n$ and $n \leq m$.

Definition 2.5 *We say a vector space V is of dimension n if it has a basis consisting of n vectors. This is well defined thanks to Corollary 2.4. We assume here that $n < \infty$ and say such a vector space is finite dimensional.*

Theorem 2.6 *If $V = span(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ then some subset of $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is a basis for V. Also, if $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\} \subseteq V$ is linearly independent and the vector space is finite dimensional, then the set, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$, can be enlarged to obtain a basis of V.*

Proof: Let

$$S = \{E \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\} \text{ such that } \operatorname{span}(E) = V\}.$$

For $E \in S$, let $|E|$ denote the number of elements of E. Let

$$m \equiv \min\{|E| \text{ such that } E \in S\}.$$

Thus there exist vectors

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_m\} \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$$

such that

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

and m is as small as possible for this to happen. If this set is linearly independent, it follows it is a basis for V and the theorem is proved. On the other hand, if the set is not linearly independent, then there exist scalars,

$$c_1, \cdots, c_m$$

such that

$$\mathbf{0} = \sum_{i=1}^{m} c_i \mathbf{v}_i$$

and not all the c_i are equal to zero. Suppose $c_k \neq 0$. Then we can solve for the vector, $\mathbf{v}_k$ in terms of the other vectors. Consequently,

$$V = \operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_m)$$

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 $\{\mathbf{v}_1, \cdots, v_n\}$. If

$$\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k) = V,$$

then $k = n$ and we are done. If not, there exists a vector,

$$\mathbf{u}_{k+1} \notin \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k).$$

Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is also linearly independent. Continue adding vectors in this way until n linearly independent vectors have been obtained. Then $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_n) = V$ because if it did not do so, we could obtain $\mathbf{u}_{n+1}$ as just described and then $\{\mathbf{u}_1, \cdots, \mathbf{u}_{n+1}\}$ would be a linearly independent set of vectors having $n+1$ elements even though $\{\mathbf{v}_1, \cdots, v_n\}$ is a basis. This would contradict Theorem 2.3. Therefore, this list is a basis and this proves the theorem.

Basic topics about matrices

We have already reviewed the notion of matrix multiplication and taking a transpose and an inverse of a matrix in the context of determinants. Now we consider a few other topics of a basic nature.

3.1 The rank of a matrix

The rows of an $m \times n$ matrix are vectors in $\mathbb{F}^n$ and the columns of such a matrix are vectors in $\mathbb{F}^m$. The rank of a matrix has to do with the dimension of the span of the rows or columns as described below. Also, there is a definition of rank in terms of determinants.

Definition 3.1 *Let A be an $m \times n$ matrix. Then the row rank is the dimension of the span of the rows in $\mathbb{F}^n$, the column rank is the dimension of the span of the columns, and the determinant rank equals r where r is the largest number such that some $r \times r$ submatrix of A has a non zero determinant.*

Theorem 3.2 *The determinant rank, row rank, and column rank coincide.*

Proof: Suppose the determinant rank of $A = (a_{ij})$ equals r. First note that if rows and columns are interchanged, the row, column, and determinant ranks of the modified matrix are unchanged. Thus we may assume without loss of generality that there is an $r \times r$ matrix in the upper left corner of the matrix which has non zero determinant. Consider the matrix

$$\begin{pmatrix} a_{11} & \cdots & a_{1r} & a_{1p} \\ \vdots & & \vdots & \vdots \\ a_{r1} & \cdots & a_{rr} & a_{rp} \\ a_{l1} & \cdots & a_{lr} & a_{lp} \end{pmatrix}$$

where we will denote by C the $r \times r$ matrix which has non zero determinant. The above matrix has determinant equal to zero. There are two cases to consider in verifying this claim. First, suppose $p > r$. Then the claim follows from the assumption that A has determinant rank r. On the other hand, if $p < r$, then the determinant is zero because there are two identical columns. Expand the determinant along the last column and divide by $\det(C)$ to obtain

$$a_{lp} = -\sum_{i=1}^{r} \frac{C_{ip}}{\det(C)} a_{ip},$$

where C_{ip} is the cofactor of a_{ip}. Now note that C_{ip} does not depend on p. Therefore the above sum is of the form

$$a_{lp} = \sum_{i=1}^{r} m_i a_{ip}$$

which shows the l^{th} row is a linear combination of the first r rows of A. Thus the first r rows of A are linearly independent and span the row space so the row rank equals r. It follows from this that

$$\text{column rank of } A = \text{ row rank of } A^T =$$

$$= \text{ determinant rank of } A^T \ = \text{ determinant rank of } A = \text{ row rank of } A.$$

This proves the theorem.

3.2 Eigenvalues and eigenvectors of a matrix

We discuss the concept of eigenvectors and eigenvalues of a matrix in this section. We will assume the field of scalars is $\mathbb{C}$. To begin with, we give a definition of what is meant by an eigenvalue and an eigenvector.

Definition 3.3 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a nonzero vector for which*

$$M\mathbf{x} = \lambda\mathbf{x} \tag{3.1}$$

for some scalar, λ. Then we say $\mathbf{x}$ is an eigenvector and λ is an eigenvalue (characteristic value) of the matrix, M. ***Eigenvectors are never equal to zero!***

Eigenvectors are vectors which are shrunk, stretched or reflected upon multiplication by a matrix. We need to find out how to identify possible eigenvalues and eigenvectors. Suppose $\mathbf{x}$ satisfies 3.1. Then we see that

$$(\lambda I - M)\,\mathbf{x} = \mathbf{0}$$

for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the matrix $M - \lambda I$ cannot have an inverse and so by Theorem 1.14 we must have

$$\det(\lambda I - M) = 0. \tag{3.2}$$

This is known as the characteristic equation. 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, $\lambda \in \mathbb{C}$. Can we show that such a solution is actually an eigenvalue? Is there a vector, $\mathbf{x}$ such that $(\lambda I - M)\,\mathbf{x} = \mathbf{0}$? This follows from the next theorem.

Theorem 3.4 *Let M be an $n \times n$ matrix whose entries are in $\mathbb{C}$. Then the following are equivalent.*

1. *M is one to one.*
2. *M maps a basis to a basis.*
3. *M is onto.*
4. *$\det(M) \neq \mathbf{0}$*
5. *If $M\mathbf{v} = \mathbf{0}$ then $\mathbf{v} = \mathbf{0}$.*

Proof: Suppose first M is one to one and let $\{\mathbf{v}_i\}_{i=1}^n$ be a basis of $\mathbb{C}^n$. Then if $\sum_{i=1}^n c_i M\mathbf{v}_i = 0$ it follows $M\left(\sum_{i=1}^n c_i\mathbf{v}_i\right) = 0$ which means that since $M(\mathbf{0}) = \mathbf{0}$, and M is one to one, it must be the case that $\sum_{i=1}^n c_i\mathbf{v}_i = \mathbf{0}$. Since $\{\mathbf{v}_i\}$ is a basis, each $c_i = 0$ which shows $\{M\mathbf{v}_i\}$ is a linearly independent set. Since there are n of these, it must be that $\{M\mathbf{v}_i\}$ is a basis. Now suppose 2.). Then letting $\{\mathbf{v}_i\}$ be a basis, and $\mathbf{y} \in \mathbb{C}^n$, we know from part 2.) that there are constants, $\{c_i\}$ such that $\mathbf{y} = \sum_{i=1}^n c_i M\mathbf{v}_i = M\left(\sum_{i=1}^n c_i\mathbf{v}_i\right)$. Thus M is onto and we have shown that 2.) implies 3.). Now suppose 3.). Then the function, $\mathbf{x} \to M\mathbf{x}$ is onto. However, the vectors in $\mathbb{C}^n$ so obtained, consist of linear combinations of the columns of M. Therefore, the column rank of M is n. By Theorem 3.2 this equals the determinant rank and so $\det(M) \neq 0$. Now assume 4.) If $M\mathbf{v} = \mathbf{0}$ for some $\mathbf{v} \neq \mathbf{0}$, then the columns of M are linearly dependent and so by Theorem 3.2 we would have $\det(M) = 0$ contrary to 4.). Therefore, we obtain 4.) implies 5.). Now suppose 5.) and suppose $M\mathbf{v} = M\mathbf{w}$. Then $M(\mathbf{v} - \mathbf{w}) = \mathbf{0}$ and so by 5.), $\mathbf{v} - \mathbf{w} = \mathbf{0}$ showing that M is one to one.

Corollary 3.5 *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 $(M - \lambda I)\mathbf{x} = \mathbf{0}$.*

There may be repeated roots to the characteristic equation, 3.2 and we do not know whether the eigenvectors corresponding to a given eigenvalue form an independent set. However, we do know the following theorem.

Theorem 3.6 *Suppose $M\mathbf{v}_i = \lambda_i \mathbf{v}_i, i = 1, \cdots, r$, $\mathbf{v}_i \neq 0$, and that if $i \neq j$, then $\lambda_i \neq \lambda_j$. Then the set of eigenvectors, $\{\mathbf{v}_i\}_{i=1}^r$ is linearly independent.*

Proof: If the conclusion of this theorem is not true, then there exist non zero scalars, c_{k_j} such that

$$\sum_{j=1}^{m} c_{k_j} \mathbf{v}_{k_j} = \mathbf{0}. \tag{3.3}$$

Since any nonempty set of non negative integers has a smallest integer in the set, we may take m is as small as possible for this to take place. Then solving for $\mathbf{v}_{k_1}$ we obtain

$$\mathbf{v}_{k_1} = \sum_{k_j \neq k_1} d_{k_j} \mathbf{v}_{k_j} \tag{3.4}$$

where $d_{k_j} = c_{k_j}/c_{k_1} \neq 0$. Multiplying both sides by M, we obtain

$$\lambda_{k_1} \mathbf{v}_{k_1} = \sum_{k_j \neq k_1} d_{k_j} \lambda_{k_j} \mathbf{v}_{k_j},$$

which from 3.4 yield

$$\sum_{k_j \neq k_1} d_{k_j} \lambda_{k_1} \mathbf{v}_{k_j} = \sum_{k_j \neq k_1} d_{k_j} \lambda_{k_j} \mathbf{v}_{k_j}$$

and therefore,

$$0 = \sum_{k_j \neq k_1} \left(d_{k_j} \lambda_{k_1} - d_{k_j} \lambda_{k_j} \right) \mathbf{v}_{k_j}.$$

However, from the assumption that m is as small as possible for 3.3 to hold with all the scalars, c_{k_j} non zero, we must conclude that for some $j \neq 1$,

$$\left(d_{k_j} \lambda_{k_1} - d_{k_j} \lambda_{k_j} \right) = 0$$

which implies $\lambda_{k_1} = \lambda_{k_j}$, a contradiction.

In words, this theorem says that eigenvectors associated with distinct eigenvalues are linearly independent.

3.3 Block matrices

Suppose A is a matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix}$$

where each A_{ij} is itself a matrix. Such a matrix is called a block matrix.

Suppose also that B is also a block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & & \vdots \\ B_{m1} & \cdots & B_{mp} \end{pmatrix}$$

and that for all i, j, it makes sense to multiply $A_{ik}B_{kj}$ for all $k \in \{1, \cdots, m\}$ (That is the two matrices are conformable.) and that for each $k, A_{ik}B_{kj}$ is the same size. Then we can obtain AB as a block matrix as follows. $AB = C$ where C is a block matrix having $r \times p$ blocks such that the ij^{th} block is of the form

$$C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}.$$

This is just like matrix multiplication for matrices whose entries are scalars except we have to worry about the order of the factors and we are summing matrices rather than numbers. Why should this formula hold? If $m = 1$, we are looking at something of the form

$$\begin{pmatrix} A_{11} \\ \vdots \\ A_{r1} \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & B_{1p} \end{pmatrix}$$

Considering the columns of B and the rows of A in the usual manner, (Recall the ij^{th} scalar entry of AB equals the dot product of the i^{th} row of A with the j^{th} column of B.) we see the formula will hold in this case. If $m = 2$, we would have something like

$$\begin{pmatrix} A_{11} & A_{12} \\ \vdots & \vdots \\ A_{r1} & A_{r2} \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ B_{21} & \cdots & B_{2p} \end{pmatrix}.$$

However, if we let $\widetilde{A_{j1}} = \begin{pmatrix} A_{j1} & A_{j2} \end{pmatrix}$ and $\widetilde{B_{1j}} = \begin{pmatrix} B_{1j} \\ B_{2j} \end{pmatrix}$ then this appears in the form

$$\begin{pmatrix} \widetilde{A_{11}} \\ \vdots \\ \widetilde{A_{r1}} \end{pmatrix} \begin{pmatrix} \widetilde{B_{11}} & \cdots & \widetilde{B_{1p}} \end{pmatrix}$$

Now it is also not hard to see that

$$\widetilde{A_{j1}}\widetilde{B_{1s}} = \begin{pmatrix} A_{j1} & A_{j2} \end{pmatrix} \begin{pmatrix} B_{1s} \\ B_{2s} \end{pmatrix} = (A_{j1}B_{1s} + A_{j2}B_{2s}) = \sum_{i=1}^{2} A_{ji}B_{is}$$

and by the first part, this would be the js^{th} entry. Continuing this way, we see the assertion about block multiplication is so. The reader is invited to give a more complete proof if desired.

This simple idea of block multiplication turns out to be very useful later. For now we give a very interesting application.

Theorem 3.7 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then*

$$p_{BA}(t) = t^{n-m} p_{AB}(t),$$

so the eigenvalues of BA and AB are the same including multiplicities except that BA has $n - m$ extra zero eigenvalues.

Proof: Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

By Problem 11 of Chapter 1, it follows that $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ and $\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}$ have the same characteristic polynomials. Therefore, noting that BA is an $n \times n$ matrix and AB is an $m \times m$ matrix,

$$t^m \det(tI - BA) = t^n \det(tI - AB)$$

and so $\det(tI - BA) = p_{BA}(t) = t^{n-m} \det(tI - AB) = t^{n-m} p_{AB}(t)$. This proves the theorem.

3.4 Exercises

1. Let A and B be $n \times n$ matrices and let the columns of B be

   $$\mathbf{b}_1, \cdots, \mathbf{b}_n$$

   and the rows of A are

   $$\mathbf{a}_1^T, \cdots, \mathbf{a}_n^T.$$

   Show the columns of AB are

   $$A\mathbf{b}_1 \cdots A\mathbf{b}_n$$

   and the rows of AB are

   $$\mathbf{a}_1^T B \cdots \mathbf{a}_n^T B.$$

2. Let M be an $n \times n$ matrix. Then we define the adjoint of M,denoted by M^* to be the transpose of the conjugate of M. For example,

   $$\begin{pmatrix} 2 & i \\ 1+i & 3 \end{pmatrix}^* = \begin{pmatrix} 2 & 1-i \\ -i & 3 \end{pmatrix}.$$

   We say a matrix, M, is self adjoint if $M^* = M$. Show the eigenvalues of a self adjoint matrix are all real. If the matrix has all real entries, we call the matrix symmetric. Show that the eigenvalues and eigenvectors of a symmetric matrix occur in conjugate pairs.

3. Find the eigenvalues and eigenvectors of the matrix

   $$\begin{pmatrix} 7 & -2 & 0 \\ 8 & -1 & 0 \\ -2 & 4 & 6 \end{pmatrix}.$$

   Can you find three independent eigenvectors?

4. Find the eigenvalues and eigenvectors of the matrix

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

Can you find three independent eigenvectors in this case?

5. 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 we can do this we say the matrix is diagonalizable.

6. Show that a matrix, M is diagonalizable if and only if it has a basis of eigenvectors. **Hint:** The first part is done in Problem 5. 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.

7. 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 $A_{22} = (3)$.

Linear Transformations

Definition 4.1 *Let V and W be two finite dimensional vector spaces. We say*

$$L \in \mathcal{L}(V, W)$$

if for all scalars α and β, and vectors $\mathbf{v}, \mathbf{w}$,

$$L(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha L(\mathbf{v}) + \beta L(\mathbf{w}).$$

An example of a linear transformation is familiar matrix multiplication. Let $A = (a_{ij})$ be an $m \times n$ matrix. Then we may define a linear transformation $L : \mathbb{F}^n \to \mathbb{F}^m$ by

$$(L\mathbf{v})_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$

Here

$$\mathbf{v} \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

Also, if V is an n dimensional vector space and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for V, there exists a linear map

$$q : \mathbb{F}^n \to V$$

defined as

$$q(\mathbf{a}) \equiv \sum_{i=1}^{n} a_i \mathbf{v}_i$$

where

$$\mathbf{a} = \sum_{i=1}^{n} a_i \mathbf{e}_i,$$

for $\mathbf{e}_i$ the standard basis vectors for $\mathbb{F}^n$ consisting of

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

where the one is in the ith slot. It is clear that q defined in this way, is one to one, onto, and linear. For $\mathbf{v} \in V$, $q^{-1}(\mathbf{v})$ is a list of scalars called the components of $\mathbf{v}$ with respect to the basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$.

Definition 4.2 *Given a linear transformation L, mapping V to W, where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis of V and $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis for W, an $m \times n$ matrix $A = (a_{ij})$is called the matrix of the transformation L with respect to the given choice of bases for V and W, if whenever $\mathbf{v} \in V$, then multiplication of the components of $\mathbf{v}$ by (a_{ij}) yields the components of $L\mathbf{v}$.*

The following diagram is descriptive of the definition. Here q_V and q_W are the maps defined above with reference to the bases, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ respectively.

$$\begin{array}{ccccc} & & L & & \\ \{\mathbf{v}_1, \cdots, \mathbf{v}_n\} & V & \rightarrow & W & \{\mathbf{w}_1, \cdots, \mathbf{w}_m\} \\ & q_V \uparrow & \circ & \uparrow q_W & \\ & \mathbb{F}^n & \rightarrow & \mathbb{F}^m & \\ & & A & & \end{array} \tag{4.1}$$

Letting $\mathbf{b} \in \mathbb{F}^n$, this requires

$$\sum_{i,j} a_{ij} b_j \mathbf{w}_i = L \sum_j b_j \mathbf{v}_j = \sum_j b_j L\mathbf{v}_j.$$

Now

$$L\mathbf{v}_j = \sum_i c_{ij} \mathbf{w}_i \tag{4.2}$$

for some choice of scalars c_{ij} because $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis for W. Hence

$$\sum_{i,j} a_{ij} b_j \mathbf{w}_i = \sum_j b_j \sum_i c_{ij} \mathbf{w}_i = \sum_{i,j} c_{ij} b_j \mathbf{w}_i.$$

It follows from the linear independence of $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ that

$$\sum_j a_{ij} b_j = \sum_j c_{ij} b_j$$

for any choice of $\mathbf{b} \in \mathbb{F}^n$ and consequently

$$a_{ij} = c_{ij}$$

where c_{ij} is defined by 4.2. It may help to write 4.2 in the form

$$\left(\begin{array}{ccc} L\mathbf{v}_1 & \cdots & L\mathbf{v}_n \end{array} \right) = \left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{array} \right) C = \left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{array} \right) A \tag{4.3}$$

where $C = (c_{ij}), A = (a_{ij})$.

Example 4.3 *Let*

$$V \equiv \{ \textit{ polynomials of degree 3 or less} \},$$

$$W \equiv \{ \textit{ polynomials of degree 2 or less} \},$$

and $L \equiv D$ where D is the differentiation operator. A basis for V is $\{1, x, x^2, x^3\}$ and a basis for W is $\{1, x, x^2\}$.

What is the matrix of this linear transformation with respect to this basis? Using 4.3,

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

It follows from this that

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

Now consider the important case where $V = \mathbb{F}^n$, $W = \mathbb{F}^m$, and the basis chosen is the standard basis of vectors $\mathbf{e}_i$ described above. Let L be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$ and let A be the matrix of the transformation with respect to these bases. In this case the coordinate maps q_V and q_W are simply the identity map and we need

$$\pi_i(L\mathbf{b}) = \pi_i(A\mathbf{b})$$

where π_i denotes the map which takes a vector in $\mathbb{F}^m$ and returns the *ith* entry in the vector, the *ith* component of the vector with respect to the standard basis vectors. Thus, if the components of the vector in $\mathbb{F}^n$ with respect to the standard basis are $(b_1, \cdots, b_n)$,

$$\mathbf{b} = \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}^T = \sum_i b_i \mathbf{e}_i,$$

then

$$\pi_i(L\mathbf{b}) \equiv (L\mathbf{b})_i = \sum_j a_{ij} b_j.$$

What about the situation where different pairs of bases are chosen for V and W? How are the two matrices with respect to these choices related? Consider the following diagram which illustrates the situation.

$$\begin{array}{ccc} \mathbb{F}^n & \xrightarrow{A_2} & \mathbb{F}^m \\ q_2 \downarrow & \circ & p_2 \downarrow \\ V & \xrightarrow{L} & W \\ q_1 \uparrow & \circ & p_1 \uparrow \\ \mathbb{F}^n & \xrightarrow{A_1} & \mathbb{F}^m \end{array}$$

In this diagram q_i and p_i are coordinate maps as described above. We see from the diagram that

$$p_1^{-1} p_2 A_2 q_2^{-1} q_1 = A_1,$$

where $q_2^{-1} q_1$ and $p_1^{-1} p_2$ are one to one, onto, and linear maps.

Definition 4.4 *In the special case where $V = W$ and only one basis is used for $V = W$, this becomes*

$$q_1^{-1} q_2 A_2 q_2^{-1} q_1 = A_1.$$

Letting S be the matrix of the linear transformation $q_2^{-1} q_1$ with respect to the standard basis vectors in $\mathbb{F}^n$, we get

$$S^{-1} A_2 S = A_1. \tag{4.4}$$

When this occurs, we say that A_1 is similar to A_2 and we call $A \to S^{-1} A S$ a similarity transformation.

Theorem 4.5 *In the vector space of $n \times n$ matrices, we say*

$$A \sim B$$

if there exists an invertible matrix S such that

$$A = S^{-1}BS.$$

Then $\sim$ is an equivalence relation and $A \sim B$ if and only if whenever V is an n dimensional vector space, there exists $L \in \mathcal{L}(V,V)$ and bases $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ such that A is the matrix of L with respect to $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and B is the matrix of L with respect to $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$.

Proof: $A \sim A$ because $S = I$ works in the definition. If $A \sim B$, then $B \sim A$, because

$$A = S^{-1}BS$$

implies

$$B = SAS^{-1}.$$

If $A \sim B$ and $B \sim C$, then

$$A = S^{-1}BS,\ B = T^{-1}CT$$

and so

$$A = S^{-1}T^{-1}CTS = (TS)^{-1}\,CTS$$

which implies $A \sim C$. This verifies the first part of the conclusion.

Now let V be an n dimensional vector space, $A \sim B$ and pick a basis for V,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}.$$

Define $L \in \mathcal{L}(V,V)$ by

$$L\mathbf{v}_i \equiv \sum_j a_{ji}\mathbf{v}_j$$

where $A = (a_{ij})$. Then if $B = (b_{ij})$, and $S = (s_{ij})$ is the matrix which provides the similarity transformation,

$$A = S^{-1}BS,$$

between A and B, it follows that

$$L\mathbf{v}_i = \sum_{r,s,j} s_{ir}b_{rs}\left(s^{-1}\right)_{sj}\mathbf{v}_j. \tag{4.5}$$

Now define

$$\mathbf{w}_i \equiv \sum_j \left(s^{-1}\right)_{ij}\mathbf{v}_j.$$

Then from 4.5,

$$\sum_i \left(s^{-1}\right)_{ki} L\mathbf{v}_i = \sum_{i,j,r,s} \left(s^{-1}\right)_{ki} s_{ir}b_{rs}\left(s^{-1}\right)_{sj}\mathbf{v}_j$$

and so

$$L\mathbf{w}_k = \sum_s b_{ks}\mathbf{w}_s.$$

This proves the theorem because the if part of the conclusion was established earlier.

Definition 4.6 *We say 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. We also write $diag(\lambda_1 \cdots, \lambda_n)$ to denote the diagonal matrix having the λ_i down the main diagonal.*

The following theorem is of great significance.

Theorem 4.7 *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 4.6 consists of the $n \times n$ matrix whose columns are the eigenvectors of A and $D = diag(\lambda_1, \cdots, \lambda_n)$.*

Proof: Suppose first that $\mathbb{F}^n$ has a basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ where $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Then let S denote the matrix $(\mathbf{v}_1 \cdots \mathbf{v}_n)$ and let $S^{-1} \equiv \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix}$ where $\mathbf{u}_i^T \mathbf{v}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$. We know S^{-1} exists because S has rank n. Then from block multiplication, we obtain,

$$S^{-1}AS = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (A\mathbf{v}_1 \cdots A\mathbf{v}_n)$$

$$= \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (\lambda_1 \mathbf{v}_1 \cdots \lambda_n \mathbf{v}_n)$$

$$= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.$$

Next suppose A is diagonalizable so $S^{-1}AS = D \equiv diag(\lambda_1, \cdots, \lambda_n)$. Then the columns of S form a basis because S^{-1} is given to exist. It only remains to verify that these columns of A are eigenvectors. But letting $S = (\mathbf{v}_1 \cdots \mathbf{v}_n)$, we see $AS = SD$ and so $(A\mathbf{v}_1 \cdots A\mathbf{v}_n) = (\lambda_1 \mathbf{v}_1 \cdots \lambda_n \mathbf{v}_n)$ which shows that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$. This proves the theorem.

We can talk about the determinant of a linear transformation as described in the following corollary.

Corollary 4.8 *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 4.5 that

$$A = S^{-1}BS$$

and so

$$\det(A) = \det\left(S^{-1}\right) \det(B) \det(S).$$

But

$$1 = \det(I) = \det\left(S^{-1}S\right) = \det(S) \det\left(S^{-1}\right)$$

and so

$$\det(A) = \det(B)$$

which proves the corollary.

Definition 4.9 *Let $A \in \mathcal{L}(X,Y)$ where X and Y are finite dimensional vector spaces. We define 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 4.10 *Let $A \in \mathcal{L}(X,Y)$. Then rank(A) = 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: We recall the diagram which describes what we mean by the matrix of A. Here the two bases are as indicated.

$$\begin{array}{ccccc} \{v_1,\cdots,v_n\} & X & \underrightarrow{A} & Y & \{w_1,\cdots,w_m\} \\ & q_X\uparrow & \circ & \uparrow q_Y & \\ & \mathbb{F}^n & \underrightarrow{M} & \mathbb{F}^m & \end{array}$$

Let $\{z_1,\cdots,z_r\}$ be a basis for $A(X)$. Then since the linear transformation, q_Y is one to one and onto, $\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}$ is a linearly independent set of vectors in $\mathbb{F}^m$. Let $Au_i = z_i$. Then

$$Mq_X^{-1}u_i = q_Y^{-1}z_i$$

and so the dimension of $M(\mathbb{F}^n) \geq r$. Now if $M(\mathbb{F}^n) < r$ then there exists

$$\mathbf{y} \in M(\mathbb{F}^n) \setminus \text{span}\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}.$$

But then there exists $\mathbf{x} \in \mathbb{F}^n$ with $M\mathbf{x} = \mathbf{y}$. Hence

$$\mathbf{y} = M\mathbf{x} = q_Y^{-1}Aq_X\mathbf{x} \in \text{span}\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}$$

a contradiction. This proves the theorem.

The following result is of fundamental importance.

Theorem 4.11 *Let $L \in \mathcal{L}(V,V)$ where V is a finite dimensional vector space. Then the following are equivalent.*

1. *L is one to one.*
2. *L maps a basis to a basis.*
3. *L is onto.*
4. *$\det(L) \neq 0$*
5. *If $Lv = 0$ then $v = 0$.*

Proof: Suppose first L is one to one and let $\{v_i\}_{i=1}^n$ be a basis. Then if $\sum_{i=1}^n c_i Lv_i = 0$ 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 $y \in V$, we know from part 2.) that there are constants, $\{c_i\}$ such that $y = \sum_{i=1}^n c_i Lv_i = L\left(\sum_{i=1}^n c_i v_i\right)$. Thus L is onto and we have shown that 2.) implies 3.). Now suppose 3.). Then the operation consisting of multiplication by the matrix of L, M_L, must be onto. However, the vectors in $\mathbb{F}^n$ so obtained, consist of linear combinations of the columns of M_L. Therefore, the column rank of M_L is n. By Theorem 3.2 this equals the determinant rank and so $\det(M_L) \equiv \det(L) \neq 0$. Now assume 4.) If $Lv = 0$ for some $v \neq 0$, it follows that $M_L\mathbf{x} = 0$ for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the columns of M_L are linearly dependent and so by Theorem 3.2 we would have $\det(M_L) = \det(L) = 0$ contrary to 4.). Therefore, we obtain 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.

Inner Product Spaces

We will assume the field of scalars, $\mathbb{F}$, is either $\mathbb{R}$ or $\mathbb{C}$ in this section although much, if not all, of what will be presented could be extended to arbitrary fields.

Definition 5.1 *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. $|x+y| \leq |x| + |y|$.

Note that we are using the same notation for the norm as for the absolute value. This is because the norm is just a generalization to vector spaces of the concept of absolute value. However, 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 5.2 *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 Formula 2 and Formula 3 imply $(x, ay+bz) = \overline{a}(x,y) + \overline{b}(x,z)$.
We will show that if $(\cdot,\cdot)$ is an inner product, then

$$(x,x)^{1/2} \equiv |x|$$

defines a norm.

Definition 5.3 *A normed linear space in which the norm comes from an inner product as just described is called an inner product space.*

Example 5.4 *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.

Example 5.5 *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 5.6 *Let V be any finite dimensional vector space and let $\{v_1, \cdots, v_n\}$ be a basis. We can make this into an inner product space as follows. We decree that*

$$(v_i, v_j) \equiv \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}.$$

Now we define the inner product by

$$(x, y) \equiv \sum_{i=1}^{n} x^i \overline{y^i}$$

where

$$x = \sum_{i=1}^{n} x^i v_i, \ y = \sum_{i=1}^{n} y^i v_i.$$

This example shows there is no loss of generality when studying finite dimensional vector spaces in assuming the vector space is actually an inner product space.

Theorem 5.7 *(Cauchy Schwartz) In any inner product space*

$$|(x, y)| \leq |x||y|.$$

where $|x| \equiv (x, x)^{1/2}$.

Proof: Let $\omega \in \mathbb{C}$, $|\omega| = 1$, and $\overline{\omega}(x, y) = |(x, y)| = \mathrm{Re}(x, y\omega)$. Let

$$F(t) = (x + ty\omega, x + t\omega y).$$

If $y = 0$ there is nothing to prove because

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

and so $(x, 0) = 0$. Thus, we may assume $y \neq 0$. Then from the axioms of the inner product,

$$F(t) = |x|^2 + 2t\mathrm{Re}(x, \omega y) + t^2|y|^2 \geq 0.$$

This yields

$$|x|^2 + 2t|(x, y)| + t^2|y|^2 \geq 0.$$

Since this inequality holds for all $t \in \mathbb{R}$, it follows from the quadratic formula that

$$4|(x, y)|^2 - 4|x|^2|y|^2 \leq 0.$$

This yields the conclusion and proves the theorem.

Earlier it was claimed that the inner product defines a norm. In this next proposition this claim is proved.

Proposition 5.8 *For an inner product space, $|x| \equiv (x,x)^{1/2}$ does specify a norm.*

Proof: All the axioms are obvious except the triangle inequality. To verify this,

$$\begin{aligned} |x+y|^2 &\equiv (x+y, x+y) \equiv |x|^2 + |y|^2 + 2\mathrm{Re}\,(x,y) \\ &\leq |x|^2 + |y|^2 + 2\,|(x,y)| \\ &\leq |x|^2 + |y|^2 + 2\,|x|\,|y| = (|x| + |y|)^2. \end{aligned}$$

The best norms of all are those which come from an inner product because of the following identity which is known as the parallelogram identity.

Proposition 5.9 *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 5.10 *We say 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, we obtain

$$0 = \sum_{j=1}^{n} c^j\,(u_j, u_k) = \sum_{j=1}^{n} c^j \delta_{jk} = c^k.$$

Lemma 5.11 *Let X be a finite dimensional inner product space of dimension n whose basis is $\{x_1, \cdots, x_n\}$. Then there exists an orthonormal basis for X, $\{u_1, \cdots, u_n\}$ which has the property that for each $k \leq n$, span$(x_1, \cdots, x_k) =$ span$\,(u_1, \cdots, u_k)$.*

Proof: Let $\{x_1, \cdots, x_n\}$ be a basis for X. Let $u_1 \equiv x_1/\,|x_1|$. Thus for $k = 1$, we have span $(u_1) =$ 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 span $(x_1, \cdots, x_k) =$ span $(u_1, \cdots, u_k)$. Then we define

$$u_{k+1} \equiv \frac{x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j)\, u_j}{\left| x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j)\, u_j \right|},$$

where the denominator is not equal to zero because the x_j form a basis. Thus $u_{k+1} \in$ span $(x_1, \cdots, x_k, x_{k+1})$ and $x_{k+1} \in$ span $(u_1, \cdots, u_k, u_{k+1})$ so by induction, the two spans are equal. Then 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.

The process by which these vectors were generated is called the Gramm Schmidt process.

Lemma 5.12 *Suppose $\{u_j\}_{j=1}^n$ is an orthonormal basis for an inner product space X. Then for all $x \in X$,*

$$x = \sum_{j=1}^{n} (x, u_j)\, u_j.$$

Proof: By assumption that this is an orthonormal basis,

$$\sum_{j=1}^{n} (x, u_j)\, (u_j, u_l) = (x, u_l)\,.$$

Letting $y = \sum_{j=1}^n (x, u_j)\, u_j$, it follows $(x - y, u_j) = 0$ for all j. Hence, for any choice of scalars, $c^1, \cdots, c^n$,

$$\left(x - y, \sum_{j=1}^{n} c^j u_j \right) = 0$$

and so $(x - y, z) = 0$ for all $z \in X$. Thus this holds in particular for $z = x - y$. Therefore, $x = y$ and this proves the theorem.

The following theorem is of fundamental importance. First we note that a subspace of an inner product space is also an inner product space itself.

Theorem 5.13 *Let M be a subspace of X, a finite dimensional inner product space and let $\{x_i\}_{i=1}^m$ be an orthonormal basis for M. Then if $y \in X$ and $w \in M$,*

$$|y - w|^2 = \inf \left\{ |y - z|^2 : z \in M \right\} \tag{5.1}$$

if and only if

$$(y - w, z) = 0 \tag{5.2}$$

for all $z \in M$. Furthermore,

$$w = \sum_{i=1}^{m} (y, x_i)\, x_i \tag{5.3}$$

has this property.

Proof: Let $t \in \mathbb{R}$. Then from the properties of the inner product,

$$|y - (w + t\,(z - w))|^2 = |y - w|^2 + 2t\mathrm{Re}\,(y - w, w - z) + t^2\,|z - w|^2\,. \tag{5.4}$$

If $(y - w, z) = 0$ for all $z \in M$, then letting $t = 1$, the middle term in the above expression vanishes and so we see that $|y - z|^2$ is minimized when $z = w$. Conversely, if 5.1 holds, then the middle term of 5.4 must also vanish since otherwise, we could choose small real t such that

$$|y - w|^2 > |y - (w + t\,(z - w))|^2\,.$$

It follows, letting $z_1 = w - z$ that

$$\mathrm{Re}\,(y - w, z_1) = 0$$

for all $z_1 \in M$. Now letting $\omega \in \mathbb{C}$ be such that $\omega (y-w, z_1) = |(y-w, z_1)|$, we see

$$|(y-w, z_1)| = (y-w, \overline{\omega} z_1) = 0,$$

which proves the first part of the theorem since z_1 is arbitrary. It only remains to verify that w given in 5.3 satisfies 5.2. To do so, we note that if c_i, d_i are scalars, then the properties of the inner product imply

$$\left(\sum_{i=1}^m c_i x_i, \sum_{j=1}^m d_j x_j \right) = \sum_i c_i \overline{d_i}.$$

By Lemma 5.12,

$$z = \sum_i (z, x_i)\, x_i$$

and so

$$\left(y - \sum_{i=1}^m (y, x_i)\, x_i, z \right) = \left(y - \sum_{i=1}^m (y, x_i)\, x_i, \sum_{i=1}^m (z, x_i)\, x_i \right)$$

$$= \sum_{i=1}^m \overline{(z, x_i)}\, (y, x_i) - \left(\sum_{i=1}^m (y, x_i)\, x_i, \sum_{j=1}^m (z, x_j)\, x_j \right)$$

$$= \sum_{i=1}^m \overline{(z, x_i)}\, (y, x_i) - \sum_{i=1}^m (y, x_i)\, \overline{(z, x_i)} = 0.$$

This proves the 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 5.14 *Let $f \in \mathcal{L}(X, \mathbb{F})$ where X is an inner product space of dimension n. Then there exists a unique $z \in X$ such that for all $x \in X$,*

$$f(x) = (x, z).$$

Proof: First we 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 5.11, there exists an orthonormal basis, $\{u_j\}_{j=1}^n$. Define

$$z \equiv \sum_{j=1}^n \overline{f(u_j)} u_j.$$

Then using Lemma 5.12,

$$\begin{aligned} (x, z) &= \left(x, \sum_{j=1}^n \overline{f(u_j)} u_j \right) = \sum_{j=1}^n f(u_j)\, (x, u_j) \\ &= f\left(\sum_{j=1}^n (x, u_j)\, u_j \right) = f(x). \end{aligned}$$

This proves the theorem.

Corollary 5.15 *Let $A \in \mathcal{L}(X,Y)$ where X and Y are two inner product spaces of finite dimension. Then there exists a unique $A^* \in \mathcal{L}(Y,X)$ such that*

$$(Ax,y)_Y = (x,A^*y)_X \tag{5.5}$$

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 $x \in X$,

$$(x,A^*(ay_1+by_2))_X \equiv \overline{a}(Ax,y_1) + \overline{b}(Ax,y_2)$$

$$\overline{a}(x,A^*(y_1)) + \overline{b}(x,A^*(y_2)) = (x,aA^*(y_1)+bA^*(y_2)) .$$

By uniqueness, $A^*(ay_1+by_2) = aA^*(y_1)+bA^*(y_2)$ which shows A^* is linear as claimed. The last assertion about the map which sends a linear transformation, A to A^*follows from

$$(x,A^*y+A^*y) = (Ax,y)+(Bx,y) = ((A+B)x,y) \equiv \left(x,(A+B)^*y\right)$$

and for α a scalar,

$$\left(x,(\alpha A)^*y\right) = (\alpha Ax,y) = \alpha(x,A^*y) = (x,\overline{\alpha}A^*y) .$$

This proves the corollary.

The linear map, A^* is called the adjoint of A. In the case when $A : X \to X$ and $A = A^*$, we call A a self adjoint map. In the case where $X = Y = \mathbb{C}^n$, we have the following simple description of the adjoint linear transformation. In this case, we consider an $n \times n$ matrix as a linear transformation mapping $\mathbb{C}^n$ to $\mathbb{C}^n$ according to the usual rules of matrix multiplication. Thus

$$(M\mathbf{x})_i \equiv \sum_j M_{ij}x_j$$

where $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. In order to do this we recall from high school algebra that for $z,w \in \mathbb{C}$, $\overline{zw} = \overline{zw}$ and also $\overline{\sum_j z_j} = \sum_j \overline{z_j}$. In words, the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.

Theorem 5.16 *Let M be an $n \times n$ matrix. Then $M^* = \left(\overline{M}\right)^T$ in words, the transpose of the conjugate of M is equal to the adjoint.*

Proof: We consider the following using the definition of the inner product in $\mathbb{C}^n$.

$$(M\mathbf{x},\mathbf{y}) = (\mathbf{x},M^*\mathbf{y}) \equiv \sum_i x_i \overline{\sum_j (M^*)_{ij}y_j} = \sum_{i,j} x_i \overline{(M^*)_{ij}}\,\overline{y_j}.$$

Also

$$(M\mathbf{x},\mathbf{y}) = \sum_j \sum_i M_{ji} x_i \overline{y_j}.$$

Since $\mathbf{x},\mathbf{y}$ are arbitrary vectors in $\mathbb{C}^n$, it follows that $M_{ji} = \overline{(M^*)_{ij}}$ and so, taking conjugates of both sides, we see

$$M^*_{ij} = \overline{M_{ji}}$$

which gives the conclusion of the theorem.

The next theorem is interesting.

Theorem 5.17 *Suppose V is a subspace of $\mathbb{F}^n$ having dimension $p \leq n$. Then there exists a $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ such that $QV \subseteq \mathbb{F}^p$ and $|Q\mathbf{x}| = |\mathbf{x}|$ for all $\mathbf{x}$. Also*

$$Q^*Q = QQ^* = I.$$

Proof: By Lemma 5.11 there exists an orthonormal basis for $V, \{\mathbf{v}_i\}_{i=1}^p$. By using the Gram Schmidt process this may be extended to an orthonormal basis of the whole space, $\mathbb{F}^n$,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{v}_{p+1}, \cdots, \mathbf{v}_n\}.$$

Now define $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ by $Q(\mathbf{v}_i) \equiv \mathbf{e}_i$ and extend linearly. If $\sum_{i=1}^n x_i \mathbf{v}_i$ is an arbitrary element of $\mathbb{F}^n$,

$$\left| Q\left(\sum_{i=1}^n x_i \mathbf{v}_i \right) \right|^2 = \left| \sum_{i=1}^n x_i \mathbf{e}_i \right|^2 = \sum_{i=1}^n |x_i|^2 = \left| \sum_{i=1}^n x_i \mathbf{v}_i \right|^2.$$

It remains to verify that $Q^*Q = QQ^* = I$. To do so, let $\mathbf{x},\mathbf{y} \in \mathbb{F}^p$. Then

$$(Q(\mathbf{x}+\mathbf{y}), Q(\mathbf{x}+\mathbf{y})) = (\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}).$$

Thus

$$|Q\mathbf{x}|^2 + |Q\mathbf{y}|^2 + 2\mathrm{Re}\,(Q\mathbf{x},Q\mathbf{y}) = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathrm{Re}\,(\mathbf{x},\mathbf{y})$$

and since Q preserves norms, it follows that for all $\mathbf{x},\mathbf{y} \in \mathbb{F}^n$,

$$\mathrm{Re}\,(Q\mathbf{x},Q\mathbf{y}) = \mathrm{Re}\,(\mathbf{x},Q^*Q\mathbf{y}) = \mathrm{Re}\,(\mathbf{x},\mathbf{y}).$$

Therefore, since this holds for all $\mathbf{x}$, it follows that $Q^*Q\mathbf{y} = \mathbf{y}$ showing that $Q^*Q = I$. Now

$$Q = Q(Q^*Q) = (QQ^*)Q.$$

Since Q is one to one, this implies

$$I = QQ^*$$

and proves the theorem.

Definition 5.18 *If Q is an $n \times n$ matrix, with the property that $Q^*Q = I$, we call Q a unitary matrix. Note that if Q is a real matrix, then $Q^* = Q^T$ and the matrix is unitary means in this case that $Q^TQ = I$. Such real matrices are also called orthogonal.*

Lemma 5.19 *If $Q^*Q = I$, then $QQ^* = I$ also. Thus $Q^{-1} = Q^*$.*

Proof: First we note that since $Q^*Q = I$, it is also the case that Q must be one to one. Therefore, by Theorem 4.11 Q is also onto and has an inverse. Hence by the fact that matrix multiplication is associative,

$$(QQ^*)\,Q = Q\,(Q^*Q) = QI = Q.$$

Therefore,

$$I = QQ^{-1} = ((QQ^*)\,Q)\,Q^{-1} = (QQ^*)\left(QQ^{-1}\right) = QQ^*.$$

This proves the lemma.

Definition 5.20 *Let X and Y be inner product spaces and let $x \in X$ and $y \in Y$. We 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.

We leave the verification that this is a linear map to the reader. Be sure to verify this! The following lemma has some of the most important properties of this linear transformation.

Lemma 5.21 *Let X, Y, Z be inner product spaces. Then for α a scalar,*

$$(\alpha\,(y \otimes x))^* = \overline{\alpha} x \otimes y \tag{5.6}$$

$$(z \otimes y_1)\,(y_2 \otimes x) = (y_2, y_1)\,z \otimes x \tag{5.7}$$

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 5.6.

To verify 5.7, 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 5.7 give the same answer for every $u \in X$, it follows the two transformations are the same. This proves the lemma.

Definition 5.22 *Let X, Y be two vector spaces. Then we 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 5.23 *Let X and Y be finite dimensional inner product spaces. Then $\mathcal{L}(X,Y)$ is a vector space with the above definition of what it means to multiply by a scalar and add. Let $\{v_1, \cdots, v_n\}$ be an orthonormal basis for X and $\{w_1, \cdots, w_m\}$ be an orthonormal basis for Y. Then a basis for $\mathcal{L}(X,Y)$ is $\{v_i \otimes w_j : i = 1, \cdots, n, j = 1, \cdots, m\}$.*

Proof: It is obvious that $\mathcal{L}(X,Y)$ is a vector space. It remains to verify the given set is a basis. We consider the following:

$$\left(\left(A-\sum_{k,l}\left(Av_k,w_l\right)w_l\otimes v_k\right)v_p,w_r\right)=\left(Av_p,w_r\right)-$$

$$\sum_{k,l}\left(Av_k,w_l\right)\left(v_p,v_k\right)\left(w_l,w_r\right)$$

$$=\left(Av_p,w_r\right)-\sum_{k,l}\left(Av_k,w_l\right)\delta_{pk}\delta_{rl}$$

$$=\left(Av_p,w_r\right)-\left(Av_p,w_r\right)=0.$$

Letting $A-\sum_{k,l}\left(Av_k,w_l\right)w_l\otimes v_k=B$, this shows that $Bv_p=0$ since w_r is an arbitrary element of the basis for Y. Since v_p is an arbitrary element of the basis for X, it follows $B=0$ as hoped. This has shown $\{v_i\otimes w_j : i=1,\cdots,n, j=1,\cdots,m\}$ spans $\mathcal{L}(X,Y)$.

It only remains to verify the $v_i\otimes w_j$ are linearly independent. Suppose then that

$$\sum_{i,j}c_{ij}v_i\otimes w_j=0$$

Then,

$$0=\left(v_s,\sum_{i,j}c_{ij}v_i\otimes w_j\left(w_r\right)\right)=\left(v_s,\sum_{i,j}c_{ij}v_i\left(w_r,w_j\right)\right)$$

$$=\sum_{i,j}\left(v_s,c_{ij}v_i\right)\left(w_r,w_j\right)=\sum_{i,j}\overline{c_{ij}}\delta_{si}\delta_{rj}=\overline{c_{sr}}$$

showing all the coefficients equal zero. This proves independence.

Note this shows the dimension of $\mathcal{L}(X,Y)=nm$. The theorem is also of enormous importance because it shows we 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 5.24 *Let $A=\sum_{i,j}c_{ij}w_i\otimes v_j\in\mathcal{L}(X,Y)$ where as before, the vectors, $\{w_i\}$ are an orthonormal basis for Y and the vectors, $\{v_j\}$ are an orthonormal basis for X. Then if the matrix of A has components, M_{ij}, it follows that $M_{ij}=c_{ij}$.*

Proof: We recall the diagram which describes what the matrix of a linear transformation is.

$$\begin{array}{ccccc}\{v_1,\cdots,v_n\} & X & \underrightarrow{A} & Y & \{w_1,\cdots,w_m\}\\ & q_V\uparrow & \circ & \uparrow q_W & \\ & \mathbb{F}^n & \underrightarrow{M} & \mathbb{F}^m & \end{array}$$

Thus, multiplication by the matrix, M followed by the map, q_W is the same as q_V followed by the linear transformation, A. Denoting by M_{ij} the components of the matrix, M, and letting $\mathbf{x}=(x_1,\cdots,x_n)\in\mathbb{F}^n$,

$$\sum_i w_i\sum_j M_{ij}x_j=A\left(\sum_k x_kv_k\right)$$

$$= \sum_{i,j} \sum_{k} c_{ij} x_k \delta_{kj} w_i = \sum_i w_i \sum_j c_{ij} x_j.$$

It follows from the linear independence of the w_i that for any $\mathbf{x} \in \mathbb{F}^n$,

$$\sum_j M_{ij} x_j = \sum_j c_{ij} x_j$$

which establishes the theorem.

5.1 Least squares

A common problem in experimental work is to find a straight line which approximates as well as possible a collection of points in the plane $\{(x_i, y_i)\}_{i=1}^p$. The most 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. First we give the following lemma. In this lemma, you can let V and W equal $\mathbb{F}^n$ and $\mathbb{F}^m$ if you like and you can take A to be a matrix.

Lemma 5.25 *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 5.13 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, we know from this theorem that this happens if and only if $Ax_0 - y$ is perpendicular to every $Ax \in A(V)$. Therefore, our 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 our solution is obtained by the solution to $A^*Ax_0 = A^*y$. Furthermore, the argument just given shows there exists a unique solution to this system of equations.

With this Lemma, we can now discuss the problem of finding the least squares regression line in statistics. Suppose we are given points in the plane, $\{(x_i, y_i)\}_{i=1}^n$ and we would really 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, we try to find m, b such that we do the best we 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 we 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 we possibly can. According to what we just showed, our solution is the unique solution to the system

$$A^*A \begin{pmatrix} m \\ b \end{pmatrix} = A^* \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

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, we see

$$m = \frac{-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + \left(\sum_{i=1}^n x_i y_i\right) n}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}$$

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 we would want to solve as well as possible for a, b, and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

We can solve this in the same way.

Definition 5.26 *Let S be a subset of an inner product space, X. We 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 5.27 *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 $y \in \ker(A^*)^\perp$. We need to show there exists x such that $Ax = y$. Since we don't know such an x exists right now, we will take the least squares solution to the problem. Thus we take $A^*Ax = A^*y$. It follows $A^*(y - Ax) = 0$ and so $y - Ax \in \ker(A^*)$ which implies from our assumption about y that $(y - Ax, y) = 0$. Also, since Ax is the closest point to y in $A(V)$, we know by Theorem 5.13 that $(y - Ax, Ax_1) = 0$ for all $x_1 \in V$. In particular this is true for $x_1 = x$ and so $0 = (y - Ax, y) - (y - Ax, Ax) = |y - Ax|^2$, showing that $y = Ax$. Thus $A(V) \supseteq \ker(A^*)^\perp$ and this proves the Theorem.

Corollary 5.28 *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 $A(V) = W$.

5.2 Exercises

1. Find the best solution to the system

$$\begin{aligned} x + 2y &= 6 \\ 2x - y &= 5 \\ 3x + 2y &= 0 \end{aligned}$$

2. Suppose you are given the data, $(1,2), (2,4), (3,8), (0,0)$. Find the linear regression line using your formulas derived above. Then graph your data along with your regression line.

3. Generalize the least squares procedure to the situation in which data is given and you desire to fit it with an expression of the form $y = af(x) + bg(x) + c$ where the problem would be to find a, b and c in order to minimize the error. Could this be generalized to higher dimensions? How about more functions?

4. Let $A \in \mathcal{L}(X, Y)$ where X and Y are finite dimensional vector spaces with the dimension of X equal to n. We define $\text{rank}(A) \equiv$ dimension of (AX) and $null(A) \equiv$ dimension of $\{x : Ax = 0\} \equiv$ dimension of $\ker(A)$. Show that $null(A) + \text{rank}(A) = \dim(X)$. **Hint:** Let $\{x_i\}_{i=1}^r$ be a basis for $\ker(A)$ and let $\{x_i\}_{i=1}^r \cup \{y_i\}_{i=1}^{n-r}$ be a basis for X. Then show that $\{Ay_i\}_{i=1}^{n-r}$ is linearly independent and spans AX.

5. Let A be an $m \times n$ matrix. Show the column rank of A equals the column rank of A^*A. Next verify column rank of A^*A is no larger than column rank of A^*. Next justify the following inequality to conclude the column rank of A equals the column rank of A^*.

$$\text{rank } (A) = \text{rank } (A^*A) \leq \text{rank } (A^*) \leq$$

$$= \text{rank } (AA^*) \leq \text{rank } (A).$$

Hint: Start with an orthonormal basis, $\{A\mathbf{x}_j\}_{j=1}^r$ of $A(\mathbb{F}^n)$ and verify $\{A^*A\mathbf{x}_j\}_{j=1}^r$ is a basis for $A^*A(\mathbb{F}^n)$.

Eigenvalues, Eigenvectors, and Schur's theorem

In this section we continue assuming we are in an inner product space and the field of scalars is $\mathbb{C}$. More generally, when we consider eigenvalues and eigenvectors, we will always assume the field of scalars is $\mathbb{C}$ unless another field is specified. First we define the concept of an eigenvector and an eigenvalue. Geometrically, an eigenvector for a linear transformation, A is a special vector, y, which has the property that Ay is a multiple of y.

Definition 6.1 *A non zero vector, y is said to be an eigenvector for $A \in \mathcal{L}(X,X)$ if there exists a scalar, λ, called an eigenvalue, such that*

$$Ay = \lambda y.$$

Lemma 6.2 *Let $A \in \mathcal{L}(X,X)$ where X is a finite dimensional vector space. Then A has an eigenvalue and eigenvector.*

Proof: Consider the non constant polynomial of degree n determined by $p(\lambda) = \det(A - \lambda I)$. By the fundamental theorem of algebra there exists λ such that $p(\lambda) = 0$. Therefore, by Theorem 4.11 there exists $v \neq 0$ such that $(A - \lambda I)(v) = 0$.

Note that here it is important that every nonconstant polynomial has a zero. This is why we said the field of scalars will be the complex numbers, $\mathbb{C}$.

The important thing to remember about eigenvectors is that they are **never** equal to zero.

Now we give an interesting generalization, following [6]. To state this generalization, we give the following definition.

Definition 6.3 *We say $\mathcal{F} \subseteq \mathcal{L}(X,X)$ is a commuting family if every pair of elements of $\mathcal{F}$ commutes. Thus, if $A, B \in \mathcal{F}$, it follows that $AB = BA$.*

With this definition, we give the following interesting corollary to Lemma 6.2.

Corollary 6.4 *Let $\mathcal{F}$ be a commuting family in $\mathcal{L}(X,X)$ for X a finite dimensional vector space. Then there exists a single nonzero vector, w such that w is an eigenvector for every $A \in \mathcal{F}$.*

Proof: Let W be a subspace of X which satisfies the following properties. For all $A \in \mathcal{F}$,

$$A(W) \subseteq W, W \neq \{\mathbf{0}\} \tag{6.1}$$

$$\text{The dimension of } W \text{ is as small as possible with 6.1 holding.} \tag{6.2}$$

To see that such a subspace exists, note that X satisfies 6.1. If no proper nonzero subspace satisfies 6.1, then we are done and we let $W = X$. Otherwise, there must be a proper nonzero subspace having smaller dimension than X for which 6.1 holds. If there is no proper nonzero subspace of this one for which 6.1 holds, then stop. Otherwise continue in this manner. The process must stop after finitely many iterations because the space, X is given to be finite dimensional. We will show every nonzero vector in W is an eigenvector for all $A \in \mathcal{F}$. If this is not so, there exists $x \neq 0$ and for some $A \in \mathcal{F}$, it happens that x is not an eigenvector of A. However, by Lemma 6.2, there exists $y \in W$ such that y is an eigenvector of A, say $Ay = \lambda y$. Now define

$$W_0 \equiv \{w \in W : (A - \lambda I) w = 0\}.$$

Then $y \in W_0$ but W_0does not contain any nonzero multiple of x and so it is a proper nonzero subspace of W having smaller dimension that W. We derive a contradiction by showing that $B(W_0) \subseteq W_0$ for all $B \in \mathcal{F}$.

Since $\mathcal{F}$ is a commuting family, if $B \in \mathcal{F}$ and $w \in W_0$,

$$(A - \lambda I) Bw = B(A - \lambda I) w = 0$$

showing that $Bw \in W_0$. Therefore, we have obtained a contradiction because the dimension of W_0 is smaller than the dimension of W but 6.1 holds.

Probably the most important theorem related to eigenvalues and eigenvectors is Schur's theorem which we present next.

Theorem 6.5 *Let $\mathcal{F}$ be a commuting family in $\mathcal{L}(X, X)$ for X a finite dimensional vector space. Then for each $A \in \mathcal{F}$, there exist constants, $c_{ij}(A)$ for $i \leq j$ and an orthonormal basis, $\{w_i\}_{i=1}^n$ such that for each $A \in \mathcal{F}$,*

$$A = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij}(A) w_i \otimes w_j$$

The constants, $c_{ii}(A)$ are the eigenvalues of A.

Proof: If $\dim(X) = 1$ let $X = \operatorname{span}(u)$ where $|u| = 1$. Then $A = Au \otimes u$ because $A(ku) = Au(ku, u) = kAu$. But $Au = \alpha(A) u$ for some $\alpha(A)$ since X is one dimensional. Therefore, $A = \alpha(A) u \otimes u$ which is of the desired form in the case where $n = 1$.

Now suppose the theorem holds for $n - 1 = \dim(X)$. By Corollary 6.4, there exists w_n, an eigenvector for each A^* for $A \in \mathcal{F}$ with $|w_n| = 1$. Using the Gramm Schmidt process, we can obtain an orthonormal basis for X of the form $\{v_1, \cdots, v_{n-1}, w_n\}$. Then for each $A \in \mathcal{F}$

$$(Av_k, w_n) = (v_k, A^* w_n) = (v_k, \mu(A^*) w_n) = 0,$$

which shows that for each $A \in \mathcal{F}$, $A : X_1 \equiv \operatorname{span}(v_1, \cdots, v_{n-1}) \to \operatorname{span}(v_1, \cdots, v_{n-1})$. For $A \in \mathcal{F}$ denote by A_1 the restriction of A to X_1. The collection of A_1 where $A \in \mathcal{F}$ is a commuting family and so, since X_1 has dimension $n - 1$, we may use the induction hypothesis to obtain an orthonormal basis, $\{w_1, \cdots, w_{n-1}\}$ for X_1 such that

$$A_1 = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij}(A_1) w_i \otimes w_j.$$

Then $\{w_1, \cdots, w_n\}$ is an orthonormal basis for X. Define for $A \in \mathcal{F}$, the scalars, $c_{in}(A)$ by

$$Aw_n \equiv \sum_{i=1}^{n} c_{in}(A) w_i, \; c_{ij}(A) \equiv c_{ij}(A_1) \text{ for } j < n.$$

Now consider for a fixed $A \in \mathcal{F}$,

$$B \equiv \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij}(A)\, w_i \otimes w_j.$$

Then

$$Bw_n = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij}(A)\, w_i \delta_{nj} = \sum_{j=1}^{n} c_{in}(A)\, w_i = Aw_n.$$

If $1 \leq k \leq n-1$,

$$Bw_k = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij}(A)\, w_i \delta_{kj} = \sum_{i=1}^{k} c_{ik}(A)\, w_i$$

while

$$Aw_k = A_1 w_k = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij}(A)\, w_i \delta_{jk} = \sum_{i=1}^{k} c_{ik}(A)\, w_i.$$

Therefore, A and B are two linear transformations which agree on a basis. Therefore, they must be equal.

We now verify using Theorem 4.11, that the numbers, c_{ii} are the eigenvalues of A. Let

$$\widetilde{c_{ij}(A)} = \begin{cases} c_{ij}(A) \text{ if } i \leq j \\ 0 \text{ if } i > j \end{cases}$$

By Theorem 5.24 the matrix of $A - c_{rr}(A) I$ is given by $(\widetilde{c_{ij}} - c_{rr}\delta_{ij})$, an upper triangular matrix having at least one zero on the main diagonal. Therefore, by Corollary 1.16, the determinant of this matrix equals zero and by Theorem 4.11, and letting $\widetilde{c}$ denote the matrix whose entries are $\widetilde{c_{ij}}$, there exists $\mathbf{x} \in \mathbb{C}^n$ such that $(\widetilde{c} - c_{rr} I)(\mathbf{x}) = 0$. It follows there exists $v \in X$ such that $(A - c_{rr} I)(v) = 0$. Hence c_{rr} is an eigenvalue as claimed. This proves the theorem.

The following corollary in which $X = \mathbb{C}^n$ and the linear transformations are matrices is also called Schur's theorem.

Corollary 6.6 *Let $\mathcal{F}$ be a commuting family of $n \times n$ matrices. Then there exists a unitary matrix, Q such that for all $A \in \mathcal{F}$, it follows that $Q^* A Q = T$ where T is upper triangular.*

Proof: We think of each $A \in \mathcal{F}$ as an element of $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ according to the rule $\mathbf{x} \to A\mathbf{x}$ where the product is the usual matrix multiplication. We review the diagram which describes what we mean by the matrix of a linear transformation.

$$\begin{array}{ccccc} \{\mathbf{w}_1, \cdots, \mathbf{w}_n\} & \mathbb{C}^n & \underrightarrow{A} & \mathbb{C}^n & \{\mathbf{w}_1, \cdots, \mathbf{w}_n\} \\ & q \uparrow & \circ & \uparrow q & \\ & \mathbb{C}^n & \underrightarrow{T} & \mathbb{C}^m & \end{array}$$

Here T is the matrix of the transformation just described which is determined by the basis, $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$. We choose for a basis the one of Theorem 6.5. Thus T is upper triangular. Now we note that $q\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{w}_i$ and so, writing this in terms of matrix multiplication,

$$q\mathbf{x} = Q\mathbf{x}$$

where Q is a matrix whose columns are the vectors, $\mathbf{w}_1, \cdots, \mathbf{w}_n$. Since these vectors are an orthonormal set, it follows from Theorem 5.16 that $Q^*Q = I$ and so from Lemma 5.19 $Q^{-1} = Q^*$. Thus from the diagram above, if $A \in \mathcal{F}$

$$T = Q^{-1}AQ = Q^*AQ$$

and this proves the corollary.

We will refer to this upper triangular matrix as a Schur matrix for A. The usual form of these theorems involves a single matrix, or linear transformation, A which clearly constitutes a commuting family.

Corollary 6.7 *In the situation of the above corollary, if every matrix in $\mathcal{F}$ is real and has all real eigenvalues, then we can assume the Q is a real matrix.*

Proof: In the proof of Theorem 6.5, $A^* = A^T$ and so A^* has all real eigenvalues. Therefore, we repeat the argument using $\mathbb{R}^n$ for the underlying inner product space using the real inner product or dot product.

The Schur form of a matrix or linear transformation is so fundamental we give another proof of this theorem. This proof is based directly on block multiplication and has the advantage of leading easily to an interesting result for real matrices which is not obtained so readily by the preceding approach. Recall from Corollary 6.4, a commuting family of matrices has an eigenvector.

Theorem 6.8 *Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices. Then there exists a unitary matrix, U, such that for every $A \in \mathcal{F}$,*

$$U^*AU = T, \tag{6.3}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal listed according to multiplicity as roots of the characteristic equation.

Proof: Let $\mathbf{v}_1$ be a unit eigenvector for every $A \in \mathcal{F}$ and let $A \in \mathcal{F}$. Then there exists λ_1 such that

$$A\mathbf{v}_1 = \lambda_1\mathbf{v}_1, \ |\mathbf{v}_1| = 1$$

and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis in $\mathbb{C}^n$. Let U_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then for every $A \in \mathcal{F}$, $U_0^*AU_0$ is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is an $n-1 \times n-1$ matrix and λ_1 is some number which may depend on A. We note the set, $\mathcal{F}_1$, of matrices, A_1 resulting in the above construction commute. This follows from a simple exercise in block multiplication after observing that the matrices of the above form commute. Therefore, we may repeat the process for the matrices, A_1 above. There exists a unitary matrix $\widetilde{U}_1$ such that for all $A_1 \in \mathcal{F}_1$, $\widetilde{U}_1^*A_1\,\widetilde{U}_1$ is of the form

$$\begin{pmatrix} \lambda_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix}.$$

Now let U_1 be the $n \times n$ matrix of the form

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}.$$

Then using block multiplication, whenever $A \in \mathcal{F}$, $U_1^* U_0^* A U_0 U_1$ is of the form

$$\begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$$

where A_2 is an $n-2 \times n-2$ matrix. Continuing in this way, we see there exists a unitary matrix, U given as the product of the U_i in the above construction such that for any $A \in \mathcal{F}$, we have

$$U^* A U = T$$

where T is some upper triangular matrix. Since the matrix is upper triangular, the characteristic equation is $\prod_{i=1}^{n} (\lambda - \lambda_i)$ where the λ_i are the diagonal entries of T. Therefore, the λ_i are the eigenvalues.

Now we adapt this proof to give an interesting result in the case where A is a real matrix.

Theorem 6.9 *Let A be a real $n \times n$ matrix. Then there exists a real unitary matrix, Q and a matrix T of the form*

$$T = \begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix} \tag{6.4}$$

where P_i equals either a real 1×1 matrix or P_i equals a real 2×2 matrix having two complex eigenvalues of A such that $Q^T A Q = T$. The matrix, T is called the real Schur form of the matrix A.

Proof: Suppose

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \ |\mathbf{v}_1| = 1$$

where λ_1 is real. Then let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis. 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 an $n-1 \times n-1$ matrix. This is just like the proof of Theorem 6.8 up to this point.

Now in case $\lambda_1 = \alpha + i\beta$, it follows since A is real that $\mathbf{v}_1 = \mathbf{z}_1 + i\mathbf{w}_1$ and that $\overline{\mathbf{v}}_1 = \mathbf{z}_1 - i\mathbf{w}_1$ is an eigenvector for the eigenvalue, $\alpha - i\beta$. Here $\mathbf{z}_1$ and $\mathbf{w}_1$ are real vectors. It is clear that $\{\mathbf{z}_1, \mathbf{w}_1\}$ is an independent set of vectors in $\mathbb{R}^n$. Indeed,$\{\mathbf{v}_1, \overline{\mathbf{v}}_1\}$ is an independent set and we see immediately that $\mathrm{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1) = \mathrm{span}(\mathbf{z}_1, \mathbf{w}_1)$. Now using the Gram Schmitt theorem in $\mathbb{R}^n$, we can get $\{\mathbf{u}_1, \mathbf{u}_2\}$ an orthonormal set of real vectors such that $\mathrm{span}(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1)$. Now we let $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ be an orthonormal basis and let Q_0 be a unitary matrix whose i^{th} column is $\mathbf{u}_i$. Then $A\mathbf{u}_j$ are both in $\mathrm{span}(\mathbf{u}_1, \mathbf{u}_2)$ for $j = 1, 2$ and so $\mathbf{u}_k^T A \mathbf{u}_j = 0$ whenever $k \geq 3$. It follows that $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} * & * & \cdots & * \\ * & * & & \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is now an $n-2\times n-2$ matrix. In this case, we find $\widetilde{Q}_1$ an $n-2\times n-2$ matrix to put A_1 in an appropriate form as we just did 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 we 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 6.4 where Q is the product of real unitary matrices Q_i above. Finally, we can verify with a little work that $\det(\lambda I - T) = \prod_{k=1}^{r} \det(\lambda I_2 - P_k)$. Therefore, λ is an eigenvalue of T if and only if it is an eigenvalue of some P_k. This proves the theorem since the eigenvalues of T are the same as those of A because they have the same characteristic polynomial.

Definition 6.10 *When a linear transformation, A, mapping an linear space, V to V has a basis of eigenvectors, we say the linear transformation is non defective. Otherwise it is called defective. We say $A \in \mathcal{L}(X,X)$ for X an inner product space is normal if $AA^* = A^*A$.*

Normal linear transformations have a very interesting property which is contained in the statement of the next lemma.

Lemma 6.11 *If A is normal and*

$$A = \sum_{j=1}^{n}\sum_{i=1}^{j} c_{ij} w_i \otimes w_j$$

then

$$A = \sum_{j=1}^{n} c_{jj} w_j \otimes w_j.$$

Proof: By Theorem 5.21,

$$A^* = \sum_{j=1}^{n}\sum_{i=1}^{j} \overline{c_{ij}} w_j \otimes w_i.$$

Now since A is normal,

$$(A^*w_p, A^*w_p) = (AA^*w_p, w_p) = (A^*Aw_p, w_p) = (Aw_p, Aw_p)$$

and we consider the ends.

$$A^*w_p = \sum_{j=1}^{n}\sum_{i=1}^{j} \overline{c_{ij}} w_j \otimes w_i (w_p)$$

$$= \sum_{i=1}^{n}\sum_{j=i}^{n} \overline{c_{ij}} w_j \otimes w_i (w_p)$$

$$= \sum_{i=1}^{n} \sum_{j=i}^{n} \overline{c_{ij}} w_j \delta_{ip} = \sum_{j=p}^{n} \overline{c_{pj}} w_j.$$

Therefore, using the properties of the inner product along with the orthogonality of the unit vectors, w_i,

$$(A^* w_p, A^* w_p) = \sum_{j=p}^{n} |c_{pj}|^2 . \tag{6.5}$$

Similarly, we find

$$(Aw_p, Aw_p) = \sum_{i=1}^{p} |c_{ip}|^2 . \tag{6.6}$$

To begin with we let $p = n$. Then

$$\sum_{i=1}^{n} |c_{in}|^2 = |c_{nn}|^2$$

showing that $c_{in} = 0$ for all $i < n$. Now with this information, let $p = n - 1$. then

$$\sum_{i=1}^{n-1} \left|c_{i(n-1)}\right|^2 = \sum_{j=n-1}^{n} \left|c_{(n-1)j}\right|^2 .$$

From what was just shown, the right sum reduces to $\left|c_{(n-1)(n-1)}\right|^2$ and so this shows that $c_{i(n-1)} = 0$ for all $i < n - 1$. We continue in this way eventually showing that if $i < j$, then $c_{ij} = 0$. It follows that

$$A = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} w_i \otimes w_j = \sum_{j=1}^{n} c_{jj} w_j \otimes w_j$$

It is important to know which linear transformations are non defective. In this regard, we have the following important theorem.

Theorem 6.12 *Suppose A is normal. Then there exists an orthonormal basis of vectors, $\{w_i\}_{i=1}^n$ such that*

$$A = \sum_{i=1}^{n} \lambda_i w_i \otimes w_i$$

The scalars, λ_i are eigenvalues and $Aw_i = \lambda_i w_i$.

Proof: By Theorem 6.5 we may write

$$A = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} w_i \otimes w_j$$

From Lemma 6.11,

$$A = \sum_{j=1}^{n} c_{jj} w_j \otimes w_j$$

We already showed in Theorem 6.5 that the c_{jj} are eigenvalues. Thus let $\lambda_j \equiv c_{jj}$. Now

$$Aw_r = \sum_{j=1}^{n} \lambda_j w_j \otimes w_j (w_r) = \sum_{j=1}^{n} \lambda_j w_j \delta_{jr} = \lambda_r w_r$$

and so each w_r is an eigenvector. This proves the theorem.

The following corollary shows that normal matrices are non defective.

Corollary 6.13 *Let $A \in \mathcal{L}(X,X)$ where X is a complex inner product space. Then A has an orthonormal basis of eigenvectors if and only if A is normal.*

Proof: If A is normal, then the above theorem, implies that there exists an orthonormal basis of eigenvectors, $\{w_k\}$. Now suppose A has such an orthonormal basis. Then

$$A = \sum_{i=1}^{n} \lambda_i w_i \otimes w_i.$$

because both sides give the same answer when acting on anything in the orthonormal basis $\{w_k\}$ and so

$$A^* = \sum_{i=1}^{n} \overline{\lambda_i} w_i \otimes w_i.$$

From this it is completely routine to verify $AA^* = A^*A = \sum_{i=1}^{n} |\lambda_i|^2 w_i \otimes w_i$ and so A is normal.

An important sort of normal transformation is one that is self adjoint.

Definition 6.14 *We say $A \in \mathcal{L}(X,X)$ is self adjoint if $A = A^*$. This is also called Hermitian.*

These ideas are imortant enough that we present them without the tensor product notation. The next lemma is equivalent to Lemma 6.11 above.

Lemma 6.15 *If T is upper triangular and normal, then T is a diagonal matrix.*

Proof: Since T is normal, we know $T^*T = TT^*$. Writing this in terms of components and using the description of the adjoint as the transpose of the conjugate, we obtain the following for the ik^{th} entry of $T^*T = TT^*$.

$$\sum_j t_{ij} t^*_{jk} = \sum_j t_{ij} \overline{t_{kj}} = \sum_j t^*_{ij} t_{jk} = \sum_j \overline{t_{ji}} t_{jk}.$$

Now we use the fact that T is upper triangular and let $i = k = 1$ to obtain

$$\sum_j |t_{1j}|^2 = \sum_j |t_{j1}|^2 = |t_{11}|^2$$

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 we do the same thing only this time we 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 we see that T is a diagonal matrix.

Theorem 6.16 *Let $\mathcal{F}$ denote a commuting family of normal matrices. Then there exists a unitary matrix, U such that U^*AU is a diagonal matrix for all $A \in \mathcal{F}$.*

Proof: We know from Theorem 6.8 there exists a unitary matrix, U such that U^*AU equals an upper triangular matrix for every $A \in \mathcal{F}$. The theorem is now proved if we can show that the property of being normal is preserved under unitary similarity transformations. Thus we 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, if $A \in \mathcal{F}$, we know U^*AU is a normal and upper triangular matrix. Therefore, by Lemma 6.15 it must be a diagonal matrix. This proves the theorem.

Corollary 6.17 *If A is Hermitian, then all the eigenvalues of A are real.*

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, $D^* = U^*A^*U = U^*AU = D$. Therefore, D is real. This proves the corollary.

6.1 Quadratic forms

Definition 6.18 *A quadratic form in three dimensions is an expression of the form*

$$\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.7}$$

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, we consider

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.8}$$

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, we look for new variables, x', y', and z' and a unitary matrix, U such that

$$U \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{6.9}$$

and if we write the quadratic form in terms of the primed variables, we hope there will be no mixed terms. We know that if A is symmetric and real, there exists a real unitary matrix, U, (an orthogonal matrix) such that $U^TAU = D$ a diagonal. Thus in the quadratic form, 6.7 we could write

$$\begin{aligned} \begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} x' & y' & z' \end{pmatrix} U^TAU \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \\ &= \begin{pmatrix} x' & y' & z' \end{pmatrix} D \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}. \end{aligned}$$

Thus in terms of these new variables, the quadratic form becomes

$$\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2$$

where $D = diag(\lambda_1, \lambda_2, \lambda_3)$. Similar considerations apply equally well in any other dimension. For the given example,

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

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

and so if our 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.

6.2 Second derivative test

Here we consider a second derivative test for functions of n variables.

Theorem 6.19 *Suppose $f : U \subseteq X \to Y$ where X and Y are inner product spaces, $D^2\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x} \in U$ and $D^2\mathbf{f}$ is continuous at $\mathbf{x} \in U$. Then*

$$D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}) = D^2\mathbf{f}(\mathbf{x})(\mathbf{v})(\mathbf{u}).$$

Proof: Let $B(\mathbf{x},r) \subseteq U$ and let $t, s \in (0, r/2]$. Now let $\mathbf{z} \in Y$ and define

$$\Delta(s,t) \equiv \mathrm{Re}\left(\frac{1}{st}\{\mathbf{f}(\mathbf{x}+t\mathbf{u}+s\mathbf{v}) - \mathbf{f}(\mathbf{x}+t\mathbf{u}) - (\mathbf{f}(\mathbf{x}+s\mathbf{v}) - \mathbf{f}(\mathbf{x}))\}, \mathbf{z}\right). \tag{6.10}$$

Let $h(t) = \mathrm{Re}(\mathbf{f}(\mathbf{x}+s\mathbf{v}+t\mathbf{u}) - \mathbf{f}(\mathbf{x}+t\mathbf{u}), \mathbf{z})$. Then by the mean value theorem,

$$\begin{aligned} \Delta(s,t) &= \frac{1}{st}(h(t) - h(0)) = \frac{1}{st}h'(\alpha t)\, t \\ &= \frac{1}{s}\mathrm{Re}(D\mathbf{f}(\mathbf{x}+s\mathbf{v}+\alpha t\mathbf{u})\mathbf{u} - D\mathbf{f}(\mathbf{x}+\alpha t\mathbf{u})\mathbf{u}, \mathbf{z}). \end{aligned}$$

Applying the mean value theorem again,

$$\Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x}+\beta s\mathbf{v}+\alpha t\mathbf{u})(\mathbf{v})(\mathbf{u}), \mathbf{z}\right)$$

where $\alpha, \beta \in (0,1)$. If the terms $\mathbf{f}(\mathbf{x}+t\mathbf{u})$ and $\mathbf{f}(\mathbf{x}+s\mathbf{v})$ are interchanged in 6.10, $\Delta(s,t)$ is also unchanged and the above argument shows there exist $\gamma, \delta \in (0,1)$ such that

$$\Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x}+\gamma s\mathbf{v}+\delta t\mathbf{u})(\mathbf{u})(\mathbf{v}), \mathbf{z}\right).$$

Letting $(s,t) \to (0,0)$ and using the continuity of $D^2\mathbf{f}$ at $\mathbf{x}$,

$$\lim_{(s,t)\to(0,0)} \Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}),\mathbf{z}\right) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x})(\mathbf{v})(\mathbf{u}),\mathbf{z}\right).$$

Since $\mathbf{z}$ is arbitrary, this demonstrates the conclusion of the theorem.

Consider the important special case when $X = \mathbb{R}^n$ and $Y = \mathbb{R}$. If $\mathbf{e}_i$ are the standard basis vectors, what is

$$D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j)?$$

To see what this is, use the definition to write

$$D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j) = t^{-1}s^{-1}D^2 f(\mathbf{x})(t\mathbf{e}_i)(s\mathbf{e}_j)$$

$$= t^{-1}s^{-1}\left(Df(\mathbf{x}+t\mathbf{e}_i) - Df(\mathbf{x}) + o(t)\right)(s\mathbf{e}_j)$$

$$= t^{-1}s^{-1}\left(f(\mathbf{x}+t\mathbf{e}_i + s\mathbf{e}_j) - f(\mathbf{x}+t\mathbf{e}_i)\right.$$

$$\left.+o(s) - \left(f(\mathbf{x}+s\mathbf{e}_j) - f(\mathbf{x}) + o(s)\right) + o(t)\,s\right).$$

First let $s \to 0$ to get

$$t^{-1}\left(\frac{\partial f}{\partial x_j}(\mathbf{x}+t\mathbf{e}_i) - \frac{\partial f}{\partial x_j}(\mathbf{x}) + o(t)\right)$$

and then let $t \to 0$ to obtain

$$D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \tag{6.11}$$

Thus the theorem asserts that in this special case the mixed partial derivatives are equal at $\mathbf{x}$ if they are defined near $\mathbf{x}$ and continuous at $\mathbf{x}$.

Definition 6.20 *The matrix,* $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})\right)$ *is called the Hessian matrix.*

Now we recall the Taylor formula with the Lagrange form of the remainder. See any good non reformed calculus book for a proof of this theorem. Ellis and Gulleck has a good proof. Since we will only need this on a specific interval, we will state it for this interval.

Theorem 6.21 *Let* $h : (-\delta, 1+\delta) \to \mathbb{R}$ *have* $m+1$ *derivatives. Then there exists* $t \in [0,1]$ *such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Now let $f : U \to \mathbb{R}$ where $U \subseteq X$ a normed linear space and suppose $f \in C^m(U)$. Let $\mathbf{x} \in U$ and let $r > 0$ be such that

$$B(\mathbf{x},r) \subseteq U.$$

Then for $||\mathbf{v}|| < r$ we consider

$$f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) \equiv h(t)$$

for $t \in [0,1]$. Then

$$h'(t) = Df(\mathbf{x}+t\mathbf{v})(\mathbf{v}), h''(t) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v})$$

and continuing in this way, we see that

$$h^{(k)}(t) = D^{(k)} f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v})\cdots(\mathbf{v}) \equiv D^{(k)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^k.$$

It follows from Taylor's formula for a function of one variable that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{k=1}^{m} \frac{D^{(k)} f(\mathbf{x})\,\mathbf{v}^k}{k!} + \frac{D^{(m+1)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^{m+1}}{(m+1)!}. \tag{6.12}$$

This proves the following theorem.

Theorem 6.22 *Let $f : U \to \mathbb{R}$ and let $f \in C^{m+1}(U)$. Then if*

$$B(\mathbf{x},r) \subseteq U,$$

and $||\mathbf{v}|| < r$, there exists $t \in (0,1)$ such that 6.12 holds.

Now we consider the case where $U \subseteq \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is $C^2(U)$. Then from Taylor's theorem, if $\mathbf{v}$ is small enough, there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\,\mathbf{v} + \frac{D^2 f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^2}{2}.$$

Letting

$$\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{e}_i,$$

where $\mathbf{e}_i$ are the usual basis vectors, the second derivative term reduces to

$$\frac{1}{2}\sum_{i,j} D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j)\,v_i v_j = \frac{1}{2}\sum_{i,j} H_{ij}(\mathbf{x}+t\mathbf{v})\,v_i v_j$$

where

$$H_{ij}(\mathbf{x}+t\mathbf{v}) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j) = \frac{\partial^2 f(\mathbf{x}+t\mathbf{v})}{\partial x_j \partial x_i},$$

the Hessian matrix. From Theorem 6.19, this is a symmetric matrix. By the continuity of the second partial derivative and this,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} +$$

$$\frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right). \tag{6.13}$$

where the last two terms involve ordinary matrix multiplication and

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

for v_i the components of $\mathbf{v}$ relative to the standard basis.

Theorem 6.23 *In the above situation, suppose $Df(\mathbf{x}) = 0$. Then if $H(\mathbf{x})$ has all positive eigenvalues, $\mathbf{x}$ is a local minimum. If $H(\mathbf{x})$ has all negative eigenvalues, then $\mathbf{x}$ is a local maximum. If $H(\mathbf{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\mathbf{x}$, while if $H(x)$ has a negative eigenvalue, there exists a direction in which $H(\mathbf{x})$ has a local maximum at $\mathbf{x}$.*

Proof: Since $Df(\mathbf{x}) = 0$, formula 6.13 holds and by continuity of the second derivative, we know $H(\mathbf{x})$ is a symmetric matrix. Thus, by Corollary 6.17 $H(\mathbf{x})$ 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, we can write $\mathbf{u} = \sum_{j=1}^n u_j \mathbf{v}_j$ 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 6.13 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. Then from 6.13,

$$f(\mathbf{x}+t\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}t^2 \mathbf{v}^T H(\mathbf{x})\,\mathbf{v} +$$

$$\frac{1}{2}t^2 \left(\mathbf{v}^T \left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)$$

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 proves the theorem.

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1(x,y) = x^4 + y^2,\ f_2(x,y) = -x^4 + y^2.$$

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.

6.3 Vibrating masses

Imagine two vertical walls and two identical masses of mass m free to slide in one dimension between these walls and suppose these two masses are connected to each other and to the two walls by identical springs having spring constant, k. We denote by x_1 the displacement from equilibrium of the first mass and by x_2 the displacement of the other mass from equilibrium. Assume these springs are of the sort that only pull, like garage door springs or slinkys. Then using some fussy arguments involving Newton's second law, we can write the equations of motion of the two springs in the following form.

$$\begin{aligned} mx_1'' &= k(x_2 - 2x_1) \\ mx_2'' &= k(x_1 - 2x_2) \end{aligned}$$

Thus

$$m\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}'' = k\begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{6.14}$$

It is troublesome to find the solution to this system of equations because they are coupled. However, there is an easy way to solve this system using the diagonalization of the matrix above.

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

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

Therefore,

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

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

and so we can substitute and write

$$m\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}'' = k\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

$$\begin{pmatrix} -1 & 0 \\ 0 & -3 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Therefore, letting

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{6.15}$$

in terms of the new variables, the differential equations become

$$m\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}'' = k\begin{pmatrix} -1 & 0 \\ 0 & -3 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

This is known as decoupling the equations because in terms of the new variables,

$$my_1'' + ky_1 = 0,\ my_2'' + 3ky_2 = 0.$$

Now it is a routine matter to find y_1 and y_2.

$$\begin{aligned} y_1(t) &= C_1 \cos\left(\sqrt{\frac{k}{m}}t\right) + C_2 \sin\left(\sqrt{\frac{k}{m}}t\right) \\ y_2(t) &= D_1 \cos\left(\sqrt{\frac{3k}{m}}t\right) + D_2 \sin\left(\sqrt{\frac{3k}{m}}t\right) \end{aligned}$$

Now we can use 6.15 to obtain the solution in terms of the original variables.

6.4 A rigid body rotating about a point

The above theorems about diagonalization have interesting applications to rigid body motion. Imagine a rigid body which is rotating about a point fixed in space. For example, you could consider a bicycle wheel rotating about its axis which is held still. More generally, we let the point about which the body rotates move also. In this case, the point is usually the center of mass of the body. We let $B(t)$ denote the set of points in three dimensional space which the body occupies at time t. We will refer to the points in three dimensional space occupied by the body at time $t = 0$ as the material points of the body. To begin with a proof is given of the existence of an angular velocity vector. This proof depends on the following lemma.

Lemma 6.24 *Let Q be a real $n \times n$ matrix which has the property that $|Q\mathbf{x}| = |\mathbf{x}|$ for all $\mathbf{x} \in \mathbb{R}^n$. Then Q is unitary. (Recall we also call such a matrix, orthogonal.)*

Proof: Let $\mathbf{x}, \mathbf{y}$ be two vectors in $\mathbb{R}^n$. Then

$$\begin{aligned} |Q(\mathbf{x}+\mathbf{y})|^2 &= |\mathbf{x}+\mathbf{y}|^2 \\ &= |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2(\mathbf{x},\mathbf{y}) \\ &= |Q\mathbf{x}|^2 + |Q\mathbf{y}|^2 + 2(Q\mathbf{x},Q\mathbf{y}) \\ &= |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\left(\mathbf{x},Q^TQ\mathbf{y}\right) \end{aligned}$$

and so for all $\mathbf{x}, \mathbf{y}$, it follows that

$$\left(\mathbf{x},Q^TQ\mathbf{y}\right) = (\mathbf{x},\mathbf{y}).$$

Since $\mathbf{x}$ is arbitrary, it follows that $Q^TQ\mathbf{y} = \mathbf{y}$. Since $\mathbf{y}$ is arbitrary, it follows that $Q^TQ = I$. Similar reasoning shows $QQ^T = I$. Since Q is real, this shows it is unitary as claimed.

Denote the fixed point as $\mathbf{0}$ and $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3$ denote a right handed ($\mathbf{e}_1 \times \mathbf{e}_2 \cdot \mathbf{e}_3 = 1$) orthonormal set of position vectors from $\mathbf{0}$ which are fixed in space, the spacial basis. Also denote by $\mathbf{e}_1(t), \mathbf{e}_2(t)$, and $\mathbf{e}_3(t)$ a right handed system of orthonormal vectors from $\mathbf{0}$ which is fixed with the body as it moves, the material basis, and for the sake of simplicity, we assume $\mathbf{e}_j(0) = \mathbf{e}_j$. The coordinates of a point of the body, $\mathbf{x}$ with respect to the system $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are called the spacial coordinates and the coordinates of the point taken with respect to $\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)$ are called the material coordinates of the point. We denote the first set of coordinates by $\mathbf{x}(t, \mathbf{x})$ and the second set by $\mathbf{x}$. Note that since the coordinate axes move with the body the material coordinates do not depend on t. Since the body is moving, the spacial coordinates will depend on t. We regard all of $\mathbb{R}^3$ as an extension of the rotating rigid body and let $\mathbf{x}(t, \mathbf{x})$ denote the position vector in $\mathbb{R}^3$ at time t of the point which was at $\mathbf{x}$ at time $t = 0$. Then we can consider a map, $\mathbf{x} \to \mathbf{x}(t, \mathbf{x})$. Since the motion is rigid, a little thought will verify that this map is linear and that $|\mathbf{x}| = |\mathbf{x}(t, \mathbf{x})|$. Therefore, if we

let $Q(t)\mathbf{x} \equiv \mathbf{x}(t,\mathbf{x})$, where $Q(t)$ is the matrix of this linear trasformation taken with respect to the spacial coordinates, it follows from Lemma 6.24 that $Q(t)$ is an orthogonal matrix.

Now letting $\mathbf{x}$ denote the position vector of a point of the body at time $t=0$, we want the velocity of this point at time t. From what was just discussed this velocity equals

$$\mathbf{x}_t(t,\mathbf{x}) = Q'(t)\mathbf{x}.$$

Now $QQ^T = I$ and so by the product rule,

$$Q'Q^T + QQ'^T = I' = 0 \tag{6.16}$$

so

$$\begin{aligned} Q'Q^T &= -QQ'^T \\ &= -\left(Q'Q^T\right)^T \end{aligned} \tag{6.17}$$

showing the matrix $Q'Q^T$ is skew symmetric. Clearly by the associative law for matrix multiplication,

$$Q' = \left(Q'Q^T\right)Q$$

and so we have shown

$$\begin{aligned} \mathbf{x}_t(t,\mathbf{x}) &= \left(Q'Q^T\right)Q\mathbf{x} \\ &= \left(Q'Q^T\right)\mathbf{x}(t,\mathbf{x}) \end{aligned}$$

where by 6.17, $\left(Q'Q^T\right)$ is skew symmetric and so is of the form

$$\left(Q'Q^T\right) = \begin{pmatrix} 0 & -\omega_3(t) & \omega_2(t) \\ \omega_3(t) & 0 & -\omega_1(t) \\ -\omega_2(t) & \omega_1(t) & 0 \end{pmatrix}$$

for some time dependent vector, $\omega = (\omega_1, \omega_2, \omega_3)$. We can write the vector $\mathbf{x}(t,\mathbf{x})$ in either of two ways, in terms of the material coordinates,

$$\mathbf{x}(t,\mathbf{x}) = x_1\mathbf{e}_1(t) + x_2\mathbf{e}_2(t) + x_3\mathbf{e}_3(t)$$

where $\mathbf{x} = (x_1, x_2, x_3)$, or in terms of the spacial coordinates,

$$\mathbf{x}(t,\mathbf{x}) = x_1(t,\mathbf{x})\mathbf{e}_1 + x_2(t,\mathbf{x})\mathbf{e}_2 + x_3(t,\mathbf{x})\mathbf{e}_3$$

but we are doing everything up to now in terms of the spacial coordinates. Thus we have obtained the vector,

$$\mathbf{x}_t(t,\mathbf{x}) = \left(x_3(t,\mathbf{x})\omega_2 - x_2(t,\mathbf{x})\omega_3(t)\right)\mathbf{e}_1 + \left(x_1(t,\mathbf{x})\omega_3(t) - x_3(t,\mathbf{x})\omega_1(t)\right)\mathbf{e}_2 +$$

$$\left(x_2(t,\mathbf{x})\omega_1(t) - x_1(t,\mathbf{x})\omega_2(t)\right)\mathbf{e}_3 \tag{6.18}$$

Since the basis vectors form a right handed system, we can apply the familiar formula for the cross product from calculus,

$$\omega \times \mathbf{x}(t,\mathbf{x}) = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \omega_1 & \omega_2 & \omega_3 \\ x_1(t,\mathbf{x}) & x_2(t,\mathbf{x}) & x_3(t,\mathbf{x}) \end{vmatrix}$$

and verify this yields 6.18. Therefore, we have proved the following interesting lemma.

Lemma 6.25 *For a body which undergoes rigid body motion about a fixed point in three dimensional space, if we let* $\mathbf{x}(t,\mathbf{x})$ *denote the position vector of the point,* $\mathbf{x}$ *at time* t*,then there exists a time dependent vector* $\omega(t)$ *such that the velocity of this point at time* t*,* $\mathbf{x}_t(t,\mathbf{x})$ *is given by*

$$\mathbf{x}_t(t,\mathbf{x}) = \omega(t) \times \mathbf{x}(t,\mathbf{x}).$$

In particular, letting $\mathbf{x} = \mathbf{e}_i$*, we see that* $\mathbf{e}_i'(t) = \omega(t) \times \mathbf{e}_i(t)$.

Definition 6.26 *The vector,* $\omega(t)$ *whose existence is given by the above lemma is called the angular velocity vector.*

We are now ready to write the total angular momentum of the rigid body. In doing so, we assume the density equals $\rho(\mathbf{x})$. Thus at time t the total angular momentum, Ω, would be given by the three dimensional integral,

$$\begin{aligned}\Omega &= \int_{B(0)} \mathbf{x}(t,\mathbf{x}) \times \rho(\mathbf{x})\, \mathbf{x}_t(t,\mathbf{x})\, dx \\ &= \int_{B(0)} \rho(\mathbf{x})\, \mathbf{x}(t,\mathbf{x}) \times (\omega(t) \times \mathbf{x}(t,\mathbf{x}))\, dx. \end{aligned} \tag{6.19}$$

At this point we begin to use the material coordinates for the vectors involved. In terms of the material basis, $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$

$$(\omega(t) \times \mathbf{x}(t,\mathbf{x})) = \begin{vmatrix} \mathbf{e}_1(t) & \mathbf{e}_2(t) & \mathbf{e}_3(t) \\ \omega_1 & \omega_2 & \omega_3 \\ x_1 & x_2 & x_3 \end{vmatrix}$$

where the ω_i are the components of ω taken with respect to $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$ and as we observed earlier, $\{x_1, x_2, x_3\}$ are the coordinates of the vector $\mathbf{x}(t,\mathbf{x})$ taken with respect to the $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$. To simplify the integrand in 6.19 we use the following lemma from calculus.

Lemma 6.27 *Let* $\mathbf{a}, \mathbf{b}, \mathbf{c}$ *be three dimensional vectors. Then* $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$.

Proof: This is a royal pain if you do not use tensor notation. Using tensor notation including the repeated index summation convention, and the usual reduction formula for the permutation symbol, ε_{ijk}, it is pretty easy to see, however. Let an orthonormal right handed coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ be given. Then

$$\begin{aligned}\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \varepsilon_{ijk} a_j (\mathbf{b} \times \mathbf{c})_k \mathbf{e}_i \\ &= \varepsilon_{ijk}\varepsilon_{kpq} a_j b_p c_q \mathbf{e}_i \\ &= \varepsilon_{kij}\varepsilon_{kpq} a_j b_p c_q \mathbf{e}_i \\ &= (\delta_{ip}\delta_{jq} - \delta_{jp}\delta_{iq}) a_j b_p c_q \mathbf{e}_i \\ &= (a_j b_i c_j - a_j b_j c_i) \mathbf{e}_i \\ &= (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}. \end{aligned}$$

This proves the lemma.

Now we can simplify the integrand using this lemma.

$$\mathbf{x}(t,\mathbf{x}) \times (\omega(t) \times \mathbf{x}(t,\mathbf{x})) = (\mathbf{x}(t,\mathbf{x}) \cdot \mathbf{x}(t,\mathbf{x}))\, \omega(t) - (\mathbf{x}(t,\mathbf{x}) \cdot \omega(t))\, \mathbf{x}(t,\mathbf{x}).$$

Writing $\mathbf{x}(t,\mathbf{x})$ and $\omega(t)$ in terms of the material coordinates,

$$\begin{aligned}\omega(t) &= \omega_1(t)\mathbf{e}_1(t) + \omega_2(t)\mathbf{e}_2(t) + \omega_3(t)\mathbf{e}_3(t), \\ \mathbf{x}(t,\mathbf{x}) &= x_1\mathbf{e}_1(t) + x_2\mathbf{e}_2(t) + x_3\mathbf{e}_3(t), \end{aligned}$$

and so

$$\mathbf{x}(t,\mathbf{x}) \times (\omega(t) \times \mathbf{x}(t,\mathbf{x})) = \sum_i |\mathbf{x}|^2 \omega_i(t) \mathbf{e}_i(t) - \left(\sum_{j,i} x_j \omega_j(t) x_i \mathbf{e}_i(t) \right). \tag{6.20}$$

Thus, listing the components of $\mathbf{x}(t,\mathbf{x}) \times (\omega(t) \times \mathbf{x}(t,\mathbf{x}))$ with respect to the material basis in the usual way the above yields

$$\begin{pmatrix} x_2^2 + x_3^2 & -x_1 x_2 & -x_1 x_3 \\ -x_1 x_2 & x_1^2 + x_3^2 & -x_2 x_3 \\ -x_1 x_3 & -x_2 x_3 & x_1^2 + x_2^2 \end{pmatrix} \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \\ \omega_3(t) \end{pmatrix}. \tag{6.21}$$

Therefore, the components of angular momentum taken with respect to the material basis are

$$\begin{pmatrix} \Omega_1(t) \\ \Omega_2(t) \\ \Omega_3(t) \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix} \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \\ \omega_3(t) \end{pmatrix}. \tag{6.22}$$

Where

$$\begin{aligned} I_{kk} &= \int_{B(0)} \left(\sum_{j \neq k} x_j^2 \right) \rho(x_1, x_2, x_3)\, dx \\ I_{ij} &= -\int_{B(0)} x_i x_j \rho(x_1, x_2, x_3)\, dx,\ i \neq j. \end{aligned}$$

Thus the matrix in 6.22 is symmetric. Because of the choice of coordinates, this matrix is also time independent. It is called the moment of inertia tensor and the off diagonal terms are called the products of inertia.

Now recall that

$$\Omega = \int_{B(0)} \mathbf{x}(t,\mathbf{x}) \times \rho(\mathbf{x}) \mathbf{x}_t(t,\mathbf{x})\, dx.$$

Taking the time derivative on both sides, (We do not worry about mathematical details related to differentiating under the integral sign here.) we obtain

$$\begin{aligned} \Omega' &= \int_{B(0)} \mathbf{x}_t(t,\mathbf{x}) \times \rho(\mathbf{x}) x_t(t,\mathbf{x})\, dx + \int_{B(0)} \mathbf{x}(t,\mathbf{x}) \times \frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t,\mathbf{x}))\, dx \\ &= \int_{B(0)} \mathbf{x}(t,\mathbf{x}) \times \frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t,\mathbf{x}))\, dx. \end{aligned}$$

Now from Newton's second law, the force on the chunk of mass, $\rho(\mathbf{x})\, dx$ at time t, denoted here by $\mathbf{F}(\mathbf{x}(t,\mathbf{x}))\, dx$ is just $\frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t,\mathbf{x}))\, dx$. Therefore,

$$\Gamma(t) \equiv \Omega'(t) = \int_{B(0)} \mathbf{x}(t,\mathbf{x}) \times \mathbf{F}(\mathbf{x}(t,\mathbf{x}))\, dx$$

which is what we define as the total torque acting on the body at time t. Note it has units of distance times units of force. Now we want to differentiate the angular momentum to find the torque. There is a slight complication due to the fact that we have the angular momentum expressed in terms of a basis which is time dependent. Therefore, when we take the derivative of this vector we must include this fact. From 6.22 we see

$$\Omega(t) = \sum_i \sum_j I_{ij} \omega_j(t) \mathbf{e}_i(t).$$

Now recall by Lemma 6.25, $\mathbf{e}_i'(t) = \omega(t) \times \mathbf{e}_i(t)$. Therefore,

$$\Gamma(t) = \Omega'(t) = \sum_i \sum_j I_{ij}\omega_j'(t)\,\mathbf{e}_i(t) + \sum_i \sum_j I_{ij}\omega_j(t)\,(\omega(t) \times \mathbf{e}_i(t)). \tag{6.23}$$

This is called Euler's equation for the torque. There are three equations hidden in the above formula, one for each $\mathbf{e}_i$ for $i = 1, 2$, and 3. If you want, you can write them down but there is a simpler way to proceed. Recall the matrix, (I_{ij}) is symmetric and real. Therefore, it can be diagonalized by a unitary real matrix. If we let the columns of this unitary matrix be the $\mathbf{e}_i$, it follows the moment of inertia tensor is a diagonal matrix, $diag(I_1, I_2, I_3)$ and 6.23 becomes

$$\Gamma(t) = \sum_i I_i\omega_i'(t)\,\mathbf{e}_i(t) + \sum_i I_i\omega_i(t)\,(\omega(t) \times \mathbf{e}_i(t))$$

Writing the right side out we get $\Gamma(t) =$

$$I_1\omega_1'\mathbf{e}_1 + I_2\omega_2'\mathbf{e}_2 + I_3\omega_3'\mathbf{e}_3 + I_1\omega_1\left(\overbrace{\omega_3\mathbf{e}_2 - \omega_2\mathbf{e}_3}^{\omega\times\mathbf{e}_1}\right) +$$
$$I_2\omega_2\left(\overbrace{\omega_1\mathbf{e}_3 - \omega_3\mathbf{e}_1}^{\omega\times\mathbf{e}_2}\right) + I_3\omega_3\left(\overbrace{\omega_2\mathbf{e}_1 - \omega_1\mathbf{e}_2}^{\omega\times\mathbf{e}_3}\right)$$

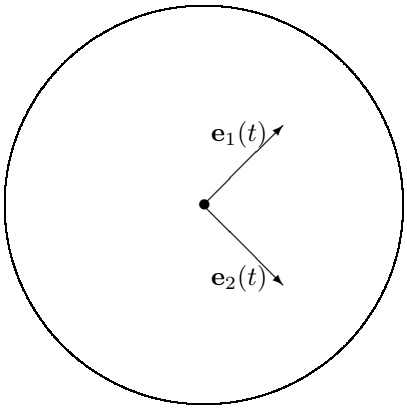
and now, collecting terms we obtain $\Gamma(t) = \Gamma_1(t)\,\mathbf{e}_1(t) + \Gamma_2(t)\,\mathbf{e}_2(t) + \Gamma_3(t)\,\mathbf{e}_3(t)$ where

$$\begin{aligned}
\Gamma_1(t) &= I_1\omega_1' + \omega_3\omega_2(I_3 - I_2) \\
\Gamma_2(t) &= I_2\omega_2' + \omega_1\omega_3(I_1 - I_3) \\
\Gamma_3(t) &= I_3\omega_3' + \omega_1\omega_2(I_2 - I_1).
\end{aligned} \tag{6.24}$$

These are called Euler's equations for the torque. Although we invoked the theorem that normal matrices can be diagonalized by a unitary transformation in order to get axes with respect to which the moment of inertia tensor is diagonal, it is usually much easier than this. Often there are symmetry considerations which make it obvious how to choose these axes and when this is done 6.24 allows us to compute the torque which results from a given angular velocity.

Example 6.28 *Consider a disc having negligible thickness and radius R with constant density ρ taken with respect to area which spins around its center. How should we choose the material bases to get a nice diagonal moment of inertia tensor?*

Consider the following picture in which the vectors $\mathbf{e}_1(t)$ and $\mathbf{e}_2(t)$ are shown fixed with the disc which is assumed to be rotating.



We let $\mathbf{e}_3(t) = \mathbf{e}_1(t) \times \mathbf{e}_2(t)$ so that we have a right handed orthonormal system of basis vectors. We calculate the moment of inertia tensor first.

$$\begin{aligned} I_{11} &\equiv \rho \int_{B(0)} x_2^2 dx \\ &= \rho \int_0^{2\pi} \int_0^R (r\sin\theta)^2 \, r dr d\theta \\ &= \frac{1}{4} R^4 \pi \rho \end{aligned}$$

By symmetry, we see that $I_{22} = \frac{1}{4}R^4\pi\rho$ also. Now

$$\begin{aligned} I_{33} &\equiv \rho \int_{B(0)} \left(x_2^2 + x_1^2\right) dx \\ &= \rho \int_0^{2\pi} \int_0^R r^3 dr d\theta \\ &= \frac{1}{2} \rho \pi R^4. \end{aligned}$$

Now by symmetry considerations, $I_{12} = 0$ as are all the other off diagonal terms. Those that have a 3 in the subscript are zero because we are assuming for the sake of simplicity that the disc has negligible thickness. However, if we didn't assume this we would still get zero for these terms by the symmetry of the shape with respect to the other variable. Therefore, the moment of inertia tensor is

$$\begin{pmatrix} \frac{1}{4}\rho\pi R^4 & 0 & 0 \\ 0 & \frac{1}{4}\rho\pi R^4 & 0 \\ 0 & 0 & \frac{1}{2}\rho\pi R^4 \end{pmatrix}.$$

It follows that for $\omega = \omega_1(t)\,\mathbf{e}_1(t) + \omega_2(t)\,\mathbf{e}_2(t) + \omega_3(t)\,\mathbf{e}_3(t)$ we can find the Torque by Euler's equations.

$$\begin{aligned} \Gamma_1(t) &= \frac{1}{4}\rho\pi R^4 \omega_1' + \omega_3\omega_2 \left(\frac{1}{4}\rho\pi R^4\right) \\ \Gamma_2(t) &= \frac{1}{4}\rho\pi R^4 \omega_2' + \omega_1\omega_3 \left(-\frac{1}{4}\rho\pi R^4\right) \\ \Gamma_3(t) &= I_3\omega_3'. \end{aligned} \tag{6.25}$$

The physical interpretation of ω given above is that the term $\omega_3(t)\,\mathbf{e}_3(t)$ represents the angular velocity about the axis determined by $\mathbf{e}_3(t)$. Thus it is a measure of how fast and in what direction the disc is spinning about this axis. If the disc were spinning very fast we would have $\omega_3(t)$ very large. The other terms of angular velocity, $\omega_1(t)\,\mathbf{e}_1(t)+\omega_2(t)\,\mathbf{e}_2(t)$, yield a vector which is in the plane determined by $\mathbf{e}_1(t)$ and $\mathbf{e}_2(t)$ and so it is a measure of the angular velocity about this axis. If we assumed $\omega_i'(t)=0$ for each $i=1,2,3$, and ω_2 and ω_1 are moderate, note that we would still have substantial components of torque, $\Gamma_2(t)$ and $\Gamma_1(t)$. Much more could be said about this problem and more examples could be given but this much will suffice for these notes. You see how the notions of a basis and the theory of diagonalization of matrices is used to solve a fairly difficult problem.

6.5 Exercises

1. Show every self adjoint linear transformation is normal.
2. Show the eigenvalues of a self adjoint linear transformation defined on an inner product space are real.
3. Show that if A is self adjoint and if $Ax=\lambda x$ and $Ay=\mu y$ for $\lambda\neq\mu$, show $(x,y)=0$.
4. Suppose A is an $n\times n$ matrix with all real entries and A is normal. Can it be concluded that A is symmetric? (A is symmetric means $A=A^T$.)
5. If possible, give an example of an $n\times n$ matrix which is Hermitian but not symmetric. Find if possible one which is symmetric but not Hermitian.
6. Show every unitary transformation is normal.
7. Write the polynomial $x^2-4xy-2xz-2y^2-4yz+z^2+x$ in terms of new variables, x',y',z' such that with respect to these new variables there are no mixed product terms in the resulting expression.
8. Give an example of a normal matrix which is not Hermitian.
9. Show A^*A has all nonnegative eigenvalues.
10. Give an example of a matrix which is not normal.
11. If A is such that there exists a unitary matrix, Q such that
$$Q^*AQ = \text{ diagonal matrix},$$
is it necessary that A be normal?
12. Suppose that
$$T=\begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix}$$
where P_i is either a 1×1 matrix or a 2×2 matrix. Show that $\det(T)=\prod_{j=1}^{r}\det(P_j)$. Can this be generalized for P_j different sizes?
13. Take any upper triangular $n\times n$ matrix, T and let
$$D_\varepsilon=\begin{pmatrix} 1 & & & 0 \\ & \varepsilon & & \\ & & \ddots & \\ 0 & & & \varepsilon^{n-1} \end{pmatrix}.$$
Consider $D_\varepsilon^{-1}TD_\varepsilon$. Show that every matrix is similar to one which has very small non diagonal entries.

14. If A and B are $n \times n$ matrices and A^{-1} exists, show that AB and BA must be similar. **Hint:** Consider $A^{-1}(AB)A$.

Generalized eigenspace and block diagonal matrices

7.1 Simultaneous Diagonalization

In this section we will consider the problem of finding a single similarity transformation which diagonalizes all the elements of a set of $n \times n$ matrices. First we give a simple technical lemma.

Lemma 7.1 *Let A be an $n \times n$ matrix and let B be an $m \times m$ matrix. Denote by C the matrix,*

$$C \equiv \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Then C is diagonalizable if and only if both A and B are diagonalizable.

Proof: Suppose $S_A^{-1} A S_A = D_A$ and $S_B^{-1} B S_B = D_B$ where D_A and D_B are diagonal matrices. Then we leave it to the reader to use block multiplication to verify that $S \equiv \begin{pmatrix} S_A & 0 \\ 0 & S_B \end{pmatrix}$ is such that $S^{-1} C S = D_C$, a diagonal matrix.

Conversely, suppose C is diagonalized by $S = (\mathbf{s}_1, \cdots, \mathbf{s}_{n+m})$. Thus S has columns $\mathbf{s}_i$. For each of these columns, write in the form

$$\mathbf{s}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$$

where $\mathbf{x}_i \in \mathbb{F}^n$ and where $\mathbf{y}_i \in \mathbb{F}^m$. Now we leave it as an exercise in block multiplication to verify that each of the $\mathbf{x}_i$ is an eigenvector of A and that each of the $\mathbf{y}_i$ is an eigenvector of B. If we can show that there are n linearly independent $\mathbf{x}_i$, then we will know that A is diagonalizable by Theorem 4.7. We know the row rank of the matrix, $(\mathbf{x}_1, \cdots, \mathbf{x}_{n+m})$ must be n because if this is not so, the rank of S would be less than $n + m$ which would mean S^{-1} does not exist. Therefore, since the column rank equals the row rank, this matrix has column rank equal to n and this means there are n linearly independent eigenvectors of A implying that A is diagonalizable. Similar reasoning applies to B. This proves the lemma.

The following corollary follows from the same type of argument as the above.

Corollary 7.2 *Let A_k be an $n_k \times n_k$ matrix and let C denote the block diagonal $(\sum_{k=1}^r n_k) \times (\sum_{k=1}^r n_k)$ matrix given below.*

$$C \equiv \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_r \end{pmatrix}.$$

Then C is diagonalizable if and only if each A_k is diagonalizable.

Definition 7.3 *We say a set, $\mathcal{F}$ of $n \times n$ matrices is simultaneously diagonalizable if and only if there exists a single invertible matrix, S such that $S^{-1}AS = D$, a diagonal matrix for all $A \in \mathcal{F}$.*

Lemma 7.4 *If $\mathcal{F}$ is a set of $n \times n$ matrices which is simultaneously diagonalizable, then $\mathcal{F}$ is a commuting family of matrices.*

Proof: Let $A, B \in \mathcal{F}$ and let S be a matrix which has the property that $S^{-1}AS$ is a diagonal matrix for all $A \in \mathcal{F}$. Then $S^{-1}AS = D_A$ and $S^{-1}BS = D_B$ where D_A and D_B are diagonal matrices. We have, 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 7.5 *Let D be a diagonal matrix of the form*

$$D \equiv \begin{pmatrix} \lambda_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{n_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r I_{n_r} \end{pmatrix}, \tag{7.1}$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix and suppose B is a matrix which commutes with D. Then B is a block diagonal matrix of the form

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_r \end{pmatrix} \tag{7.2}$$

where B_i is an $n_i \times n_i$ matrix.

Proof: Suppose $B = (b_{ij})$. Then since it is given to commute with D we must have $\lambda_i b_{ij} = b_{ij}\lambda_j$. But this shows that if $\lambda_i \neq \lambda_j$, then this could not occur unless $b_{ij} = 0$. Therefore, B must be of the claimed form.

Lemma 7.6 *Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathcal{F}$ is diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable.*

Proof: We prove this by induction on n. If $n = 1$, there is nothing to prove because all the 1×1 matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n-1$ where $n \geq 2$ and let $\mathcal{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathcal{F}$ and let S be an invertible matrix such that $S^{-1}AS = D$ where D is of the form given in 7.1. Now denote by $\widetilde{\mathcal{F}}$ the collection of matrices, $\left\{S^{-1}BS : B \in \mathcal{F}\right\}$. It follows easily that $\widetilde{\mathcal{F}}$ is also a commuting family of diagonalizable matrices. By Lemma 7.5 every $B \in \widetilde{\mathcal{F}}$ is of the form given in 7.2 and by block multiplication, the B_i corresponding to different $B \in \widetilde{\mathcal{F}}$ commute. Therefore, by the induction hypothesis, the knowledge that each $B \in \widetilde{\mathcal{F}}$ is diagonalizable, and Corollary 7.2, there exist invertible $n_i \times n_i$ matrices, T_i such that $T_i^{-1}B_iT_i$ is a diagonal matrix whenever B_i is one of the matrices making up the block diagonal of any $B \in \mathcal{F}$. It follows that if we let

$$T \equiv \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix},$$

then $T^{-1}BT =$ a diagonal matrix for every $B \in \widetilde{\mathcal{F}}$ including D. But now if we consider ST, it follows that for all $B \in \mathcal{F}$,

$$T^{-1}S^{-1}BST = (ST)^{-1} B (ST) = \text{ a diagonal matrix.}$$

This proves the lemma.

Theorem 7.7 *Let $\mathcal{F}$ denote a family of matrices which are diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if $\mathcal{F}$ is a commuting family.*

Proof: If $\mathcal{F}$ is a commuting family, it follows from Lemma 7.6 that it is simultaneously diagonalizable. If it is simultaneously diagonalizable, then it follows from Lemma 7.4 that it is a commuting family. This proves the theorem.

7.2 Generalized Eigenspace

Let V be a finite dimensional vector space. For example, it could be a subspace of $\mathbb{C}^n$.

Theorem 7.8 *Let V be a nonzero finite dimensional complex vector space of dimension n. Suppose also the field of scalars equals $\mathbb{C}$.*[1] *Suppose $A \in \mathcal{L}(V,V)$. Then there exists $v \neq 0$ and $\lambda \in \mathbb{C}$ such that*

$$Av = \lambda v.$$

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 10.9 the set is linearly dependent. Thus there exist constants, $c_i \in \mathbb{C}$ such that

$$c_0 I + \sum_{k=1}^{n^2} c_k A^k = 0.$$

This implies there exists a polynomial, $q(\lambda)$ which has the property that $q(A) = 0$. In fact, $q(\lambda) \equiv c_0 + \sum_{k=1}^{n^2} c_k \lambda^k$. Dividing by the leading term, it can be assumed this polynomial is of the form $\lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$, a monic polynomial. Now consider all such monic polynomials, q such that $q(A) = 0$ and pick one which has the smallest degree. This is called the minimial polynomial and will be denoted here by $p(\lambda)$. By the fundamental theorem of algebra, $p(\lambda)$ is of the form

$$p(\lambda) = \prod_{k=1}^{p} (\lambda - \lambda_k).$$

Thus, since p has minimial degree,

$$\prod_{k=1}^{p} (A - \lambda_k I) = 0, \text{ but } \prod_{k=1}^{p-1} (A - \lambda_k I) \neq 0.$$

Therefore, there exists $u \neq 0$ such that

$$v \equiv \left(\prod_{k=1}^{p-1} (A - \lambda_k I) \right)(u) \neq 0.$$

But then

$$(A - \lambda_p I) v = (A - \lambda_p I) \left(\prod_{k=1}^{p-1} (A - \lambda_k I) \right)(u) = 0.$$

This proves the theorem.

[1] All that is really needed is that the characteristic polynomial can be completely factored in the given field. Therefore, I will continue to write $\mathbb{F}$ for the field of scalars.

Corollary 7.9 *In the above theorem, each of the scalars, λ_k has the property that there exists a nonzero v such that $(A-\lambda_i I)\,v=0$. Furthermore the λ_i are the only scalars with this property.*

Proof: For the first claim, just factor out $(A-\lambda_i I)$ instead of $(A-\lambda_p I)$. Next suppose $(A-\mu I)\,v=0$ for some μ and $v\neq 0$. Then

$$\begin{aligned} 0 &= \prod_{k=1}^{p}(A-\lambda_k I)\,v=\prod_{k=1}^{p-1}(A-\lambda_k I)\,(Av-\lambda_p v)\\ &= (\mu-\lambda_p)\left(\prod_{k=1}^{p-1}(A-\lambda_k I)\right)v. \end{aligned}$$

continuing this way yields

$$=\prod_{k=1}^{p}(\mu-\lambda_k)\,v,$$

a contradiction unless $\mu=\lambda_k$ for some k.

Definition 7.10 *For $A\in\mathcal{L}(V,V)$ where $\dim(V)=n$, the numbers, λ_k in the minimal polynomial,*

$$p(\lambda)=\prod_{k=1}^{p}(\lambda-\lambda_k)$$

are called the eigenvalues of A. The collection of eigenvalues of A is denoted by $\sigma(A)$. For λ an eigenvalue of $A\in\mathcal{L}(V,V)$, the generalized eigenspace is defined as

$$V_\lambda\equiv\{x\in V:(A-\lambda I)^m x=0 \text{ for some } m\in\mathbb{N}\}$$

and the eigenspace is defined as

$$\{x\in V:(A-\lambda I)\,x=0\}=\ker(A-\lambda I).$$

Also, for subspaces of V, $V_1,V_2,\cdots,V_r$, the symbol, $V_1+V_2+\cdots+V_r$ or the shortened version, $\sum_{i=1}^{r}V_i$ will denote the set of all vectors of the form $\sum_{i=1}^{r}v_i$ where $v_i\in V_i$.

It will be shown that in the case where A is an $n\times n$ matrix, these eigenvalues coincide with those determined using the determinant. The following theorem is of major importance and will be the basis for the very important theorems concerning block diagonal matrices.

Lemma 7.11 *The generalized eigenspace for $\lambda\in\sigma(A)$ is indeed a subspace satisfying*

$$A:V_\lambda\to V_\lambda,$$

and there exists a smallest integer, m with the property that

$$\ker(A-\lambda I)^m=\{x\in V:(A-\lambda I)^m x=0 \text{ for some } m\in\mathbb{N}\}.$$

Proof: The claim that the generalized eigenspace is a subspace is obvious. To establish the second part, note that

$$\left\{\ker(A-\lambda I)^k\right\}$$

yields an increasing sequence of subspaces. Therefore, there must exist $m\leq n$ such that

$$\ker(A-\lambda I)^m=\ker(A-\lambda I)^{m+1}=\cdots=\ker(A-\lambda I)^{m+k}$$

for all $k\in\mathbb{N}$ because if this were not so, V would not be of dimension n.

Theorem 7.12 *Let V be a complex vector space of dimension n and suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_k\}$ where the λ_i are the distinct eigenvalues of A. Denote by V_i the generalized eigenspace for λ_i and let r_i be the multiplicity of λ_i. Thus*

$$V_i = \ker (A - \lambda_i I)^{r_i} \tag{7.3}$$

and r_i is the smallest integer with this property. Then

$$V = \sum_{i=1}^{k} V_i. \tag{7.4}$$

Proof: This is proved by induction on k. First suppose there is only one eigenvalue, λ_1 of multiplicity m. Then by the definition of eigenvalues given in Definition 7.10, A satisfies an equation of the form

$$(A - \lambda_1 I)^r = 0$$

where r is as small as possible for this to take place. Thus $\ker (A - \lambda_1 I)^r = V$ and the theorem is proved in the case of one eigenvalue.

Now suppose the theorem is true for any $i \leq k-1$ where $k \geq 2$ and suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_k\}$.

Claim 1: Let $\mu \neq \lambda_i$, Then $(A - \mu I)^m : V_i \to V_i$ and is one to one and onto for every $m \in \mathbb{N}$.

Proof: Let $w \in V_i$ and suppose $(A - \mu I) w = 0$ so that $Aw = \mu w$. Then for some $m \in \mathbb{N}$, $(A - \lambda_i I)^m w = 0$ and so by the binomial theorem,

$$(\mu - \lambda_i)^m w = \sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} \mu^l w$$

$$\sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} A^l w = (A - \lambda_i I)^m w = 0.$$

Therefore, since $\mu \neq \lambda_i$, it follows $w = 0$ and this verifies $(A - \mu I)$ is one to one. Thus $(A - \mu I)^m$ is also one to one on V_i Letting $\{u_1^i, \cdots, u_{r_k}^i\}$ be a basis for V_i, it follows $\{(A - \mu I)^m u_1^i, \cdots, (A - \mu I)^m u_{r_k}^i\}$ is also a basis and so $(A - \mu I)^m$ is also onto.

Let p be the smallest integer such that $\ker (A - \lambda_k I)^p = V_k$ and define

$$W \equiv (A - \lambda_k I)^p (V).$$

Claim 2: $A : W \to W$ and λ_k is not an eigenvalue for A restricted to W.

Proof: Suppose $A (A - \lambda_k I)^p u = \lambda_k (A - \lambda_k I)^p u$ where $(A - \lambda_k I)^p u \neq 0$. Then subtracting $\lambda_k (A - \lambda_k I)^p u$ from both sides yields

$$(A - \lambda_k I)^{p+1} u = 0$$

and so $u \in \ker ((A - \lambda_k I)^p)$ from the definition of p. But this requires $(A - \lambda_k I)^p u = 0$ contrary to $(A - \lambda_k I)^p u \neq 0$.

It follows from this claim that the eigenvalues of A restricted to W are a subset of $\{\lambda_1, \cdots, \lambda_{k-1}\}$. Letting

$$V_i' \equiv \left\{ w \in W : (A - \lambda_i)^l w = 0 \text{ for some } l \in \mathbb{N} \right\},$$

it follows from the induction hypothesis that

$$W = \sum_{i=1}^{k-1} V_i' \subseteq \sum_{i=1}^{k-1} V_i.$$

From Claim 1, $(A-\lambda_k I)^p$ maps V_i one to one and onto V_i. Therefore, if $x \in V$, then $(A-\lambda_k I)^p x \in W$. It follows there exist $x_i \in V_i$ such that

$$(A-\lambda_k I)^p x = \sum_{i=1}^{k-1} \overbrace{(A-\lambda_k I)^p x_i}^{\in V_i}.$$

Consequently

$$(A-\lambda_k I)^p \left(\overbrace{x - \sum_{i=1}^{k-1} x_i}^{\in V_k} \right) = 0$$

and so there exists $x_k \in V_k$ such that

$$x - \sum_{i=1}^{k-1} x_i = x_k$$

and this proves the theorem.

Definition 7.13 *Let $\{V_i\}_{i=1}^r$ be subspaces of V which have the property that if $v_i \in V_i$ and*

$$\sum_{i=1}^{r} v_i = 0, \tag{7.5}$$

then $v_i = 0$ for each i. Under this condition,

$$V_1 \oplus \cdots \oplus V_r \equiv \sum_{i=1}^{r} V_i$$

This is called a direct sum of subspaces.

Theorem 7.14 *Let $\{V_i\}_{i=1}^m$ be subspaces of V which have the property 7.5 and let $B_i = \{u_1^i, \cdots, u_{r_i}^i\}$ be a basis for V_i. Then $\{B_1, \cdots, B_m\}$ is a basis for $V_1 \oplus \cdots \oplus V_m = \sum_{i=1}^m V_i$.*

Proof: It is clear that span $(B_1, \cdots, B_m) = V_1 \oplus \cdots \oplus V_m$. It only remains to verify that $\{B_1, \cdots, B_m\}$ is linearly independent. Arbitrary elements of span $(B_1, \cdots, B_m)$ are of the form

$$\sum_{k=1}^{m} \sum_{i=1}^{r_i} c_i^k u_i^k.$$

Suppose then that

$$\sum_{k=1}^{m} \sum_{i=1}^{r_i} c_i^k u_i^k = 0.$$

Since $\sum_{i=1}^{r_i} c_i^k u_i^k \in V_k$ it follows $\sum_{i=1}^{r_i} c_i^k u_i^k = 0$ for each k. But then $c_i^k = 0$ for each $i = 1, \cdots, r_i$. This proves the theorem.

The following corollary is the main result.

Corollary 7.15 *Let V be a complex vector space of dimension, n and let $A \in \mathcal{L}(V,V)$. Also suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_s\}$ where the λ_i are distinct. Then letting V_{λ_i} denote the generalized eigenspace for λ_i,*

$$V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s}$$

and if B_i is a basis for V_{λ_i}, then $\{B_1, B_2, \cdots, B_s\}$ is a basis for V.

Proof: It is necessary to verify that the V_{λ_i} satisfy condition 7.5. Let $V_{\lambda_i} = \ker(A - \lambda_i I)^{r_i}$ and suppose $v_i \in V_{\lambda_i}$ and $\sum_{i=1}^{k} v_i = 0$ where $k \leq s$. It is desired to show this implies each $v_i = 0$. It is clearly true if $k = 1$. Suppose then that the condition holds for $k-1$ and

$$\sum_{i=1}^{k} v_i = 0$$

and not all the $v_i = 0$. By Claim 1 in the proof of Theorem 7.12, multiplying by $(A - \lambda_k I)^{r_k}$ yields

$$\sum_{i=1}^{k-1} (A - \lambda_k I)^{r_k} v_i = \sum_{i=1}^{k-1} v_i' = 0$$

where $v_i' \in V_{\lambda_i}$. Now by induction, each $v_i' = 0$ and so each $v_i = 0$ for $i \leq k-1$. Therefore, the sum, $\sum_{i=1}^{k} v_i$ reduces to v_k and so $v_k = 0$ also.

By Theorem 7.12, $\sum_{i=1}^{s} V_{\lambda_i} = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s} = V$ and by Theorem 7.14 $\{B_1, B_2, \cdots, B_s\}$ is a basis for V. This proves the corollary.

Next are improved results on representing a linear transformation in terms of an upper triangular matrix. This will lead to another proof of the Cayley Hamilton theorem and will also make possible the proof of some significant theorems about the convergence of Stochastic matrices. As explained earlier on Page 34, there is no loss of generality in assuming an inner product space.

Let $A \in \mathcal{L}(X,X)$ where X is a complex n dimensional inner product space. You can think $\mathbb{C}^n$ if you desire. By Schur's theorem there is an orthonormal basis $\{w_i\}_{i=1}^{n}$ such that

$$A = \sum_{j=1}^{n} \sum_{i=1}^{j} a_{ij} w_i \otimes w_j$$

and the matrix of A with respect to this basis is the upper triangular matrix, $(\widetilde{a_{ij}})$ where $\widetilde{a_{ij}} = a_{ij}$ if $i \leq j$ and $\widetilde{a_{ij}} = 0$ if $i > j$. We also saw that the eigenvalues of A are the numbers a_{ii}. Let the distinct values of these be denoted by$\{\alpha_1, \cdots, \alpha_r\}$ and α_i occurs m_i times in the list $\{a_{11}, \cdots, a_{nn}\}$. Thus from Corollary 1.16 the characteristic equation for A is of the form

$$p(\lambda) \equiv \prod_{i=1}^{r} (\alpha_i - \lambda)^{m_i} = 0$$

and α_i is a root of multiplicity m_i. In words, the number of times the α_i occurs on the diagonal of a Schur matrix of A equals the multiplicity of the eigenvalue as a root of the characteristic polynomial. The following lemma is a review of much of the above.

Lemma 7.16 *Let the distinct eigenvalues of A be $\{\alpha_1, \cdots, \alpha_m\}$.*

1.) The generalized eigenspaces are subspaces of X such that $A : X_i \to X_i$.

2.)If $v \neq 0$, $v \in X_p$, and $Av = \mu v$, then $\mu = \alpha_p$.

3.)If $0 = \sum_{i=1}^{m} v_i$ where $v_i \in X_i$, then for all i, $v_i = 0$.

Proof: The first claim of the lemma is obvious and follows from the observation that A commutes with itself. It is the second claim which is the most interesting.

Suppose $Av = \mu v$ where $v \in X_p$ and $v \neq 0$. Suppose $\mu \neq \alpha_p$. Since $v \in X_p$, we know $(A - \alpha_p I)^m v = 0$ for some $m = 1, 2, \cdots$. Considering all such pairs (v, m) where $v \in X_p$, $Av = \mu v$ and $(A - \alpha_p I)^m v = 0$, take the one for which m is as small as possible. There are two cases.

Case 1: $m = 1$.

This case yields a contradiction immediately because it requires $Av = \alpha_p v$ and $Av = \mu v$ for $v \neq 0$.

Case 2: $m > 1$.

In this case,

$$0 = (A - \alpha_p I)^m v = (A - \alpha_p I)^{m-1} (A - \alpha_p I) v,$$

and since we are not in Case 1, $(A - \alpha_p I) v \neq 0$. But then

$$(A - \mu I)(A - \alpha_p I) v = (A - \alpha_p I)(A - \mu I) v = (A - \alpha_p I)(0) = 0,$$

which contradicts the assumption that m was as small as possible since we could have obtained a smaller value for m by using the vector $(A - \alpha_p I) v$. Therefore, $\mu = \alpha_p$.

It follows from claim 2 that if $i \neq j$, then $X_i \cap X_j = \{0\}$ because if this intersection has a nonzero vector in it, it must be a subspace for which $A : X_i \cap X_j \to X_i \cap X_j$ and so there would be an eigenvector, $v \in X_i \cap X_j$. But then for some μ, we would have $Av = \mu v$ and by the second claim of the lemma, $\mu = \alpha_i$ and $\mu = \alpha_j$ showing that $X_i = X_j$. Now suppose $0 = \sum_{i=1}^m v_i$ where $v_i \in X_i$ and not all the v_i are equal to zero. Considering all such sums in which not all the v_i are equal to zero, we may assume that we have the one for which the number of nonzero vectors in the sum is as small as possible. Thus $\sum_{k=1}^r v_{i_k} = 0$, each v_{i_k} is nonzero, $v_{i_k} \in X_{i_k}$ and r is as small as possible for this to occur. Thus $r \geq 2$. There exists l, a positive integer such that $(A - \alpha_{i_r} I)^l v_{i_r} = 0$. Therefore, if we do $(A - \alpha_{i_r} I)^l$ to both sides of the sum, we see that $(A - \alpha_{i_r} I)^l v_{i_k} \in X_{i_k}$ and $\sum_{k=1}^{r-1} (A - \alpha_{i_r} I)^l v_{i_k} = 0$. Therefore, since r was as small as possible, we must have $(A - \alpha_{i_r} I)^l v_{i_k} = 0$ for all $k = 1, \cdots, r$. But this shows that $v_{i_k} \in X_{i_r}$ which implies $v_{i_k} \in X_{i_k} \cap X_{i_r} = \{0\}$, a contradiction to the assumption that all the $v_{i_k} \neq 0$. This proves the lemma.

The following corollary was proved earlier.

Corollary 7.17 *Suppose the $X_i, i = 1, \cdots, r$ are subspaces of X having property 3 of Lemma 7.16. Suppose also that $X_1 \oplus \cdots \oplus X_r = X$ and that $\left\{v_j^i\right\}_{j=1}^{m_i}$ is a basis for X_i. Then $B \equiv \cup_{i=1}^r \left\{v_j^i\right\}_{j=1}^{m_i}$ is a basis for X.*

Restating Corollary 7.15 on Page 73 yields the following theorem.

Theorem 7.18 *Let $A \in \mathcal{L}(X, X)$ and let X_p be the generalized eigenspace associated with the eigenvalue α_p where the eigenvalues of A are $\{\alpha_1, \cdots, \alpha_r\}$. Then the following hold.*

1. *$X = X_1 \oplus \cdots \oplus X_r$.*
2. *The dimension of X_i is equal to m_i, the multiplicity of the eigenvalue, α_i as a root of the characteristic equation, and $(A - \alpha_k I)^{m_k} X_k = 0$.*

Proof: The first claim follows from Corollary 7.15 on Page 73.

Finally we verify 2) by showing that $l_k = m_k$.

Claim: $(A - \alpha_k I)^{l_k} (X_k) = 0$ where l_k is defined to be the dimension of X_k.

If the claim is not true, there exists $z \in X_k$ such that for $B \equiv (A - \alpha_k I)$, none of the $l_k + 1$ vectors, $\left\{z, Bz, \cdots, B^{l_k} z\right\}$ are equal to zero. We obtain a contradiction by showing that in this case the foregoing list of vectors is linearly independent, contradicting the definition of l_k as the dimension of X_k. Suppose therefore, that

$$\sum_{i=0}^{l_k} c_i B^i z = 0. \tag{7.6}$$

Then for some m, $B^m z = 0$. Let m be as small as possible (m is necessarily larger than l_k.) so that $B^{m-1} z \neq 0$. Then multiply 7.6 by B^{m-1} to obtain $c_0 = 0$. Next multiply by B^{m-2} to obtain that $c_1 = 0$. Continuing in this way, yields $c_i = 0$ for all i thus establishing the vectors, $\left\{z, Bz, \cdots, B^{l_k} z\right\}$ are a linearly independent set and contradicting the definition of l_k as hoped. This proves the claim.

Now by 1) $(A - \alpha_k I)^{l_k}(X) \subseteq X_1 \oplus \cdots \oplus X_{k-1} \oplus X_{k+1} \oplus \cdots \oplus X_r$. Actually these are equal. This is easily seen by showing that $(A - \alpha_k I)^{l_k}$ is one to one on $X_s, s \neq k$. If $(A - \alpha_k I)^{l_k} v = 0$ for $v \in X_s$ then $v \in X_k$ also and so by Lemma 7.16, $v = 0$. Thus $(A - \alpha_k I)^{l_k}$ is one to one as just indicated. Furthermore, it maps X_s to X_s and so by Theorem 4.11, it maps this X_s onto X_s which implies $(A - \alpha_k I)^{l_k}$ maps X onto $X_1 \oplus \cdots \oplus X_{k-1} \oplus X_{k+1} \oplus \cdots \oplus X_r$. Therefore,

$$\operatorname{rank}(A - \alpha_k I)^{l_k} = n - l_k.$$

With respect to a basis for X which gives a Schur matrix which is similar to A, we see the matrix of $(A - \alpha_k I)^{l_k}$ is upper triangular and has exactly m_k zeroes on the main diagonal. Therefore,

$$\operatorname{rank}(A - \alpha_k I)^{l_k} \geq n - m_k$$

which shows that $m_k \geq l_k$ for each k. But also we have $\sum_{k=1}^r m_k = \sum_{k=1}^r l_k = n$ and so $m_k = l_k$ for each k. This proves the theorem.

With this theorem it is very easy to prove the Cayley Hamilton theorem

Corollary 7.19 *Let $A \in \mathcal{L}(X, X)$. Then A satisfies its characteristic equation.*

Proof: By Theorem 7.18,

$$\prod_{i=1}^r (A - \alpha_i I)^{m_i} = 0$$

but the characteristic equation is

$$\prod_{i=1}^r (\lambda - \alpha_i)^{m_i} = 0.$$

This proves the corollary.

Let A_s be the restriction of A to the space, X_s and let $\left\{v_j^s\right\}_{j=1}^{m_s}$ be an orthonormal basis for X_s such that the matrix of A_s, T_s, taken with respect to this basis is upper triangular. Thus in terms of the diagram for what is meant by the matrix of a linear transformation,

$$\begin{array}{ccccc} \{v_1^s, \cdots, v_{m_s}^s\} & X_s & \underrightarrow{A_s} & X_s & \{v_1^s, \cdots, v_{m_s}^s\} \\ & q_s \uparrow & \circ & \uparrow q_s & \\ & \mathbb{C}^{m_s} & \underrightarrow{T_s} & \mathbb{C}^{m_s} & \end{array}$$

By Lemma 7.16, A_s has only one eigenvalue, α_s, and so T_s is an upper triangular matrix of size $m_s \times m_s$ having the constant α_s down the main diagonal. We now describe the upper triangular matrix, T, as follows.

$$T = \begin{pmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_r \end{pmatrix}$$

where T_s was just described. Then we obtain the following corollary which follows directly from the definition of what is meant by the matrix of a linear transformation.

Corollary 7.20 *With respect to the basis* $\left\{v_j^1\right\}_{j=1}^{m_1}, \cdots, \left\{v_j^r\right\}_{j=1}^{m_r}$ *the matrix of* A *is the upper triangular matrix,* T*, just described.*

Proof: Let $N_0 = 0$, and $N_k = \sum_{i=1}^{k} m_i$ for $k = 1, \cdots, r$. Now let $i \in (N_{s-1}, N_s]$ and consider $\mathbf{e}_i$ the standard basis vector which has a one in the i^{th} slot. In the following diagram, q is the map which takes the vector $\mathbf{x} \in \mathbb{C}^n$ to the vector $\sum_{k=1}^{n} x_k w_k$ where $\{w_k\}_{k=1}^{n}$ is the given basis for X consisting of the vectors $\left\{v_j^1\right\}_{j=1}^{m_1}$ listed first and then the vectors, $\left\{v_j^2\right\}_{j=1}^{m_2}$ listed next and so forth. Therefore, for the given i, we must have $q\mathbf{e}_i$ is one of the basis vectors, $\left\{v_j^s\right\}_{j=1}^{m_s}$. In fact, $q\mathbf{e}_i = v_{i-N_{s-1}}^s$.

$$\begin{array}{ccc} X & \underrightarrow{A} & X \\ q\uparrow & \circ & \uparrow q \\ \mathbb{C}^n & \underrightarrow{T} & \mathbb{C}^n \end{array}$$

Now for each such i, using the definition of the T_s as the matrix of A_s, and block multiplication,

$$Aq\mathbf{e}_i = A_s v_{i-N_{s-1}}^s = q_s T_s \mathbf{e}_{i-N_{s-1}} = qT\mathbf{e}_i$$

Since $Aq = qT$ on a basis, it follows the two maps are equal and so T is the matrix of the linear transformation, A.

Corollary 7.21 *In the case where* A *is a real* $n \times n$ *matrix having all real eigenvalues, the above basis can be taken to consist of all real vectors.*

The basis in Corollary 7.20, $\left\{v_j^s\right\}_{j=1}^{m_s}$ may be changed so T_s may be simplified even further to one which has α_s down the main diagonal and either zero or one on the super diagonal, one space up from the main diagonal. This is done by abandoning the requirement that $\left\{v_j^s\right\}_{j=1}^{m_s}$ be orthonormal. When this further simplification has been accomplished we say the matrix is in Jordan Canonical form. However, the above reduction is sufficient for the applications of linear algebra. The matrix, T is called an upper triangular block diagonal matrix. In the next section we will consider the Jordan Canonical form. A good source for more on the Jordan Canonical form is [8]. For a more theoretical treatment of this subject which contains rigorous proofs in greater generality than presented here, along with other types of canonical forms see [5] or [6].

What is the significance of writing such a matrix for A? The reason this is important is that we can gain great understanding of powers of A by using the matrix T. Consider the case of most interest when A is itself a matrix mapping $\mathbb{C}^n$ to $\mathbb{C}^n$. Then we know from Theorem 4.5 there exists an $n \times n$ matrix, S such that

$$A = S^{-1}TS.$$

Therefore, it is easy to verify that

$$A^k = S^{-1}T^kS.$$

We now examine T^k. By block multiplication, we see that

$$T^k = \begin{pmatrix} T_1^k & & 0 \\ & \ddots & \\ 0 & & T_r^k \end{pmatrix}.$$

The significance of all this is that we can really understand T_s^k for large values of k. The matrix, T_s is of the form

$$T_s = \begin{pmatrix} \alpha & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha \end{pmatrix}$$

and we see that this is of the form

$$T_s = D + N$$

where D is a multiple of the identity and N is upper triangular with zeros down the main diagonal. Therefore, by the Cayley Hamilton theorem, $N^{m_s} = 0$ because the characteristic equation for N is just $\lambda^{m_s} = 0$. Such a transformation is called nilpotent . Now since D is just a multiple of the identity, it follows that $DN = ND$. Therefore, we can apply the usual binomial theorem and write

$$\begin{aligned} T^k &= (D+N)^k = \sum_{j=0}^{k} \binom{k}{j} D^{k-j} N^j \\ &= \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j} N^j = \sum_{j=0}^{m_s} \binom{k}{j} \alpha^{k-j} N^j, \end{aligned} \tag{7.7}$$

the last equation holding because $N^{m_s} = 0$. Thus T^k is of the form

$$T^k = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}$$

We use this formula in the next lemma.

Lemma 7.22 *Suppose T is of the form T_s described above where the constant, α, on the main diagonal is less than one in absolute value. Then*

$$\lim_{k\to\infty} \left(T^k\right)_{ij} = 0.$$

Proof: We use the formula 7.7. 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 $0 \leq j \leq m_s$,

$$\left|\left(T^k\right)_{pq}\right| \leq m_s C \left(\frac{k\,(k-1)\cdots(k-m_s+1)}{m_s!}\right) |\alpha|^{k-m_s}$$

which converges to zero. This is most easily seen by applying the ratio test to the series

$$\sum \left(\frac{k\,(k-1)\cdots(k-m_s+1)}{m_s!}\right) |\alpha|^{k-m_s}$$

and then noting that if a series converges, then the kth term converges to zero.

7.3 The Jordan canonical form

How can we tell if two matrices are similar, assuming the field of scalars is $\mathbb{C}$? The simple answer is that we compare their Jordan Canonical forms. Of course there are harder questions which come up when we change the field of scalars and these lead to different types of canonical forms but we shall not discuss these

questions here. Given an $n \times n$ matrix, A, we have proved in Corollary 7.20 the existence of an invertible matrix, S such that $S^{-1}AS = T$, an upper triangular matrix of the form

$$T = \begin{pmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_r \end{pmatrix}$$

Where T_k is an upper triangular matrix which is the matrix of A restricted to X_k for X_k the generalized eigenspace associated with an eigenvalue, α_k of A. Recall also that the dimension of X_k was equal to the algebraic multiplicity of the eigenvalue, α_k as a zero of the characteristic equation and

$$T_k = \begin{pmatrix} \alpha_k & \cdots & * \\ & \ddots & \vdots \\ 0 & & \alpha_k \end{pmatrix}.$$

We will show that if S is modified we can obtain a special form for each of these upper triangular matrices.

Definition 7.23 *We say $J_k(\alpha)$ 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, we have an unbroken string of ones down the super diagonal and the number, α filling every space on the main diagonal and zero everywhere else. We also will say a matrix is strictly upper triangular if it is of the form

$$\begin{pmatrix} 0 & * & * \\ \vdots & \ddots & * \\ 0 & \cdots & 0 \end{pmatrix},$$

where there are zeroes on the main diagonal and below the main diagonal.

Lemma 7.24 *Let A be a strictly upper triangular $k \times k$ matrix. Then $A^k = 0$.*

The proof of this lemma is left to the reader. Now following [6] we prove the following important Lemma about strictly upper triangular matrices.

Lemma 7.25 *Let A be an $n \times n$ matrix which is strictly upper triangular. Then there exists an invertible matrix, S such that*

$$S^{-1}AS = \begin{pmatrix} J_{k_1}(0) & & & 0 \\ & J_{k_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(0) \end{pmatrix}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = n$.

Proof: We prove this theorem by induction on the dimension, n. First of all if $n = 1$, there is nothing to prove because in this case, $A = (0)$ and there is no super diagonal. Suppose the theorem is true for $n-1$ and consider A. Since A is strictly upper triangular, we see that

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

where A_1 is an $n-1 \times n-1$ matrix which is strictly upper triangular. By induction, there exists S_1 such that

$$S_1^{-1} A_1 S_1 = \begin{pmatrix} J_{k_1} & & & 0 \\ & J_{k_2} & & \\ & & \ddots & \\ 0 & & & J_{k_s} \end{pmatrix}$$

where $J_{k_r} = J_{k_r}(0)$, $\sum k_i = n-1$ and the k_i are decreasing. Therefore, if we consider the matrices,

$$S_2 = \begin{pmatrix} 1 & 0 \\ 0 & S_1 \end{pmatrix}, S_2^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & S_1^{-1} \end{pmatrix},$$

$$\begin{aligned} S_2^{-1} A S_2 &= \begin{pmatrix} 0 & \mathbf{a}^T S_1 \\ 0 & S_1^{-1} A_1 S_1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{a}_1^T & \mathbf{a}_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{pmatrix}, \end{aligned} \tag{7.8}$$

where

$$J = \begin{pmatrix} J_{k_2} & & 0 \\ & \ddots & \\ 0 & & J_{k_s} \end{pmatrix} \tag{7.9}$$

and is an $n - k_1 - 1 \times n - k_1 - 1$ matrix. Now define a matrix,

$$S \equiv \begin{pmatrix} 1 & \mathbf{a}_1^T J_{k_1}^T & 0 \\ 0 & I_{k_1} & 0 \\ 0 & 0 & I_{n-k_1-1} \end{pmatrix},$$

where the entries of this matrix are themselves matrices. Then by block multiplication,

$$S^{-1} = \begin{pmatrix} 1 & -\mathbf{a}_1^T J_{k_1}^T & 0 \\ 0 & I_{k_1} & 0 \\ 0 & 0 & I_{n-k_1-1} \end{pmatrix}.$$

For M the matrix of 7.8 which is similar to A,

$$S^{-1} M S = \begin{pmatrix} 0 & \mathbf{a}_1^T - \mathbf{a}_1^T J_{k_1}^T J_{k_1} & \mathbf{a}_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{pmatrix}$$

where J is the matrix of 7.9. Now the top middle term is quite interesting. It is an easy computation to verify that

$$\mathbf{a}_1^T - \mathbf{a}_1^T J_{k_1}^T J_{k_1} = \left(\mathbf{a}_1^T \mathbf{e}_1\right) \mathbf{e}_1^T$$

where $\mathbf{e}_1$ is the standard basis vector in $\mathbb{F}^{k_1}$ which has a one in the first slot and a zero everywhere else. Thus we have shown that our matrix A is similar to the matrix,

$$\begin{pmatrix} 0 & \left(\mathbf{a}_1^T \mathbf{e}_1\right) \mathbf{e}_1^T & \mathbf{a}_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{pmatrix}. \tag{7.10}$$

Now there are two cases depending on whether $\mathbf{a}_1^T \mathbf{e}_1 = 0$. First we assume this is nonzero. Then we let

$$S = \begin{pmatrix} \mathbf{a}_1^T \mathbf{e}_1 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \left(\mathbf{a}_1^T \mathbf{e}_1\right) I \end{pmatrix}$$

where the I matrices in S have size k_1 and $n - k_1 - 1$. Then it follows that

$$S^{-1} = \begin{pmatrix} \left(\mathbf{a}_1^T \mathbf{e}_1\right)^{-1} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \left(\mathbf{a}_1^T \mathbf{e}_1\right)^{-1} I \end{pmatrix}$$

and for M the matrix of 7.10,

$$S^{-1} M S = \begin{pmatrix} 0 & \mathbf{e}_1^T & \mathbf{a}_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{pmatrix} \tag{7.11}$$

which shows this last matrix is similar to A. Now note that

$$\begin{pmatrix} 0 & \mathbf{e}_1^T \\ 0 & J_{k_1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 0 \end{pmatrix} = J_{k_1+1},$$

showing that the matrix in 7.11 is of the form

$$\begin{pmatrix} J_{k_1+1} & \mathbf{e}_1 \mathbf{a}_2^T \\ 0 & J \end{pmatrix} \tag{7.12}$$

where $\mathbf{e}_1$ is the right size $(k_1 + 1)$ such that

$$\begin{pmatrix} \mathbf{a}_2^T \\ 0 \end{pmatrix} = \left(\mathbf{e}_1 \mathbf{a}_2^T\right).$$

At this point the argument gets very interesting. We note that

$$J_{k_1+1} \mathbf{e}_{i+1} = \mathbf{e}_i \tag{7.13}$$

where $\mathbf{e}_i$ is the vector with zeroes in every slot except the i^{th} where it has a one. We have shown that the matrix in 7.12 is similar to A. Now we let

$$S = \begin{pmatrix} I & -\mathbf{e}_2 \mathbf{a}_2^T \\ 0 & I \end{pmatrix}, \text{so } S^{-1} = \begin{pmatrix} I & \mathbf{e}_2 \mathbf{a}_2^T \\ 0 & I \end{pmatrix}.$$

Then using 7.13 we see that

$$S^{-1} M S = \begin{pmatrix} J_{k_1+1} & \mathbf{e}_2 \mathbf{a}_2^T J \\ 0 & J \end{pmatrix}. \tag{7.14}$$

This is very interesting because what it did was to introduce a factor of J in the upper right corner. Our next similarity transformation will be

$$S = \begin{pmatrix} I & -\mathbf{e}_3\mathbf{a}_2^T J \\ 0 & I \end{pmatrix}, \text{so } S^{-1} = \begin{pmatrix} I & \mathbf{e}_3\mathbf{a}_2^T J \\ 0 & I \end{pmatrix}$$

and this will yield

$$\begin{pmatrix} J_{k_1+1} & \mathbf{e}_3\mathbf{a}_2^T J^2 \\ 0 & J \end{pmatrix},$$

which looks like what we started with in 7.14 except we have $\mathbf{e}_3$ and J^2 rather than J. We continue in this way, the next similarity being

$$S = \begin{pmatrix} I & -\mathbf{e}_4\mathbf{a}_2^T J^2 \\ 0 & I \end{pmatrix}$$

which will produce a power of 3 on the J in the upper right corner. After enough iterations, less than k_1+1, the upper right corner vanishes thanks to Lemma 7.24 and the observation that J is block diagonal having all blocks of size no larger than k_1. Therefore, we will be left with

$$\begin{pmatrix} J_{k_1+1} & 0 \\ 0 & J \end{pmatrix},$$

a matrix of the desired form. This proves the lemma in the case when $\mathbf{a}_1^T\mathbf{e}_1 \neq 0$ in 7.10.

Now we must consider the case when this quantity equals zero. In this case the matrix of 7.10 becomes

$$\begin{pmatrix} 0 & 0 & \mathbf{a}_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{pmatrix}. \tag{7.15}$$

Now we consider the matrix, P and its inverse,

$$P = \begin{pmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{pmatrix}, \; P^{-1} = \begin{pmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{pmatrix},$$

in which the identity matrices are the right size so we can form the following product for M the matrix of 7.15.

$$P^{-1}MP = \begin{pmatrix} J_{k_1} & 0 & 0 \\ 0 & 0 & \mathbf{a}_2^T \\ 0 & 0 & J \end{pmatrix} \tag{7.16}$$

Now we use induction to get the existence of S_1 of the appropriate size, less than n, such that

$$S_1^{-1}\begin{pmatrix} 0 & \mathbf{a}_2^T \\ 0 & J \end{pmatrix} S_1 = \widetilde{J}.$$

Therefore, letting

$$S = \begin{pmatrix} I & 0 \\ 0 & S_1 \end{pmatrix},$$

$$S^{-1}MS = \begin{pmatrix} J_{k_1} & 0 \\ 0 & \widetilde{J} \end{pmatrix},$$

which is a matrix of the right sort except that the Jordan blocks might not be getting smaller as you go from the top left to the lower right. If this is the case, it can be fixed by using a similarity involving matrices like P above which permute the order of the blocks. This proves the lemma.

Corollary 7.26 *Suppose A is an upper triangular $n \times n$ matrix having α in every position on the main diagonal. Then there exists an invertible matrix, S such that*

$$S^{-1}AS = \begin{pmatrix} J_{k_1}(\alpha) & & & 0 \\ & J_{k_2}(\alpha) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\alpha) \end{pmatrix}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = n$.

Proof: The matrix, $A - \alpha I$ is strictly upper triangular and so by Lemma 7.25 there exists S such that

$$S^{-1}(A - \alpha I)S = \begin{pmatrix} J_{k_1}(0) & & & 0 \\ & J_{k_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(0) \end{pmatrix} \equiv J$$

Therefore, $S^{-1}AS = \alpha I + J$, which is of the form claimed in the conclusion of the corollary. This proves the corollary.

With this corollary, we can now present the following theorem which gives the existence of the Jordan canonical form.

Theorem 7.27 *Let A be an $n \times n$ matrix having eigenvalues $\alpha_1, \cdots, \alpha_r$ where the multiplicity of α_i as a zero of the characteristic polynomial equals m_i. Then there exists an invertible matrix, S such that*

$$S^{-1}AS = \begin{pmatrix} J(\alpha_1) & & 0 \\ & \ddots & \\ 0 & & J(\alpha_r) \end{pmatrix} \tag{7.17}$$

where $J(\alpha_k)$ is an $m_k \times m_k$ matrix of the form

$$\begin{pmatrix} J_{k_1}(\alpha_k) & & & 0 \\ & J_{k_2}(\alpha_k) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\alpha_k) \end{pmatrix} \tag{7.18}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = m_k$.

Proof: By Corollary 7.20,

$$T = \begin{pmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_r \end{pmatrix}$$

where T_k is an upper triangular $m_k \times m_k$ matrix having α_k down the main diagonal. Now by Corollary 7.26, there exists S_k such that $S_k^{-1} T_k S_k$ is of the form given in 7.18. Therefore, we let S be given by

$$S = \begin{pmatrix} S_1 & & 0 \\ & \ddots & \\ 0 & & S_r \end{pmatrix}.$$

This proves the theorem.

Corollary 7.28 *If the matrix, A is real and has all real eigenvalues, then the similarity matrix can be taken to be real.*

Proof: This follows from the above process and the corresponding theorem for the Schur form of a matrix.

Corollary 7.29 *Let $\varepsilon > 0$ be given. Any matrix is similar to one of the form*

$$\begin{pmatrix} J(\alpha_1, \varepsilon) & & 0 \\ & \ddots & \\ 0 & & J(\alpha_r, \varepsilon) \end{pmatrix}$$

where $J(\alpha_k, \varepsilon)$ is an $m_k \times m_k$ matrix of the form

$$\begin{pmatrix} J_{k_1}(\alpha_k, \varepsilon) & & & 0 \\ & J_{k_2}(\alpha_k, \varepsilon) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\alpha_k, \varepsilon) \end{pmatrix}$$

where $J_k(\alpha, \varepsilon)$ is of the form

$$J_k(\alpha, \varepsilon) = \begin{pmatrix} \alpha & \varepsilon & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \varepsilon \\ 0 & \cdots & 0 & \alpha \end{pmatrix}$$

Thus we replaced the ones on the superdiagonal in the Jordan form with ε.

Proof: Take the Jordan form of a matrix and do another similarity transformation with the diagonal matrix,

$$D_\varepsilon = \begin{pmatrix} 1 & & 0 \\ & \varepsilon & \\ & & \ddots \\ 0 & & \varepsilon^{n-1} \end{pmatrix}.$$

For example

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

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

Obviously if we change the order in which we list the eigenvalues, we get a different matrix for the Jordan canonical form. However, if we keep the order of the eigenvalues the same, we can show the Jordan canonical form for any two similar matrices is the same.

Theorem 7.30 *Let A and B be two similar matrices. Let J_A and J_B be Jordan forms of A and B respectively made up of the blocks $J_A(\alpha_i)$ and $J_B(\alpha_i)$ respectively. Then J_A and J_B are identical except possibly for the order of the $J(\alpha_i)$ where the α_i are defined above.*

Proof: First note that for α_i an eigenvalue, the matrices $J_A(\alpha_i)$ and $J_B(\alpha_i)$ are both of size $m_i \times m_i$ because the two matrices A and B, being similar, have exactly the same characteristic equation. We only need to worry about the number and size of the Jordan blocks making up $J_A(\alpha_i)$ and $J_B(\alpha_i)$. Let the eigenvalues of A and B be $\{\alpha_1, \cdots, \alpha_r\}$. Consider the two sequences of numbers $\{\operatorname{rank}(A-\lambda I)^m\}$ and $\{\operatorname{rank}(B-\lambda I)^m\}$. Since A and B are similar, there exists a linear transformation, L for which both A and B are matrices. Therefore, the above sequences both coincide with the sequence $\{\operatorname{rank}(L-\lambda I)^m\}$. Also, for the same reason, we must have $\{\operatorname{rank}(J_A-\lambda I)^m\}$ coincides with $\{\operatorname{rank}(J_B-\lambda I)^m\}$. Now pick α_k an eigenvalue and consider $\{\operatorname{rank}(J_A-\alpha_k I)^m\}$ and $\{\operatorname{rank}(J_B-\alpha_k I)^m\}$. Then

$$J_A - \alpha_k I = \begin{pmatrix} J_A(\alpha_1-\alpha_k) & & & & 0 \\ & \ddots & & & \\ & & J_A(0) & & \\ & & & \ddots & \\ 0 & & & & J_A(\alpha_r-\alpha_k) \end{pmatrix}$$

and a similar formula holds for $J_B - \alpha_k I$. Here

$$J_A(0) = \begin{pmatrix} J_{k_1}(0) & & & 0 \\ & J_{k_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(0) \end{pmatrix}$$

and

$$J_B(0) = \begin{pmatrix} J_{l_1}(0) & & & 0 \\ & J_{l_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{l_p}(0) \end{pmatrix}$$

and we need to verify that $l_i = k_i$ for all i. We already know, as noted above, that $\sum k_i = \sum l_i$. Now from the above formulas, we see that

$$\begin{aligned} \operatorname{rank}(J_A-\alpha_k I)^m &= \sum_{i\neq k} m_i + \operatorname{rank}(J_A(0)^m) \\ &= \sum_{i\neq k} m_i + \operatorname{rank}(J_B(0)^m) \\ &= \operatorname{rank}(J_B-\alpha_k I)^m, \end{aligned}$$

which shows that we must have $\text{rank}\left(J_A\left(0\right)^m\right) = \text{rank}\left(J_B\left(0\right)^m\right)$ for all m. However,

$$J_B\left(0\right)^m = \begin{pmatrix} J_{l_1}\left(0\right)^m & & & 0 \\ & J_{l_2}\left(0\right)^m & & \\ & & \ddots & \\ 0 & & & J_{l_p}\left(0\right)^m \end{pmatrix}$$

with a similar formula holding for $J_A\left(0\right)^m$ and $\text{rank}\left(J_B\left(0\right)^m\right) = \sum_{i=1}^{p} \text{rank}\left(J_{l_i}\left(0\right)^m\right)$, similar for $\text{rank}\left(J_A\left(0\right)^m\right)$. In going from m to $m+1$,

$$\text{rank}\left(J_{l_i}\left(0\right)^m\right) - 1 = \text{rank}\left(J_{l_i}\left(0\right)^{m+1}\right)$$

until $m = l_i$ at which time there is no further change. Therefore, we see $p = r$ since otherwise, we would introduce a discrepancy right away in going from $m = 1$ to $m = 2$. Now suppose the sequence $\{l_i\}$ is not equal to the sequence, $\{k_i\}$. Then $l_{r-b} \neq k_{r-b}$ for some b a nonnegative integer and we will take b to be a small as possible. Say $l_{r-b} > k_{r-b}$. Then, letting $m = k_{r-b}$,

$$\sum_{i=1}^{r} \text{rank}\left(J_{l_i}\left(0\right)^m\right) = \sum_{i=1}^{r} \text{rank}\left(J_{k_i}\left(0\right)^m\right)$$

and in going to $m+1$ a discrepancy must occur because the sum on the right will contribute less to the decrease in rank than the sum on the left. This proves the theorem.

The proof of this theorem gives a way to find the Jordan canonical form of a matrix without finding the corresponding similarity transformation. We illustrate with the following example.

Example 7.31 *Consider the matrix*

$$\begin{pmatrix} 3 & 1 & 1 & -1 \\ -6 & 3 & -6 & 12 \\ -6 & 1 & -2 & 11 \\ -3 & 1 & -2 & 8 \end{pmatrix}$$

Some work will show it has a repeated eigenvalue of 3. Now we will calculate the various ranks.

$$\begin{pmatrix} 3 & 1 & 1 & -1 \\ -6 & 3 & -6 & 12 \\ -6 & 1 & -2 & 11 \\ -3 & 1 & -2 & 8 \end{pmatrix} - 3\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

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

Recall that we can easily determine rank by finding the row rank, a process that can be accomplished through the standard row operations which do not change the rank of the matrix. We do row operations till we obtain the row reduced echelon form and then it is obvious what the row rank is. Thus for our above matrix, the row reduced echelon form is

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

and so

$$\begin{pmatrix} 0 & 1 & 1 & -1 \\ -6 & 0 & -6 & 12 \\ -6 & 1 & -5 & 11 \\ -3 & 1 & -2 & 5 \end{pmatrix}, \text{ rank: } 2$$

Similarly,

$$\begin{pmatrix} 0 & 1 & 1 & -1 \\ -6 & 0 & -6 & 12 \\ -6 & 1 & -5 & 11 \\ -3 & 1 & -2 & 5 \end{pmatrix}^2 = \begin{pmatrix} -9 & 0 & -9 & 18 \\ 0 & 0 & 0 & 0 \\ -9 & 0 & -9 & 18 \\ -9 & 0 & -9 & 18 \end{pmatrix}, \text{ rank: } 1$$

$$\begin{pmatrix} -9 & 0 & -9 & 18 \\ 0 & 0 & 0 & 0 \\ -9 & 0 & -9 & 18 \\ -9 & 0 & -9 & 18 \end{pmatrix}^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \text{ rank: } 0$$

Because of this sequence of numbers we see $A - 3I$ has Jordan form

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

Therefore, the Jordan form of this matrix must be

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

This did not give us a similarity transformation which will put the given matrix in Jordan form. Although the above proof of the existence of the Jordan form of a matrix might seem to provide an algorithm for producing the Jordan form along with the similarity transformation, this is really an illusion. It was based first of all on being able to obtain a similar matrix to the given one which is upper triangular and block diagonal as described above. A first step was a Schur form for the matrix. This required us to find eigenvalues and eigenvectors which is arguably the hardest problem in algebra. No substantial improvement occurs if we don't try to find the similarity transformation but only the Jordan form as in the above because we still need to find the eigenvalues. Theoretically if we do not have precise knowledge of the eigenvalues, everything fails even in the above process because $(A - \lambda I)^m$ will always have rank n unless λ is exactly an eigenvalue. However, in principle, one can always use the above methods to find the Jordan form if the eigenvalues are known exactly.

7.4 Applications to differential equations

In elementary courses in differential equations, a procedure is presented for finding the solutions to simple linear first order systems of ordinary differential equations

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

in terms of the eigenvectors and generalized eigenvectors of the matrix, A. It is a procedure which is entrenched in most of the current books on ordinary differential equations even though it is hard to imagine anything more completely asinine than to declare one has made progress by taking one of the easiest problems in analysis, the initial value problem, and transforming it into arguably the hardest problem in Algebra, the eigenvalue problem. The reason it is done this way is that people desire desperately to have closed form

expressions for the solutions to differential equations. They want these closed form expressions more than they want to understand what is really going on and this is why the frantic introduction of algebra in an area where it does not belong. Furthermore, there is a monumental loose end in the presentations given in most elementary books on the subject which amounts to the question of the existence of the Jordan canonical form. We now know a theorem on the existence of this matrix so we can now justify the elementary differential equations approach. The procedure does supply a nice way to try to calculate the Jordan form of a matrix, however, and so we present it here. (In fairness, linear algebra can be used to give a correct treatment of linear differential equations without the loose ends. See Apostol [1] or Chapter 11.)

Consider the first order system,

$$\mathbf{x}' = A\mathbf{x},$$

where A is an $n \times n$ matrix. Then we see that any vector valued function of the form $\mathbf{x}(t) = e^{\lambda t}\mathbf{v}$ is a solution of this system provided $\mathbf{v}$ is an eigenvector and λ is the eigenvalue which goes with λ. If we can find a basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, then we can successfully argue that the general solution to the system above is

$$\sum_{i=1}^{n} C_i e^{\lambda_i t}\mathbf{v}_i$$

where $(\lambda_i, \mathbf{v}_i)$ is an eigen pair for the matrix, A. The difficulty occurs when we cannot find a basis of eigenvectors. In other words, the difficulty occurs when A is defective. In this case we look for generalized eigenvectors. Starting with an eigenvector, $\mathbf{v}$, we look for a vector, $\mathbf{w}$ which has the property that $(A - \lambda I)\mathbf{w} = \mathbf{v}$. Thus both $\mathbf{v}$ and $\mathbf{w}$ are in the generalized eigenspace but $\mathbf{w}$ is in $\ker(A - \lambda I)^2$ although it is not in $\ker(A - \lambda I)$. Then we can show that

$$t \to t\mathbf{v}e^{\lambda t} + \mathbf{w}e^{\lambda t}$$

is a solution to our system and by Problem 3, $\mathbf{v}$ and $\mathbf{w}$ form an independent set. We then find all solutions of this form and if we can't find enough of these we look for vectors in $\ker(A - \lambda I)^3$ which are not in $\ker(A - \lambda I)^2$ and then look for solutions of the form

$$\frac{t^2}{2}\mathbf{v}e^{\lambda t} + t\mathbf{w}e^{\lambda t} + \mathbf{z}e^{\lambda t},$$

which can be shown to solve the differential equation provided $(A - \lambda I)\mathbf{z} = \mathbf{w}$ and $(A - \lambda I)\mathbf{w} = \mathbf{v}$. In general, we look for solutions of the form

$$\sum_{k=0}^{p} \frac{t^{p-k}}{(p-k)!} e^{\lambda t}\mathbf{v}_k \tag{7.19}$$

where $(A - \lambda I)\mathbf{v}_{k+1} = \mathbf{v}_k$, and $\mathbf{v}_0$ is an eigenvector for the eigenvalue, λ. Then using existence and uniqueness theorems we see we have the general solution exactly when we can find a basis consisting of generalized eigenvectors and eigenvectors of the sort just described. However, this turns out to be exactly equivalent to asking whether we can reduce the matrix to Jordan canonical form. We illustrate with a simple problem.

Example 7.32 *Consider the matrix*

$$\begin{pmatrix} 3 & 1 & 1 & -1 \\ -6 & 3 & -6 & 12 \\ -6 & 1 & -2 & 11 \\ -3 & 1 & -2 & 8 \end{pmatrix}$$

and the problem of finding the general solution to $\mathbf{x}' = A\mathbf{x}$.

We first look for eigenvectors and eigenvalues. We find there are two eigenvectors and that every eigenvector is a linear combination of these two. Thus the eigenspace is of dimension 2. There vectors are $\mathbf{v}_1 = (-1,-1,1,0)^T$ and $\mathbf{v}_2 = (2,1,0,1)^T$ and the eigenvalue is 3. Therefore, anything of the form

$$C_1 \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} e^{3t} + C_2 \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} e^{3t}$$

is a solution. However we need two more solutions so we look for generalized eigenvectors. Thus we look for a constants, a and b such that there is a solution, $\mathbf{w}$ to $(A-3I)\,\mathbf{w} = a\mathbf{v}_1 + b\mathbf{v}_2$. The reason for the constants, a and b is that we may be able to find a generalized eigenvector, $\mathbf{w}$ which maps to $a\mathbf{v}_1 + b\mathbf{v}_2$ for some values of a and b and not for others. Thus we need to solve the following

$$\begin{pmatrix} 0 & 1 & 1 & -1 \\ -6 & 0 & -6 & 12 \\ -6 & 1 & -5 & 11 \\ -3 & 1 & -2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = a \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

Placing this in augmented matrix form we have

$$\begin{pmatrix} 0 & 1 & 1 & -1 & -a+2b \\ -6 & 0 & -6 & 12 & -a+b \\ -6 & 1 & -5 & 11 & a \\ -3 & 1 & -2 & 5 & b \end{pmatrix}$$

Doing row operations,

$$\begin{pmatrix} 1 & 0 & 1 & -2 & \frac{-1}{3}(a-b) \\ 0 & 1 & 1 & -1 & -a+2b \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3a+3b \end{pmatrix}$$

and so we see that we get a solution exactly when $a = b$. Thus we must solve the system of equations whose augmented matrix is

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

This shows we have a generalized eigenvector $(2w-z, a+w-z, z, w)$ where we can pick a, w and z. However, we can't take $a = 0$ because then this would mean we would not be getting anything but an eigenvector. This will provide us with another solution but we will still need one more. We try to find yet another generalized eigenvector. Thus we look for a solution to the system whose augmented matrix is

$$\begin{pmatrix} 0 & 1 & 1 & -1 & 2w-z \\ -6 & 0 & -6 & 12 & a+w-z \\ -6 & 1 & -5 & 11 & z \\ -3 & 1 & -2 & 5 & w \end{pmatrix}$$

The row echelon form of this is

$$\begin{pmatrix} 1 & 0 & 1 & -2 & \frac{-1}{3}(z-w) \\ 0 & 1 & 1 & -1 & 2w-z \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3w+a-3z \end{pmatrix}$$

and we see we should take $3w + a - 3z = 0$ in order to obtain a solution. Therefore, we take $a = 3, w = 1$, and $z = 2$.

Now we put all this wretched stuff together and find our general solution. The third solution is

$$t\begin{pmatrix} 3 \\ 0 \\ 3 \\ 3 \end{pmatrix} e^{3t} + \begin{pmatrix} 0 \\ 2 \\ 2 \\ 1 \end{pmatrix} e^{3t}$$

and the fourth solution is

$$\frac{t^2}{2}\begin{pmatrix} 3 \\ 0 \\ 3 \\ 3 \end{pmatrix} e^{3t} + t\begin{pmatrix} 0 \\ 2 \\ 2 \\ 1 \end{pmatrix} e^{3t} + \begin{pmatrix} -(1/3) \\ -1 \\ 2 \\ 1 \end{pmatrix} e^{3t}.$$

Therefore, a fundamental solution matrix is

$$F(t) = e^{3t}\begin{pmatrix} -1 & 2 & 3t & \frac{3}{2}t^2 - \frac{1}{3} \\ -1 & 1 & 2 & 2t - 1 \\ 1 & 0 & 3t + 2 & \frac{3}{2}t^2 + 2t + 2 \\ 0 & 1 & 3t + 1 & \frac{3}{2}t^2 + t + 1 \end{pmatrix}$$

where we simply let the j^{th} solution equal the j^{th} column of this matrix. Thus the general solution of the system of differential equations is

$$F(t)\,\mathbf{C}$$

where $\mathbf{C} \in \mathbb{C}^n$. If you like to do lots of symbol pushing, you can verify that this does indeed work. Furthermore, it is a general solution because when we let $t = 0$ we get

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

a matrix which has nonzero determinant. Therefore, by the theorems of differential equations based on existence and uniqueness, we see this is the general solution as claimed. A more interesting aspect of this problem in terms of finding the Jordan form is as follows. We consider the matrix,

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

which we got by listing the two eigenvectors first, but instead of $\begin{pmatrix} 2 & 1 & 0 & 1 \end{pmatrix}^T$ for the second column, we put the eigenvector which was at the base of the chain of generalized eigenvectors. We will consider the similarity transformation determined by this matrix. Thus, the inverse of this matrix is

$$\begin{pmatrix} 3 & -\frac{1}{3} & \frac{11}{3} & -\frac{20}{3} \\ 1 & -\frac{1}{9} & \frac{8}{9} & -\frac{14}{9} \\ 0 & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ -3 & 0 & -3 & 6 \end{pmatrix}$$

and we take the similarity

$$\begin{pmatrix} 3 & -\frac{1}{3} & \frac{11}{3} & -\frac{20}{3} \\ 1 & -\frac{1}{9} & \frac{8}{9} & -\frac{14}{9} \\ 0 & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ -3 & 0 & -3 & 6 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 & -1 \\ -6 & 3 & -6 & 12 \\ -6 & 1 & -2 & 11 \\ -3 & 1 & -2 & 8 \end{pmatrix}.$$

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

which is essentially the Jordan form of the matrix, A except that the blocks are out of order. This could be fixed by using a similarity involving a permutation matrix.

7.5 Exercises

1. Let V be a finite dimensional inner product space and let M be a subspace of V. We define $M^{\perp} \equiv \{x \in V : (x,u) = 0 \text{ for all } u \in M\}$. Show that $M^{\perp}$ is a subspace of V even if M is only a subset and then verify that in the case when M is a subspace, $V = M^{\perp} \oplus M$.

2. Prove Lemma 7.24.

3. Suppose A is an $n \times n$ matrix and consider the chain of vectors, $\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \cdots, A^p\mathbf{v}$ where none of $A^k\mathbf{v}$ are equal to zero for $k \leq p$ but $A^{p+1}\mathbf{v} = \mathbf{0}$. Show $\left\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \cdots, A^p\mathbf{v}\right\}$ is linearly independent.

4. Verify that 7.19 is a solution to the equation, $\mathbf{x}' = A\mathbf{x}$ as claimed.

5. In the finding of the Jordan form and similarity matrix for

$$\begin{pmatrix} 3 & 1 & 1 & -1 \\ -6 & 3 & -6 & 12 \\ -6 & 1 & -2 & 11 \\ -3 & 1 & -2 & 8 \end{pmatrix}$$

   find a similarity matrix which will result in a Jordan form having the size of the blocks decreasing from the upper left to the lower right.

6. Prove that a similarity transformation of the sort discussed above in Corollary 7.29 has the desired effect of replacing the ones on the super diagonal with ε.

7. By the Cayley Hamilton theorem, we know every matrix, A, satisfies its characteristic equation. The minimal polynomial is a polynomial, $p(t) = t^m + \cdots + a_1 t + a_0$ such that $p(A) = 0$ and m is as small as possible. Show the minimal polynomial divides the characteristic polynomial. **Hint:** Consider the matrix, $J(\alpha)$ which is block diagonal. What if the largest block is $k \times k$ and you considered the factor $(t-\alpha)^k$?

8. The eigenspace of a matrix associated with an eigenvalue, α is defined as $\ker(A - \alpha I)$. Show the eigenspace associated with an eigenvalue, α has dimension equal to the number of Jordan blocks in $J(\alpha)$.

9. Find the Jordan canonical form for the following matrices and give the corresponding similarity transformations.

(a) $\begin{pmatrix} 0 & 4 & 2 \\ -1 & -4 & -1 \\ 0 & 0 & -2 \end{pmatrix}$

(b) $\begin{pmatrix} 8 & 12 & -12 \\ 6 & 24 & 11 \\ 8 & -8 & 28 \end{pmatrix}$

10. Find the Jordan form and similarity transformation for the matrix

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

11. Suppose X_i are subspaces of X. Show that the condition: "If $\sum_{i=1}^{m} v_i = 0$ for $v_i \in X_i$, it follows that $v_i = 0$ for all i." is equivalent to the condition $X_i \cap \left(\sum_{j \neq i} X_j\right) = \{0\}$.

12. Let A be a linear transformation mapping an inner product space, X of dimension n to itself and let X_i be subspaces of X such that $X = X_1 \oplus \cdots \oplus X_r$. Let $\left\{v_j^i\right\}_{j=1}^{m_i}$ be any basis for X_i. Suppose also that $A : X_i \to X_i$ and let M_i be the matrix of A restricted to X_i. Now let $\{w_j\}_{j=1}^{n}$ be the basis for X consisting of $\left\{v_j^1\right\}_{j=1}^{m_1}$ first and then $\left\{v_j^2\right\}_{j=1}^{m_2}$ and so forth. show that with respect to $\{w_j\}_{j=1}^{n}$, the matrix of A is the block diagonal matrix,

$$\begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_r \end{pmatrix}.$$

The Perron Frobenius Theorem

In this section we use the theory of similar block diagonal matrices to prove a very significant theorem important in the study of Markov processes.

Definition 8.1 *We say an $n \times n$ matrix, $A = (a_{ij})$, is a Markov matrix (Stochastic matrix) if $a_{ij} \geq 0$ for all i, j and*

$$\sum_i a_{ij} = 1.$$

A Markov matrix is called regular if some power of A has all entries strictly positive. We also call a vector, $\mathbf{v} \in \mathbb{R}^n$, a steady state if $A\mathbf{v} = \mathbf{v}$.

Lemma 8.2 *Suppose $A = (a_{ij})$ is a Markov matrix in which $a_{ij} > 0$ for all i, j. Then if μ is an eigenvalue of A, either $|\mu| < 1$ or $\mu = 1$. In addition to this, if $A\mathbf{v} = \mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^n$, then $v_j v_i \geq 0$ for all i, j so the components of $\mathbf{v}$ have the same sign.*

Proof: Let $\sum_j a_{ij} v_j = \mu v_i$ where $\mathbf{v} \equiv (v_1, \cdots, v_n)^T \neq \mathbf{0}$. Then

$$\sum_j a_{ij} v_j \overline{\mu v_i} = |\mu|^2 |v_i|^2$$

and so

$$|\mu|^2 |v_i|^2 = \sum_j a_{ij} \mathrm{Re}\left(v_j \overline{\mu v_i}\right) \leq \sum_j a_{ij} |v_j| |\mu| |v_i| \tag{8.1}$$

so

$$|\mu| |v_i| \leq \sum_j a_{ij} |v_j| .$$

Summing on i, we obtain

$$|\mu| \sum_i |v_i| \leq \sum_j \sum_i a_{ij} |v_j| = \sum_j |v_j| . \tag{8.2}$$

Therefore, $|\mu| \leq 1$. If $|\mu| = 1$, then equality must hold in 8.1 for each i and so $v_j \overline{\mu v_i}$ must be real and nonnegative for all j. In particular, for $j = i$, we must have $|v_i|^2 \overline{\mu} \geq 0$ for each i. Hence μ must be real and non negative. Thus $\mu = 1$. If $A\mathbf{v} = \mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^n$ so $\mu = 1$, we must have equality in 8.1 since otherwise, we would have strict inequality in 8.2 which cannot occur. Therefore, $v_j v_i \geq 0$ for all i, j showing the sign of v_i is constant. This proves the lemma.

Lemma 8.3 *If A is any Markov matrix, there exists $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ with $A\mathbf{v} = \mathbf{v}$. Also, if A is a Markov matrix in which $a_{ij} > 0$ for all i, j, and*

$$X_1 \equiv \{\mathbf{x} : (A - I)^m \mathbf{x} = \mathbf{0} \text{ for some } m\},$$

then the dimension of $X_1 = 1$.

Proof: Let $\mathbf{u} = (1, \cdots, 1)^T$ be the vector in which there is a one for every entry. Then since A is a Markov matrix,

$$\mathbf{u}^T A = \mathbf{u}^T.$$

Therefore, $A^T \mathbf{u} = \mathbf{u}$ showing that 1 is an eigenvalue for A^T. It follows that 1 must also be an eigenvalue for A since A and A^T have the same characteristic equation due to the fact the determinant of a matrix equals the determinant of its transpose. Since A is a real matrix, it follows there exists $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that $(A - I)\mathbf{v} = 0$. By Lemma 8.2, v_i has the same sign for all i. Without loss of generality assume $\sum_i v_i = 1$ and so $v_i \geq 0$ for all i.

Now suppose A is a Markov matrix in which $a_{ij} > 0$ for all i, j and suppose $\mathbf{w} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ satisfies $A\mathbf{w} = \mathbf{w}$. Then some $w_p \neq 0$ and we must have equality in 8.1. Therefore,

$$w_j \overline{w_p} \equiv r_j \geq 0.$$

Then letting $\mathbf{r} \equiv (r_1, \cdots, r_n)^T$,

$$A\mathbf{r} = A\mathbf{w}\overline{w_p} = \mathbf{w}\overline{w_p} = \mathbf{r}.$$

Now defining $||\mathbf{r}||_1 \equiv \sum_i |r_i|$, we see that $\sum_i \left(\frac{r_i}{||\mathbf{r}||_1}\right) = 1$. Also,

$$A\left(\frac{\mathbf{r}}{||\mathbf{r}||_1} - \mathbf{v}\right) = \frac{\mathbf{r}}{||\mathbf{r}||_1} - \mathbf{v}$$

and so, since all eigenvectors for $\lambda = 1$ have all entries the same sign, and

$$\sum_i \left(\frac{r_i}{||\mathbf{r}||_1} - v_i\right) = 1 - 1 = 0,$$

it follows that for all i,

$$\frac{r_i}{||\mathbf{r}||_1} = v_i$$

and so $\frac{\mathbf{r}}{||\mathbf{r}||_1} = \frac{\mathbf{w}\overline{w_p}}{||\mathbf{r}||_1} = \mathbf{v}$ showing that

$$\mathbf{w} = \frac{||\mathbf{r}||_1}{\overline{w_p}}\mathbf{v}.$$

This shows that all eigenvectors for the eigenvalue 1 are multiples of the single eigenvector, $\mathbf{v}$, described above.

Now suppose that

$$(A - I)\mathbf{w} = \mathbf{z}$$

where $\mathbf{z}$ is an eigenvector. Then we know $\mathbf{z} = \alpha\mathbf{v}$ where $v_j \geq 0$ for all j, and $\sum_j v_j = 1$. It follows that

$$\sum_j a_{ij} w_j - w_i = z_i = \alpha v_i.$$

Then summing on i,

$$\sum_j w_j - \sum_i w_i = 0 = \alpha \sum_i v_i.$$

But $\sum_i v_i = 1$. Therefore, $\alpha = 0$ and so $\mathbf{z}$ is not an eigenvector. Therefore, if $(A-I)^2 \mathbf{w} = \mathbf{0}$, then it follows $(A-I)\mathbf{w} = \mathbf{0}$ and so in fact,

$$X_1 = \{\mathbf{w} : (A-I)\mathbf{w} = \mathbf{0}\}$$

and this was just shown to be one dimensional. This proves the lemma.

The following lemma is fundamental to what follows.

Lemma 8.4 *Let A be a Markov matrix in which $a_{ij} > 0$ for all i, j. Then there exists a basis for $\mathbb{C}^n$ such that with respect to this basis, the matrix for A is the upper triangular, block diagonal matrix,*

$$T = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & T_1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix}$$

where T_s is an upper triangular matrix of the following form.

$$T_s = \begin{pmatrix} \mu_s & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu_s \end{pmatrix}.$$

Here $|\mu_s| < 1$.

Proof: This follows from Lemma 8.3 and Corollary 7.20. The assertion about $|\mu_s|$ follows from Lemma 8.2.

Lemma 8.5 *Let A be any Markov matrix and let $\mathbf{v}$ be a vector having all its components non negative and having $\sum_i v_i = 1$. Then if $\mathbf{w} = A\mathbf{v}$, it follows $w_i \geq 0$ for all i and $\sum_i w_i = 1$.*

Proof: From the definition of $\mathbf{w}$,

$$w_i \equiv \sum_j a_{ij} v_j \geq 0.$$

Also

$$\sum_i w_i = \sum_i \sum_j a_{ij} v_j = \sum_j \sum_i a_{ij} v_j = \sum_j v_j = 1.$$

The following theorem, a special case of the Perron Frobenius theorem can now be proved.

Theorem 8.6 *Suppose A is a Markov matrix in which $a_{ij} > 0$ for all i, j and suppose $\mathbf{w}$ is a vector. Then for each i,*

$$\lim_{k\to\infty} \left(A^k \mathbf{w}\right)_i = v_i$$

where $A\mathbf{v} = \mathbf{v}$. In words, $A^k\mathbf{w}$ always converges to a steady state. In addition to this, if the vector, $\mathbf{w}$ satisfies $w_i \geq 0$ for all i and $\sum_i w_i = 1$, Then the vector, $\mathbf{v}$ will also satisfy the conditions, $v_i \geq 0$, $\sum_i v_i = 1$.

Proof: We know there exists a matrix, S such that

$$A = S^{-1}TS$$

where T is defined above. Therefore,

$$A^k = S^{-1}T^kS$$

By Lemma 7.22, the components of the matrix, A^k converge to the components of the matrix

$$S^{-1}US$$

where U is an $n \times n$ matrix of the form

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

a matrix with a one in the upper left corner and zeros elsewhere. It follows that there exists a vector $\mathbf{v} \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} \left(A^k\mathbf{w}\right)_i = v_i.$$

It follows

$$v_i = \lim_{k\to\infty} \left(AA^k\mathbf{w}\right)_i = \lim_{k\to\infty} \sum_j a_{ij} \left(A^k\mathbf{w}\right)_j = \sum_j a_{ij}v_j = (A\mathbf{v})_i$$

Since i is arbitrary, this implies $A\mathbf{v} = \mathbf{v}$ as claimed.

Now if $w_i \geq 0$ and $\sum_i w_i = 1$, then by Lemma 8.5, $\left(A^k\mathbf{w}\right)_i \geq 0$ and

$$\sum_i \left(A^k\mathbf{w}\right)_i = 1.$$

Therefore, $\mathbf{v}$ also satisfies these conditions. This proves the theorem.

The following corollary is called the Perron Frobenius theorem and it is the fundamental result of this section. This corollary is a simple consequence of the following interesting lemma.

Lemma 8.7 *Let A be an $n \times n$ matrix having distinct eigenvalues $\{\mu_1, \cdots, \mu_r\}$. Then letting X_i denote the generalized eigenspace corresponding to μ_i,*

$$X_i \subseteq X_i^k$$

where

$$X_i^k \equiv \left\{\mathbf{x} : \left(A^k - \mu_i^k I\right)^m \mathbf{x} = \mathbf{0} \text{ for some } m \in \mathbb{N}\right\}$$

Proof: Let $\mathbf{x} \in X_i$ so that $(A - \mu_i I)^m \mathbf{x} = \mathbf{0}$ for some positive integer, m. Then multiplying both sides by

$$\left(A^{k-1} + \mu_i A \cdots + \mu_i^{k-2}A + \mu_i^{k-1}I\right)^m,$$

We obtain $\left(A^k - \mu_i^k I\right)^m \mathbf{x} = \mathbf{0}$ showing that $\mathbf{x} \in X_i^k$ as claimed.

Corollary 8.8 *Suppose A is a regular Markov matrix. Then the conclusions of Theorem 8.6 holds.*

Proof: In the proof of Theorem 8.6 the only thing needed was that A was similar to an upper triangular, block diagonal matrix of the form described in Lemma 8.4. We prove this corollary by showing that A is similar to such a matrix. We know from the assumption that A is regular, that some power of A, say A^k is similar to a matrix of this form, having a one in the upper left position and having the diagonal blocks of the form described in Lemma 8.4 where the diagonal entries on these blocks have absolute value less than one. Now we observe that if A and B are two Markov matrices such that the entries of A are all positive, then AB is also a Markov matrix having all positive entries. Thus A^{k+r} is a Markov matrix having all positive entries for every $r \in \mathbb{N}$. Therefore, we know each of these Markov matrices has 1 as an eigenvalue and that the generalized eigenspace associated with 1 is of dimension 1. By Lemma 8.3, we know 1 is an eigenvalue for A. By Lemma 8.7 and Lemma 8.3 we see the generalized eigenspace for 1 is of dimension 1. If μ is an eigenvalue of A, then it is clear that μ^{k+r} is an eigenvalue for A^{k+r} and since these are all Markov matrices having all positive entries, Lemma 8.2 implies that for all $r \in \mathbb{N}$, either $\mu^{k+r} = 1$ or $\left|\mu^{k+r}\right| < 1$. Therefore, since r is arbitrary, we have that either $\mu = 1$ or in the case that $\left|\mu^{k+r}\right| < 1$, we have $|\mu| < 1$. Therefore, A is similar to an upper triangular matrix described in Lemma 8.4 and this proves the corollary.

A good source for more on general Stochastic processes including the Markov processes is [7]. This book gives a very different proof of the Perron Frobenius theorem which is less dependent on algebra.

Definition 8.9 *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, we would obtain the population of location i one year later by computing $\sum_j a_{ij}x_j = (A\mathbf{v})_i$. Therefore, we obtain the population of location i after k years by $\left(A^k\mathbf{v}\right)_i$. Furthermore, we can predict, using Corollary 8.8 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 ith 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 we say that the total population at the beginning is 100,000, then we would need to solve

$$\begin{gathered} y = x \\ z = 4x \\ x + y + z = 150000 \end{gathered}$$

which is easily seen to be $\{x = 25\,000, z = 100\,000, y = 25\,000\}$. Thus, after a long time we would have about four times as many people in the third location as in either of the other two.

Remark 8.10 *Sometimes the Markov matrix is not regular. Let there be n points, x_i, and suppose we are concerned with a process in which we move from one point to another. Let a_{ij} be the probability that a point moves from x_j to x_i. Thus, since every application of the process results in the point being at some point, x_i, we should have $\sum_i a_{ij} = 1$ and so $A = (a_{ij})$ is a Markov matrix. If we know that p_i is the probability that the point is at the point, x_i, then the rules of probability require that the probability the point ends up at point x_i would be $\sum_j a_{ij} p_j = (A\mathbf{p})_i$. Therefore, the probability the point is at x_i after k applications of the process would be $\left(A^k \mathbf{p}\right)_i$.*

As an example, let x_1, x_2, x_3, x_4 be four points taken in order on $\mathbb{R}$ and suppose that once a point has reached either x_1 or x_4, the point remains forever at this point. We say these points are absorbing. Otherwise, it can move either to the left or to the right with probability $\frac{1}{2}$. Therefore, the Markov matrix associated with this situation would be

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

Now this is an example of a Markov matrix which is not regular. You can verify this by noticing that the eigenspace for the eigenvalue 1 is of dimension 2 contrary to what occurs for a regular Markov matrix. Suppose the point starts off at x_2. Then we know $p_2 = 1$ and $p_i = 0$ if $i \neq 2$. Then consider the probabilities the point is at x_i after k applications of the process. By the above discussion, we obtain this as the column vector $A^k \mathbf{p}$ where $\mathbf{p} = (0, 1, 0, 0)^T$. Lets consider this for 25 applications of the process.

$$\begin{pmatrix} 1 & .5 & 0 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & .5 & 1 \end{pmatrix}^{25} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} .666\,666\,657 \\ 0 \\ 2.980\,232\,24 \times 10^{-8} \\ .333\,333\,313 \end{pmatrix}.$$

We see that after 25 applications of the process, it is almost certain that the point is not at either x_2 or x_3 and so it has been absorbed by either x_1 or x_4. We also see it is about twice as likely to have been absorbed by x_1 as by x_2. This is indicative of the fact that it started off with its distance to x_1 half its distance to x_2. This is an example of a random walk.

8.1 Exercises

1. Suppose $X_i, i = 1, \cdots, r$ is a subspace of a vector space, X and $X_i \cap X_j = \{\mathbf{0}\}$ whenever $i \neq j$. Then show that the dimension of $X_1 \oplus \cdots \oplus X_r$ equals the sum of the dimensions of the spaces X_i. Also show that if $x \in X_1 \oplus \cdots \oplus X_r$ with $x = \sum_i x_i$ for $x_i \in X_i$, the vectors, x_i are unique.

2. In obtaining the block diagonal upper triangular matrix, T, which is similar to A, we used orthonormal bases for the generalized eigenspaces, X_i. Find a condition on these subspaces which will ensure that the union of these bases is an orthonormal basis for X.

3. Suppose $B \in \mathcal{L}(X,X)$ where X is a finite dimensional vector space and $B^m = 0$ for some m a positive integer. Letting $v \in X$, consider the string of vectors, $v, Bv, B^2v, \cdots, B^kv$. Show this string of vectors is a linearly independent set if and only if $B^kv \neq 0$.

4. 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.

5. Let $B \in \mathcal{L}(X,X)$ where X is a vector space of dimension n. Show directly, using arguments of dimension that B satisfies some polynomial of degree n^2. **Hint:** Consider the transformations, $I, B, B^2, \cdots, B^{n^2}$. There are a few too many of these for them to be linearly independent.

6. For $A \in \mathcal{L}(V,V)$ and V an inner product space, we can define an inner product of elements of $\mathcal{L}(V,V)$ as follows.

$$(A,B) \equiv trace(AB^*)$$

   Show this is actually an inner product. The norm associated with this inner product is known as the Frobenius norm. **Hint:** First recall that the trace of a linear transformation is the sum of the eigenvalues. Then use the theorems about self adjoint linear transformations.

7. We saw that every normal transformation, A, could be written as

$$A = \sum_i \lambda_i w_i \otimes w_i.$$

   for some orthonormal basis, $\{w_i\}$. In particular this holds for every self adjoint transformation, $A = A^*$. We say a self adjoint matrix is positive if the eigenvalues are all nonnegative. If A is a self adjoint positive linear transformation mapping X to X, show that

$$B = \sum_i \lambda_i^{1/2} w_i \otimes w_i$$

   deserves to be known as the square root of the operator, A. Also define

$$A^+ \equiv \sum_i \lambda_i^+ w_i \otimes w_i$$

   where $\lambda^+ \equiv \frac{\lambda + |\lambda|}{2}$ is the positive part of λ. Show that just as in the scalar case, $|A^+ - B^+| \leq |A - B|$ where $|\cdot|$ is the Frobenius norm defined in Problem 6. Are there other functions of A which can be defined in this way? Give some examples.

8. Give an easy proof of the Cayley Hamilton theorem for normal transformations using the fact that any normal transformation is similar to a diagonal matrix.

Self Adjoint Operators

The following theorem is about the eigenvectors and eigenvalues of a self adjoint operator. The proof given generalizes to the situation of a compact self adjoint operator on a Hilbert space and leads to many very useful results. It is also a very elementary proof because it does not use the fundamental theorem of algebra and it contains a way, very important in applications, of finding the eigenvalues. This proof depends more directly on the methods of analysis than the preceding material. We will use the following notation.

Definition 9.1 *Let X be an inner product space and let $S \subseteq X$. Then*

$$S^{\perp} \equiv \{x \in X : (x,s) = 0 \text{ for all } s \in S\}.$$

Note that even if S is not a subspace, $S^{\perp}$ is.

Definition 9.2 *A Hilbert space is a complete inner product space. Recall this means that every Cauchy sequence,$\{x_n\}$, one which satisfies*

$$\lim_{n,m\to\infty} |x_n - x_m| = 0,$$

converges. It can be shown, although we will not do so here, that for the field of scalars either $\mathbb{R}$ or $\mathbb{C}$, any finite dimensional inner product space is automatically complete.

Theorem 9.3 *Let $A \in \mathcal{L}(X,X)$ be self adjoint where X is a finite dimensional Hilbert space. Thus $A = A^*$. Then there exists an orthonormal basis of eigenvectors, $\{u_j\}_{j=1}^n$.*

Proof: Consider (Ax,x). This quantity is always a real number because

$$\overline{(Ax,x)} = (x,Ax) = (x,A^*x) = (Ax,x)$$

thanks to the assumption that A is self adjoint. Now define

$$\lambda_1 \equiv \inf\{(Ax,x) : |x| = 1, x \in X_1 \equiv X\}.$$

Claim: λ_1 is finite and there exists $v_1 \in X$ with $|v_1| = 1$ such that $(Av_1, v_1) = \lambda_1$.

Proof of claim: Let $\{u_j\}_{j=1}^n$ be an orthonormal basis for X and for $x \in X$, let $(x_1, \cdots, x_n)$ be defined as the components of the vector x. Thus,

$$x = \sum_{j=1}^{n} x_j u_j.$$

Since this is an orthonormal basis, it follows from the axioms of the inner product that

$$|x|^2 = \sum_{j=1}^{n} |x_j|^2.$$

Thus

$$(Ax,x)=\left(\sum_{k=1}^{n}x_kAu_k,\sum_{j=1}^{n}x_ju_j\right)=\sum_{k,j}x_k\overline{x_j}\left(Au_k,u_j\right),$$

a continuous function of $(x_1,\cdots,x_n)$. Thus this function achieves its minimum on the closed and bounded subset of $\mathbb{F}^n$ given by

$$\{(x_1,\cdots,x_n)\in\mathbb{F}^n:\sum_{j=1}^{n}|x_j|^2=1\}.$$

Then $v_1\equiv\sum_{j=1}^{n}x_ju_j$ where $(x_1,\cdots,x_n)$ is the point of $\mathbb{F}^n$ at which the above function achieves its minimum. This proves the claim.

Continuing with the proof of the theorem, let $X_2\equiv\{v_1\}^{\perp}$ and let

$$\lambda_2\equiv\inf\{(Ax,x):|x|=1,x\in X_2\}$$

As before, there exists $v_2\in X_2$ such that $(Av_2,v_2)=\lambda_2$. Now let $X_3\equiv\{v_1,v_2\}^{\perp}$ and continue in this way. This leads to an increasing sequence of real numbers, $\{\lambda_k\}_{k=1}^{n}$ and an orthonormal set of vectors, $\{v_1,\cdots,v_n\}$. It only remains to show these are eigenvectors and that the λ_j are eigenvalues.

Consider the first of these vectors. Letting $w\in X_1\equiv X$, the function of the real variable, t, given by

$$f(t)\equiv\frac{\left(A\left(v_1+tw\right),v_1+tw\right)}{|v_1+tw|^2}$$

$$=\frac{(Av_1,v_1)+2t\mathrm{Re}\left(Av_1,w\right)+t^2\left(Aw,w\right)}{|v_1|^2+2t\mathrm{Re}\left(v_1,w\right)+t^2|w|^2}$$

achieves its minimum when $t=0$. Therefore, the derivative of this function evaluated at $t=0$ must equal zero. Using the quotient rule, this implies

$$2\mathrm{Re}\left(Av_1,w\right)-2\mathrm{Re}\left(v_1,w\right)\left(Av_1,v_1\right)$$

$$=2\left(\mathrm{Re}\left(Av_1,w\right)-\mathrm{Re}\left(v_1,w\right)\lambda_1\right)=0.$$

Thus $\mathrm{Re}\left(Av_1-\lambda_1v_1,w\right)=0$ for all $w\in X$. This implies $Av_1=\lambda_1v_1$. To see this, let $w\in X$ be arbitrary and let θ be a complex number with $|\theta|=1$ and

$$|(Av_1-\lambda_1v_1,w)|=\theta\left(Av_1-\lambda_1v_1,w\right).$$

Then

$$|(Av_1-\lambda_1v_1,w)|=\mathrm{Re}\left(Av_1-\lambda_1v_1,\overline{\theta}w\right)=0.$$

Since this holds for all w, $Av_1=\lambda_1v_1$. Now suppose $Av_k=\lambda_kv_k$ for all $k<m$. We observe that $A:X_m\to X_m$ because if $y\in X_m$ and $k<m$,

$$(Ay,v_k)=(y,Av_k)=(y,\lambda_kv_k)=0,$$

showing that $Ay\in\{v_1,\cdots,v_{m-1}\}^{\perp}\equiv X_m$. Thus the same argument just given shows that for all $w\in X_m$,

$$(Av_m-\lambda_mv_m,w)=0. \tag{9.1}$$

For arbitrary $w \in X$.

$$w = \left(w - \sum_{k=1}^{m-1} (w, v_k)\, v_k \right) + \sum_{k=1}^{m-1} (w, v_k)\, v_k \equiv w_\perp + w_m$$

and the term in parenthesis is in $\{v_1, \cdots, v_{m-1}\}^\perp \equiv X_m$ while the other term is contained in the span of the vectors, $\{v_1, \cdots, v_{m-1}\}$. Thus by 9.1,

$$\begin{aligned}(Av_m - \lambda_m v_m, w) &= (Av_m - \lambda_m v_m, w_\perp + w_m) \\ &= (Av_m - \lambda_m v_m, w_m) = 0\end{aligned}$$

because

$$A : X_m \to X_m \equiv \{v_1, \cdots, v_{m-1}\}^\perp$$

and $w_m \in \operatorname{span}(v_1, \cdots, v_{m-1})$. Therefore, $Av_m = \lambda_m v_m$ for all m. This proves the theorem.

Contained in the proof of this theorem is the following important corollary.

Corollary 9.4 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then all the eigenvalues are real and for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore, we have

$$\lambda_k \equiv \inf \{(Ax, x) : |x| = 1, x \in X_k\}$$

where

$$X_k \equiv \{u_1, \cdots, u_{k-1}\}^\perp, X_1 \equiv X.$$

Corollary 9.5 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then the largest eigenvalue of A is given by*

$$\max \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\} \tag{9.2}$$

and the minimum eigenvalue of A is given by

$$\min \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\}. \tag{9.3}$$

Proof: The proof of this is just like the proof of Theorem 9.3. Simply replace inf with sup and obtain a decreasing list of eigenvalues. This establishes 9.2. The claim 9.3 follows from Theorem 9.3.

Another important observation is found in the following corollary.

Corollary 9.6 *Let $A \in \mathcal{L}(X, X)$ where A is self adjoint. Then $A = \sum_i \lambda_i v_i \otimes v_i$ where $Av_i = \lambda_i v_i$ and $\{v_i\}_{i=1}^n$ is an orthonormal basis.*

Proof : If v_k is one of the orthonormal basis vectors, we have $Av_k = \lambda_k v_k$. Also,

$$\begin{aligned}\sum_i \lambda_i v_i \otimes v_i (v_k) &= \sum_i \lambda_i v_i (v_k, v_i) \\ &= \sum_i \lambda_i \delta_{ik} v_i = \lambda_k v_k.\end{aligned}$$

Since the two linear transformations agree on a basis, it follows they must coincide. This proves the corollary.

We can also establish the result of Courant and Fischer which resembles Corollary 9.4 but which is more useful because it does not depend on a knowledge of the eigenvectors.

Theorem 9.7 *Let $A \in \mathcal{L}(X,X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore, we have

$$\lambda_k \equiv \max_{w_1,\cdots,w_{k-1}} \left\{ \min \left\{ (Ax,x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \tag{9.4}$$

where if $k = 1, \{w_1, \cdots, w_{k-1}\}^{\perp} \equiv X$.

Proof: We know from Theorem 9.3 the existence of the eigenvalues and eigenvectors with $\{u_1, \cdots, u_n\}$ orthonormal and $\lambda_i \leq \lambda_{i+1}$. Therefore, by Corollary 9.6

$$A = \sum_{j=1}^{n} \lambda_j u_j \otimes u_j$$

Fix $\{w_1, \cdots, w_{k-1}\}$.

$$\begin{aligned} (Ax,x) &= \sum_{j=1}^{n} \lambda_j (x,u_j)(u_j,x) \\ &= \sum_{j=1}^{n} \lambda_j |(x,u_j)|^2 \end{aligned}$$

Then let $Y = \{w_1, \cdots, w_{k-1}\}^{\perp}$

$$\inf \{(Ax,x) : |x| = 1, x \in Y\}$$

$$= \inf \left\{ \sum_{j=1}^{n} \lambda_j |(x,u_j)|^2 : |x| = 1, x \in Y \right\}$$

$$\leq \inf \left\{ \sum_{j=1}^{k} \lambda_j |(x,u_j)|^2 : |x| = 1, (x,u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\}. \tag{9.5}$$

The reason this is so is that we have taken the infimum over a smaller set. Therefore, the infimum gets larger. Now 9.5 is no larger than

$$\inf \left\{ \lambda_k \sum_{j=1}^{k} |(x,u_j)|^2 : |x| = 1, (x,u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\} = \lambda_k$$

because since $\{u_1, \cdots, u_n\}$ is an orthonormal basis, $|x|^2 = \sum_{j=1}^{n} |(x,u_j)|^2$. It follows since $\{w_1, \cdots, w_{k-1}\}$ is arbitrary, that we have

$$\sup_{w_1,\cdots,w_{k-1}} \left\{ \inf \left\{ (Ax,x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \leq \lambda_k. \tag{9.6}$$

However, we know that for each $w_1, \cdots, w_{k-1}$, the infimum is achieved so we can replace the inf in the above with min. In addition to this, we know from Corollary 9.4 that there exists a set, $\{w_1, \cdots, w_{k-1}\}$ for which

$$\inf\left\{(Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp}\right\} = \lambda_k.$$

In fact, we pick $\{w_1, \cdots, w_{k-1}\} = \{u_1, \cdots, u_{k-1}\}$. Therefore, the sup in 9.6 is achieved and equals λ_k and so we may write 9.4. This proves the theorem.

The following corollary is immediate.

Corollary 9.8 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore, we have

$$\lambda_k \equiv \max_{w_1, \cdots, w_{k-1}} \left\{\min\left\{\frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{k-1}\}^{\perp}\right\}\right\} \tag{9.7}$$

where if $k = 1, \{w_1, \cdots, w_{k-1}\}^{\perp} \equiv X$.

Also, we can give a version of this for which we reverse the roles of max and min.

Corollary 9.9 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore, we have

$$\lambda_k \equiv \min_{w_1, \cdots, w_{n-k}} \left\{\max\left\{\frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{n-k}\}^{\perp}\right\}\right\} \tag{9.8}$$

where if $k = n, \{w_1, \cdots, w_{n-k}\}^{\perp} \equiv X$.

9.1 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. We are mainly interested in the case of a linear transformation which is an $n \times n$ matrix but we state and prove things using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

Lemma 9.10 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X, X)$. Then if $\{v_1, \cdots, v_n\}$ is an orthonormal basis for X and $M(A)$ denotes the matrix of the linear transformation, A then $M(A^*) = M(A)^*$. In particular, A is self adjoint, if and only if $M(A)$ is.*

Proof: Consider the following picture

$$\begin{array}{ccc} & A & \\ X & \rightarrow & X \\ q \uparrow & \circ & \uparrow q \\ \mathbb{F}^n & \rightarrow & \mathbb{F}^n \\ & M(A) & \end{array}$$

where q is the coordinate map which satisfies $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Therefore, since $\{v_1, \cdots, v_n\}$ is orthonormal, it is clear that $|\mathbf{x}| = |q(\mathbf{x})|$. Therefore,

$$\begin{aligned} |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathrm{Re}(\mathbf{x}, \mathbf{y}) &= |\mathbf{x}+\mathbf{y}|^2 = |q(\mathbf{x}+\mathbf{y})|^2 \\ &= |q(\mathbf{x})|^2 + |q(\mathbf{y})|^2 + 2\mathrm{Re}(q(\mathbf{x}), q(\mathbf{y})) \end{aligned} \tag{9.9}$$

Now in any inner product space,

$$(x, iy) = \mathrm{Re}(x, iy) + i\mathrm{Im}(x, iy).$$

Also

$$(x, iy) = (-i)(x, y) = (-i)\,\mathrm{Re}(x, y) + \mathrm{Im}(x, y).$$

Therefore, equating the real parts we see that $\mathrm{Im}(x, y) = \mathrm{Re}(x, iy)$ and so

$$(x, y) = \mathrm{Re}(x, y) + i\mathrm{Re}(x, iy) \tag{9.10}$$

Now from 9.9 we see that, since q preserves distances, $.\mathrm{Re}(q(\mathbf{x}), q(\mathbf{y})) = \mathrm{Re}(\mathbf{x}, \mathbf{y})$ which implies from 9.10 that

$$(\mathbf{x}, \mathbf{y}) = (q(\mathbf{x}), q(\mathbf{y})). \tag{9.11}$$

Now consulting the diagram which gives the meaning for the matrix of a linear transformation, we observe that $q \circ M(A) = A \circ q$ and $q \circ M(A^*) = A^* \circ q$. Therefore, from 9.11

$$(A(q(\mathbf{x})), q(\mathbf{y})) = (q(\mathbf{x}), A^* q(\mathbf{y})) = (q(\mathbf{x}), q(M(A^*)(\mathbf{y}))) = (\mathbf{x}, M(A^*)(\mathbf{y}))$$

but also

$$(A(q(\mathbf{x})), q(\mathbf{y})) = (q(M(A)(\mathbf{x})), q(\mathbf{y})) = (M(A)(\mathbf{x}), \mathbf{y}) = \left(\mathbf{x}, M(A)^*(\mathbf{y})\right).$$

Since $\mathbf{x}, \mathbf{y}$ are arbitrary, this shows that $M(A^*) = M(A)^*$ as claimed. Therefore, if A is self adjoint, $M(A) = M(A^*) = M(A)^*$ and so $M(A)$ is also self adjoint. If $M(A) = M(A)^*$ then $M(A) = M(A^*)$ and so $A = A^*$. This proves the lemma.

We state as a corollary one of the items which was shown in the proof.

Corollary 9.11 *Let X be a finite dimensional Hilbert space and let $\{v_1, \cdots, v_n\}$ be an orthonormal basis for X. Also, let q be the coordinate map associated with this basis satisfying $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Then $(\mathbf{x}, \mathbf{y})_{\mathbb{F}^n} = (q(\mathbf{x}), q(\mathbf{y}))_X$. Also, if $A \in \mathcal{L}(X, X)$, and $M(A)$ is the matrix of A with respect to this basis, we have*

$$(Aq(\mathbf{x}), q(\mathbf{y}))_X = (M(A)\mathbf{x}, \mathbf{y})_{\mathbb{F}^n}.$$

Definition 9.12 *We say a self adjoint $A \in \mathcal{L}(X, X)$, is positive definite if whenever $\mathbf{x} \neq \mathbf{0}$, we have $(A\mathbf{x}, \mathbf{x}) > 0$ and we say A is negative definite if for all $\mathbf{x} \neq \mathbf{0}$, we have $(A\mathbf{x}, \mathbf{x}) < 0$. We say A is positive semidefinite or just nonnegative for short if for all $\mathbf{x}$, we have $(A\mathbf{x}, \mathbf{x}) \geq 0$. We say A is negative semidefinite or nonpositive for short if for all $\mathbf{x}$, we have $(A\mathbf{x}, \mathbf{x}) \leq 0$.*

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

Lemma 9.13 *Let X be a finite dimensional Hilbert space. A self adjoint $A \in \mathcal{L}(X, X)$ is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.*

Proof: Suppose first that A is positive definite and let λ be an eigenvalue. Then for $\mathbf{x}$ an eigenvector corresponding to λ, we have, $\lambda(\mathbf{x},\mathbf{x}) = (\lambda\mathbf{x},\mathbf{x}) = (A\mathbf{x},\mathbf{x}) > 0$. Therefore, $\lambda > 0$ as claimed.

Now suppose all the eigenvalues of A are positive. We know from Theorem 9.3 and Corollary 9.6 that $A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i$ where the λ_i are the positive eigenvalues and $\{\mathbf{u}_i\}$ are an orthonormal set of eigenvectors. Therefore, letting $\mathbf{x} \neq \mathbf{0}$, we can write $(A\mathbf{x},\mathbf{x}) = ((\sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i)\mathbf{x},\mathbf{x}) = (\sum_{i=1}^{n} \lambda_i (\mathbf{x},\mathbf{u}_i)(\mathbf{u}_i,\mathbf{x})) = \sum_{i=1}^{n} \lambda_i \left|(\mathbf{u}_i,\mathbf{x})^2\right| > 0$ because, since $\{\mathbf{u}_i\}$ is an orthonormal basis, $|\mathbf{x}|^2 = \sum_{i=1}^{n} \left|(\mathbf{u}_i,\mathbf{x})^2\right|$.

To establish the claim about negative definite, it suffices to note that A is negative definite if and only if $-A$ is positive definite and the eigenvalues of A are (-1) times the eigenvalues of $-A$. The claims about positive semidefinite and negative semidefinite are obtained similarly. This proves the lemma.

The next theorem is about a way to recognize whether a self adjoint $A \in \mathcal{L}(X,X)$ is positive or negative definite without having to find the eigenvalues. In order to state this theorem, we must first give some notation.

Definition 9.14 *Let A be an $n \times n$ matrix. We 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.*

With this definition, we can now state the following theorem, the proof found in [2]

Theorem 9.15 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X,X)$ be self adjoint. Then A is positive definite if and only if $\det(M(A)_k) > 0$ for every $k = 1,\cdots,n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: We prove this theorem by induction on n. It is clearly true if $n = 1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\det(M(A)) > 0$, it follows that all the eigenvalues are nonzero. We need to show they are all positive. Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\det(M(A))$. Pick two, λ_1 and λ_2 and let $M(A)\mathbf{u}_i = \lambda_i \mathbf{u}_i$ where $\mathbf{u}_i \neq \mathbf{0}$ for $i = 1,2$ and $(\mathbf{u}_1,\mathbf{u}_2) = 0$. Now if $\mathbf{y} \equiv \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$ is an element of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$, then since these are eigenvalues and $(\mathbf{u}_1,\mathbf{u}_2) = 0$, a short computation shows

$$(M(A)(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2), \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2)$$

$$= |\alpha_1|^2 \lambda_1 |\mathbf{u}_1|^2 + |\alpha_2|^2 \lambda_2 |\mathbf{u}_2|^2 < 0.$$

Now letting $\mathbf{x} \in \mathbb{C}^{n-1}$, we can use the induction hypothesis to write

$$(\mathbf{x}^*, 0) M(A) \begin{pmatrix} \mathbf{x} \\ 0 \end{pmatrix} = \mathbf{x}^* M(A)_{n-1} \mathbf{x} = (M(A)\mathbf{x},\mathbf{x}) > 0.$$

Now the dimension of $\{\mathbf{z} \in \mathbb{C}^n : z_n = 0\}$ is $n-1$ and the dimension of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2) = 2$ and so there must be some nonzero $\mathbf{x} \in \mathbb{C}^n$ which is in both of these subspaces of $\mathbb{C}^n$. However, the first computation would require that $(M(A)\mathbf{x},\mathbf{x}) < 0$ while the second would require that $(M(A)\mathbf{x},\mathbf{x}) > 0$. This contradiction shows that all the eigenvalues must be positive. This proves the if part of the theorem. The only if part is left to the reader.

Corollary 9.16 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X,X)$ be self adjoint. Then A is negative definite if and only if $\det(M(A)_k)(-1)^k > 0$ for every $k = 1,\cdots,n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: This is immediate from the above theorem when we notice, as in the proof of Lemma 9.13 that A is negative definite if and only if $-A$ is positive definite. Therefore, if $\det(-M(A)_k) > 0$ for all $k = 1,\cdots,n$, it follows that A is negative definite. However, $\det(-M(A)_k) = (-1)^k \det(M(A)_k)$. This proves the corollary.

9.2 Fractional powers

With the above theory, it is possible to take fractional powers of certain elements of $\mathcal{L}(X,X)$ where X is a finite dimensional Hilbert space. The main result is the following theorem.

Theorem 9.17 *Let $A \in \mathcal{L}(X,X)$ be self adjoint and nonnegative and let k be a positive integer. Then there exists a unique self adjoint nonnegative $B \in \mathcal{L}(X,X)$ such that $B^k = A$.*

Proof: We know by Theorem 9.3 there exists an orthonormal basis of eigenvectors of A, say $\{v_i\}_{i=1}^n$ such that $Av_i = \lambda_i v_i$. Therefore, by Corollary 9.6, $A = \sum_i \lambda_i v_i \otimes v_i$.

Now by Lemma 9.13, each $\lambda_i \geq 0$. Therefore, it makes sense to define

$$B \equiv \sum_i \lambda_i^{1/k} v_i \otimes v_i.$$

It is easy to verify that

$$(v_i \otimes v_i)(v_j \otimes v_j) = \begin{cases} 0 \text{ if } i \neq j \\ v_i \otimes v_i \text{ if } i = j \end{cases}.$$

Therefore, a short computation verifies that $B^k = \sum_i \lambda_i v_i \otimes v_i = A$. This proves existence.

In order to prove uniqueness, let $p(t)$ be a polynomial which has the property that $p(\lambda_i) = \lambda_i^{1/k}$. Then a similar short computation shows

$$p(A) = \sum_i p(\lambda_i) v_i \otimes v_i = \sum_i \lambda_i^{1/k} v_i \otimes v_i = B.$$

Now suppose $C^k = A$ where $C \in \mathcal{L}(X,X)$ is self adjoint and nonnegative. We have

$$CB = Cp(A) = Cp\left(C^k\right) = p\left(C^k\right) C = BC.$$

Therefore, $\{B, C\}$ is a commuting family of linear transformations which are both self adjoint. Letting $M(B)$ and $M(C)$ denote matrices of these linear transformations taken with respect to some fixed orthonormal basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, it follows that $M(B)$ and $M(C)$ commute and that both can be diagonalized (Lemma 9.10). See the diagram for a short verification of the claim the two matrices commute..

$$\begin{array}{ccccc} & B & & C & \\ X & \rightarrow & X & \rightarrow & X \\ q \uparrow & \circ & \uparrow q & \circ & \uparrow q \\ \mathbb{F}^n & \rightarrow & \mathbb{F}^n & \rightarrow & \mathbb{F}^n \\ & M(B) & & M(C) & \end{array}$$

Therefore, by Theorem 7.7, these two matrices can be simultaneously diagonalized. Thus

$$U^{-1} M(B) U = D_1,\ U^{-1} M(C) U = D_2$$

where the D_i is a diagonal matrix consisting of the eigenvalues of B. Then raising these to powers, we have

$$U^{-1} M(A) U = U^{-1} M(B)^k U = D_1^k$$

and

$$U^{-1} M(A) U = U^{-1} M(C)^k U = D_2^k.$$

Therefore, $D_1^k = D_2^k$ and since the diagonal entries of D_i are nonnegative, this requires that $D_1 = D_2$. Therefore, $M(B) = M(C)$ and so $B = C$. This proves the theorem.

9.3 Polar decompositions

As an application of Theorem 9.3, we give 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 [3]. Don't confuse the U in this theorem with a unitary transformation. It is not so. When the following theorem is applied in continuum mechanics, F is normally the deformation gradient, the derivative of a nonlinear map from some subset of three dimensional space to three dimensional space. In this context, U is called the right Cauchy Green strain tensor. It is a measure of how a body is stretched independent of rigid motions.

Theorem 9.18 *Let X be a Hilbert space of dimension n and let Y be a Hilbert space of dimension $m \geq n$ and let $F \in \mathcal{L}(X,Y)$. Then there exists $R \in \mathcal{L}(X,Y)$ and $U \in \mathcal{L}(X,X)$ such that*

$$F = RU,\ U = U^*,$$

all eigenvalues of U are non negative,

$$U^2 = F^*F, R^*R = I,$$

and $|R\mathbf{x}| = |\mathbf{x}|$.

Proof: $(F^*F)^* = F^*F$ and so by Theorem 9.3, there is an orthonormal basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$F^*F\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

It is also clear that $\lambda_i \geq 0$ because

$$\lambda_i (\mathbf{v}_i, \mathbf{v}_i) = (F^*F\mathbf{v}_i, \mathbf{v}_i) = (F\mathbf{v}_i, F\mathbf{v}_i) \geq 0.$$

Let

$$U \equiv \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{v}_i \otimes \mathbf{v}_i.$$

Then $U^2 = F^*F$, $U = U^*$, and the eigenvalues of U, $\left\{\lambda_i^{1/2}\right\}_{i=1}^{n}$ are all non negative.

Now R is defined on $U(X)$ by

$$RU\mathbf{x} \equiv F\mathbf{x}.$$

This is well defined because if $U\mathbf{x}_1 = U\mathbf{x}_2$, then $U^2(\mathbf{x}_1 - \mathbf{x}_2) = 0$ and so

$$0 = (F^*F(\mathbf{x}_1 - \mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2) = |F(\mathbf{x}_1 - \mathbf{x}_2)|^2.$$

Now $|RU\mathbf{x}|^2 = |U\mathbf{x}|^2$ because

$$|RU\mathbf{x}|^2 = |F\mathbf{x}|^2 = (F\mathbf{x},F\mathbf{x})$$

$$= (F^*F\mathbf{x}, \mathbf{x}) = \left(U^2\mathbf{x}, \mathbf{x}\right) = (U\mathbf{x},U\mathbf{x}) = |U\mathbf{x}|^2.$$

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be an orthonormal basis for

$$U(X)^{\perp} \equiv \{\mathbf{x} \in X : (\mathbf{x}, \mathbf{z}) = 0 \text{ for all } \mathbf{z} \in U(X)\}$$

and let $\{\mathbf{y}_1, \cdots, \mathbf{y}_p\}$ be an orthonormal basis for $F(X)^\perp$. Then $p \geq r$ because if $\{F(\mathbf{z}_i)\}_{i=1}^s$ is an orthonormal basis for $F(X)$, it follows that $\{U(\mathbf{z}_i)\}_{i=1}^s$ is orthonormal in $U(X)$ because

$$(U\mathbf{z}_i, U\mathbf{z}_j) = \left(U^2\mathbf{z}_i, \mathbf{z}_j\right) = (F^*F\mathbf{z}_i, \mathbf{z}_j) = (F\mathbf{z}_i, F\mathbf{z}_j).$$

Therefore,

$$p + s = m \geq n = r + \dim U(X) \geq r + s.$$

Now define $R \in \mathcal{L}(X, Y)$ by $R\mathbf{x}_i \equiv \mathbf{y}_i$, $i = 1, \cdots, r$. Note that R is already defined on $U(X)$. We have only extended this by telling what R does to a basis for $U(X)^\perp$. Thus

$$\left|R\left(\sum_{i=1}^r c_i\mathbf{x}_i + U\mathbf{v}\right)\right|^2 = \left|\sum_{i=1}^r c_i\mathbf{y}_i + F\mathbf{v}\right|^2 = \sum_{i=1}^r |c_i|^2 + |F\mathbf{v}|^2$$

$$= \sum_{i=1}^r |c_i|^2 + |U\mathbf{v}|^2 = \left|\sum_{i=1}^r c_i\mathbf{x}_i + U\mathbf{v}\right|^2,$$

and so $|R\mathbf{z}| = |\mathbf{z}|$ which implies that for all $\mathbf{x}, \mathbf{y}$,

$$|\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathrm{Re}(\mathbf{x}, \mathbf{y}) = |\mathbf{x} + \mathbf{y}|^2$$

$$= |R(\mathbf{x} + \mathbf{y})|^2 = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathrm{Re}(R\mathbf{x}, R\mathbf{y}).$$

Therefore, as in Lemma 9.10,

$$(\mathbf{x}, \mathbf{y}) = (R\mathbf{x}, R\mathbf{y}) = (R^*R\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y}$ and so $R^*R = I$ as claimed. This proves the theorem.

The following corollary follows as a simple consequence of this theorem. It is called the left polar decomposition.

Corollary 9.19 *Let $F \in \mathcal{L}(X, Y)$ and suppose $n \geq m$ where X is a Hilbert space of dimension n and Y is a Hilbert space of dimension m. Then there exists a symmetric nonnegative element of $\mathcal{L}(X, X)$, U, and an element of $\mathcal{L}(X, Y)$, R, such that*

$$F = UR,\ RR^* = I.$$

Proof: We recall that $L^{**} = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 9.18 to $F^* \in \mathcal{L}(X, Y)$. Thus,

$$F^* = R^*U$$

where R^* and U satisfy the conditions of that theorem. Then

$$F = UR$$

and $RR^* = R^{**}R^* = I$. This proves the corollary.

As a corollary we obtain the following existence theorem for the polar decomposition of an element of $\mathcal{L}(X, X)$.

Corollary 9.20 *Let $F \in \mathcal{L}(X, X)$. Then there exists a symmetric nonnegative element of $\mathcal{L}(X, X)$, W, and a unitary matrix, Q such that $F = WQ$, and there exists a symmetric nonnegative element of $\mathcal{L}(X, X)$, U, and a unitary R, such that $F = RU$.*

This corollary has a fascinating relation to the question whether a given linear transformation is normal. Recall that an $n \times n$ matrix, A, is normal if $AA^* = A^*A$. We retain the same definition for an element of $\mathcal{L}(X,X)$.

Theorem 9.21 *Let $F \in \mathcal{L}(X,X)$. Then F is normal if and only if in Corollary 9.20 $RU = UR$ and $QW = WQ$.*

Proof: We prove the statement about $RU = UR$ and leave the other part as an exercise. First we suppose that $RU = UR$ and show F is normal. To begin with we have

$$UR^* = (RU)^* = (UR)^* = R^*U.$$

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. We need to verify $RU = UR$. Since F is normal,

$$FF^* = RUUR^* = RU^2R^*$$

and

$$F^*F = UR^*RU = U^2.$$

Therefore, $RU^2R^* = U^2$, and we see 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.

9.4 The singular value decomposition

In this section, A will be an $m \times n$ matrix. To begin with, we state a simple lemma.

Lemma 9.22 *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 9.23 *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 we can state the main theorem on the singular value decomposition.

Theorem 9.24 *Let A be an $m \times n$ matrix. Then there exist unitary matrices, U and V of the appropriate size such that*

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is of the form

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

for the σ_i the singular values of A.

Proof: By the above lemma and Theorem 9.3 there exists an orthonormal basis, $\{\mathbf{v}_i\}_{i=1}^n$ such that $A^*A\mathbf{v}_i = \sigma_i^2\mathbf{v}_i$ where $\sigma_i^2 > 0$ for $i = 1, \cdots, k$ and equals zero if $i > k$. Thus for $i > k$, we must have $A\mathbf{v}_i = \mathbf{0}$ because

$$(A\mathbf{v}_i, A\mathbf{v}_i) = (A^*A\mathbf{v}_i, \mathbf{v}_i) = (\mathbf{0}, \mathbf{v}_i) = 0.$$

For $i = 1, \cdots, k$, define $\mathbf{u}_i \in \mathbb{F}^m$ by

$$\mathbf{u}_i \equiv \sigma_i^{-1} A\mathbf{v}_i.$$

Thus $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$. Now

$$\begin{aligned}(\mathbf{u}_i, \mathbf{u}_j) &= \left(\sigma_i^{-1}A\mathbf{v}_i, \sigma_j^{-1}A\mathbf{v}_j\right) = \left(\sigma_i^{-1}\mathbf{v}_i, \sigma_j^{-1}A^*A\mathbf{v}_j\right) \\ &= \left(\sigma_i^{-1}\mathbf{v}_i, \sigma_j^{-1}\sigma_j^2\mathbf{v}_j\right) = \frac{\sigma_j}{\sigma_i}(\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij}.\end{aligned}$$

Thus $\{\mathbf{u}_i\}_{i=1}^k$ is an orthonormal set of vectors in $\mathbb{F}^m$. Also,

$$AA^*\mathbf{u}_i = AA^*\sigma_i^{-1}A\mathbf{v}_i = \sigma_i^{-1}AA^*A\mathbf{v}_i = \sigma_i^{-1}A\sigma_i^2\mathbf{v}_i = \sigma_i^2\mathbf{u}_i.$$

Now extend $\{\mathbf{u}_i\}_{i=1}^k$ to an orthonormal basis for all of $\mathbb{F}^m, \{\mathbf{u}_i\}_{i=1}^m$ and let $U \equiv (\mathbf{u}_1 \cdots \mathbf{u}_m)$ while $V \equiv (\mathbf{v}_1 \cdots \mathbf{v}_n)$. Thus U is the matrix which has the $\mathbf{u}_i$ as columns and V is defined as the matrix which has the $\mathbf{v}_i$ as columns. Then

$$\begin{aligned}U^*AV &= \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} A(\mathbf{v}_1 \cdots \mathbf{v}_n) \\ &= \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} (\sigma_1\mathbf{u}_1 \cdots \sigma_k\mathbf{u}_k\mathbf{0} \cdots \mathbf{0}) \\ &= \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\end{aligned}$$

where σ is given in the statement of the theorem.

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 9.25 *Let A be an $m \times n$ matrix. Then the rank of A and A^* equals the number of singular values.*

Proof: Since V and U are unitary, it follows that $\text{rank}(A) = \text{rank}(U^*AV) = \text{rank}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \text{number}$ of singular values. A similar argument holds for A^*.

The singular value decomposition also has a very interesting connection to the problem of least squares solutions. Recall that we wanted to find $\mathbf{x}$ such that $|A\mathbf{x} - \mathbf{y}|$ is as small as possible. We showed in Lemma 5.25 that there was a solution to this problem and that we could find it 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 we 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^*.

$$\begin{pmatrix}\sigma^2 & 0\\ 0 & 0\end{pmatrix}V^*\mathbf{x} = \begin{pmatrix}\sigma & 0\\ 0 & 0\end{pmatrix}U^*\mathbf{y}. \tag{9.12}$$

One solution to this equation which is very easy to spot is

$$\mathbf{x} = V\begin{pmatrix}\sigma^{-1} & 0\\ 0 & 0\end{pmatrix}U^*\mathbf{y}. \tag{9.13}$$

This particular solution is important enough that it motivates the following definintion.

Definition 9.26 *Let A be an $m\times n$ matrix. Then the Moore Penrose inverse of A, denoted by A^+ is defined as*

$$A^+ \equiv V\begin{pmatrix}\sigma^{-1} & 0\\ 0 & 0\end{pmatrix}U^*.$$

Thus $A^+\mathbf{y}$ is a solution to the minimization problem to find $\mathbf{x}$ which minimizes $|A\mathbf{x}-\mathbf{y}|$. In fact, one can say more about this.

Proposition 9.27 *$A^+\mathbf{y}$ is the solution to the problem of minimizing $|A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$ which has smallest norm. Thus*

$$\left|AA^+\mathbf{y}-\mathbf{y}\right| \le |A\mathbf{x}-\mathbf{y}| \text{ for all } \mathbf{x}$$

and if $\mathbf{x}_1$ satisfies $|A\mathbf{x}_1-\mathbf{y}| \le |A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$, then $|A^+\mathbf{y}| \le |\mathbf{x}_1|$.

Proof: If we look at $\mathbf{x}$ satisfying 9.12 and seek the one 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 9.12 we have

$$\begin{pmatrix}\sigma^2\left(V^*\mathbf{x}\right)_k\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\sigma\left(U^*\mathbf{y}\right)_k\\ \mathbf{0}\end{pmatrix}$$

and so we must have $\left(V^*\mathbf{x}\right)_k = \sigma^{-1}\left(U^*\mathbf{y}\right)_k$. Thus the first k entries of $V^*\mathbf{x}$ are determined. In order to make $|V^*\mathbf{x}|$ as small as possible, we should take the remaining $n-k$ entries equal to zero. Therefore, we must have

$$V^*\mathbf{x} = \begin{pmatrix}\sigma^{-1} & 0\\ 0 & 0\end{pmatrix}U^*\mathbf{y}$$

which shows that $A^+\mathbf{y} = \mathbf{x}$. This proves the proposition.

Lemma 9.28 *The matrix, A^+ satisfies the following conditions.*

$$AA^+A = A,\ A^+AA^+ = A^+,\ A^+A \text{ and } AA^+ \text{ are Hermitian.} \tag{9.14}$$

Proof: The proof is completely routine and is left to the reader.

A much more interesting observation is that A^+ is characterized as being the unique matrix which satisfies 9.14. This is the content of the following Theorem.

Theorem 9.29 *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*

$$AA_0A = A,\ A_0AA_0 = A_0,\ A_0A \text{ and } AA_0 \text{ are Hermitian.} \tag{9.15}$$

Proof: From the above lemma we know that the Moore Penrose inverse satisfies 9.15. Suppose then that A_0 satisfies 9.15. We need to verify that $A_0 = A^+$. Recall that from the singular value decomposition, there exist unitary matrices, U and V such that

$$U^*AV = \Sigma \equiv \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},\ A = U\Sigma V^*.$$

Let

$$V^*A_0U = \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \tag{9.16}$$

where P is $k \times k$.

Next we use the first equation of 9.15 to write

$$\overbrace{U\Sigma V^*}^{A}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}\overbrace{U\Sigma V^*}^{A} = \overbrace{U\Sigma V^*}^{A}.$$

Then multiplying both sides on the left by V^* and on the right by U, we obtain

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

Now this requires

$$\begin{pmatrix} \sigma P\sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \tag{9.17}$$

Therefore, we must have $P = \sigma^{-1}$. Now from the requirement that AA_0 is Hermitian, we have

$$\overbrace{U\Sigma V^*}^{A}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0} = U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*$$

must be Hermitian. Therefore, we need

$$\begin{aligned}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix} &= \begin{pmatrix} \sigma P & \sigma Q \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix}\end{aligned}$$

to be Hermitian. We must have

$$\begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ \left(\overline{Q}\right)^T\sigma & 0 \end{pmatrix}$$

which requires that $Q = 0$. From the requirement that $A_0 A$ is Hermitian, we need

$$\begin{aligned} \overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}\overbrace{U\Sigma V^*}^{A} &= V\begin{pmatrix} P\sigma & 0 \\ R\sigma & 0 \end{pmatrix}V^* \\ &= V\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}V^* \end{aligned}$$

is Hermitian. Therefore, also

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}$$

is Hermitian. Thus $R = 0$ by reasoning similar to that used to show $Q = 0$.

We may use 9.16 and the second equation of 9.15 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

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}.$$

Using that which we have shown, this yields

$$\begin{aligned} \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \quad (9.18) \\ &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}. \quad (9.19) \end{aligned}$$

Therefore, $S = 0$ also and so we have shown that

$$V^* A_0 U = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

which says

$$A_0 = V\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}U^* \equiv A^+.$$

This proves the theorem.

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in 9.15. It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 4.

9.5 Exercises

1. Show $(A^*)^* = A$ and $(AB)^* = B^* A^*$.
2. Suppose $A : X \to X$, an inner product space, and $A \geq 0$. By this we mean $(Ax, x) \geq 0$ for all $x \in X$ and $A = A^*$. Show that A has a square root, U, such that $U^2 = A$. **Hint:** Let $\{u_k\}_{k=1}^n$ be an orthonormal basis of eigenvectors with $Au_k = \lambda_k u_k$. Show each $\lambda_k \geq 0$ and consider

$$U \equiv \sum_{k=1}^{n} \lambda_k^{1/2} u_k \otimes u_k$$

3. Prove Corollary 9.9.

4. Show that if A is an $n \times n$ matrix which has an inverse then $A^+ = A^{-1}$.

5. Using the singular value decomposition, show that for any square matrix, A, it follows that A^*A is unitarily similar to AA^*.

6. Prove that Theorem 9.15 and Corollary 9.16 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 $\mathbf{x} \in \mathbb{F}^k$.

7. Show directly that if A is an $n \times n$ matrix and $A = A^*$ (A is Hermitian) then all the eigenvalues and eigenvectors are real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).

8. 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
$$Q^*Q = QQ^* = I.$$

9. Show that a matrix, Q is unitary if and only if it preserves distances. By this we mean $|Q\mathbf{v}| = |\mathbf{v}|$.

10. 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 $\mathbf{e}_1$.

11. Finish the proof of Theorem 9.21.

12. 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
$$(\mathbf{v}, \mathbf{u}) \equiv \sum_{j=1}^{n} v_j \overline{u_j}.$$

13. Let X be an inner product space. Show $|x+y|^2 + |x-y|^2 = 2|x|^2 + 2|y|^2$. This is called the parallelogram identity.

Norms for finite dimensional vector spaces

In this chapter, X and Y are finite dimensional vector spaces which have a norm. The following is a definition.

Definition 10.1 *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. We will say a set, $U \subseteq X$, a normed linear space is open if for every $p \in U$, there exists $\delta > 0$ such that

$$B(p, \delta) \equiv \{x : ||x - p|| < \delta\} \subseteq U.$$

Thus, a set is open if every point of the set is an interior point.

To begin with we give an important inequality known as the Cauchy Schwartz inequality.

Theorem 10.2 *The following inequality holds for a_i and $b_i \in \mathbb{C}$.*

$$\left| \sum_{i=1}^{n} a_i \overline{b}_i \right| \leq \left(\sum_{i=1}^{n} |a_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |b_i|^2 \right)^{1/2}. \tag{10.1}$$

Proof: Let $t \in \mathbb{R}$ and define

$$h(t) \equiv \sum_{i=1}^{n} (a_i + t b_i) \overline{(a_i + t b_i)} = \sum_{i=1}^{n} |a_i|^2 + 2t \mathrm{Re} \sum_{i=1}^{n} a_i \overline{b}_i + t^2 \sum_{i=1}^{n} |b_i|^2.$$

Now $h(t) \geq 0$ for all $t \in \mathbb{R}$. If all b_i equal 0, then the inequality 10.1 clearly holds so assume this does not happen. Then the graph of $y = h(t)$ is a parabola which opens up and intersects the t axis in at most one point. Thus there is either one real zero or none. Therefore, from the quadratic formula,

$$4 \left(\mathrm{Re} \sum_{i=1}^{n} a_i \overline{b}_i \right)^2 \leq 4 \left(\sum_{i=1}^{n} |a_i|^2 \right) \left(\sum_{i=1}^{n} |b_i|^2 \right)$$

which shows

$$\left|\mathrm{Re}\sum_{i=1}^{n} a_i\overline{b}_i\right| \leq \left(\sum_{i=1}^{n}|a_i|^2\right)^{1/2}\left(\sum_{i=1}^{n}|b_i|^2\right)^{1/2} \tag{10.2}$$

To get the desired result, let $\omega \in \mathbb{C}$ be such that $|\omega| = 1$ and

$$\sum_{i=1}^{n}\omega a_i\overline{b}_i = \omega\sum_{i=1}^{n}a_i\overline{b}_i = \left|\sum_{i=1}^{n}a_i\overline{b}_i\right|.$$

Then apply 10.2 replacing a_i with ωa_i. Then

$$\begin{aligned}\left|\sum_{i=1}^{n}a_i\overline{b}_i\right| &= \mathrm{Re}\sum_{i=1}^{n}\omega a_i\overline{b}_i \leq \left(\sum_{i=1}^{n}|\omega a_i|^2\right)^{1/2}\left(\sum_{i=1}^{n}|b_i|^2\right)^{1/2}\\ &= \left(\sum_{i=1}^{n}|a_i|^2\right)^{1/2}\left(\sum_{i=1}^{n}|b_i|^2\right)^{1/2}.\end{aligned}$$

This proves the theorem.

Definition 10.3 *We say a normed linear space, $(X, ||\cdot||)$ is a Banach space if it is complete. Thus, whenever, $\{\mathbf{x}_n\}$ is a Cauchy sequence,*

$$\lim_{m,n\to\infty}||x_n - x_m|| = 0,$$

there exists $\mathbf{x} \in X$ such that $\lim_{n\to\infty}||\mathbf{x} - \mathbf{x}_n|| = 0$.

Let X be a finite dimensional normed linear space with norm $||\cdot||$ where the field of scalars is denoted by $\mathbb{F}$ and is understood to be either $\mathbb{R}$ or $\mathbb{C}$. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X. If $\mathbf{x} \in X$, we will 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 10.4 *For $\mathbf{x} \in X$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ a basis, we define a new norm by*

$$|\mathbf{x}| \equiv \left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}.$$

Similarly, for $\mathbf{y} \in Y$ with basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$, and y_i its components with respect to this basis,

$$|\mathbf{y}| \equiv \left(\sum_{i=1}^{m}|y_i|^2\right)^{1/2}$$

For $A \in \mathcal{L}(X, Y)$, the space of linear mappings from X to Y,

$$||A|| \equiv \sup\{|A\mathbf{x}| : |\mathbf{x}| \leq 1\}. \tag{10.3}$$

The first thing we will show is that these two norms, $||\cdot||$ and $|\cdot|$, are equivalent. This means the conclusion of the following theorem holds.

Theorem 10.5 *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*

$$\delta ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta ||\mathbf{x}||. \tag{10.4}$$

Proof: All of the above properties of a norm are obvious except the second, the triangle inequality. To establish this inequality, we use the Cauchy Schwartz inequality to write

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

and this proves the second property above.

It remains to show the equivalence of the two norms. By the Cauchy Schwartz inequality again,

$$\begin{aligned} ||\mathbf{x}|| &\equiv \left|\left|\sum_{i=1}^n x_i \mathbf{v}_i\right|\right| \leq \sum_{i=1}^n |x_i| \, ||\mathbf{v}_i|| \leq |\mathbf{x}| \left(\sum_{i=1}^n ||\mathbf{v}_i||^2\right)^{1/2} \\ &\equiv \delta^{-1} |\mathbf{x}|. \end{aligned}$$

This proves the first half of the inequality.

Suppose the second half of the inequality is not valid. Then there exists a sequence $\mathbf{x}^k \in X$ such that

$$|\mathbf{x}^k| > k ||\mathbf{x}^k||, \ k = 1, 2, \cdots.$$

Then define

$$\mathbf{y}^k \equiv \frac{\mathbf{x}^k}{|\mathbf{x}^k|}.$$

It follows

$$|\mathbf{y}^k| = 1, \quad |\mathbf{y}^k| > k ||\mathbf{y}^k||. \tag{10.5}$$

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

$$\left(y_1^k, \cdots, y_n^k\right) \to (y_1, \cdots, y_n).$$

It follows from 10.5 and this that for

$$\mathbf{y} = \sum_{i=1}^n y_i \mathbf{v}_i,$$

$$0 = \lim_{k\to\infty} ||\mathbf{y}^k|| = \lim_{k\to\infty} \left|\left|\sum_{i=1}^n y_i^k \mathbf{v}_i\right|\right| = \left|\left|\sum_{i=1}^n y_i \mathbf{v}_i\right|\right|$$

but not all the y_i equal zero. This contradicts the assumption that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis and this proves the second half of the inequality.

Corollary 10.6 *If $(X, ||\cdot||)$ is a finite dimensional normed linear space with the field of scalars $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, then X is complete.*

Proof: Let $\{\mathbf{x}^k\}$ be a Cauchy sequence. Then letting the components of $\mathbf{x}^k$ with respect to the given basis be

$$x_1^k, \cdots, x_n^k,$$

it follows from Theorem 10.5, that

$$\left(x_1^k, \cdots, x_n^k\right)$$

is a Cauchy sequence in $\mathbb{F}^n$ and so

$$\left(x_1^k, \cdots, x_n^k\right) \to (x_1, \cdots, x_n) \in \mathbb{F}^n.$$

Thus,

$$\mathbf{x}^k = \sum_{i=1}^n x_i^k \mathbf{v}_i \to \sum_{i=1}^n x_i \mathbf{v}_i \in X.$$

This proves the corollary.

Corollary 10.7 *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.

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 10.5, there are positive constants $\delta_1, \Delta_1, \delta_2, \Delta_2$, all independent of $\mathbf{x} \in X$ such that

$$\delta_2 \, |||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_2 \, |||\mathbf{x}||| \, ,$$

$$\delta_1 \, ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta_1 \, ||\mathbf{x}|| \, .$$

Then

$$\delta_2 \, |||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_1 \, ||\mathbf{x}|| \leq \frac{\Delta_1}{\delta_1} \, |\mathbf{x}| \leq \frac{\Delta_1 \Delta_2}{\delta_1} \, |||\mathbf{x}|||$$

and so

$$\frac{\delta_2}{\Delta_1} \, |||\mathbf{x}||| \leq ||\mathbf{x}|| \leq \frac{\Delta_2}{\delta_1} \, |||\mathbf{x}|||$$

which proves the corollary.

Definition 10.8 *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.

We leave it as an easy exercise to verify that $||\cdot||$ is a norm on $\mathcal{L}(X,Y)$ and it is always the case that

$$||Ax||_Y \leq ||A||\,||x||_X.$$

Theorem 10.9 *Let X and Y be finite dimensional normed linear spaces of dimension n and m respectively and denote by $||\cdot||$ the norm on either X or Y. Then if A is any linear function mapping X to Y, then $A \in \mathcal{L}(X,Y)$ and $(\mathcal{L}(X,Y), ||\cdot||)$ is a complete normed linear space of dimension nm with*

$$||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}||.$$

Proof: We need to show the norm defined on linear transformations really is a norm. Again the first and third properties listed above for norms are obvious. We need to show the second and verify $||A|| < \infty$. Letting $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis and $|\cdot|$ defined with respect to this basis as above, there exist constants $\delta, \Delta > 0$ such that

$$\delta\,||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta\,||\mathbf{x}||.$$

Then,

$$\begin{aligned} ||A+B|| &\equiv \sup\{||(A+B)(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\} \\ &\leq \sup\{||A\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} + \sup\{||B\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \\ &\equiv ||A|| + ||B||. \end{aligned}$$

Next we verify 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 we verify the assertion about the dimension of $\mathcal{L}(X,Y)$. Let the two sets of bases be

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \text{ and } \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$$

for X and Y respectively. Let $\mathbf{w}_i \otimes \mathbf{v}_k \in \mathcal{L}(X,Y)$ be defined by

$$\mathbf{w}_i \otimes \mathbf{v}_k \mathbf{v}_l \equiv \begin{cases} \mathbf{0} \text{ if } l \neq k \\ \mathbf{w}_i \text{ if } l = k \end{cases}$$

and let $L \in \mathcal{L}(X,Y)$. Then

$$L\mathbf{v}_r = \sum_{j=1}^m d_{jr} \mathbf{w}_j$$

for some d_{jk}. Also

$$\sum_{j=1}^m \sum_{k=1}^n d_{jk} \mathbf{w}_j \otimes \mathbf{v}_k(\mathbf{v}_r) = \sum_{j=1}^m d_{jr} \mathbf{w}_j.$$

It follows that

$$L = \sum_{j=1}^{m}\sum_{k=1}^{n} d_{jk}\mathbf{w}_j \otimes \mathbf{v}_k$$

because the two linear transformations agree on a basis. Since L is arbitrary this shows

$$\{\mathbf{w}_i \otimes \mathbf{v}_k : i = 1, \cdots, m,\ k = 1, \cdots, n\}$$

spans $\mathcal{L}(X, Y)$. If

$$\sum_{i,k} d_{ik}\mathbf{w}_i \otimes \mathbf{v}_k = \mathbf{0},$$

then

$$\mathbf{0} = \sum_{i,k} d_{ik}\mathbf{w}_i \otimes \mathbf{v}_k(\mathbf{v}_l) = \sum_{i=1}^{m} d_{il}\mathbf{w}_i$$

and so, since $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows the dimension of $\mathcal{L}(X, Y)$ is mn as claimed. By Corollary 10.6 $(\mathcal{L}(X, Y), ||\cdot||)$ is complete. If $\mathbf{x} \neq \mathbf{0}$,

$$||A\mathbf{x}|| \frac{1}{||\mathbf{x}||} = \left|\left|A\frac{\mathbf{x}}{||\mathbf{x}||}\right|\right| \leq ||A||$$

This proves the theorem.

Note by Corollary 10.7 we can define a norm any way we want 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 we use for a finite dimensional space. In particular in the space of $m \times n$ matrices, we can use the operator norm defined above or some other way of giving this space a norm. In particular, we could use the Frobenius norm.

Definition 10.10 *We make the space of $m \times n$ matrices into a Hilbert space by defining*

$$(A, B) \equiv tr(AB^*).$$

Another way of describing a norm for an $n \times n$ matrix is as follows.

Definition 10.11 *Let A be an $n \times n$ matrix. We define the spectral norm of A, written as $||A||_2$ to be*

$$\max\left\{|\lambda|^{1/2} : \lambda \text{ is an eigenvalue of } A^*A\right\}.$$

Actually, this is nothing new. It turns out that $||\cdot||_2$ is nothing more than the operator norm for A taken with respect to the usual Euclidean norm,

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

Proposition 10.12 *The following holds.*

$$||A||_2 = \sup\{|A\mathbf{x}| : |\mathbf{x}| = 1\} \equiv ||A||.$$

Proof: We note that A^*A is Hermitian and so by Corollary 9.5,

$$\begin{aligned} ||A||_2 &= \max\left\{(A^*A\mathbf{x},\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &= \max\left\{(A\mathbf{x},A\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &\leq ||A||. \end{aligned}$$

Now to go the other direction, let $|\mathbf{x}| \leq 1$. Then

$$|A\mathbf{x}| = \left|(A\mathbf{x},A\mathbf{x})^{1/2}\right| = (A^*A\mathbf{x},\mathbf{x})^{1/2} \leq ||A||_2,$$

and so, taking the sup over all $|\mathbf{x}| \leq 1$, we obtain $||A|| \leq ||A||_2$.

An interesting application of the notion of equivalent norms on $\mathbb{R}^n$ is the process of giving a norm on a finite Cartesian product of normed linear spaces.

Definition 10.13 *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 $\mathbb{R}^n$, we 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 10.7.

Theorem 10.14 *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, we may define

$$||\mathbf{x}||_1 \equiv \sum_{i=1}^n |x_i|,$$

$$||\mathbf{x}||_\infty \equiv \max\{|x_i|, i = 1, \cdots, n\},$$

or

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$$

and all three are equivalent norms on $\prod_{i=1}^n X_i$.

In addition to $||\cdot||_1$ and $||\cdot||_\infty$ mentioned above, it is common to consider the so called p norms for $\mathbf{x} \in \mathbb{C}^n$.

Definition 10.15 *Let $\mathbf{x} \in \mathbb{C}^n$. Then we 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 10.16 *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 10.17 *If* $a, b \geq 0$ *and* p' *is defined by* $\frac{1}{p} + \frac{1}{p'} = 1$, *we have the inequality*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

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 10.17,

$$\begin{aligned} \sum_{i=1}^{n} \frac{|x_i|}{A} \frac{|y_i|}{B} &\leq \sum_{i=1}^{n} \left[\frac{1}{p} \left(\frac{|x_i|}{A} \right)^p + \frac{1}{p'} \left(\frac{|y_i|}{B} \right)^{p'} \right] \\ &= 1 \end{aligned}$$

and so

$$\sum_{i=1}^{n} |x_i| \, |y_i| \leq AB = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'}.$$

This proves the proposition.

Theorem 10.18 *The p norms do indeed satisfy the axioms of a norm.*

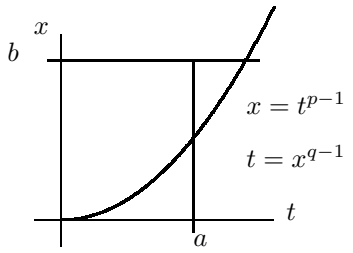
Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation we write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $\frac{p}{p'} = p - 1$. Then using the Holder inequality,

$$\begin{aligned} ||\mathbf{x} + \mathbf{y}||^p &= \sum_{i=1}^{n} |x_i + y_i|^p \\ &\leq \sum_{i=1}^{n} |x_i + y_i|^{p-1} \, |x_i| + \sum_{i=1}^{n} |x_i + y_i|^{p-1} \, |y_i| \\ &= \sum_{i=1}^{n} |x_i + y_i|^{\frac{p}{p'}} \, |x_i| + \sum_{i=1}^{n} |x_i + y_i|^{\frac{p}{p'}} \, |y_i| \\ &\leq \left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p'} \left[\left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p} \right] \\ &= ||\mathbf{x} + \mathbf{y}||^{p/p'} \left(||\mathbf{x}||_p + ||\mathbf{y}||_p \right) \end{aligned}$$

so $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$. This proves the theorem.

It only remains to prove Lemma 10.17.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1}dt + \int_0^b x^{q-1}dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$.

Now we can refer to $||A||_p$ as the operator norm of A taken with respect to $||\cdot||_p$. In the case when $p = 2$, and A is an $n \times n$ matrix, we showed above that this is just the spectral norm.

10.1 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 10.19 *We define $\sigma(A)$ to be the eigenvalues of A. Also,*

$$\rho(A) \equiv \max(|\lambda| : \lambda \in \sigma(A))$$

The number, $\rho(A)$ is known as the spectral radius of A.

Before beginning this discussion, we need to define what we mean by convergence in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$.

Definition 10.20 *Let $\{A_k\}_{k=1}^{\infty}$ be a sequence in $\mathcal{L}(X,Y)$ where X, Y are finite dimensional normed linear spaces. We say $\lim_{n\to\infty} A_k = A$ 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 10.7 and Theorem 10.9 it doesn't matter which one we use. We say

$$\sum_{k=1}^{\infty} A_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} A_k$$

in the usual way.

Lemma 10.21 *Suppose $\{A_k\}_{k=1}^{\infty}$ be 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

$$\sum_{k=1}^{\infty} A_k \tag{10.6}$$

exists. In words, absolute convergence implies convergence.

Proof: For $p \leq m \leq n$,

$$\left|\left|\sum_{k=1}^{n} A_k - \sum_{k=1}^{m} A_k\right|\right| \leq \sum_{k=p}^{\infty} ||A_k||$$

and so for p large enough, this term on the right in the above inequality is less than ε. Since ε is arbitrary, this shows the partial sums of 10.6 are a Cauchy sequence. Therefore by Corollary 10.6 it follows that these partial sums converge.

The next lemma is normally discussed in advanced calculus courses but we prove it here for the convenience of the reader. It is known as the root test.

Lemma 10.22 *Let $\{a_p\}$ be a sequence of nonnegative terms and let*

$$r = \limsup_{p\to\infty} a_p^{1/p}.$$

Then if $r < 1$, it follows the series, $\sum_{k=1}^{\infty} a_k$ converges and if $r > 1$, then a_p fails to converge to 0 so the series diverges. If A is an $n \times n$ matrix and

$$1 < \limsup_{p\to\infty} ||A^p||^{1/p}, \tag{10.7}$$

then $\sum_{k=0}^{\infty} A^k$ fails to converge.

Proof: Suppose $r < 1$. Then there exists N such that if $p > N$,

$$a_p^{1/p} < R$$

where $r < R < 1$. Therefore, for all such p, we have $a_p < R^p$ and so by comparison with the geometric series, $\sum R^p$, it follows $\sum_{p=1}^{\infty} a_p$ converges.

Next suppose $r > 1$. Then letting $1 < R < r$, it follows there are infinitely many values of p at which

$$R < a_p^{1/p}$$

which implies $R^p < a_p$, showing that a_p cannot converge to 0.

To see the last claim, if 10.7 holds, then from the first part of this lemma, $||A^p||$ fails to converge to 0 and so we can conclude that $\left\{\sum_{k=0}^{m} A^k\right\}_{m=0}^{\infty}$ is not a Cauchy sequence. Hence $\sum_{k=0}^{\infty} A^k = \lim_{m\to\infty} \sum_{k=0}^{m} A^k$ cannot exist.

In this section we give a way to estimate $\rho(A)$ which is of great significance. It is based on the following lemma.

Lemma 10.23 *If $|\lambda| > \rho(A)$, for A an $n \times n$ matrix, then the series,*

$$\frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{A^k}{\lambda^k}$$

converges.

Proof: Let J denote the Jordan canonical form of A. Also, let $||A|| \equiv \max\{|a_{ij}|, i,j = 1,2,\cdots,n\}$. Then for some invertible matrix, S, we have $A = S^{-1}JS$. Therefore, it is routine to show

$$\frac{1}{\lambda}\sum_{k=0}^{p}\frac{A^k}{\lambda^k} = S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k}\right)S.$$

Now from the structure of the Jordan form, we see that $J = D+N$ where D is the diagonal matrix consisting of the eigenvalues of A listed according to algebraic multiplicity and N is a nilpotent matrix which commutes with D. Say $N^m = 0$. Therefore, for k much larger than m, say $k > 2m$,

$$J^k = (D+N)^k = \sum_{l=0}^{m}\binom{k}{l}D^{k-l}N^l.$$

It follows that

$$\left|\left|J^k\right|\right| \le C(m,N)\,k(k-1)\cdots(k-m+1)\,||D||^k$$

and so

$$\limsup_{k\to\infty}\left|\left|\frac{J^k}{\lambda^k}\right|\right|^{1/k} \le \lim_{k\to\infty}\left(\frac{C(m,N)\,k(k-1)\cdots(k-m+1)\,||D||^k}{|\lambda|^k}\right)^{1/k} = \frac{||D||}{|\lambda|} < 1.$$

Therefore, this shows by the root test that $\sum_{k=0}^{\infty}\left|\left|\frac{J^k}{\lambda^k}\right|\right|$ converges. Therefore, by Lemma 10.21 it follows that

$$\lim_{k\to\infty}\frac{1}{\lambda}\sum_{l=0}^{k}\frac{J^l}{\lambda^l}$$

exists. In particular this limit exists in every norm placed on $\mathcal{L}(\mathbb{F}^n,\mathbb{F}^n)$, and in particular for every operator norm. Now for any operator norm, $||AB|| \le ||A||\,||B||$. Therefore,

$$\left|\left|S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k} - \frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)S\right|\right| \le \left|\left|S^{-1}\right|\right|\,||S||\left|\left|\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k} - \frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)\right|\right|$$

and this converges to 0 as $p\to\infty$. Therefore,

$$\frac{1}{\lambda}\sum_{k=0}^{p}\frac{A^k}{\lambda^k} \to S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)S$$

and this proves the lemma.

Actually this lemma is usually accomplished using the theory of functions of a complex variable but the theory involving the Laurent series is not assumed for these notes.

Lemma 10.24 *Let A be an $n\times n$ matrix. Then for any $||\cdot||$, $\rho(A) \ge \limsup_{p\to\infty}||A^p||^{1/p}$.*

Proof: By Lemma 10.23 and Lemma 10.22, we know that if $|\lambda| > \rho(A)$,

$$\limsup\left|\left|\frac{A^k}{\lambda^k}\right|\right|^{1/k} \le 1,$$

and it doesn't matter which norm we use because they are all equivalent. Therefore, $\limsup_{k\to\infty}\left|\left|A^k\right|\right|^{1/k} \le |\lambda|$. Therefore, since this holds for all $|\lambda| > \rho(A)$, this proves the lemma.

Now we denote by $\sigma(A)^p$ the collection of all numbers of the form λ^p where $\lambda\in\sigma(A)$.

Lemma 10.25 $\sigma(A^p) = \sigma(A)^p$

Proof: In dealing with $\sigma(A^p)$, we can just as well 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 we must have $\lambda = \alpha$ where α is one of the entries on the main diagonal of J^p. Thus $\lambda \in \alpha(A)^p$ and this shows $\sigma(A^p) \subseteq \sigma(A)^p$.

Now take $\alpha \in \sigma(A)$ and consider α^p.

$$\alpha^p I - A^p = \left(\alpha^{p-1} I + \cdots + \alpha A^{p-2} + A^{p-1}\right)(\alpha I - A)$$

and so $\alpha^p I - A^p$ fails to be one to one which shows that $\alpha^p \in \sigma(A^p)$ which shows that $\sigma(A)^p \subseteq \sigma(A^p)$. This proves the lemma.

Lemma 10.26 *Let A be an $n \times n$ matrix and suppose $|\lambda| > ||A||_2$. Then $(\lambda I - A)^{-1}$ exists.*

Proof: Suppose $(\lambda I - A)\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \neq \mathbf{0}$. Then

$$|\lambda|\,||\mathbf{x}||_2 = ||A\mathbf{x}||_2 \leq ||A||\,||\mathbf{x}||_2 < |\lambda|\,||\mathbf{x}||_2,$$

a contradiction. Therefore, $(\lambda I - A)$ is one to one and this proves the lemma.

The main result is the following theorem due to Gelfand in 1941.

Theorem 10.27 *Let A be an $n \times n$ matrix. Then for any $||\cdot||$ defined on $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$*

$$\rho(A) = \lim_{p\to\infty} ||A^p||^{1/p}.$$

Proof: If $\lambda \in \sigma(A)$, then by Lemma 10.25 $\lambda^p \in \sigma(A^p)$ and so by Lemma 10.26, it follows that

$$|\lambda|^p \leq ||A^p||$$

and so $|\lambda| \leq ||A^p||^{1/p}$. Since this holds for every $\lambda \in \sigma(A)$, it follows that for each p,

$$\rho(A) \leq ||A^p||^{1/p}.$$

Now using Lemma 10.24,

$$\rho(A) \geq \limsup_{p\to\infty} ||A^p||^{1/p} \geq \liminf_{p\to\infty} ||A^p||^{1/p} \geq \rho(A)$$

which proves the theorem.

Example 10.28 *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. We will take $\left|\left|A^7\right|\right|^{1/7}$ 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

$$\rho(A) \approx 8015\,625^{1/7} = 9.\,688\,951\,236\,71.$$

Of course the interest lies primarily in matrices for which we cannot find the exact roots to the characteristic equation.

10.2 Functions of matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \tag{10.8}$$

for all $|\lambda| < R$. We will give a formula for $f(A) \equiv \sum_{n=0}^{\infty} a_n A^n$ and show that it makes sense whenever $\rho(A) < R$. Thus we will be able to 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 $k < P$

$$\begin{aligned} f_P^{(k)}(\lambda) &= \sum_{n=k}^{P} a_n n \cdots (n-k+1) \lambda^{n-k} \\ &= \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}. \end{aligned} \tag{10.9}$$

To begin with we will examine $f(J_m(\lambda))$ where $J_m(\lambda)$ is an $m \times m$ Jordan block. Thus $J_m(\lambda) = D + N$ where $N^m = 0$ and N commutes with D. Therefore, letting $P > m$

$$\begin{aligned} \sum_{n=0}^{P} a_n J_m(\lambda)^n &= \sum_{n=0}^{P} a_n \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{P} \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{m-1} N^k \sum_{n=k}^{P} \binom{n}{k} D^{n-k}. \end{aligned} \tag{10.10}$$

Now for $k = 0, \cdots, m-1$, define $diag_k(a_1, \cdots, a_{m-k})$ the $m \times m$ matrix which equals zero everywhere except on the k^{th} super diagonal where this diagonal is filled with the numbers, $\{a_1, \cdots, a_{m-k}\}$ from the upper left to the lower right. Thus in 4×4 matrices, $diag_2(1,2)$ 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 in 10.10, we see with the aid of 10.9 that

$$\sum_{n=0}^{P} a_n J_m(\lambda)^n = \sum_{k=0}^{m-1} diag_k\left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right).$$

Therefore, $\sum_{n=0}^{P} a_n J_m(\lambda)^n =$

$$\begin{pmatrix} f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \frac{f_P^{(2)}(\lambda)}{2!} & \cdots & \frac{f_P^{(m-1)}(\lambda)}{(m-1)!} \\ & f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \ddots & \vdots \\ & & f_P(\lambda) & \ddots & \frac{f_P^{(2)}(\lambda)}{2!} \\ & & & \ddots & \frac{f_P'(\lambda)}{1!} \\ 0 & & & & f_P(\lambda) \end{pmatrix} \tag{10.11}$$

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

$$J = \begin{pmatrix} J_1 & & 0 \\ & J_2 & \\ & & \ddots \\ 0 & & J_r \end{pmatrix} \tag{10.12}$$

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 10.11 $\sum_{n=0}^{P} a_n J_k^n$ converges as $P \to \infty$ to the $m_k \times m_k$ matrix,

$$\begin{pmatrix} f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \frac{f^{(2)}(\lambda_k)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \ddots & \vdots \\ 0 & 0 & f(\lambda_k) & \ddots & \frac{f^{(2)}(\lambda_k)}{2!} \\ \vdots & & \ddots & \ddots & \frac{f'(\lambda_k)}{1!} \\ 0 & 0 & \cdots & 0 & f(\lambda_k) \end{pmatrix} \tag{10.13}$$

There is no convergence problem because we are given that $|\lambda| < R$ for all $\lambda \in \sigma(A)$. This has proved the following theorem.

Theorem 10.29 *Let f be given by 10.8 and suppose $\rho(A) < R$ where R is the radius of convergence of the power series in 10.8. Then the series,*

$$\sum_{k=0}^{\infty} a_n A^n \tag{10.14}$$

converges in the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ with respect to any of the norms on this space and furthermore, we have the formula,

$$\sum_{k=0}^{\infty} a_n A^n = S^{-1} \begin{pmatrix} \sum_{n=0}^{\infty} a_n J_1^n & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \sum_{n=0}^{\infty} a_n J_r^n \end{pmatrix} S$$

where $\sum_{n=0}^{\infty} a_n J_k^n$ is an $m_k \times m_k$ matrix of the form given in 10.13 where $A = S^{-1}JS$ and the Jordan form of A, J is given by 10.12. Therefore, we may define $f(A)$ by the series in 10.14.

Now we give a simple example.

Example 10.30 *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, we can find the Jordan canonical form of the matrix without too much trouble.

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

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

Then from the above theorem we can immediately compute $\sin(A)$ as follows.

$$\sin\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix}.$$

Therefore, $\sin(A) =$

$$\begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix} \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} =$$

$$\begin{pmatrix} \sin 4 & \sin 4 - \sin 2 - \cos 2 & -\cos 2 & \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 & \frac{1}{2}\sin 4 + \frac{3}{2}\sin 2 - 2\cos 2 & \sin 2 & \frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 - 2\cos 2 \\ 0 & -\cos 2 & \sin 2 - \cos 2 & -\cos 2 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 & -\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 + 3\cos 2 & \cos 2 - \sin 2 & -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}.$$

Perhaps this isn't the first thing you would think of. Of course the ability to get this nice closed form description of $\sin(A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

We also have the following interesting corollary to the above theorem which is known as the spectral mapping theorem.

Corollary 10.31 *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 to us 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 when we considered 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 we define what we mean by the positive part of an Hermitian matrix.

Definition 10.32 *Let A be an Hermitian matrix. Thus we may consider A as an element of $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ according to the usual notion of matrix multiplication. Then we know 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. We 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 we mean but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix, A. We have the following interesting theorem.

Theorem 10.33 *If A, B be Hermitian matrices, then for $|\cdot|$ the Frobenius norm,*

$$\left|A^+ - B^+\right| \leq |A - B|.$$

Proof: Let $A = \sum_i \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i$ and let $B = \sum_j \mu_j \mathbf{w}_j \otimes \mathbf{w}_j$ where $\{\mathbf{v}_i\}$ and $\{\mathbf{w}_j\}$ are orthonormal bases of eigenvectors.

$$\left|A^+ - B^+\right|^2 = trace\left(\sum_i \lambda_i^+ \mathbf{v}_i \otimes \mathbf{v}_i - \sum_j \mu_j^+ \mathbf{w}_j \otimes \mathbf{w}_j\right)^2 =$$

$$trace\left[\sum_i \left(\lambda_i^+\right)^2 \mathbf{v}_i \otimes \mathbf{v}_i + \sum_j \left(\mu_j^+\right)^2 \mathbf{w}_j \otimes \mathbf{w}_j\right.$$

$$\left. - \sum_{i,j} \lambda_i^+ \mu_j^+ \left(\mathbf{w}_j, \mathbf{v}_i\right) \mathbf{v}_i \otimes \mathbf{w}_j - \sum_{i,j} \lambda_i^+ \mu_j^+ \left(\mathbf{v}_i, \mathbf{w}_j\right) \mathbf{w}_j \otimes \mathbf{v}_i\right]$$

Since the trace of $\mathbf{v}_i \otimes \mathbf{w}_j$ is $(\mathbf{v}_i, \mathbf{w}_j)$, a fact which follows from $(\mathbf{v}_i, \mathbf{w}_j)$ being the only possibly nonzero eigenvalue,

$$= \sum_i \left(\lambda_i^+\right)^2 + \sum_j \left(\mu_j^+\right)^2 - 2\sum_{i,j} \lambda_i^+ \mu_j^+ \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2. \tag{10.15}$$

Since these are orthonormal bases,

$$\sum_i |(\mathbf{v}_i, \mathbf{w}_j)|^2 = 1 = \sum_j |(\mathbf{v}_i, \mathbf{w}_j)|^2$$

and so 10.15 equals

$$= \sum_i \sum_j \left(\left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+ \mu_j^+ \right) |(\mathbf{v}_i, \mathbf{w}_j)|^2 .$$

Similarly,

$$|A - B|^2 = \sum_i \sum_j \left((\lambda_i)^2 + (\mu_j)^2 - 2\lambda_i \mu_j \right) |(\mathbf{v}_i, \mathbf{w}_j)|^2 .$$

Now it is easy to check that $(\lambda_i)^2 + (\mu_j)^2 - 2\lambda_i\mu_j \geq \left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+\mu_j^+$ and so this proves the theorem.

10.3 The estimation of eigenvalues

We have already seen a way to estimate eigenvalues for Hermitian matrices. However, there are ways to estimate the eigenvalues for general matrices. The first discussed here is known as Gerschgorin's theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

Theorem 10.34 *Let A be an $n \times n$ matrix. Consider the n Gerschgorin discs defined as*

$$D_i \equiv \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\} .$$

Then every eigenvalue is contained in some Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the i^{th} row which are off the main diagonal and form the disc centered at a_{ii} having this radius. The union of these discs contains $\sigma(A)$.

Proof: Suppose $A\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$. Then for $A = (a_{ij})$

$$\sum_{j \neq i} a_{ij} x_j = (\lambda - a_{ii}) x_i .$$

Therefore, if we pick k such that $|x_k| \geq |x_j|$ for all x_j, it follows that $|x_k| \neq 0$ since $|\mathbf{x}| \neq 0$ and

$$|x_k| \sum_{j \neq i} |a_{kj}| \geq \sum_{j \neq i} |a_{kj}| |x_j| \geq |\lambda - a_{ii}| |x_k| .$$

Now dividing by $|x_k|$ we see that λ is contained in the k^{th} Gerschgorin disc.

More can be said but this requires some theory from complex variables. We give the following fundamental theorem about counting zeros.

Theorem 10.35 *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: We are given $f(z) = \prod_{j=1}^{m} (z - a_j)\, g(z)$ where $g(z) \neq 0$ on U. Hence using the product rule, we obtain

$$\frac{f'(z)}{f(z)} = \sum_{j=1}^{m} \frac{1}{z - a_j} + \frac{g'(z)}{g(z)}$$

where $\frac{g'(z)}{g(z)}$ is analytic on U and so

$$\begin{aligned} \frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} dz &= \sum_{j=1}^{m} n(\gamma, a_j) + \frac{1}{2\pi i} \int_\gamma \frac{g'(z)}{g(z)} dz \\ &= \sum_{j=1}^{m} n(\gamma, a_j). \end{aligned}$$

Therefore, this proves the theorem.

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. Using the Frobenius norm or any other matrix norm, we 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 10.35. 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 if we agree to count the zeros 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. We now state the following conclusion about continuous dependence of eigenvalues.

Theorem 10.36 *If λ is an eigenvalue of A, then if $||B - A||$ is small enough, some eigenvalue of B will be within ε of λ.*

We now 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 10.37 *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 we can say 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 that $A(s)$ 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

$$H = \bigcup_{s \in [t, t+\eta]} \sigma(A(s)) \cap D(\lambda(t), \delta).$$

We will show H is connected. Suppose not. Then $H = P \cup Q$ where P, Q are separated and $\lambda(t) \in P$. Let $s_0 \equiv \inf\{s : \lambda(s) \in Q \text{ for some } \lambda(s) \in \sigma(A(s))\}$. We know 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 $s_n \uparrow s_0$. We know $\lambda(s_n) \in P$ for any $\lambda(s_n) \in \sigma(A(s_n)) \cap D(\lambda(t), \delta)$ and we also know from the above discussion that for some choice of $s_n \to s_0$, we have $\lambda(s_n) \to \lambda(s_0)$ which contradicts P and Q separated and nonempty. Since P is nonempty, this shows $Q = \emptyset$. Therefore, H is connected as claimed. But $K_t \supseteq H$ and so $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$. This proves the lemma.

Now we are ready to prove the theorem we need.

Theorem 10.38 *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 $\sigma(A(t_0)) = \lambda_1(t_0), \cdots, \lambda_r(t_0)$. We claim at least one of these is a limit point of K_0 and consequently must be in K_0 which will show 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 we know from Theorem 10.35 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 10.37, if $t_0 < 1$, then $K_0 \cup K_t$ would be a strictly larger connected set containing $\lambda(0)$. (The reason this would be strictly larger is that $K_0 \cap \sigma(A(s)) = \emptyset$ for some $s \in (t, t+\eta)$ while $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$.) Therefore, $t_0 = 1$ and this proves the theorem.

Now we can prove the following interesting corollary of the Gerschgorin theorem.

Corollary 10.39 *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 $A(t)$ we multiply all non diagonal terms by t. We let $t \in [0,1]$. Then $A(0) = diag(a_{11}, \cdots, a_{nn})$ 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. We see that 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 10.38 which is determined by a_{ii}, we know by Gerschgorin's theorem 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, we cannot have points of K in both D_i and $(\cup_{j \neq i} D_j)$. Since we know at least one point of K which is in D_i,(a_{ii}) it follows all of K must be contained in D_i. Now by Theorem 10.38 this shows there are points of $K \cap \sigma(A)$ in D_i. The last assertion follows immediately.

Actually, we can improve the conclusion in this corollary slightly. It involves the following lemma.

Lemma 10.40 *In the situation of Theorem 10.38 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, we can use similar arguments to those of Lemma 10.37 to conclude that

$$H_i \equiv \cup_{s\in[t_0-\eta,t_0]}\sigma\left(A\left(s\right)\right)\cap D_i$$

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 $A\left(s\right)$ contained in K_0, contrary to the definition of t_0. Therefore, there is at most one eigenvalue, $\lambda\left(t_0\right) \in K_0 \cap \sigma\left(A\left(t_0\right)\right)$. We now need to rule out the possibility that it could be a repeated root of the characteristic equation. Suppose then that $\lambda\left(t_0\right)$ is a repeated root of the characteristic equation. As before, we can choose a small disc, D centered at $\lambda\left(t_0\right)$ and η small enough that

$$H \equiv \cup_{s\in[t_0-\eta,t_0]}\sigma\left(A\left(s\right)\right)\cap D$$

is a nonempty connected set containing either multiple eigenvalues of $A\left(s\right)$ or else a single repeated root to the characteristic equation of $A\left(s\right)$. But since H is connected and contains $\lambda\left(t_0\right)$ 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 we hoped. If $t_0 < 1$, there exists a small disc centered at $\lambda\left(t_0\right)$ and $\eta > 0$ such that for all $s \in [t_0, t_0 + \eta]$, $A\left(s\right)$ has only simple eigenvalues in D and the only eigenvalues of $A\left(s\right)$ which could be in K_0 are in D. (This last assertion follows from noting that $\lambda\left(t_0\right)$ is the only eigenvalue of $A\left(t_0\right)$ in K_0 and so the others are at a positive distance from K_0. For s close enough to t_0, we know the eigenvalues of $A\left(s\right)$ are either close to these eigenvalues of $A\left(t_0\right)$ at a positive distance from K_0 or they are close to the eigenvalue, $\lambda\left(t_0\right)$ in which case we can assume 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, we can now sharpen the conclusion of the above corollary.

Corollary 10.41 *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 10.39, we first note that a_{ii} is a simple root of $A\left(0\right)$ 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, $K \cap \sigma\left(A\left(t\right)\right)$ must be contained in the union of the Gerschgorin discs. Since all the other eigenvalues of $A\left(0\right)$, the a_{jj}, are outside D_i, it follows that $K \cap \sigma\left(A\left(0\right)\right) = a_{ii}$. Therefore, by Lemma 10.40, $K \cap \sigma\left(A\left(1\right)\right) = K \cap \sigma\left(A\right)$ consists of a single simple eigenvalue. This proves the corollary.

Example 10.42 *Consider the matrix,*

$$\begin{pmatrix} 5 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

The Gerschgorin discs are $D\left(5,1\right), D\left(1,2\right)$, and $D\left(0,1\right)$. We see that $D\left(5,1\right)$ is disjoint from the other discs. Therefore, there should be an eigenvalue in $D\left(5,1\right)$. 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.

10.4 Exercises

1. Show the inner product on $m \times n$ matrices given by $(A, B) \equiv tr\left(AB^*\right)$ is a genuine inner product satisfying all the necessary axioms. Explain why the space of $m \times n$ matrices is indeed a Hilbert space with this inner product.

2. We say a normed linear space, $(X, ||\cdot||)$ is uniformly convex if whenever $||x_n||, ||y_n|| \leq 1$ and $||x_n + y_n|| \to 2$, it follows that $||x_n - y_n|| \to 0$. Show that every Hilbert space is uniformly convex. **Hint:** Use the parallelogram identity.
3. A normed linear space is strictly convex if whenever $||x||, ||y|| = 1$, it follows that $\left|\left|\frac{x+y}{2}\right|\right| < 1$.
4. Show that $||\cdot||_1$ and $||\cdot||_\infty$ are not strictly convex on $\mathbb{F}^n$.
5. Show that uniform convexity implies strict convexity.
6. Show that for $||\cdot||$ an operator norm and A, B two elements in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$, $||AB|| \leq ||A|| \, ||B||$.
7. Find e^A where A is the matrix given in Example 10.30.
8. State and prove some generalizations to Theorem 10.33 by considering other functions.

An application to differential equations

Here we give a very important application of linear algebra to the theory of first order systems of linear differential equations. We will not use the Jordan canonical form in this presentation. First we need some definitions.

Definition 11.1 *Suppose $t \to M(t)$ is a matrix valued function of t. Thus $M(t) = (m_{ij}(t))$. Then we define*

$$M'(t) \equiv \left(m'_{ij}(t)\right).$$

In words, the derivative of $M(t)$ is the matrix whose entries consist of the derivatives of the entries of $M(t)$. We may define integrals of matrices the same way. Thus

$$\int_a^b M(t)\,di \equiv \left(\int_a^b m_{ij}(t)\,dt\right).$$

In words, the integral of $M(t)$ is the matrix obtained by replacing each entry of $M(t)$ by the integral of that entry. The definition of what we mean by integrals or derivatives of vector valued functions is exactly analogous. Indeed, we may think of a vector in $\mathbb{C}^n$ as an $n \times 1$ matrix.

With this definition, it is easy to prove the following theorem whose proof we leave as an exercise.

Theorem 11.2 *Suppose $M(t)$ and $N(t)$ are matrices for which $M(t)N(t)$ makes sense. Then if $M'(t)$ and $N'(t)$ both exist, it follows that*

$$(M(t)N(t))' = M'(t)N(t) + M(t)N'(t).$$

In the study of differential equations, one of the most important theorems is Gronwall's inequality which we present next.

Theorem 11.3 *Suppose $u(t) \geq 0$ and for all $t \in [0,T]$,*

$$u(t) \leq u_0 + \int_0^t Ku(s)\,ds. \tag{11.1}$$

where K is some constant. Then

$$u(t) \leq u_0 e^{Kt}. \tag{11.2}$$

Proof: Let $w(t) = \int_0^t u(s)\,ds$. Then using the fundamental theorem of calculus, 11.1 $w(t)$ satisfies the following.

$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0,\ w(0) = 0. \tag{11.3}$$

Multiply both sides of this inequality by e^{-Kt} and using the product rule and the chain rule, we obtain

$$e^{-Kt}(w'(t) - Kw(t)) = \frac{d}{dt}\left(e^{-Kt}w(t)\right) \leq u_0 e^{-Kt}.$$

Integrating this from 0 to t, we obtain,

$$e^{-Kt}w(t) \leq u_0 \int_0^t e^{-Ks}ds = u_0\left(-\frac{e^{-tK}-1}{K}\right).$$

Now multiply through by e^{Kt} to obtain

$$w(t) \leq u_0\left(-\frac{e^{-tK}-1}{K}\right)e^{Kt} = -\frac{u_0}{K} + \frac{u_0}{K}e^{tK}.$$

Therefore, 11.3 implies

$$u(t) \leq u_0 + K\left(-\frac{u_0}{K} + \frac{u_0}{K}e^{tK}\right) = u_0 e^{Kt}.$$

This proves the theorem.

With Gronwall's inequality, we establish the following fundamental theorem on uniqueness of solutions to the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t),\ \mathbf{x}(a) = \mathbf{x}_a, \tag{11.4}$$

in which A is an $n \times n$ matrix and $\mathbf{f}$ is a continuous function having values in $\mathbb{C}^n$.

Theorem 11.4 *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy 11.4. Then* $\mathbf{x}(t) = \mathbf{y}(t)$ *for all* t.

Proof: Let $\mathbf{z}(t) = \mathbf{x}(t+a) - \mathbf{y}(t+a)$. Then for $t \geq 0$,

$$\mathbf{z}' = A\mathbf{z},\ \mathbf{z}(0) = \mathbf{0}. \tag{11.5}$$

We note that for $K = \max\{|a_{ij}|\}$, where $A = (a_{ij})$, we can write the following.

$$|(A\mathbf{z},\mathbf{z})| = \left|\sum_{ij} a_{ij}z_j\overline{z_i}\right| \leq K\sum_{ij}|z_i|\,|z_j|$$

$$\leq K\sum_{ij}\left(\frac{|z_i|^2}{2} + \frac{|z_j|^2}{2}\right) = nK|\mathbf{z}|^2.$$

(For x and y real numbers, $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ because this is equivalent to saying $(x-y)^2 \geq 0$.)Similarly,

$$|(\mathbf{z},A\mathbf{z})| \leq nK|\mathbf{z}|^2$$

Thus,

$$|(\mathbf{z},A\mathbf{z})|, |(A\mathbf{z},\mathbf{z})| \leq nK|\mathbf{z}|^2. \tag{11.6}$$

Now multiplying 11.5 by $\mathbf{z}$ and observing that

$$\frac{d}{dt}\left(|\mathbf{z}|^2\right) = (\mathbf{z}', \mathbf{z}) + (\mathbf{z}, \mathbf{z}') = (A\mathbf{z}, \mathbf{z}) + (\mathbf{z}, A\mathbf{z}),$$

it follows from 11.6 and the observation that $\mathbf{z}(0) = 0$,

$$|\mathbf{z}(t)|^2 \leq \int_0^t 2nK |\mathbf{z}(s)|^2 \, ds$$

and so by Gronwall's inequality, $|\mathbf{z}(t)|^2 = 0$ for all $t \geq 0$. Thus,

$$\mathbf{x}(t) = \mathbf{y}(t)$$

for all $t \geq a$.

Now let $\mathbf{w}(t) = \mathbf{x}(a-t) - \mathbf{y}(a-t)$ for $t \geq 0$. Then $\mathbf{w}'(t) = (-A)\mathbf{w}(t)$ and we may repeat the argument which was just given to conclude that $\mathbf{x}(t) = \mathbf{y}(t)$ for all $t \leq a$. This proves the theorem.

Definition 11.5 *Let A be an $n \times n$ matrix. We say $\Phi_A(t)$ is a fundamental matrix for A if*

$$\Phi_A'(t) = A\Phi_A(t), \ \Phi_A(0) = I, \tag{11.7}$$

and $\Phi_A(t)^{-1}$ exists for all $t \in \mathbb{R}$.

Why should we care about a fundamental matrix? The reason is that such a matrix valued function enables us to solve the first order linear differential equation and initial condition, known as the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t), \ \mathbf{x}(0) = \mathbf{x}_0, \tag{11.8}$$

on the interval, $[0, T]$. We wish to establish the existence and uniqueness of fundamental matrices. In order to do so, it is necessary to first consider the special case where $n = 1$. In particular, we need to solve the first order ordinary differential equation and initial condition,

$$r' = \lambda r + g, \ r(0) = r_0, \tag{11.9}$$

where g is a continuous scalar valued function, but even before this we need to consider the case where $g = 0$.

Lemma 11.6 *There exists a unique solution to the initial value problem,*

$$r' = \lambda r, \ r(0) = 1, \tag{11.10}$$

and the solution for $\lambda = a + ib$ is given by

$$r(t) = e^{at}(\cos bt + i \sin bt). \tag{11.11}$$

We denote this solution to the initial value problem as $e^{\lambda t}$. (If λ is real, $e^{\lambda t}$ as defined here reduces to the usual exponential function so there is no contradiction between this and earlier notation seen in Calculus.)

Proof: We know from the uniqueness theorem presented above, Theorem 11.4, applied to the case where $n = 1$ that there can be no more than one solution to the initial value problem, 11.10. Therefore, it only remains to verify 11.11 is a solution to 11.10. However, this is an easy calculus exercise. This proves the Lemma.

Note the differential equation in 11.10 says

$$\frac{d}{dt}\left(e^{\lambda t}\right) = \lambda e^{\lambda t}. \tag{11.12}$$

With this lemma, it becomes possible to easily solve the case in which $g \neq 0$.

Theorem 11.7 *There exists a unique solution to 11.9 and this solution is given by the formula,*

$$r(t) = e^{\lambda t} r_0 + e^{\lambda t} \int_0^t e^{-\lambda s} g(s)\, ds. \tag{11.13}$$

Proof: By the uniqueness theorem, Theorem 11.4, we know there is no more than one solution. It only remains to verify that 11.13 is a solution and we are done. But $r(0) = e^{\lambda 0} r_0 + \int_0^0 e^{-\lambda s} g(s)\, ds = r_0$ and so the initial condition is satisfied. Next we must differentiate this expression to verify the differential equation is also satisfied. Using 11.12, the product rule and the fundamental theorem of calculus,

$$\begin{aligned} r'(t) &= \lambda e^{\lambda t} r_0 + \lambda e^{\lambda t} \int_0^t e^{-\lambda s} g(s)\, ds + e^{\lambda t} e^{-\lambda t} g(t) \\ &= \lambda r(t) + g(t). \end{aligned}$$

This proves the Theorem.

Now we are ready to consider the question of finding a fundamental matrix for A. When this is done, it will be easy to give a formula for the general solution to 11.8 known as the variation of constants formula, arguably the most important result in differential equations.

The next theorem gives a way of obtaining a solution to the system, 11.7. It is known as Putzer's method [1].

Theorem 11.8 *Let A be an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Define*

$$P_k(A) \equiv \prod_{m=1}^{k} (A - \lambda_m I), P_0(A) \equiv I,$$

and let the scalar valued functions, $r_k(t)$ be defined as follows.

$$r_0(t) \equiv 0,$$

$$r'_{k+1} = \lambda_{k+1} r_{k+1} + r_k,$$

$$r_{k+1}(0) = b_{k+1},$$

where $b_1 = 1$ and $b_k = 0$ if $k > 1$. Now define

$$\Phi(t) \equiv \sum_{k=0}^{n-1} r_{k+1}(t) P_k(A).$$

Then

$$\Phi'(t) = A\Phi(t).$$

Furthermore, for all t, we have that $\Phi(t)^{-1}$ exists and $\Phi(t)$ is the unique fundamental matrix for A.

Proof: The first part of this follows from a computation.

$$\Phi'(t) = \sum_{k=0}^{n-1} r'_{k+1}(t) P_k(A) = \sum_{k=0}^{n-1} (\lambda_{k+1} r_{k+1}(t) + r_k(t)) P_k(A) =$$

$$\sum_{k=0}^{n-1}\lambda_{k+1}r_{k+1}(t)P_k(A)+\sum_{k=0}^{n-1}r_k(t)P_k(A)=\sum_{k=0}^{n-1}(\lambda_{k+1}I-A)r_{k+1}(t)P_k(A)+$$

$$\sum_{k=0}^{n-1}r_k(t)P_k(A)+\sum_{k=0}^{n-1}Ar_{k+1}(t)P_k(A)$$

$$=-\sum_{k=0}^{n-1}r_{k+1}(t)P_{k+1}(A)+\sum_{k=0}^{n-1}r_k(t)P_k(A)+A\sum_{k=0}^{n-1}r_{k+1}(t)P_k(A). \tag{11.14}$$

Now

$$\sum_{k=0}^{n-1}r_{k+1}(t)P_{k+1}(A)=\sum_{k=1}^{n}r_k(t)P_k(A)=\sum_{k=1}^{n-1}r_k(t)P_k(A),$$

by the Cayley Hamilton theorem.(The Cayley Hamilton theorem says that $P_n(A)=0$.) Also,

$$\sum_{k=0}^{n-1}r_k(t)P_k(A)=\sum_{k=1}^{n-1}r_k(t)P_k(A)$$

because $r_0(t)\equiv 0$. Therefore, the first two terms in 11.14 cancel and we are left with

$$\Phi'(t)=A\sum_{k=0}^{n-1}r_{k+1}(t)P_k(A)=A\Phi(t).$$

It remains to verify that $\Phi(t)^{-1}$ exists for all t. To do so, consider $\mathbf{v}\neq\mathbf{0}$ and suppose for some t_0, we have $\Phi(t_0)\mathbf{v}=\mathbf{0}$. Then let $\mathbf{x}(t)\equiv\Phi(t_0+t)\mathbf{v}$. Then we have that

$$\mathbf{x}'(t)=A\Phi(t_0+t)\mathbf{v}=A\mathbf{x}(t),\ \mathbf{x}(0)=\Phi(t_0)\mathbf{v}=\mathbf{0}.$$

But also $\mathbf{z}(t)\equiv\mathbf{0}$ also satisfies

$$\mathbf{z}'(t)=A\mathbf{z}(t),\ \mathbf{z}(0)=\mathbf{0},$$

and so by the theorem on uniqueness, it must be the case that $\mathbf{z}(t)=\mathbf{x}(t)$ for all t, showing that $\Phi(t+t_0)\mathbf{v}=\mathbf{0}$ for all t, and in particular for $t=-t_0$. Therefore, we see that

$$\Phi(-t_0+t_0)\mathbf{v}=I\mathbf{v}=\mathbf{0}$$

and so $\mathbf{v}=\mathbf{0}$, a contradiction. It follows that $\Phi(t)$ must be one to one for all t and so, $\Phi(t)^{-1}$ exists for all t.

It only remains to verify that the fundamental matrix is unique. Suppose Ψ is another fundamental matrix. Then letting $\mathbf{v}$ be an arbitrary vector, we have

$$\mathbf{z}(t)\equiv\Phi(t)\mathbf{v},\ \mathbf{y}(t)\equiv\Psi(t)\mathbf{v}$$

both solve the initial value problem,

$$\mathbf{x}'=A\mathbf{x},\ \mathbf{x}(0)=\mathbf{v},$$

and so we must have $\mathbf{z}(t)=\mathbf{y}(t)$ for all t showing that $\Phi(t)\mathbf{v}=\Psi(t)\mathbf{v}$ for all t. Since $\mathbf{v}$ is arbitrary, this shows that $\Phi(t)=\Psi(t)$ for every t. This proves the theorem.

The following theorem gives the variation of constants formula,.

Theorem 11.9 *Let* $\mathbf{f}$ *be continuous on* $[0,T]$ *and let* A *be an* $n \times n$ *matrix and* $\mathbf{x}_0$ *a vector in* $\mathbb{C}^n$. *Then there exists a unique solution to 11.8,* $\mathbf{x}$, *given by the variation of constants formula,*

$$\mathbf{x}(t) = \Phi_A(t)\,\mathbf{x}_0 + \Phi_A(t)\int_0^t \Phi_A(s)^{-1}\,\mathbf{f}(s)\,ds. \tag{11.15}$$

Also, $\Phi_A(t)^{-1} = \Phi_A(-t)$.

Proof: We know from the uniqueness theorem there is at most one solution to 11.8. Therefore, if we can verify 11.15 solves 11.8, we are done. The verification that the given formula works is identical with the verification that the scalar formula given in Theorem 11.7 solves the initial value problem given there. We know $\Phi_A(s)^{-1}$ is continuous because of the formula for the inverse of a matrix in terms of the transpose of the cofactor matrix. Therefore, the integrand in 11.15 is continuous and we may use the fundamental theorem of calculus as in this earlier Theorem. To verify the formula for the inverse, fix s and consider $\mathbf{x}(t) = \Phi_A(s+t)\,\mathbf{v}$, and $\mathbf{y}(t) = \Phi_A(t)\,\Phi_A(s)\,\mathbf{v}$. Then

$$\mathbf{x}'(t) = A\Phi_A(t+s)\,\mathbf{v} = A\mathbf{x}(t),\ \mathbf{x}(0) = \Phi_A(s)\,\mathbf{v}$$

$$\mathbf{y}'(t) = A\Phi_A(t)\,\Phi_A(s)\,\mathbf{v} = A\mathbf{y}(t),\ \mathbf{y}(0) = \Phi_A(s)\,\mathbf{v}.$$

By the uniqueness theorem, we have $\mathbf{x}(t) = \mathbf{y}(t)$ for all t. Since s and $\mathbf{v}$ are arbitrary, this shows $\Phi_A(t+s) = \Phi_A(t)\,\Phi_A(s)$ for all t, s. Thus the last claim is verified from $\Phi_A(0) = I$.

Theorem 11.9 is general enough to include all constant coefficient linear differential equations or any order. Thus it includes as a special case the main topics of an entire elementary differential equations class. We illustrate in the following example. Basically, we reduce an arbitrary linear differential equation to a first order system and then apply the above theory to solve the problem. The next example is a differential equation of damped vibration.

Example 11.10 *The differential equation is* $y'' + 2y' + 2y = \cos t$ *and we give the initial conditions,* $y(0) = 1$ *and* $y'(0) = 0$.

To solve this equation, we let $x_1 = y$ and $x_2 = x_1' = y'$. Then, writing this in terms of these new variables, we obtain the following system.

$$\begin{aligned} x_2' + 2x_2 + 2x_1 &= \cos t \\ x_1' &= x_2 \end{aligned}$$

We can write this system in the form studied above.

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' &= \begin{pmatrix} x_2 \\ -2x_2 - 2x_1 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix}. \end{aligned}$$

In a similar way other scalar differential equations with initial conditions can be reduced to systems of the form just discussed. Now Putzer's method is not a practical way to solve this problem. This should be done by other methods taught in ordinary differential equations classes. However, the theory of first order systems is well understood from the above discussion.

11.1 Exercises

1. Suppose $\mathbf{x}(t) = \mathbf{v}e^{\lambda t}$ and that $\mathbf{x}$ is a solution to the first order system, $\mathbf{x}' = A\mathbf{x}$. Show that $A\mathbf{v} = \lambda\mathbf{v}$.

2. Suppose $\mathbf{x}' = A\mathbf{x}$ and suppose Φ is the fundamental matrix for A. Show the columns of Φ are solutions to $\mathbf{x}' = A\mathbf{x}$ and show that whenever $\mathbf{x}' = A\mathbf{x}$, there exists a constant vector, $\mathbf{c}$, such that $\mathbf{x}(t) = \Phi(t)\mathbf{c}$. Use this to verify that the set of solutions to $\mathbf{x}' = A\mathbf{x}$ is an n dimensional vector space.

3. Let A be an $n \times n$ matrix and suppose $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are n solutions to the equation $\mathbf{x}' = A\mathbf{x}$. Let $\Psi(t)$ denote the $n \times n$ matrix such that the k^{th} column of $\Psi(t)$ is the vector, $\mathbf{x}_k(t)$. Show that $\Psi'(t) = A\Psi(t)$. Also show that $\det(\Psi(t))$ is either always equal to zero or never equal to zero. **Hint:** $\Psi(t) = (\Phi(t)\mathbf{c}_1, \cdots, \Phi(t)\mathbf{c}_n) = \Phi(t)C$, where C is the matrix which has $\mathbf{c}_k$ as the k^{th} column. Now ask whether $\det C = 0$. ($\det \Psi(t)$ is known as the Wronskian.)

4. Suppose that $\mathbf{x}_p' = A\mathbf{x}_p + \mathbf{f}$. We call $\mathbf{x}_p$ a particular solution. Show that if $\mathbf{y}' = A\mathbf{y} + \mathbf{f}$, then there exists a constant vector, $\mathbf{c}$ such that $\mathbf{y}(t) = \Phi(t)\mathbf{c} + \mathbf{x}_p(t)$. **Hint:** Consider $\mathbf{y} - \mathbf{x}_p$ and show it satisfies $\mathbf{x}' = A\mathbf{x}$.

5. Look for a particular solution to $\mathbf{x}' = A\mathbf{x} + \mathbf{f}$ in the form $\mathbf{x}_p(t) = \Phi(t)\mathbf{v}(t)$ where $\mathbf{v}(t)$ is a differentiable function. Obtain the variation of constants formula by choosing $\mathbf{v}$ appropriately. This approach is why this formula is called the variation of constants formula. We "vary the constants" by replacing them with a function, $\mathbf{v}(t)$.

6. Let A be an $n \times n$ matrix and recall the generalized eigenspaces of Section 7, X_s associated with the eigenvalue, λ_s, and how $X_1 \oplus \cdots \oplus X_p = \mathbb{C}^n$. Also recall that if m_s was the algebraic multiplicity of λ_s we had

$$(A - \lambda_s I)^{m_s} \mathbf{v} = \mathbf{0}$$

for any $\mathbf{v} \in X_s$. Let $\mathbf{v} \in X_1$. We can use Putzer's method to find a convenient formula for $\Phi(t)\mathbf{v}$. We note that

$$\Phi(t)\mathbf{v} = \sum_{k=0}^{m_1 - 1} r_{k+1}(t)(A - \lambda_1 I)^k \mathbf{v}$$

because $P_k(A)\mathbf{v} = \mathbf{0}$ whenever $k \geq m_1$ due to the fact that for such k, $P_k(A)$ contains the factor,

$$(A - \lambda_1 I)^{m_1} \mathbf{v} = \mathbf{0}.$$

Therefore, we only need to find a formula for $r_{k+1}(t)$. Show by solving the first order linear scalar equations in the definition of Putzer's method that

$$r_{k+1}(t) = \frac{t^k}{k!} e^{\lambda_1 t}.$$

Conclude that

$$\Phi(t)\mathbf{v} = \sum_{k=0}^{m_1 - 1} \frac{t^k}{k!} e^{\lambda_1 t}(A - \lambda_1 I)^k \mathbf{v}.$$

If $\mathbf{v} \in X_s$, show that the same kind of formula holds for $\Phi(t)\mathbf{v}$. That is,

$$\Phi(t)\mathbf{v} = \sum_{k=0}^{m_s - 1} \frac{t^k}{k!} e^{\lambda_s t}(A - \lambda_s I)^k \mathbf{v}.$$

Hint: Simply reorder the eigenvalues and use uniqueness of $\Phi(t)$. Now suppose we consider the initial value problem, $\mathbf{x}' = A\mathbf{x}$, $\mathbf{x}(0) = \mathbf{x}_0$. By $X_1 \oplus \cdots \oplus X_p = \mathbb{C}^n$, we write $\mathbf{x}_0$ in a unique way as

$$\mathbf{x}_0 = \mathbf{v}_1 + \cdots + \mathbf{v}_p$$

where $\mathbf{v}_s \in X_s$. Write an interesting formula for the solution to this initial value problem, $\Phi(t)\mathbf{x}_0$ using the first part of this problem. The reason this is interesting is that, depending on the sign of the real part of the eigenvalue, we can determine the qualitative behavior of the solution as $t \to \infty$ and as $t \to -\infty$.

The Binet Cauchy formula

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 which whose i^{th} column equals $\mathbf{v}_i$. Define

$$d(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \det(M(\mathbf{v}_1, \cdots, \mathbf{v}_n)).$$

Then by Problem 4 in Chapter 1,

$$\mathbf{v}^i \to d_n\left(\mathbf{v}^1 \cdots \mathbf{v}^i \cdots \mathbf{v}^n\right)$$

is linear and is alternating because

$$d_n\left(\mathbf{v}^1 \cdots \mathbf{v}^i \cdots \mathbf{v}^j \cdots \mathbf{v}^n\right) = -d_n\left(\mathbf{v}^1 \cdots \mathbf{v}^j \cdots \mathbf{v}^i \cdots \mathbf{v}^n\right).$$

It maps an ordered list of n vectors in $\mathbb{F}^n$ to $\mathbb{F}$. This can be generalized to consider alternating m linear forms on $\mathbb{F}^n$ where $m \leq n$.

We say M is an alternating m linear form on $\mathbb{F}^n$ if

$$M : (\mathbb{F}^n)^m \to \mathbb{F}$$

is multilinear and alternating. Note the determinant d_n is an example in the case where $m = n$.

Now we define $\Lambda(n, m)$ to be the set of increasing lists of m numbers in $\{1, \cdots, n\}$. These will be denoted by $\mathbf{i}, \mathbf{j}$,or $\mathbf{k}$. Thus $\mathbf{i} \in \Lambda(n, m)$ means

$$\mathbf{i} = (i_1, \cdots, i_m),\ i_1 < \cdots < i_m.$$

Also for $\mathbf{i} \in \Lambda(n, m)$, we will define

$$\mathbf{e}^{\mathbf{i}} \equiv \left(\mathbf{e}^{i_1}, \cdots, \mathbf{e}^{i_m}\right).$$

Now let $\mathbf{v}^1, \cdots, \mathbf{v}^m$ be vectors in $\mathbb{F}^n$ and $\mathbf{i} \in \Lambda(n, m)$. Let

$$d_{\mathbf{i}}\left(\mathbf{v}^1 \cdots \mathbf{v}^m\right) \equiv \det \begin{bmatrix} v^1_{i_1} & \cdots & v^m_{i_1} \\ \vdots & & \vdots \\ v^1_{i_m} & \cdots & v^m_{i_m} \end{bmatrix}.$$

Thus $d_{\mathbf{i}}\left(\mathbf{v}^1 \cdots \mathbf{v}^m\right)$ is the determinant of the $m \times n$ matrix obtained from retaining only the $i_1, \cdots, i_m$ rows of the $n \times m$ matrix

$$\left[\mathbf{v}^1 \cdots \mathbf{v}^m\right].$$

Lemma 12.1 *The following holds for* $\mathbf{i}, \mathbf{j} \in \Lambda(n,m)$.

$$d_{\mathbf{i}}\left(\mathbf{e}^{\mathbf{j}}\right) = \begin{cases} 1 \text{ if } \mathbf{i} = \mathbf{j} \\ 0 \text{ if } \mathbf{i} \neq \mathbf{j} \end{cases}.$$

Proof: If $\mathbf{i} = \mathbf{j}$, this equals

$$\det \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = 1$$

If $d_{\mathbf{i}}\left(\mathbf{e}^{\mathbf{j}}\right) \neq 0$, then for each r, the r *th* column must contain a 1 and not be a column of zeroes. Thus $j_r = i_s$ for some s. Hence

$$\{j_1, \cdots, j_m\} = \{i_1, \cdots, i_m\}$$

but both lists are increasing and so $j_r = i_r$ for all r. This proves the lemma.

Clearly the set of alternating m linear functions on $\mathbb{F}^n$ is an vector space. We denote this space by $A(n,m)$.

Theorem 12.2 $\{d_{\mathbf{i}} : \mathbf{i} \in \Lambda(n,m)\}$ *is a basis for* $A(n,m)$.

Proof: Let $M \in A(n,m)$ and let $\mathbf{v}^1, \cdots, \mathbf{v}^m$ be vectors in $\mathbb{F}^n$. Since M is m linear,

$$M\left(\mathbf{v}^1 \cdots \mathbf{v}^m\right) = \sum_{\{j_1, \cdots j_m\}} M\left(\mathbf{e}^{j_1} \cdots \mathbf{e}^{j_m}\right) \mathbf{v}^1_{j_1} \cdots \mathbf{v}^m_{j_m}.$$

Since $d_{\mathbf{i}}$ is also in $A(n,m)$, then if the set $\{j_1, \cdots, j_m\} = \{i_1, \cdots, i_m\}$ with no regard to order on $\{j_1, \cdots, j_m\}$, and

$$i_1 < \cdots < i_m,$$

then

$$M\left(\mathbf{e}^{j_1} \cdots \mathbf{e}^{j_m}\right)$$

$$= d_{\mathbf{i}}\left(\mathbf{e}^{j_1} \cdots \mathbf{e}^{j_m}\right) M\left(\mathbf{e}^{i_1} \cdots \mathbf{e}^{i_m}\right) = d_{\mathbf{i}}\left(\mathbf{e}^{j_1} \cdots \mathbf{e}^{j_m}\right) M\left(\mathbf{e}^{\mathbf{i}}\right).$$

Therefore, the above sum equals

$$= \sum_{\mathbf{i} \in \Lambda(n,m)} \sum_{\{j_1, \cdots, j_m\} = \{\mathbf{i}\}} M\left(\mathbf{e}^{\mathbf{i}}\right) d_{\mathbf{i}}\left(\mathbf{e}^{j_1} \cdots \mathbf{e}^{j_m}\right) \mathbf{v}^1_{j_1} \cdots \mathbf{v}^m_{j_m}$$

$$= \sum_{\mathbf{i} \in \Lambda(n,m)} M\left(\mathbf{e}^{\mathbf{i}}\right) d_{\mathbf{i}}\left(\mathbf{v}^1 \cdots \mathbf{v}^m\right).$$

Since $\{\mathbf{v}^1, \cdots, \mathbf{v}^m\}$ is arbitrary, this shows

$$M = \sum_{\mathbf{i} \in \Lambda(n,m)} M\left(\mathbf{e}^{\mathbf{i}}\right) d_{\mathbf{i}}.$$

If

$$0 = \sum_{\mathbf{i} \in \Lambda(n,m)} a^{\mathbf{i}} d_{\mathbf{i}},$$

then

$$0 = \sum_{\mathbf{i} \in \Lambda(n,m)} a^{\mathbf{i}} d_{\mathbf{i}} \left(\mathbf{e}^{\mathbf{j}} \right) = a^{\mathbf{j}}.$$

This proves the theorem.

Now let A be an $m \times n$ matrix and let B be an $n \times m$ matrix. Thus the ij^{th}component of AB is $\mathbf{a}^i \cdot \mathbf{b}^j$ where $\mathbf{a}^i$ is the i^{th} row of A and $\mathbf{b}^j$ is the j^{th} column of B. Here $\mathbf{v} \cdot \mathbf{w} \equiv \sum_{i=1}^{n} v_i w_i$.

$$AB = \begin{bmatrix} \mathbf{a}^1 \cdot \mathbf{b}^1 & \cdots & \mathbf{a}^1 \cdot \mathbf{b}^m \\ \vdots & & \vdots \\ \mathbf{a}^m \cdot \mathbf{b}^1 & \cdots & \mathbf{a}^m \cdot \mathbf{b}^m \end{bmatrix}.$$

Fixing A,

$$\left(\mathbf{b}^1 \cdots \mathbf{b}^m \right) \to \det (AB)$$

is an alternating m linear form on $\mathbb{F}^n$. Thus from Theorem 12.2,

$$\det (AB) = \sum_{\mathbf{i} \in \Lambda(n,m)} \alpha^{\mathbf{i}} d_{\mathbf{i}} \left(\mathbf{b}^1 \cdots \mathbf{b}^m \right)$$

where

$$\alpha^{\mathbf{i}} = \det \left(A \mathbf{e}^{\mathbf{i}} \right)$$

$$= \det \begin{bmatrix} a^1_{i_1} & a^1_{i_2} & \cdots & a^1_{i_m} \\ a^2_{i_1} & a^2_{i_2} & \cdots & a^2_{i_m} \\ \vdots & \vdots & & \vdots \\ a^m_{i_1} & a^m_{i_2} & \cdots & a^m_{i_m} \end{bmatrix}.$$

Thus, using the property that $\det (A) = \det \left(A^T \right)$, and letting

$$A^T = \left[\mathbf{a}^{1T} \cdots \mathbf{a}^{mT} \right],$$

it follows

$$\alpha^{\mathbf{i}} = d_{\mathbf{i}} \left(\mathbf{a}^{1T} \cdots \mathbf{a}^{mT} \right).$$

This proves the Binet Cauchy formula.

Theorem 12.3 *Let A be an $m \times n$ matrix and B is an $n \times m$ matrix. Denoting the columns of B by $\mathbf{b}^1, \cdots, \mathbf{b}^m$ and the rows of A by $\mathbf{a}^1, \cdots, \mathbf{a}^m$,*

$$\det (AB) = \sum_{\mathbf{i} \in \Lambda(n,m)} d_{\mathbf{i}} \left(\mathbf{a}^{1T} \cdots \mathbf{a}^{mT} \right) d_{\mathbf{i}} \left(\mathbf{b}^1 \cdots \mathbf{b}^m \right).$$

Note that if $n = m$, then there is only one term in the sum and the formula reduces to

$$\det (AB) = \det \left(A^T \right) \det (B) = \det (A) \det (B).$$

Corollary 12.4 *Let A be an $m \times n$ matrix. Then $\det \left(AA^T \right)$ equals the sum of the squares of the determinants of all possible $m \times m$ matrices obtained by deleting $n - m$ columns of A.*

The Fundamental Theorem Of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in $\mathbb{C}$ has a zero in $\mathbb{C}$. If $\mathbb{C}$ is replaced by $\mathbb{R}$, this is not true because of the example, $x^2 + 1 = 0$. This theorem is a very remarkable result and notwithstanding its title, all the best proofs of it depend on either analysis or topology. It was first proved by Gauss in 1797. The proof given here follows Rudin [9]. See also Hardy [4] for another proof, more discussion and references. To begin with we need a fundamental result known as De Moivre's theorem.

Theorem A.1 *Let $r > 0$ be given. Then if n is a positive integer,*

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

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

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

which by induction equals

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

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

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

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

Corollary A.2 *Let z be a non zero complex number. Then there are always exactly k kth roots of z in $\mathbb{C}$.*

We only need the part of the corollary which asserts the existence of at least one kth root.

Proof: Let $z = x + iy$. Then

$$z = |z| \left(\frac{x}{|z|} + i \frac{y}{|z|} \right)$$

and from the definition of $|z|$,

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

Thus $\left(\frac{x}{|z|}, \frac{y}{|z|}\right)$ is a point on the unit circle and so

$$\frac{y}{|z|} = \sin t, \ \frac{x}{|z|} = \cos t$$

for a unique $t \in [0, 2\pi)$. By De Moivre's theorem, a number is a kth root of z if and only if it is of the form

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

for l an integer. Since the cosine and sine are periodic of period 2π, there are exactly k of these numbers.

Lemma A.3 *Let $a_k \in \mathbb{C}$ for $k = 1, \cdots, n$ and let $p(z) \equiv \sum_{k=1}^{n} a_k z^k$. Then p is continuous.*

Proof:

$$|az^n - aw^n| \leq |a| \, |z - w| \left|z^{n-1} + z^{n-2}w + \cdots + w^{n-1}\right|.$$

Then for $|z - w| < 1$, the triangle inequality implies $|w| < 1 + |z|$ and so if $|z - w| < 1$,

$$|az^n - aw^n| \leq |a| \, |z - w| \, n \left(1 + |z|\right)^n.$$

If $\epsilon > 0$ is given, let

$$\delta < \min\left(1, \frac{\epsilon}{|a| \, n \left(1 + |z|\right)^n}\right).$$

It follows from the above inequality that for $|z - w| < \delta$, $|az^n - aw^n| < \epsilon$. The function of the lemma is just the sum of functions of this sort and so it follows that it is also continuous.

Theorem A.4 *(Fundamental theorem of Algebra) Let $p(z)$ be a non constant polynomial. Then there exists $z \in \mathbb{C}$ such that $p(z) = 0$.*

Proof: Suppose not. Then

$$p(z) = \sum_{k=0}^{n} a_k z^k$$

where $a_n \neq 0, \ n > 0$. Then

$$|p(z)| \geq |a_n| \, |z|^n - \sum_{k=0}^{n-1} |a_k| \, |z|^k$$

and so

$$\lim_{|z| \to \infty} |p(z)| = \infty. \tag{1.1}$$

Now let

$$\lambda \equiv \inf\{|p(z)| : z \in \mathbb{C}\}.$$

By 1.1, there exists an $R > 0$ such that if $|z| > R$, it follows that $|p(z)| > \lambda + 1$. Therefore,

$$\lambda \equiv \inf\{|p(z)| : z \in \mathbb{C}\} = \inf\{|p(z)| : |z| \leq R\}.$$

The set $\{z : |z| \leq R\}$ is a closed and bounded set and so this infimum is achieved at some point w with $|w| \leq R$. If $|p(w)| = 0$, we have obtained a contradiction so assume $|p(w)| > 0$. Then consider

$$q(z) \equiv \frac{p(z+w)}{p(w)}.$$

It follows $q(z)$ is of the form

$$q(z) = 1 + c_k z^k + \cdots + c_n z^n$$

where $c_k \neq 0$, because $q(0) = 1$. It is also true that $|q(z)| \geq 1$ by the assumption that $|p(w)|$ is the smallest value of $|p(z)|$. Now let $\theta \in \mathbb{C}$ be a complex number with $|\theta| = 1$ and

$$\theta c_k w^k = -|w|^k |c_k|.$$

If

$$w \neq 0, \theta = \frac{-\left|w^k\right| |c_k|}{w^k c_k}$$

and if $w = 0$, $\theta = 1$ will work. Now let $\eta^k = \theta$ and let t be a small positive number.

$$q(t\eta w) \equiv 1 - t^k |w|^k |c_k| + \cdots + c_n t^n (\eta w)^n$$

which is of the form

$$1 - t^k |w|^k |c_k| + t^k (g(t,w))$$

where $\lim_{t\to 0} g(t,w) = 0$. Letting t be small enough,

$$|g(t,w)| < |w|^k |c_k| /2$$

and so for such t,

$$|q(t\eta w)| < 1 - t^k |w|^k |c_k| + t^k |w|^k |c_k| /2 < 1,$$

a contradiction to $|q(z)| \geq 1$. This proves the theorem.

Bibliography

[1] **Apostol T.** *Calculus Volume II Second editioin,* Wiley *1969.*

[2] **Edwards C.H.** *Advanced Calculus of several Variables,* Dover 1994.

[3] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[4] **Hardy G.** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[5] **Hoffman K. and Kunze R**. *Linear Algebra* Prentice Hall 1971.

[6] **Horn R. and Johnson C.** *matrix Analysis,* Cambridge University Press, 1985.

[7] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes,* Academic Press, 1975.

[8] **Nobel B. and Daniel J**. *Applied Linear Algebra,* Prentice Hall 1977.

[9] **Rudin W.** *Principles of Mathematical Analysis*, McGraw Hill, 1976.