Kenneth Kuttler

Linear Algebra III

Advanced topics

bookboon.com
The eBook company

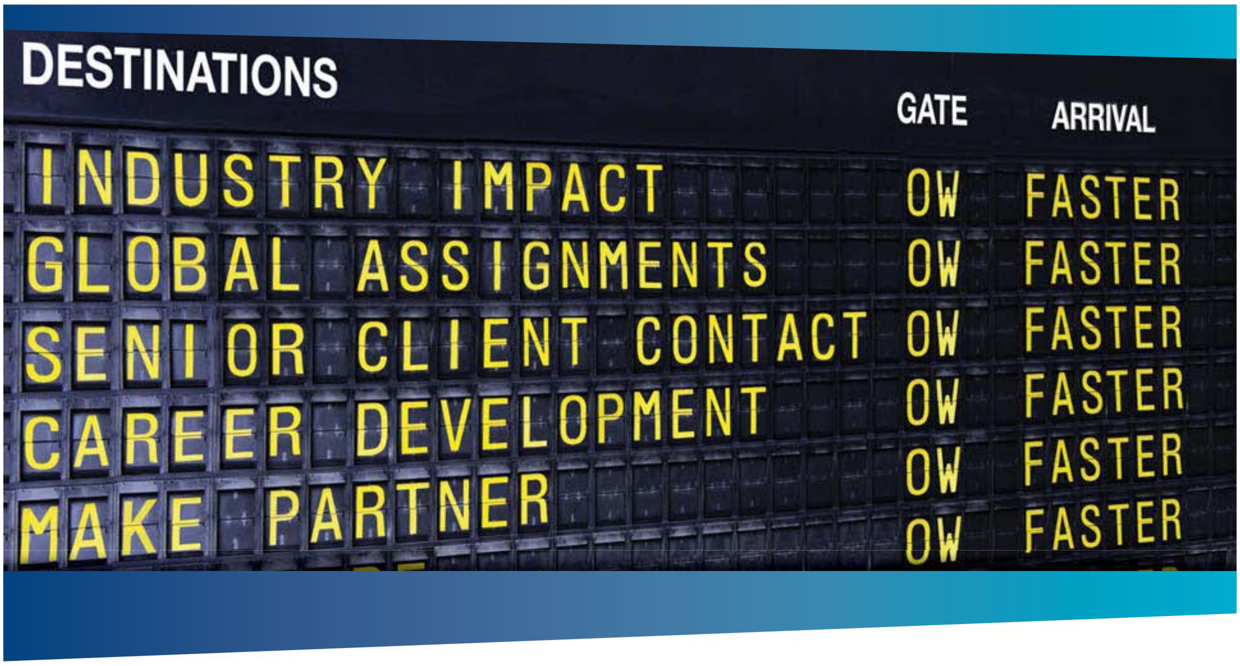
Kenneth Kuttler

Linear Algebra III Advanced topics

Linear Algebra III Advanced topics

ISBN 978-87-403-0242-4

Contents



Could you think of 101 new things
to do with eggs and oil?

To see Chapter 1-6 Download
Linear Algebra I Matrices and Row operations

To see Chapter 7-12 download
Linear Algebra II Spectral Theory and Abstract Vector Spaces

Self Adjoint Operators

13.1 Simultaneous Diagonalization

Recall the following definition of what it means for a matrix to be diagonalizable.

Definition 13.1.1 *Let A be an $n \times n$ matrix. It is said to be diagonalizable if there exists an invertible matrix S such that*

$$S^{-1}AS = D$$

where D is a diagonal matrix.

Also, here is a useful observation.

Observation 13.1.2 *If A is an $n \times n$ matrix and $AS = SD$ for D a diagonal matrix, then each column of S is an eigenvector or else it is the zero vector. This follows from observing that for $\mathbf{s}_k$ the k^{th} column of S and from the way we multiply matrices,*

$$A\mathbf{s}_k = \lambda_k \mathbf{s}_k$$

It is sometimes interesting to consider the problem of finding a single similarity transformation which will diagonalize all the matrices in some set.

Lemma 13.1.3 *Let A be an $n \times n$ matrix and let B be an $m \times m$ matrix. Denote by C the matrix*

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

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

Proof: Suppose $S_A^{-1}AS_A = D_A$ and $S_B^{-1}BS_B = D_B$ where D_A and D_B are diagonal matrices. You should use block multiplication to verify that $S \equiv \begin{pmatrix} S_A & 0 \\ 0 & S_B \end{pmatrix}$ is such that $S^{-1}CS = D_C$, a diagonal matrix.

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

$$\mathbf{s}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$$

where $\mathbf{x}_i \in \mathbb{F}^n$ and where $\mathbf{y}_i \in \mathbb{F}^m$. The result is

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where S_{11} is an $n \times n$ matrix and S_{22} is an $m \times m$ matrix. Then there is a diagonal matrix

$$D = diag\left(\lambda_1, \cdots, \lambda_{n+m}\right) = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

such that

$$\begin{aligned} & \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \\ = & \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \end{aligned}$$

Hence by block multiplication

$$AS_{11} = S_{11}D_1,\ BS_{22} = S_{22}D_2$$

$$BS_{21} = S_{21}D_1,\ AS_{12} = S_{12}D_2$$

It follows each of the $\mathbf{x}_i$ is an eigenvector of A or else is the zero vector and that each of the $\mathbf{y}_i$ is an eigenvector of B or is the zero vector. If there are n linearly independent $\mathbf{x}_i$, then A is diagonalizable by Theorem 9.3.12 on Page 9.3.12.

The row rank of the matrix $(\mathbf{x}_1, \cdots, \mathbf{x}_{n+m})$ must be n because if this is not so, the rank of S would be less than $n+m$ which would mean S^{-1} does not exist. Therefore, since the column rank equals the row rank, this matrix has column rank equal to n and this means there are n linearly independent eigenvectors of A implying that A is diagonalizable. Similar reasoning applies to B. ■

The following corollary follows from the same type of argument as the above.

Corollary 13.1.4 *Let A_k be an $n_k \times n_k$ matrix and let C denote the block diagonal*

$$\left(\sum_{k=1}^{r} n_k\right) \times \left(\sum_{k=1}^{r} n_k\right)$$

matrix given below.

$$C \equiv \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_r \end{pmatrix}.$$

Then C is diagonalizable if and only if each A_k is diagonalizable.

Definition 13.1.5 *A set, $\mathcal{F}$ of $n \times n$ matrices is said to be simultaneously diagonalizable if and only if there exists a single invertible matrix S such that for every $A \in \mathcal{F}$, $S^{-1}AS = D_A$ where D_A is a diagonal matrix.*

Lemma 13.1.6 *If $\mathcal{F}$ is a set of $n \times n$ matrices which is simultaneously diagonalizable, then $\mathcal{F}$ is a commuting family of matrices.*

Proof: Let $A, B \in \mathcal{F}$ and let S be a matrix which has the property that $S^{-1}AS$ is a diagonal matrix for all $A \in \mathcal{F}$. Then $S^{-1}AS = D_A$ and $S^{-1}BS = D_B$ where D_A and D_B are diagonal matrices. Since diagonal matrices commute,

$$\begin{aligned} AB &= SD_AS^{-1}SD_BS^{-1} = SD_AD_BS^{-1} \\ &= SD_BD_AS^{-1} = SD_BS^{-1}SD_AS^{-1} = BA. \end{aligned}$$

Lemma 13.1.7 *Let D be a diagonal matrix of the form*

$$D \equiv \begin{pmatrix} \lambda_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{n_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r I_{n_r} \end{pmatrix}, \tag{13.1}$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix and $\lambda_i \neq \lambda_j$ for $i \neq j$ and suppose B is a matrix which commutes with D. Then B is a block diagonal matrix of the form

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_r \end{pmatrix} \tag{13.2}$$

where B_i is an $n_i \times n_i$ matrix.

Proof: Let $B = (B_{ij})$ where $B_{ii} = B_i$ a block matrix as above in 13.2.

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1r} \\ B_{21} & B_{22} & \ddots & B_{2r} \\ \vdots & \ddots & \ddots & \vdots \\ B_{r1} & B_{r2} & \cdots & B_{rr} \end{pmatrix}$$

Then by block multiplication, since B is given to commute with D,

$$\lambda_j B_{ij} = \lambda_i B_{ij}$$

Therefore, if $i \neq j, B_{ij} = 0$. ■

Lemma 13.1.8 *Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathcal{F}$ is diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable.*

Proof: First note that if every matrix in $\mathcal{F}$ has only one eigenvalue, there is nothing to prove. This is because for A such a matrix,

$$S^{-1}AS = \lambda I$$

and so

$$A = \lambda I$$

Thus all the matrices in $\mathcal{F}$ are diagonal matrices and you could pick any S to diagonalize them all. Therefore, without loss of generality, assume some matrix in $\mathcal{F}$ has more than one eigenvalue.

The significant part of the lemma is proved by induction on n. If $n = 1$, there is nothing to prove because all the 1×1 matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n - 1$ where $n \geq 2$ and let $\mathcal{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathcal{F}$ which has more than one eigenvalue and let S be an invertible matrix such that $S^{-1}AS = D$ where D is of the form given in 13.1. By permuting the columns of S there is no loss of generality in assuming D has this form. Now denote by $\widetilde{\mathcal{F}}$ the collection of matrices, $\left\{S^{-1}CS : C \in \mathcal{F}\right\}$. Note $\widetilde{\mathcal{F}}$ features the single matrix S.

It follows easily that $\widetilde{\mathcal{F}}$ is also a commuting family of diagonalizable matrices. By Lemma 13.1.7 every $B \in \widetilde{\mathcal{F}}$ is of the form given in 13.2 because each of these commutes with D described above as $S^{-1}AS$ and so by block multiplication, the diagonal blocks B_i corresponding to different $B \in \widetilde{\mathcal{F}}$ commute.

By Corollary 13.1.4 each of these blocks is diagonalizable. This is because B is known to be so. Therefore, by induction, since all the blocks are no larger than $n-1 \times n-1$ thanks to the assumption that A has more than one eigenvalue, there exist invertible $n_i \times n_i$ matrices, T_i such that $T_i^{-1}B_iT_i$ is a diagonal matrix whenever B_i is one of the matrices making up the block diagonal of any $B \in \mathcal{F}$. It follows that for T defined by

$$T \equiv \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix},$$

then $T^{-1}BT =$ a diagonal matrix for every $B \in \widetilde{\mathcal{F}}$ including D. Consider ST. It follows that for all $C \in \mathcal{F}$,

$$T^{-1} \overbrace{S^{-1}CS}^{\text{something in } \widetilde{\mathcal{F}}} T = (ST)^{-1} C (ST) = \text{ a diagonal matrix.} \blacksquare$$

Theorem 13.1.9 *Let $\mathcal{F}$ denote a family of matrices which are diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if $\mathcal{F}$ is a commuting family.*

Proof: If $\mathcal{F}$ is a commuting family, it follows from Lemma 13.1.8 that it is simultaneously diagonalizable. If it is simultaneously diagonalizable, then it follows from Lemma 13.1.6 that it is a commuting family. ■

13.2 Schur's Theorem

Recall that for a linear transformation, $L \in \mathcal{L}(V, V)$ for V a finite dimensional inner product space, it could be represented in the form

$$L = \sum_{ij} l_{ij} \mathbf{v}_i \otimes \mathbf{v}_j$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is an orthonormal basis. Of course different bases will yield different matrices, (l_{ij}). Schur's theorem gives the existence of a basis in an inner product space such that (l_{ij}) is particularly simple.

Definition 13.2.1 *Let $L \in \mathcal{L}(V, V)$ where V is vector space. Then a subspace U of V is L invariant if $L(U) \subseteq U$.*

In what follows, $\mathbb{F}$ will be the field of scalars, usually $\mathbb{C}$ but maybe something else.

Theorem 13.2.2 *Let $L \in \mathcal{L}(H, H)$ for H a finite dimensional inner product space such that the restriction of $L^*$to every L invariant subspace has its eigenvalues in $\mathbb{F}$. Then there exist constants, c_{ij} for $i \leq j$ and an orthonormal basis, $\{\mathbf{w}_i\}_{i=1}^n$ such that*

$$L = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j$$

The constants, c_{ii} are the eigenvalues of L.

Proof: If $\dim(H) = 1$, let $H = \text{span}(\mathbf{w})$ where $|\mathbf{w}| = 1$. Then $L\mathbf{w} = k\mathbf{w}$ for some k. Then

$$L = k\mathbf{w} \otimes \mathbf{w}$$

because by definition, $\mathbf{w} \otimes \mathbf{w}(\mathbf{w}) = \mathbf{w}$. Therefore, the theorem holds if H is 1 dimensional.

Now suppose the theorem holds for $n - 1 = \dim(H)$. Let $\mathbf{w}_n$ be an eigenvector for L^*. Dividing by its length, it can be assumed $|\mathbf{w}_n| = 1$. Say $L^*\mathbf{w}_n = \mu\mathbf{w}_n$. Using the Gram Schmidt process, there exists an orthonormal basis for H of the form $\{\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}, \mathbf{w}_n\}$. Then

$$(L\mathbf{v}_k, \mathbf{w}_n) = (\mathbf{v}_k, L^*\mathbf{w}_n) = (\mathbf{v}_k, \mu\mathbf{w}_n) = 0,$$

which shows

$$L : H_1 \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}) \to \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}).$$

Denote by L_1 the restriction of L to H_1. Since H_1 has dimension $n-1$, the induction hypothesis yields an orthonormal basis, $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$ for H_1 such that

$$L_1 = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j. \tag{13.3}$$

Then $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ is an orthonormal basis for H because every vector in

$$\text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1})$$

has the property that its inner product with $\mathbf{w}_n$ is 0 so in particular, this is true for the vectors $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$. Now define c_{in} to be the scalars satisfying

$$L\mathbf{w}_n \equiv \sum_{i=1}^{n} c_{in} \mathbf{w}_i \tag{13.4}$$

and let

$$B \equiv \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \otimes \mathbf{w}_j.$$

Then by 13.4,

$$B\mathbf{w}_n = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{nj} = \sum_{j=1}^{n} c_{in} \mathbf{w}_i = L\mathbf{w}_n.$$

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

$$B\mathbf{w}_k = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{kj} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i$$

while from 13.3,

$$L\mathbf{w}_k = L_1 \mathbf{w}_k = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{jk} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i.$$

Since $L = B$ on the basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$, it follows $L = B$.

It remains to verify the constants, c_{kk} are the eigenvalues of L, solutions of the equation, $\det(\lambda I - L) = 0$. However, the definition of $\det(\lambda I - L)$ is the same as

$$\det(\lambda I - C)$$

where C is the upper triangular matrix which has c_{ij} for $i \leq j$ and zeros elsewhere. This equals 0 if and only if λ is one of the diagonal entries, one of the c_{kk}. ■

Now with the above Schur's theorem, the following diagonalization theorem comes very easily. Recall the following definition.

Definition 13.2.3 *Let $L \in \mathcal{L}(H, H)$ where H is a finite dimensional inner product space. Then L is Hermitian if $L^* = L$.*

Theorem 13.2.4 *Let $L \in \mathcal{L}(H, H)$ where H is an n dimensional inner product space. If L is Hermitian, then all of its eigenvalues λ_k are real and there exists an orthonormal basis of eigenvectors $\{\mathbf{w}_k\}$ such that*

$$L = \sum_k \lambda_k \mathbf{w}_k \otimes \mathbf{w}_k.$$

Proof: By Schur's theorem, Theorem 13.2.2, there exist $l_{ij} \in \mathbb{F}$ such that

$$L = \sum_{j=1}^{n} \sum_{i=1}^{j} l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j$$

Then by Lemma 12.4.2,

$$\begin{aligned} \sum_{j=1}^{n} \sum_{i=1}^{j} l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j &= L = L^* = \sum_{j=1}^{n} \sum_{i=1}^{j} \left(l_{ij} \mathbf{w}_i \otimes \mathbf{w}_j \right)^* \\ &= \sum_{j=1}^{n} \sum_{i=1}^{j} \overline{l_{ij}} \mathbf{w}_j \otimes \mathbf{w}_i = \sum_{i=1}^{n} \sum_{j=1}^{i} \overline{l_{ji}} \mathbf{w}_i \otimes \mathbf{w}_j \end{aligned}$$

By independence, if $i = j$,

$$l_{ii} = \overline{l_{ii}}$$

and so these are all real. If $i < j$, it follows from independence again that

$$l_{ij} = 0$$

because the coefficients corresponding to $i < j$ are all 0 on the right side. Similarly if $i > j$, it follows $l_{ij} = 0$. Letting $\lambda_k = l_{kk}$, this shows

$$L = \sum_k \lambda_k \mathbf{w}_k \otimes \mathbf{w}_k$$

That each of these $\mathbf{w}_k$ is an eigenvector corresponding to λ_k is obvious from the definition of the tensor product. ■

13.3 Spectral Theory Of Self Adjoint Operators

The following theorem is about the eigenvectors and eigenvalues of a self adjoint operator. Such operators are also called Hermitian as in the case of matrices. The proof given generalizes to the situation of a compact self adjoint operator on a Hilbert space and leads to many very useful results. It is also a very elementary proof because it does not use the fundamental theorem of algebra and it contains a way, very important in applications, of finding the eigenvalues. This proof depends more directly on the methods of analysis than the preceding material. The field of scalars will be $\mathbb{R}$ or $\mathbb{C}$. The following is useful notation.

Definition 13.3.1 *Let X be an inner product space and let $S \subseteq X$. Then*

$$S^{\perp} \equiv \{x \in X : (x, s) = 0 \text{ for all } s \in S\}.$$

Note that even if S is not a subspace, $S^{\perp}$ is.

Definition 13.3.2 *A Hilbert space is a complete inner product space. Recall this means that every Cauchy sequence,$\{x_n\}$, one which satisfies*

$$\lim_{n,m \to \infty} |x_n - x_m| = 0,$$

converges. It can be shown, although I will not do so here, that for the field of scalars either $\mathbb{R}$ or $\mathbb{C}$, any finite dimensional inner product space is automatically complete.

Theorem 13.3.3 *Let $A \in \mathcal{L}(X, X)$ be self adjoint (Hermitian) where X is a finite dimensional Hilbert space. Thus $A = A^*$. Then there exists an orthonormal basis of eigenvectors, $\{u_j\}_{j=1}^n$.*

Proof: Consider (Ax, x). This quantity is always a real number because

$$\overline{(Ax, x)} = (x, Ax) = (x, A^* x) = (Ax, x)$$

thanks to the assumption that A is self adjoint. Now define

$$\lambda_1 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_1 \equiv X\}.$$

Claim: λ_1 is finite and there exists $v_1 \in X$ with $|v_1| = 1$ such that $(Av_1, v_1) = \lambda_1$.

Proof of claim: Let $\{u_j\}_{j=1}^n$ be an orthonormal basis for X and for $x \in X$, let $(x_1, \cdots, x_n)$ be defined as the components of the vector x. Thus,

$$x = \sum_{j=1}^n x_j u_j.$$

Since this is an orthonormal basis, it follows from the axioms of the inner product that

$$|x|^2 = \sum_{j=1}^n |x_j|^2.$$

Thus

$$(Ax, x) = \left(\sum_{k=1}^n x_k A u_k, \sum_{j=1} x_j u_j \right) = \sum_{k,j} x_k \overline{x_j} (Au_k, u_j),$$

a real valued continuous function of $(x_1, \cdots, x_n)$ which is defined on the compact set

$$K \equiv \{(x_1, \cdots, x_n) \in \mathbb{F}^n : \sum_{j=1}^n |x_j|^2 = 1\}.$$

Therefore, it achieves its minimum from the extreme value theorem. Then define

$$v_1 \equiv \sum_{j=1}^n x_j u_j$$

where $(x_1, \cdots, x_n)$ is the point of K at which the above function achieves its minimum. This proves the claim.

Continuing with the proof of the theorem, let $X_2 \equiv \{v_1\}^{\perp}$. This is a closed subspace of X. Let

$$\lambda_2 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_2\}$$

As before, there exists $v_2 \in X_2$ such that $(Av_2, v_2) = \lambda_2, \lambda_1 \leq \lambda_2$. Now let $X_3 \equiv \{v_1, v_2\}^{\perp}$ and continue in this way. This leads to an increasing sequence of real numbers, $\{\lambda_k\}_{k=1}^n$ and an orthonormal set of vectors, $\{v_1, \cdots, v_n\}$. It only remains to show these are eigenvectors and that the λ_j are eigenvalues.

Consider the first of these vectors. Letting $w \in X_1 \equiv X$, the function of the real variable, t, given by

$$f(t) \equiv \frac{(A(v_1 + tw), v_1 + tw)}{|v_1 + tw|^2}$$

$$= \frac{(Av_1, v_1) + 2t \operatorname{Re}(Av_1, w) + t^2 (Aw, w)}{|v_1|^2 + 2t \operatorname{Re}(v_1, w) + t^2 |w|^2}$$

achieves its minimum when $t = 0$. Therefore, the derivative of this function evaluated at $t = 0$ must equal zero. Using the quotient rule, this implies, since $|v_1| = 1$ that

$$2\operatorname{Re}(Av_1, w)|v_1|^2 - 2\operatorname{Re}(v_1, w)(Av_1, v_1)$$

$$= 2(\operatorname{Re}(Av_1, w) - \operatorname{Re}(v_1, w)\lambda_1) = 0.$$

Thus $\operatorname{Re}(Av_1 - \lambda_1 v_1, w) = 0$ for all $w \in X$. This implies $Av_1 = \lambda_1 v_1$. To see this, let $w \in X$ be arbitrary and let θ be a complex number with $|\theta| = 1$ and

$$|(Av_1 - \lambda_1 v_1, w)| = \theta(Av_1 - \lambda_1 v_1, w).$$

Then

$$|(Av_1 - \lambda_1 v_1, w)| = \operatorname{Re}\left(Av_1 - \lambda_1 v_1, \overline{\theta} w\right) = 0.$$

Since this holds for all w, $Av_1 = \lambda_1 v_1$.

Now suppose $Av_k = \lambda_k v_k$ for all $k < m$. Observe that $A : X_m \to X_m$ because if $y \in X_m$ and $k < m$,

$$(Ay, v_k) = (y, Av_k) = (y, \lambda_k v_k) = 0,$$

showing that $Ay \in \{v_1, \cdots, v_{m-1}\}^{\perp} \equiv X_m$. Thus the same argument just given shows that for all $w \in X_m$,

$$(Av_m - \lambda_m v_m, w) = 0. \tag{13.5}$$

Since $Av_m \in X_m$, I can let $w = Av_m - \lambda_m v_m$ in the above and thereby conclude $Av_m = \lambda_m v_m$. ■

Contained in the proof of this theorem is the following important corollary.

Corollary 13.3.4 *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 exists an orthonormal set of vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \inf\{(Ax, x) : |x| = 1, x \in X_k\}$$

where

$$X_k \equiv \{u_1, \cdots, u_{k-1}\}^{\perp}, X_1 \equiv X.$$

Corollary 13.3.5 *Let $A \in \mathcal{L}(X, X)$ be self adjoint (Hermitian) where X is a finite dimensional Hilbert space. Then the largest eigenvalue of A is given by*

$$\max\{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\} \tag{13.6}$$

and the minimum eigenvalue of A is given by

$$\min\{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\}. \tag{13.7}$$

Proof: The proof of this is just like the proof of Theorem 13.3.3. Simply replace inf with sup and obtain a decreasing list of eigenvalues. This establishes 13.6. The claim 13.7 follows from Theorem 13.3.3.

Another important observation is found in the following corollary.

Corollary 13.3.6 *Let $A \in \mathcal{L}(X, X)$ where A is self adjoint. Then $A = \sum_i \lambda_i v_i \otimes v_i$ where $Av_i = \lambda_i v_i$ and $\{v_i\}_{i=1}^n$ is an orthonormal basis.*

Proof : If v_k is one of the orthonormal basis vectors, $Av_k = \lambda_k v_k$. Also,

$$\begin{aligned}\sum_i \lambda_i v_i \otimes v_i (v_k) &= \sum_i \lambda_i v_i (v_k, v_i) \\ &= \sum_i \lambda_i \delta_{ik} v_i = \lambda_k v_k.\end{aligned}$$

Since the two linear transformations agree on a basis, it follows they must coincide. ■

By Theorem 12.4.5 this says the matrix of A with respect to this basis $\{v_i\}_{i=1}^n$ is the diagonal matrix having the eigenvalues $\lambda_1, \cdots, \lambda_n$ down the main diagonal.

The result of Courant and Fischer which follows resembles Corollary 13.3.4 but is more useful because it does not depend on a knowledge of the eigenvectors.

Theorem 13.3.7 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1, \cdots, w_{k-1}} \left\{ \min \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^\perp \right\} \right\} \tag{13.8}$$

where if $k = 1, \{w_1, \cdots, w_{k-1}\}^\perp \equiv X$.

Proof: From Theorem 13.3.3, there exist eigenvalues and eigenvectors with $\{u_1, \cdots, u_n\}$ orthonormal and $\lambda_i \le \lambda_{i+1}$. Therefore, by Corollary 13.3.6

$$A = \sum_{j=1}^n \lambda_j u_j \otimes u_j$$

Fix $\{w_1, \cdots, w_{k-1}\}$.

$$(Ax, x) = \sum_{j=1}^n \lambda_j (x, u_j)(u_j, x) = \sum_{j=1}^n \lambda_j |(x, u_j)|^2$$

Then let $Y = \{w_1, \cdots, w_{k-1}\}^\perp$

$$\inf \{(Ax, x) : |x| = 1, x \in Y\} = \inf \left\{ \sum_{j=1}^n \lambda_j |(x, u_j)|^2 : |x| = 1, x \in Y \right\}$$

$$\le \inf \left\{ \sum_{j=1}^k \lambda_j |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\}. \tag{13.9}$$

The reason this is so is that the infimum is taken over a smaller set. Therefore, the infimum gets larger. Now 13.9 is no larger than

$$\inf \left\{ \lambda_k \sum_{j=1}^k |(x, u_j)|^2 : |x| = 1, (x, u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\} = \lambda_k$$

because since $\{u_1, \cdots, u_n\}$ is an orthonormal basis, $|x|^2 = \sum_{j=1}^n |(x, u_j)|^2$. It follows since $\{w_1, \cdots, w_{k-1}\}$ is arbitrary,

$$\sup_{w_1, \cdots, w_{k-1}} \left\{ \inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^\perp \right\} \right\} \leq \lambda_k. \tag{13.10}$$

However, for each $w_1, \cdots, w_{k-1}$, the infimum is achieved so you can replace the inf in the above with min. In addition to this, it follows from Corollary 13.3.4 that there exists a set, $\{w_1, \cdots, w_{k-1}\}$ for which

$$\inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^\perp \right\} = \lambda_k.$$

Pick $\{w_1, \cdots, w_{k-1}\} = \{u_1, \cdots, u_{k-1}\}$. Therefore, the sup in 13.10 is achieved and equals λ_k and 13.8 follows. ■

The following corollary is immediate.

Corollary 13.3.8 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1, \cdots, w_{k-1}} \left\{ \min \left\{ \frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{k-1}\}^\perp \right\} \right\} \tag{13.11}$$

where if $k = 1, \{w_1, \cdots, w_{k-1}\}^\perp \equiv X$.

Here is a version of this for which the roles of max and min are reversed.

Corollary 13.3.9 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \min_{w_1, \cdots, w_{n-k}} \left\{ \max \left\{ \frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{n-k}\}^\perp \right\} \right\} \tag{13.12}$$

where if $k = n, \{w_1, \cdots, w_{n-k}\}^\perp \equiv X$.

13.4 Positive And Negative Linear Transformations

The notion of a positive definite or negative definite linear transformation is very important in many applications. In particular it is used in versions of the second derivative test for functions of many variables. Here the main interest is the case of a linear transformation which is an $n \times n$ matrix but the theorem is stated and proved using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

Lemma 13.4.1 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X, X)$. Then if $\{v_1, \cdots, v_n\}$ is an orthonormal basis for X and $M(A)$ denotes the matrix of the linear transformation A then $M(A^*) = (M(A))^*$. In particular, A is self adjoint, if and only if $M(A)$ is.*

Proof: Consider the following picture

$$\begin{tikzcd} X \arrow[r, "A"] & X \\ \mathbb{F}^n \arrow[u, "q"] \arrow[r, "M(A)"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^n \arrow[u, "q"'] \end{tikzcd}$$

where q is the coordinate map which satisfies $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Therefore, since $\{v_1, \cdots, v_n\}$ is orthonormal, it is clear that $|\mathbf{x}| = |q(\mathbf{x})|$. Therefore,

$$\begin{aligned} |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\operatorname{Re}(\mathbf{x}, \mathbf{y}) &= |\mathbf{x} + \mathbf{y}|^2 = |q(\mathbf{x} + \mathbf{y})|^2 \\ &= |q(\mathbf{x})|^2 + |q(\mathbf{y})|^2 + 2\operatorname{Re}(q(\mathbf{x}), q(\mathbf{y})) \end{aligned} \tag{13.13}$$

Now in any inner product space,

$$(x, iy) = \operatorname{Re}(x, iy) + i\operatorname{Im}(x, iy).$$

Also

$$(x, iy) = (-i)(x, y) = (-i)\operatorname{Re}(x, y) + \operatorname{Im}(x, y).$$

Therefore, equating the real parts, $\operatorname{Im}(x, y) = \operatorname{Re}(x, iy)$ and so

$$(x, y) = \operatorname{Re}(x, y) + i\operatorname{Re}(x, iy) \tag{13.14}$$

Now from 13.13, since q preserves distances, $.\operatorname{Re}(q(\mathbf{x}), q(\mathbf{y})) = \operatorname{Re}(\mathbf{x}, \mathbf{y})$ which implies from 13.14 that

$$(\mathbf{x}, \mathbf{y}) = (q(\mathbf{x}), q(\mathbf{y})). \tag{13.15}$$

Now consulting the diagram which gives the meaning for the matrix of a linear transformation, observe that $q \circ M(A) = A \circ q$ and $q \circ M(A^*) = A^* \circ q$. Therefore, from 13.15

$$(A(q(\mathbf{x})), q(\mathbf{y})) = (q(\mathbf{x}), A^* q(\mathbf{y})) = (q(\mathbf{x}), q(M(A^*)(\mathbf{y}))) = (\mathbf{x}, M(A^*)(\mathbf{y}))$$

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 $A = A^*$. ■

The following corollary is one of the items in the above proof.

Corollary 13.4.2 *Let X be a finite dimensional Hilbert space and let $\{v_1, \cdots, v_n\}$ be an orthonormal basis for X. Also, let q be the coordinate map associated with this basis satisfying $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Then $(\mathbf{x}, \mathbf{y})_{\mathbb{F}^n} = (q(\mathbf{x}), q(\mathbf{y}))_X$. Also, if $A \in \mathcal{L}(X, X)$, and $M(A)$ is the matrix of A with respect to this basis,*

$$(Aq(\mathbf{x}), q(\mathbf{y}))_X = (M(A)\mathbf{x}, \mathbf{y})_{\mathbb{F}^n}.$$

Definition 13.4.3 *A self adjoint $A \in \mathcal{L}(X, X)$, is positive definite if whenever $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x}, \mathbf{x}) > 0$ and A is negative definite if for all $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x}, \mathbf{x}) < 0$. A is positive semidefinite or just nonnegative for short if for all $\mathbf{x}$, $(A\mathbf{x}, \mathbf{x}) \geq 0$. A is negative semidefinite or nonpositive for short if for all $\mathbf{x}$, $(A\mathbf{x}, \mathbf{x}) \leq 0$.*

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

Lemma 13.4.4 *Let X be a finite dimensional Hilbert space. A self adjoint $A \in \mathcal{L}(X, X)$ is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.*

Proof: Suppose first that A is positive definite and let λ be an eigenvalue. Then for $\mathbf{x}$ an eigenvector corresponding to λ, $\lambda(\mathbf{x}, \mathbf{x}) = (\lambda\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, \mathbf{x}) > 0$. Therefore, $\lambda > 0$ as claimed.

Now suppose all the eigenvalues of A are positive. From Theorem 13.3.3 and Corollary 13.3.6, $A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i$ where the λ_i are the positive eigenvalues and $\{\mathbf{u}_i\}$ are an orthonormal set of eigenvectors. Therefore, letting $\mathbf{x} \neq \mathbf{0}$,

$$\begin{aligned}(A\mathbf{x}, \mathbf{x}) &= \left(\left(\sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i\right)\mathbf{x}, \mathbf{x}\right) = \left(\sum_{i=1}^{n} \lambda_i \mathbf{u}_i (\mathbf{x}, \mathbf{u}_i), \mathbf{x}\right) \\ &= \left(\sum_{i=1}^{n} \lambda_i (\mathbf{x}, \mathbf{u}_i)(\mathbf{u}_i, \mathbf{x})\right) = \sum_{i=1}^{n} \lambda_i \left|(\mathbf{u}_i, \mathbf{x})\right|^2 > 0\end{aligned}$$

because, since $\{\mathbf{u}_i\}$ is an orthonormal basis, $|\mathbf{x}|^2 = \sum_{i=1}^{n} |(\mathbf{u}_i, \mathbf{x})|^2$.

To establish the claim about negative definite, it suffices to note that A is negative definite if and only if $-A$ is positive definite and the eigenvalues of A are (-1) times the eigenvalues of $-A$. The claims about positive semidefinite and negative semidefinite are obtained similarly. ■

The next theorem is about a way to recognize whether a self adjoint $A \in \mathcal{L}(X, X)$ is positive or negative definite without having to find the eigenvalues. In order to state this theorem, here is some notation.

Definition 13.4.5 *Let A be an $n \times n$ matrix. Denote by A_k the $k \times k$ matrix obtained by deleting the $k+1, \cdots, n$ columns and the $k+1, \cdots, n$ rows from A. Thus $A_n = A$ and A_k is the $k \times k$ submatrix of A which occupies the upper left corner of A. The determinants of these submatrices are called the principle minors.*

The following theorem is proved in [8]

Theorem 13.4.6 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X, X)$ be self adjoint. Then A is positive definite if and only if $\det\left(M(A)_k\right) > 0$ for every $k = 1, \cdots, n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: This theorem is proved by induction on n. It is clearly true if $n = 1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\det(M(A)) > 0$, it follows that all the eigenvalues are nonzero. Are they all positive? Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\det(M(A))$. Pick two, λ_1 and λ_2 and let $M(A)\mathbf{u}_i = \lambda_i \mathbf{u}_i$ where $\mathbf{u}_i \neq \mathbf{0}$ for $i = 1, 2$ and $(\mathbf{u}_1, \mathbf{u}_2) = 0$. Now if $\mathbf{y} \equiv \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2$ is an element of span $(\mathbf{u}_1, \mathbf{u}_2)$, then since these are eigenvalues and $(\mathbf{u}_1, \mathbf{u}_2) = 0$, a short computation shows

$$\begin{gathered}(M(A)(\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2), \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2) \\ = |\alpha_1|^2 \lambda_1 |\mathbf{u}_1|^2 + |\alpha_2|^2 \lambda_2 |\mathbf{u}_2|^2 < 0.\end{gathered}$$

Now letting $\mathbf{x} \in \mathbb{C}^{n-1}$, the induction hypothesis implies

$$(\mathbf{x}^*, 0) M(A) \begin{pmatrix} \mathbf{x} \\ 0 \end{pmatrix} = \mathbf{x}^* M(A)_{n-1} \mathbf{x} = (M(A)\mathbf{x}, \mathbf{x}) > 0.$$

Now the dimension of $\{\mathbf{z} \in \mathbb{C}^n : z_n = 0\}$ is $n-1$ and the dimension of 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 13.4.7 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X, X)$ be self adjoint. Then A is negative definite if and only if $\det(M(A)_k)(-1)^k > 0$ for every $k = 1, \cdots, n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: This is immediate from the above theorem by noting that, as in the proof of Lemma 13.4.4, A is negative definite if and only if $-A$ is positive definite. Therefore, if $\det(-M(A)_k) > 0$ for all $k = 1, \cdots, n$, it follows that A is negative definite. However, $\det(-M(A)_k) = (-1)^k \det(M(A)_k)$. ■

13.5 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. To begin with, consider the square root of a nonnegative self adjoint operator. This is easier than the general theory and it is the square root which is of most importance.

Theorem 13.5.1 *Let $A \in \mathcal{L}(X, X)$ be self adjoint and nonnegative. Then there exists a unique self adjoint nonnegative $B \in \mathcal{L}(X, X)$ such that $B^2 = A$ and B commutes with every element of $\mathcal{L}(X, X)$ which commutes with A.*

Proof: By Theorem 13.3.3, 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 Theorem 13.2.4, $A = \sum_i \lambda_i v_i \otimes v_i$ where each $\lambda_i \geq 0$.

Now by Lemma 13.4.4, each $\lambda_i \geq 0$. Therefore, it makes sense to define

$$B \equiv \sum_i \lambda_i^{1/2} v_i \otimes v_i.$$

It is easy to verify that

$$(v_i \otimes v_i)(v_j \otimes v_j) = \begin{cases} 0 \text{ if } i \neq j \\ v_i \otimes v_i \text{ if } i = j \end{cases} .$$

Therefore, a short computation verifies that $B^2 = \sum_i \lambda_i v_i \otimes v_i = A$. If C commutes with A, then for some c_{ij},

$$C = \sum_{ij} c_{ij} v_i \otimes v_j$$

and so since they commute,

$$\sum_{i,j,k} c_{ij} v_i \otimes v_j \lambda_k v_k \otimes v_k = \sum_{i,j,k} c_{ij} \lambda_k \delta_{jk} v_i \otimes v_k = \sum_{i,k} c_{ik} \lambda_k v_i \otimes v_k$$

$$
\begin{aligned}
&= \sum_{i,j,k} c_{ij}\lambda_k v_k \otimes v_k v_i \otimes v_j = \sum_{i,j,k} c_{ij}\lambda_k \delta_{ki} v_k \otimes v_j = \sum_{j,k} c_{kj}\lambda_k v_k \otimes v_j \\
&= \sum_{k,i} c_{ik}\lambda_i v_i \otimes v_k
\end{aligned}
$$

Then by independence,

$$c_{ik}\lambda_i = c_{ik}\lambda_k$$

Therefore, $c_{ik}\lambda_i^{1/2} = c_{ik}\lambda_k^{1/2}$ which amounts to saying that B also commutes with C. It is clear that this operator is self adjoint. This proves existence.

Suppose B_1 is another square root which is self adjoint, nonnegative and commutes with every matrix which commutes with A. Since both B, B_1 are nonnegative,

$$
\begin{aligned}
&(B(B-B_1)x, (B-B_1)x) \geq 0, \\
&(B_1(B-B_1)x, (B-B_1)x) \geq 0 \qquad (13.16)
\end{aligned}
$$

Now, adding these together, and using the fact that the two commute,

$$\left(\left(B^2 - B_1^2\right)x, (B-B_1)x\right) = ((A-A)x, (B-B_1)x) = 0.$$

It follows that both inner products in 13.16 equal 0. Next use the existence part of this to take the square root of B and B_1 which is denoted by $\sqrt{B}, \sqrt{B_1}$ respectively. Then

$$
\begin{aligned}
0 &= \left(\sqrt{B}(B-B_1)x, \sqrt{B}(B-B_1)x\right) \\
0 &= \left(\sqrt{B_1}(B-B_1)x, \sqrt{B_1}(B-B_1)x\right)
\end{aligned}
$$

which implies $\sqrt{B}(B-B_1)x = \sqrt{B_1}(B-B_1)x = 0$. Thus also,

$$B(B-B_1)x = B_1(B-B_1)x = 0$$

Hence

$$0 = (B(B-B_1)x - B_1(B-B_1)x, x) = ((B-B_1)x, (B-B_1)x)$$

and so, since x is arbitrary, $B_1 = B$. ■

The main result is the following theorem.

Theorem 13.5.2 *Let $A \in \mathcal{L}(X,X)$ be self adjoint and nonnegative and let k be a positive integer. Then there exists a unique self adjoint nonnegative $B \in \mathcal{L}(X,X)$ such that $B^k = A$.*

Proof: By Theorem 13.3.3, 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 13.3.6 or Theorem 13.2.4, $A = \sum_i \lambda_i v_i \otimes v_i$ where each $\lambda_i \geq 0$.

Now by Lemma 13.4.4, each $\lambda_i \geq 0$. Therefore, it makes sense to define

$$B \equiv \sum_i \lambda_i^{1/k} v_i \otimes v_i.$$

It is easy to verify that

$$(v_i \otimes v_i)(v_j \otimes v_j) = \begin{cases} 0 \text{ if } i \neq j \\ v_i \otimes v_i \text{ if } i = j \end{cases}.$$

Therefore, a short computation verifies that $B^k = \sum_i \lambda_i v_i \otimes v_i = A$. This proves existence.

In order to prove uniqueness, let $p(t)$ be a polynomial which has the property that $p(\lambda_i) = \lambda_i^{1/k}$ for each i. In other words, goes through the ordered pairs $\left(\lambda_i, \lambda_i^{1/k}\right)$. Then a similar short computation shows

$$p(A) = \sum_i p(\lambda_i) v_i \otimes v_i = \sum_i \lambda_i^{1/k} v_i \otimes v_i = B.$$

Now suppose $C^k = A$ where $C \in \mathcal{L}(X, X)$ is self adjoint and nonnegative. Then

$$CB = Cp(A) = Cp\left(C^k\right) = p\left(C^k\right)C = p(A)C = BC.$$

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 13.4.1). See the diagram for a short verification of the claim the two matrices commute..

$$\begin{tikzcd}
X \arrow[r, "B"] & X \arrow[r, "C"] & X \\
\mathbb{F}^n \arrow[u, "q"] \arrow[r, "M(B)"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^n \arrow[u, "q"'] \arrow[r, "M(C)"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^n \arrow[u, "q"']
\end{tikzcd}$$

Therefore, by Theorem 13.1.9, these two matrices can be simultaneously diagonalized. Thus

$$U^{-1}M(B)U = D_1,\ U^{-1}M(C)U = D_2 \tag{13.17}$$

where the D_i is a diagonal matrix consisting of the eigenvalues of B or C. Also it is clear that

$$M(C)^k = M(A)$$

because $M(C)^k$ is given by

$$\overbrace{q^{-1}Cqq^{-1}Cq\cdots q^{-1}Cq}^{k \text{ times}} = q^{-1}C^kq = q^{-1}Aq = M(A)$$

and similarly

$$M(B)^k = M(A).$$

Then raising these to powers,

$$U^{-1}M(A)U = U^{-1}M(B)^k U = D_1^k$$

and

$$U^{-1}M(A)U = U^{-1}M(C)^k U = D_2^k.$$

Therefore, $D_1^k = D_2^k$ and since the diagonal entries of D_i are nonnegative, this requires that $D_1 = D_2$. Therefore, from 13.17, $M(B) = M(C)$ and so $B = C$. ■

13.6 Polar Decompositions

An application of Theorem 13.3.3, is the following fundamental result, important in geometric measure theory and continuum mechanics. It is sometimes called the right polar decomposition. The notation used is that which is seen in continuum mechanics, see for example Gurtin [11]. Don't confuse the U in this theorem with a unitary transformation. It is not so. When the following theorem is applied in continuum mechanics, F is normally the deformation gradient, the derivative of a nonlinear map from some subset of three dimensional space to three dimensional space. In this context, U is called the right Cauchy Green strain tensor. It is a measure of how a body is stretched independent of rigid motions. First, here is a simple lemma.

Lemma 13.6.1 *Suppose $R \in \mathcal{L}(X,Y)$ where X, Y are Hilbert spaces and R preserves distances. Then $R^*R = I$.*

Proof: Since R preserves distances, $|R\mathbf{x}| = |\mathbf{x}|$ for every $\mathbf{x}$. Therefore from the axioms of the inner product,

$$|\mathbf{x}|^2 + |\mathbf{y}|^2 + (\mathbf{x},\mathbf{y}) + (\mathbf{y},\mathbf{x}) = |\mathbf{x}+\mathbf{y}|^2 = (R(\mathbf{x}+\mathbf{y}), R(\mathbf{x}+\mathbf{y}))$$

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

$$= |\mathbf{x}|^2 + |\mathbf{y}|^2 + (R^*R\mathbf{x},\mathbf{y}) + (\mathbf{y},R^*R\mathbf{x})$$

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

$$(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) + (\mathbf{y},R^*R\mathbf{x}-\mathbf{x}) = 0$$

Hence for all $\mathbf{x},\mathbf{y}$,

$$\operatorname{Re}(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) = 0$$

Now for $\mathbf{x},\mathbf{y}$ given, choose $\alpha \in \mathbb{C}$ such that

$$\alpha(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) = |(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})|$$

Then

$$\begin{aligned} 0 &= \operatorname{Re}(R^*R\mathbf{x}-\mathbf{x},\overline{\alpha}\mathbf{y}) = \operatorname{Re}\alpha(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) \\ &= |(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})| \end{aligned}$$

Thus $|(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})| = 0$ for all $\mathbf{x},\mathbf{y}$ because the given $\mathbf{x},\mathbf{y}$ were arbitrary. Let $\mathbf{y} = R^*R\mathbf{x}-\mathbf{x}$ to conclude that for all $\mathbf{x}$,

$$R^*R\mathbf{x}-\mathbf{x} = \mathbf{0}$$

which says $R^*R = I$ since $\mathbf{x}$ is arbitrary. ■

The decomposition in the following is called the right polar decomposition.

Theorem 13.6.2 *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*

$$F = RU,\ U = U^*, (U \text{ is Hermitian}),$$

all eigenvalues of U are non negative,

$$U^2 = F^*F, R^*R = I,$$

and $|R\mathbf{x}| = |\mathbf{x}|$.

Proof: $(F^*F)^* = F^*F$ and so by Theorem 13.3.3, there is an orthonormal basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$F^*F\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ F^*F = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i.$$

It is also clear that $\lambda_i \geq 0$ because

$$\lambda_i (\mathbf{v}_i, \mathbf{v}_i) = (F^*F\mathbf{v}_i, \mathbf{v}_i) = (F\mathbf{v}_i, F\mathbf{v}_i) \geq 0.$$

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.

Let $\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r\}$ be an orthonormal basis for $U(X)$. By the Gram Schmidt procedure there exists an extension to an orthonormal basis for X,

$$\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r, \mathbf{y}_{r+1}, \cdots, \mathbf{y}_n\}.$$

Next note that $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ is also an orthonormal set of vectors in Y because

$$(F\mathbf{x}_k, F\mathbf{x}_j) = (F^*F\mathbf{x}_k, \mathbf{x}_j) = \left(U^2\mathbf{x}_k, \mathbf{x}_j\right) = (U\mathbf{x}_k, U\mathbf{x}_j) = \delta_{jk}.$$

By the Gram Schmidt procedure, there exists an extension of $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ to an orthonormal basis for Y,

$$\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r, \mathbf{z}_{r+1}, \cdots, \mathbf{z}_m\}.$$

Since $m \geq n$, there are at least as many $\mathbf{z}_k$ as there are $\mathbf{y}_k$. Now for $\mathbf{x} \in X$, since

$$\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r, \mathbf{y}_{r+1}, \cdots, \mathbf{y}_n\}$$

is an orthonormal basis for X, there exist unique scalars

$$c_1, \cdots, c_r, d_{r+1}, \cdots, d_n$$

such that

$$\mathbf{x} = \sum_{k=1}^{r} c_k U\mathbf{x}_k + \sum_{k=r+1}^{n} d_k \mathbf{y}_k$$

Define

$$R\mathbf{x} \equiv \sum_{k=1}^{r} c_k F\mathbf{x}_k + \sum_{k=r+1}^{n} d_k \mathbf{z}_k \tag{13.18}$$

Thus

$$|R\mathbf{x}|^2 = \sum_{k=1}^{r} |c_k|^2 + \sum_{k=r+1}^{n} |d_k|^2 = |\mathbf{x}|^2.$$

Therefore, by Lemma 13.6.1 $R^*R = I$.

Then also there exist scalars b_k such that

$$U\mathbf{x} = \sum_{k=1}^{r} b_k U\mathbf{x}_k \tag{13.19}$$

and so from 13.18,

$$RU\mathbf{x} = \sum_{k=1}^{r} b_k F\mathbf{x}_k = F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right)$$

Is $F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$?

$$\begin{aligned}
&\left(F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x}), F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x})\right) \\
&= \left((F^*F)\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x}, \sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x}\right) = 0
\end{aligned}$$

Because from 13.19, $U\mathbf{x} = \sum_{k=1}^{r} b_k U\mathbf{x}_k$. Therefore, $RU\mathbf{x} = F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$. ■

The following corollary follows as a simple consequence of this theorem. It is called the left polar decomposition.

Corollary 13.6.3 *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 Hermitian $U \in \mathcal{L}(X,X)$, and an element of $\mathcal{L}(X,Y)$, R, such that*

$$F = UR,\ RR^* = I.$$

Proof: Recall that $L^{**} = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 13.6.2 to $F^* \in \mathcal{L}(Y,X)$. Thus,

$$F^* = R^*U$$

where R^* and U satisfy the conditions of that theorem. Then

$$F = UR$$

and $RR^* = R^{**}R^* = I$. ■

The following existence theorem for the polar decomposition of an element of $\mathcal{L}(X,X)$ is a corollary.

Corollary 13.6.4 *Let $F \in \mathcal{L}(X,X)$. Then there exists a Hermitian $W \in \mathcal{L}(X,X)$, and a unitary matrix Q such that $F = WQ$, and there exists a Hermitian $U \in \mathcal{L}(X,X)$ and a unitary R, such that $F = RU$.*

This corollary has a fascinating relation to the question whether a given linear transformation is normal. Recall that an $n \times n$ matrix A, is normal if $AA^* = A^*A$. Retain the same definition for an element of $\mathcal{L}(X,X)$.

Theorem 13.6.5 *Let $F \in \mathcal{L}(X,X)$. Then F is normal if and only if in Corollary 13.6.4 $RU = UR$ and $QW = WQ$.*

Proof: I will prove the statement about $RU = UR$ and leave the other part as an exercise. First suppose that $RU = UR$ and show F is normal. To begin with,

$$UR^* = (RU)^* = (UR)^* = R^*U.$$

Therefore,

$$\begin{aligned} F^*F &= UR^*RU = U^2 \\ FF^* &= RUUR^* = URR^*U = U^2 \end{aligned}$$

which shows F is normal.

Now suppose F is normal. Is $RU = UR$? Since F is normal,

$$FF^* = RUUR^* = RU^2R^*$$

and

$$F^*F = UR^*RU = U^2.$$

Therefore, $RU^2R^* = U^2$, and both are nonnegative and self adjoint. Therefore, the square roots of both sides must be equal by the uniqueness part of the theorem on fractional powers. It follows that the square root of the first, RUR^* must equal the square root of the second, U. Therefore, $RUR^* = U$ and so $RU = UR$. This proves the theorem in one case. The other case in which W and Q commute is left as an exercise. ■

13.7 An Application To Statistics

A random vector is a function $\mathbf{X} : \Omega \to \mathbb{R}^p$ where Ω is a probability space. This means that there exists a σ algebra of measurable sets $\mathcal{F}$ and a probability measure $P : \mathcal{F} \to [0,1]$. In practice, people often don't worry too much about the underlying probability space and instead pay more attention to the distribution measure of the random variable. For E a suitable subset of $\mathbb{R}^p$, this measure gives the probability that $\mathbf{X}$ has values in E. There are often excellent reasons for believing that a random vector is normally distributed. This means that the probability that $\mathbf{X}$ has values in a set E is given by

$$\int_E \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^* \Sigma^{-1}(\mathbf{x}-\mathbf{m})\right) dx$$

The expression in the integral is called the normal probability density function. There are two parameters, $\mathbf{m}$ and Σ where $\mathbf{m}$ is called the mean and Σ is called the covariance matrix. It is a symmetric matrix which has all real eigenvalues which are all positive. While it may be reasonable to assume this is the distribution, in general, you won't know $\mathbf{m}$ and Σ and in order to use this formula to predict anything, you would need to know these quantities.

What people do to estimate these is to take n independent observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and try to predict what $\mathbf{m}$ and Σ should be based on these observations. One criterion used for making this determination is the method of maximum likelihood. In this method, you seek to choose the two parameters in such a way as to maximize the likelihood which is given as

$$\prod_{i=1}^{n} \frac{1}{\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{m})^* \Sigma^{-1}(\mathbf{x}_i - \mathbf{m})\right).$$

For convenience the term $(2\pi)^{p/2}$ was ignored. This leads to the estimate for $\mathbf{m}$ as

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \equiv \overline{\mathbf{x}}.$$

This part follows fairly easily from taking the ln and then setting partial derivatives equal to 0. The estimation of Σ is harder. However, it is not too hard using the theorems presented above. I am following a nice discussion given in Wikipedia. It will make use of Theorem 7.5.2 on the trace as well as the theorem about the square root of a linear transformation given above. First note that by Theorem 7.5.2,

$$\begin{aligned}(\mathbf{x}_i-\mathbf{m})^*\,\Sigma^{-1}\,(\mathbf{x}_i-\mathbf{m}) &= \operatorname{trace}\left((\mathbf{x}_i-\mathbf{m})^*\,\Sigma^{-1}\,(\mathbf{x}_i-\mathbf{m})\right)\\ &= \operatorname{trace}\left((\mathbf{x}_i-\mathbf{m})\,(\mathbf{x}_i-\mathbf{m})^*\,\Sigma^{-1}\right)\end{aligned}$$

Therefore, the thing to maximize is

$$\prod_{i=1}^{n}\frac{1}{\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}\operatorname{trace}\left((\mathbf{x}_i-\mathbf{m})\,(\mathbf{x}_i-\mathbf{m})^*\,\Sigma^{-1}\right)\right)$$

$$= \det\left(\Sigma^{-1}\right)^{n/2}\exp\left(-\frac{1}{2}\operatorname{trace}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^*\Sigma^{-1}\right)$$

$$= \det\left(\Sigma^{-1}\right)^{n/2}\exp\left(-\frac{1}{2}\operatorname{trace}\overbrace{\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^*}^{S}\Sigma^{-1}\right)$$

$$\equiv \det\left(\Sigma^{-1}\right)^{n/2}\exp\left(-\frac{1}{2}\operatorname{trace}\left(S\Sigma^{-1}\right)\right)$$

where S is the $p\times p$ matrix indicated above. Now S is symmetric and has eigenvalues which are all nonnegative because $(S\mathbf{y},\mathbf{y})\geq 0$. Therefore, S has a unique self adjoint square root. Using Theorem 7.5.2 again, the above equals

$$\det\left(\Sigma^{-1}\right)^{n/2}\exp\left(-\frac{1}{2}\operatorname{trace}\left(S^{1/2}\Sigma^{-1}S^{1/2}\right)\right)$$

Let $B = S^{1/2}\Sigma^{-1}S^{1/2}$ and assume $\det(S)\neq 0$. Then $\Sigma^{-1} = S^{-1/2}BS^{-1/2}$. The above equals

$$\det\left(S^{-1}\right)\det(B)^{n/2}\exp\left(-\frac{1}{2}\operatorname{trace}(B)\right)$$

Of course the thing to estimate is only found in B. Therefore, $\det\left(S^{-1}\right)$ can be discarded in trying to maximize things. Since B is symmetric, it is similar to a diagonal matrix D which has $\lambda_1,\cdots,\lambda_n$ down the diagonal. Thus it is desired to maximize

$$\left(\prod_{i=1}^{p}\lambda_i\right)^{n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}\lambda_i\right)$$

Taking ln it follows that it suffices to maximize

$$\frac{n}{2}\sum_{i=1}^{p}\ln\lambda_i-\frac{1}{2}\sum_{i=1}^{p}\lambda_i$$

Taking the derivative with respect to λ_i,

$$\frac{n}{2}\frac{1}{\lambda_i}-\frac{1}{2}=0$$

and so $\lambda_i = n$. It follows from the above that

$$\Sigma = S^{1/2}B^{-1}S^{1/2}$$

where B^{-1} has only the eigenvalues $1/n$. It follows B^{-1} must equal the diagonal matrix which has $1/n$ down the diagonal. The reason for this is that B is similar to a diagonal matrix because it is symmetric. Thus $B = P^{-1}\frac{1}{n}IP = \frac{1}{n}I$ because the identity commutes with every matrix. But now it follows that

$$\Sigma = \frac{1}{n}S$$

Of course this is just an estimate and so we write $\hat{\Sigma}$ instead of Σ.

This has shown that the maximum likelihood estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^*$$

13.8 The Singular Value Decomposition

In this section, A will be an $m \times n$ matrix. To begin with, here is a simple lemma.

Lemma 13.8.1 *Let A be an $m \times n$ matrix. Then A^*A is self adjoint and all its eigenvalues are nonnegative.*

Proof: It is obvious that A^*A is self adjoint. Suppose $A^*A\mathbf{x} = \lambda\mathbf{x}$. Then $\lambda |\mathbf{x}|^2 = (\lambda\mathbf{x}, \mathbf{x}) = (A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \geq 0$. ■

Definition 13.8.2 *Let A be an $m \times n$ matrix. The singular values of A are the square roots of the positive eigenvalues of A^*A.*

With this definition and lemma here is the main theorem on the singular value decomposition. In all that follows, I will write the following partitioned matrix

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

where σ denotes an $r \times r$ diagonal matrix of the form

$$\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

and the bottom row of zero matrices in the partitioned matrix, as well as the right columns of zero matrices are each of the right size so that the resulting matrix is $m \times n$. Either could vanish completely. However, I will write it in the above form. It is easy to make the necessary adjustments in the other two cases.

Theorem 13.8.3 *Let A be an $m \times n$ matrix. Then there exist unitary matrices, U and V of the appropriate size such that*

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is of the form

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

for the σ_i the singular values of A, arranged in order of decreasing size.

Proof: By the above lemma and Theorem 13.3.3 there exists an orthonormal basis, $\{\mathbf{v}_i\}_{i=1}^n$ such that $A^*A\mathbf{v}_i = \sigma_i^2\mathbf{v}_i$ where $\sigma_i^2 > 0$ for $i = 1, \cdots, k, (\sigma_i > 0)$, and equals zero if $i > k$. Thus for $i > k$, $A\mathbf{v}_i = \mathbf{0}$ because

$$(A\mathbf{v}_i, A\mathbf{v}_i) = (A^*A\mathbf{v}_i, \mathbf{v}_i) = (\mathbf{0}, \mathbf{v}_i) = 0.$$

For $i = 1, \cdots, k$, define $\mathbf{u}_i \in \mathbb{F}^m$ by

$$\mathbf{u}_i \equiv \sigma_i^{-1}A\mathbf{v}_i.$$

Thus $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$. Now

$$\begin{aligned}(\mathbf{u}_i, \mathbf{u}_j) &= \left(\sigma_i^{-1} A\mathbf{v}_i, \sigma_j^{-1} A\mathbf{v}_j\right) = \left(\sigma_i^{-1}\mathbf{v}_i, \sigma_j^{-1} A^* A\mathbf{v}_j\right) \\ &= \left(\sigma_i^{-1}\mathbf{v}_i, \sigma_j^{-1}\sigma_j^2 \mathbf{v}_j\right) = \frac{\sigma_j}{\sigma_i}(\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij}.\end{aligned}$$

Thus $\{\mathbf{u}_i\}_{i=1}^k$ is an orthonormal set of vectors in $\mathbb{F}^m$. Also,

$$AA^*\mathbf{u}_i = AA^*\sigma_i^{-1} A\mathbf{v}_i = \sigma_i^{-1} AA^* A\mathbf{v}_i = \sigma_i^{-1} A\sigma_i^2 \mathbf{v}_i = \sigma_i^2 \mathbf{u}_i.$$

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

$$U \equiv \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{pmatrix}$$

while

$$V \equiv \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}.$$

Thus U is the matrix which has the $\mathbf{u}_i$ as columns and V is defined as the matrix which has the $\mathbf{v}_i$ as columns. Then

$$U^* AV = \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} A \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} \begin{pmatrix} \sigma_1 \mathbf{u}_1 & \cdots & \sigma_k \mathbf{u}_k & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is given in the statement of the theorem. ■

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 13.8.4 *Let A be an $m \times n$ matrix. Then the rank of A and A^* equals the number of singular values.*

Proof: Since V and U are unitary, they are each one to one and onto and so it follows that

$$\text{rank}\,(A) = \text{rank}\,(U^* AV) = \text{rank}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \text{number of singular values.}$$

Also since U, V are unitary,

$$\text{rank}\,(A^*) = \text{rank}\,(V^* A^* U) = \text{rank}\left((U^* AV)^*\right)$$

$$= \text{rank}\left(\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}^*\right) = \text{number of singular values. ■}$$

13.9 Approximation In The Frobenius Norm

The Frobenius norm is one of many norms for a matrix. It is arguably the most obvious of all norms. Here is its definition.

Definition 13.9.1 *Let A be a complex $m \times n$ matrix. Then*

$$||A||_F \equiv (\text{trace}\,(AA^*))^{1/2}$$

Also this norm comes from the inner product

$$(A,B)_F \equiv \text{trace}\,(AB^*)$$

Thus $||A||_F^2$ is easily seen to equal $\sum_{ij} |a_{ij}|^2$ so essentially, it treats the matrix as a vector in $\mathbb{F}^{m\times n}$.

Lemma 13.9.2 *Let A be an $m \times n$ complex matrix with singular matrix*

$$\Sigma = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

with σ as defined above. Then

$$||\Sigma||_F^2 = ||A||_F^2 \tag{13.20}$$

and the following hold for the Frobenius norm. If U, V are unitary and of the right size,

$$||UA||_F = ||A||_F\,,\ ||UAV||_F = ||A||_F\,. \tag{13.21}$$

Proof: From the definition and letting U, V be unitary and of the right size,

$$||UA||_F^2 \equiv \text{trace}\,(UAA^*U^*) = \text{trace}\,(AA^*) = ||A||_F^2$$

Also,

$$||AV||_F^2 \equiv \text{trace}\,(AVV^*A^*) = \text{trace}\,(AA^*) = ||A||_F^2\,.$$

It follows

$$||UAV||_F^2 = ||AV||_F^2 = ||A||_F^2\,.$$

Now consider 13.20. From what was just shown,

$$||A||_F^2 = ||U\Sigma V^*||_F^2 = ||\Sigma||_F^2\,. \blacksquare$$

Of course, this shows that

$$||A||_F^2 = \sum_i \sigma_i^2,$$

the sum of the squares of the singular values of A.

Why is the singular value decomposition important? It implies

$$A = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*$$

where σ is the diagonal matrix having the singular values down the diagonal. Now sometimes A is a huge matrix, 1000×2000 or something like that. This happens in applications to situations where the entries of A describe a picture. What also happens is that most of the singular values are very small. What if you deleted those which were very small, say for all $i \geq l$ and got a new matrix

$$A' \equiv U \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^*?$$

Then the entries of A' would end up being close to the entries of A but there is much less information to keep track of. This turns out to be very useful. More precisely, letting

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}, \ U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},$$

$$||A - A'||_F^2 = \left|\left| U \begin{pmatrix} \sigma - \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^* \right|\right|_F^2 = \sum_{k=l+1}^{r} \sigma_k^2$$

Thus A is approximated by A' where A' has rank $l < r$. In fact, it is also true that out of all matrices of rank l, this A' is the one which is closest to A in the Frobenius norm. Here is why.

Let B be a matrix which has rank l. Then from Lemma 13.9.2

$$||A - B||_F^2 = ||U^* (A - B) V||_F^2 = ||U^*AV - U^*BV||_F^2 = \left|\left| \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} - U^*BV \right|\right|_F^2$$

and since the singular values of A decrease from the upper left to the lower right, it follows that for B to be closest as possible to A in the Frobenius norm,

$$U^*BV = \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}$$

which implies $B = A'$ above. This is really obvious if you look at a simple example. Say

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

for example. Then what rank 1 matrix would be closest to this one in the Frobenius norm? Obviously

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

13.10 Least Squares And Singular Value Decomposition

The singular value decomposition also has a very interesting connection to the problem of least squares solutions. Recall that it was desired to find $\mathbf{x}$ such that $|A\mathbf{x} - \mathbf{y}|$ is as small as possible. Lemma 12.5.1 shows that there is a solution to this problem which can be found by solving the system $A^*A\mathbf{x} = A^*\mathbf{y}$. Each $\mathbf{x}$ which solves this system solves the minimization problem as was shown in the lemma just mentioned. Now consider this equation for the solutions of the minimization problem in terms of the singular value decomposition.

$$\overbrace{V \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*}^{A^*} \overbrace{U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*}^{A} \mathbf{x} = \overbrace{V \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*}^{A^*} \mathbf{y}.$$

Therefore, this yields the following upon using block multiplication and multiplying on the left by V^*.

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^*\mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}. \tag{13.22}$$

One solution to this equation which is very easy to spot is

$$\mathbf{x} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}. \tag{13.23}$$

13.11 The Moore Penrose Inverse

The particular solution of the least squares problem given in 13.23 is important enough that it motivates the following definition.

Definition 13.11.1 *Let A be an $m \times n$ matrix. Then the Moore Penrose inverse of A, denoted by A^+ is defined as*

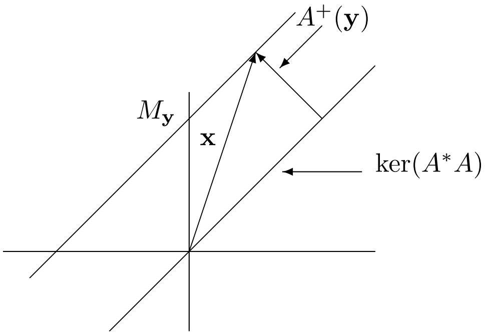
$$A^+ \equiv V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*.$$

Here

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

as above.

Thus $A^+\mathbf{y}$ is a solution to the minimization problem to find $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{y}|$. In fact, one can say more about this. In the following picture $M_\mathbf{y}$ denotes the set of least squares solutions $\mathbf{x}$ such that $A^*A\mathbf{x} = A^*\mathbf{y}$.



Then $A^{+}(\mathbf{y})$ is as given in the picture.

Proposition 13.11.2 *$A^{+}\mathbf{y}$ is the solution to the problem of minimizing $|A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$ which has smallest norm. Thus*

$$\left|AA^{+}\mathbf{y}-\mathbf{y}\right| \leq |A\mathbf{x}-\mathbf{y}| \text{ for all } \mathbf{x}$$

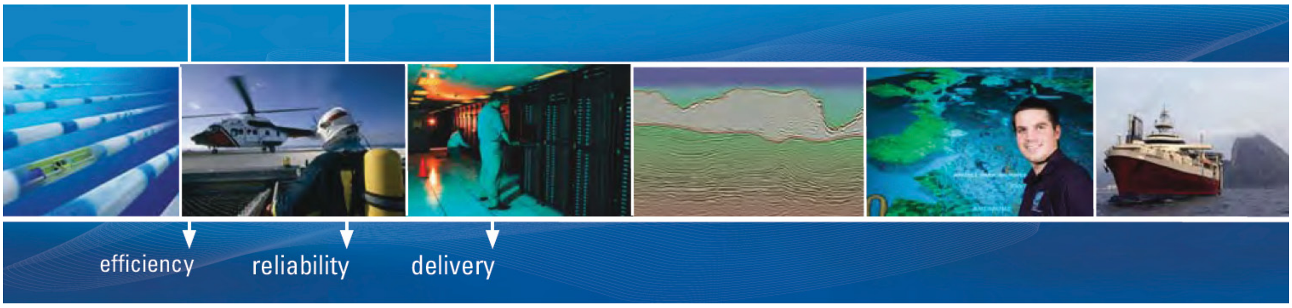
and if $\mathbf{x}_1$ satisfies $|A\mathbf{x}_1-\mathbf{y}| \leq |A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$, then $|A^{+}\mathbf{y}| \leq |\mathbf{x}_1|$.

Proof: Consider $\mathbf{x}$ satisfying 13.22, equivalently $A^{*}A\mathbf{x} = A^{*}\mathbf{y}$,

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^{*}\mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^{*}\mathbf{y}$$

which has smallest norm. This is equivalent to making $|V^{*}\mathbf{x}|$ as small as possible because V^{*} is unitary and so it preserves norms. For $\mathbf{z}$ a vector, denote by $(\mathbf{z})_k$ the vector in $\mathbb{F}^k$ which consists of the first k entries of $\mathbf{z}$. Then if $\mathbf{x}$ is a solution to 13.22

$$\begin{pmatrix} \sigma^2 (V^{*}\mathbf{x})_k \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sigma (U^{*}\mathbf{y})_k \\ \mathbf{0} \end{pmatrix}$$



and so $(V^*\mathbf{x})_k = \sigma^{-1}(U^*\mathbf{y})_k$. Thus the first k entries of $V^*\mathbf{x}$ are determined. In order to make $|V^*\mathbf{x}|$ as small as possible, the remaining $n-k$ entries should equal zero. Therefore,

$$V^*\mathbf{x} = \begin{pmatrix} (V^*\mathbf{x})_k \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma^{-1}(U^*\mathbf{y})_k \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

and so

$$\mathbf{x} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y} \equiv A^+\mathbf{y} \blacksquare$$

Lemma 13.11.3 *The matrix A^+ satisfies the following conditions.*

$$AA^+A = A,\ A^+AA^+ = A^+,\ A^+A \text{ and } AA^+ \text{ are Hermitian.} \tag{13.24}$$

Proof: This is routine. Recall

$$A = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} V^*$$

and

$$A^+ = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$$

so you just plug in and verify it works. ■

A much more interesting observation is that A^+ is characterized as being the unique matrix which satisfies 13.24. This is the content of the following Theorem. The conditions are sometimes called the Penrose conditions.

Theorem 13.11.4 *Let A be an $m \times n$ matrix. Then a matrix A_0, is the Moore Penrose inverse of A if and only if A_0 satisfies*

$$AA_0A = A,\ A_0AA_0 = A_0,\ A_0A \text{ and } AA_0 \text{ are Hermitian.} \tag{13.25}$$

Proof: From the above lemma, the Moore Penrose inverse satisfies 13.25. Suppose then that A_0 satisfies 13.25. It is necessary to verify that $A_0 = A^+$. Recall that from the singular value decomposition, there exist unitary matrices, U and V such that

$$U^*AV = \Sigma \equiv \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},\ A = U\Sigma V^*.$$

Let

$$V^*A_0U = \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \tag{13.26}$$

where P is $k \times k$.

Next use the first equation of 13.25 to write

$$\overbrace{U\Sigma V^*}^{A}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}\overbrace{U\Sigma V^*}^{A} = \overbrace{U\Sigma V^*}^{A}.$$

Then multiplying both sides on the left by U^* and on the right by V,

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

Now this requires

$$\begin{pmatrix} \sigma P \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \tag{13.27}$$

Therefore, $P = \sigma^{-1}$. From the requirement that AA_0 is Hermitian,

$$\overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} = U \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*$$

must be Hermitian. Therefore, it is necessary that

$$\begin{aligned} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} &= \begin{pmatrix} \sigma P & \sigma Q \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} \end{aligned}$$

is Hermitian. Then

$$\begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ Q^* \sigma & 0 \end{pmatrix}$$

Thus

$$Q^* \sigma = 0$$

and so multiplying both sides on the right by σ^{-1}, it follows $Q^* = 0$ and so $Q = 0$.

From the requirement that $A_0 A$ is Hermitian, it is necessary that

$$\begin{aligned} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} &= V \begin{pmatrix} P\sigma & 0 \\ R\sigma & 0 \end{pmatrix} V^* \\ &= V \begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix} V^* \end{aligned}$$

is Hermitian. Therefore, also

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}$$

is Hermitian. Thus $R = 0$ because this equals

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}^* = \begin{pmatrix} I & \sigma^* R^* \\ 0 & 0 \end{pmatrix}$$

which requires $R\sigma = 0$. Now multiply on right by σ^{-1} to find that $R = 0$.

Use 13.26 and the second equation of 13.25 to write

$$\overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} = \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0}.$$

which implies

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}.$$

This yields from the above in which is was shown that R, Q are both 0

$$\begin{aligned} \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \qquad (13.28) \\ &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}. \qquad (13.29) \end{aligned}$$

Therefore, $S = 0$ also and so

$$V^* A_0 U \equiv \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

which says

$$A_0 = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \equiv A^+. \blacksquare$$

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in 13.25. It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 3.

13.12 Exercises

1. Show $(A^*)^* = A$ and $(AB)^* = B^*A^*$.

2. Prove Corollary 13.3.9.

3. Show that if A is an $n \times n$ matrix which has an inverse then $A^+ = A^{-1}$.

4. Using the singular value decomposition, show that for any square matrix A, it follows that A^*A is unitarily similar to AA^*.

5. Let A, B be a $m \times n$ matrices. Define an inner product on the set of $m \times n$ matrices by
$$(A, B)_F \equiv \text{trace}\,(AB^*)\,.$$
Show this is an inner product satisfying all the inner product axioms. Recall for M an $n \times n$ matrix, trace $(M) \equiv \sum_{i=1}^{n} M_{ii}$. The resulting norm, $||\cdot||_F$ is called the Frobenius norm and it can be used to measure the distance between two matrices.

6. Let A be an $m \times n$ matrix. Show $||A||_F^2 \equiv (A, A)_F = \sum_j \sigma_j^2$ where the σ_j are the singular values of A.

7. If A is a general $n \times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\{A_k\}$ of $n \times n$ matrices having distinct eigenvalues which has the property that the ij^{th} entry of A_k converges to the ij^{th} entry of A for all ij. **Hint:** Use Schur's theorem.

8. Prove the Cayley Hamilton theorem as follows. First suppose A has a basis of eigenvectors $\{\mathbf{v}_k\}_{k=1}^{n}, A\mathbf{v}_k = \lambda_k \mathbf{v}_k$. Let $p(\lambda)$ be the characteristic polynomial. Show $p(A)\,\mathbf{v}_k = p(\lambda_k)\,\mathbf{v}_k = \mathbf{0}$. Then since $\{\mathbf{v}_k\}$ is a basis, it follows $p(A)\,\mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$ and so $p(A) = 0$. Next in the general case, use Problem 7 to obtain a sequence $\{A_k\}$ of matrices whose entries converge to the entries of A such that A_k has n distinct eigenvalues and therefore by Theorem 7.1.7 A_k has a basis of eigenvectors. Therefore, from the first part and for $p_k(\lambda)$ the characteristic polynomial for A_k, it follows $p_k(A_k) = 0$. Now explain why and the sense in which $\lim_{k\to\infty} p_k(A_k) = p(A)$.

9. Prove that Theorem 13.4.6 and Corollary 13.4.7 can be strengthened so that the condition on the A_k is necessary as well as sufficient. **Hint:** Consider vectors of the form $\begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{x} \in \mathbb{F}^k$.

10. Show directly that if A is an $n \times n$ matrix and $A = A^*$ (A is Hermitian) then all the eigenvalues are real and eigenvectors can be assumed to be real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).

11. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be an orthonormal basis for $\mathbb{F}^n$. Let Q be a matrix whose i^{th} column is $\mathbf{v}_i$. Show
$$Q^*Q = QQ^* = I.$$

12. Show that an $n \times n$ matrix Q is unitary if and only if it preserves distances. This means $|Q\mathbf{v}| = |\mathbf{v}|$. This was done in the text but you should try to do it for yourself.

13. Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ are two orthonormal bases for $\mathbb{F}^n$ and suppose Q is an $n \times n$ matrix satisfying $Q\mathbf{v}_i = \mathbf{w}_i$. Then show Q is unitary. If $|\mathbf{v}| = 1$, show there is a unitary transformation which maps $\mathbf{v}$ to $\mathbf{e}_1$.

14. Finish the proof of Theorem 13.6.5.

15. Let A be a Hermitian matrix so $A = A^*$ and suppose all eigenvalues of A are larger than δ^2. Show
$$(A\mathbf{v}, \mathbf{v}) \geq \delta^2 |\mathbf{v}|^2$$
Where here, the inner product is $(\mathbf{v}, \mathbf{u}) \equiv \sum_{j=1}^n v_j \overline{u_j}$.

16. Suppose $A + A^*$ has all negative eigenvalues. Then show that the eigenvalues of A have all negative real parts.

17. The discrete Fourier transform maps $\mathbb{C}^n \to \mathbb{C}^n$ as follows.
$$F(\mathbf{x}) = \mathbf{z} \text{ where } z_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{-i\frac{2\pi}{n}jk} x_j.$$
Show that F^{-1} exists and is given by the formula
$$F^{-1}(\mathbf{z}) = \mathbf{x} \text{ where } x_j = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{i\frac{2\pi}{n}jk} z_k$$
Here is one way to approach this problem. Note $\mathbf{z} = U\mathbf{x}$ where
$$U = \frac{1}{\sqrt{n}} \begin{pmatrix} e^{-i\frac{2\pi}{n}0\cdot 0} & e^{-i\frac{2\pi}{n}1\cdot 0} & e^{-i\frac{2\pi}{n}2\cdot 0} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 0} \\ e^{-i\frac{2\pi}{n}0\cdot 1} & e^{-i\frac{2\pi}{n}1\cdot 1} & e^{-i\frac{2\pi}{n}2\cdot 1} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 1} \\ e^{-i\frac{2\pi}{n}0\cdot 2} & e^{-i\frac{2\pi}{n}1\cdot 2} & e^{-i\frac{2\pi}{n}2\cdot 2} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot 2} \\ \vdots & \vdots & \vdots & & \vdots \\ e^{-i\frac{2\pi}{n}0\cdot(n-1)} & e^{-i\frac{2\pi}{n}1\cdot(n-1)} & e^{-i\frac{2\pi}{n}2\cdot(n-1)} & \cdots & e^{-i\frac{2\pi}{n}(n-1)\cdot(n-1)} \end{pmatrix}$$
Now argue U is unitary and use this to establish the result. To show this verify each row has length 1 and the inner product of two different rows gives 0. Now $U_{kj} = e^{-i\frac{2\pi}{n}jk}$ and so $(U^*)_{kj} = e^{i\frac{2\pi}{n}jk}$.

18. Let f be a periodic function having period 2π. The Fourier series of f is an expression of the form
$$\sum_{k=-\infty}^{\infty} c_k e^{ikx} \equiv \lim_{n\to\infty} \sum_{k=-n}^{n} c_k e^{ikx}$$
and the idea is to find c_k such that the above sequence converges in some way to f. If
$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$$
and you formally multiply both sides by e^{-imx} and then integrate from 0 to 2π, interchanging the integral with the sum without any concern for whether this makes sense, show it is reasonable from this to expect
$$c_m = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-imx} dx.$$
Now suppose you only know $f(x)$ at equally spaced points $2\pi j/n$ for $j = 0, 1, \cdots, n$. Consider the Riemann sum for this integral obtained from using the left endpoint of the subintervals determined from the partition $\left\{\frac{2\pi}{n}j\right\}_{j=0}^n$. How does this compare with the discrete Fourier transform? What happens as $n \to \infty$ to this approximation?

19. Suppose A is a real 3×3 orthogonal matrix (Recall this means $AA^T = A^T A = I$.) having determinant 1. Show it must have an eigenvalue equal to 1. Note this shows there exists a vector $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{x}$. **Hint:** Show first or recall that any orthogonal matrix must preserve lengths. That is, $|A\mathbf{x}| = |\mathbf{x}|$.

20. Let A be a complex $m \times n$ matrix. Using the description of the Moore Penrose inverse in terms of the singular value decomposition, show that

$$\lim_{\delta \to 0+} \left(A^* A + \delta I\right)^{-1} A^* = A^+$$

where the convergence happens in the Frobenius norm. Also verify, using the singular value decomposition, that the inverse exists in the above formula.

21. Show that $A^+ = \left(A^* A\right)^+ A^*$. **Hint:** You might use the description of A^+ in terms of the singular value decomposition.

Norms

In this chapter, X and Y are finite dimensional vector spaces which have a norm. The following is a definition.

Definition 14.0.1 *A linear space X is a normed linear space if there is a norm defined on X, $||\cdot||$ satisfying*

$$||\mathbf{x}|| \geq 0, \ \ ||\mathbf{x}|| = 0 \text{ if and only if } \mathbf{x} = 0,$$

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

$$||c\mathbf{x}|| = |c| \, ||\mathbf{x}||$$

whenever c is a scalar. A set, $U \subseteq X$, a normed linear space is open if for every $p \in U$, there exists $\delta > 0$ such that

$$B(p, \delta) \equiv \{x : ||x - p|| < \delta\} \subseteq U.$$

Thus, a set is open if every point of the set is an interior point.

To begin with recall the Cauchy Schwarz inequality which is stated here for convenience in terms of the inner product space, $\mathbb{C}^n$.

Theorem 14.0.2 *The following inequality holds for a_i and $b_i \in \mathbb{C}$.*

$$\left| \sum_{i=1}^{n} a_i \bar{b}_i \right| \leq \left(\sum_{i=1}^{n} |a_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |b_i|^2 \right)^{1/2} . \tag{14.1}$$

Definition 14.0.3 *Let $(X, ||\cdot||)$ be a normed linear space and let $\{x_n\}_{n=1}^{\infty}$ be a sequence of vectors. Then this is called a Cauchy sequence if for all $\varepsilon > 0$ there exists N such that if $m, n \geq N$, then*

$$||x_n - x_m|| < \varepsilon.$$

This is written more briefly as

$$\lim_{m,n \to \infty} ||x_n - x_m|| = 0.$$

Definition 14.0.4 *A normed linear space, $(X, ||\cdot||)$ is called a Banach space if it is complete. This means that, whenever, $\{\mathbf{x}_n\}$ is a Cauchy sequence there exists a unique $\mathbf{x} \in X$ such that $\lim_{n \to \infty} ||\mathbf{x} - \mathbf{x}_n|| = 0$.*

Let X be a finite dimensional normed linear space with norm $||\cdot||$ where the field of scalars is denoted by $\mathbb{F}$ and is understood to be either $\mathbb{R}$ or $\mathbb{C}$. Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ be a basis for X. If $\mathbf{x}\in X$, denote by x_i the i^{th} component of $\mathbf{x}$ with respect to this basis. Thus

$$\mathbf{x}=\sum_{i=1}^{n}x_i\mathbf{v}_i.$$

Definition 14.0.5 *For $\mathbf{x}\in X$ and $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ a basis, define a new norm by*

$$|\mathbf{x}|\equiv\left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}.$$

where

$$\mathbf{x}=\sum_{i=1}^{n}x_i\mathbf{v}_i.$$

Similarly, for $\mathbf{y}\in Y$ with basis $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$, and y_i its components with respect to this basis,

$$|\mathbf{y}|\equiv\left(\sum_{i=1}^{m}|y_i|^2\right)^{1/2}$$

For $A\in\mathcal{L}(X,Y)$, the space of linear mappings from X to Y,

$$||A||\equiv\sup\{|A\mathbf{x}|:|\mathbf{x}|\leq 1\}. \tag{14.2}$$

The first thing to show is that the two norms, $||\cdot||$ and $|\cdot|$, are equivalent. This means the conclusion of the following theorem holds.

Theorem 14.0.6 *Let $(X,||\cdot||)$ be a finite dimensional normed linear space and let $|\cdot|$ be described above relative to a given basis, $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$. Then $|\cdot|$ is a norm and there exist constants $\delta,\Delta>0$ independent of $\mathbf{x}$ such that*

$$\delta\,||\mathbf{x}||\leq|\mathbf{x}|\leq\Delta\,||\mathbf{x}||. \tag{14.3}$$

Proof: All of the above properties of a norm are obvious except the second, the triangle inequality. To establish this inequality, use the Cauchy Schwarz inequality to write

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

and this proves the second property above.

It remains to show the equivalence of the two norms. By the Cauchy Schwarz inequality again,

$$\begin{aligned}||\mathbf{x}|| &\equiv \left|\left|\sum_{i=1}^{n}x_i\mathbf{v}_i\right|\right|\leq\sum_{i=1}^{n}|x_i|\,||\mathbf{v}_i||\leq|\mathbf{x}|\left(\sum_{i=1}^{n}||\mathbf{v}_i||^2\right)^{1/2}\\ &\equiv \delta^{-1}\,|\mathbf{x}|.\end{aligned}$$

This proves the first half of the inequality.

Suppose the second half of the inequality is not valid. Then there exists a sequence $\mathbf{x}^k \in X$ such that

$$\left|\mathbf{x}^k\right| > k\left|\left|\mathbf{x}^k\right|\right|,\ k = 1, 2, \cdots .$$

Then define

$$\mathbf{y}^k \equiv \frac{\mathbf{x}^k}{\left|\mathbf{x}^k\right|}.$$

It follows

$$\left|\mathbf{y}^k\right| = 1, \quad \left|\mathbf{y}^k\right| > k\left|\left|\mathbf{y}^k\right|\right|. \tag{14.4}$$

Letting y_i^k be the components of $\mathbf{y}^k$ with respect to the given basis, it follows the vector

$$\left(y_1^k, \cdots, y_n^k\right)$$

is a unit vector in $\mathbb{F}^n$. By the Heine Borel theorem, there exists a subsequence, still denoted by k such that

$$\left(y_1^k, \cdots, y_n^k\right) \to (y_1, \cdots, y_n).$$

It follows from 14.4 and this that for

$$\mathbf{y} = \sum_{i=1}^n y_i \mathbf{v}_i,$$

$$0 = \lim_{k\to\infty} \left|\left|\mathbf{y}^k\right|\right| = \lim_{k\to\infty} \left|\left|\sum_{i=1}^n y_i^k \mathbf{v}_i\right|\right| = \left|\left|\sum_{i=1}^n y_i \mathbf{v}_i\right|\right|$$

but not all the y_i equal zero. This contradicts the assumption that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis and proves the second half of the inequality. ■

Corollary 14.0.7 *If $(X, ||\cdot||)$ is a finite dimensional normed linear space with the field of scalars $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, then X is complete.*

Proof: Let $\{\mathbf{x}^k\}$ be a Cauchy sequence. Then letting the components of $\mathbf{x}^k$ with respect to the given basis be

$$x_1^k, \cdots, x_n^k,$$

it follows from Theorem 14.0.6, that

$$\left(x_1^k, \cdots, x_n^k\right)$$

is a Cauchy sequence in $\mathbb{F}^n$ and so

$$\left(x_1^k, \cdots, x_n^k\right) \to (x_1, \cdots, x_n) \in \mathbb{F}^n.$$

Thus,

$$\mathbf{x}^k = \sum_{i=1}^n x_i^k \mathbf{v}_i \to \sum_{i=1}^n x_i \mathbf{v}_i \in X. \ ■$$

Corollary 14.0.8 *Suppose X is a finite dimensional linear space with the field of scalars either $\mathbb{C}$ or $\mathbb{R}$ and $||\cdot||$ and $|||\cdot|||$ are two norms on X. Then there exist positive constants, δ and Δ, independent of $\mathbf{x} \in X$ such that*

$$\delta\, |||\mathbf{x}||| \leq ||\mathbf{x}|| \leq \Delta\, |||\mathbf{x}|||.$$

Thus any two norms are equivalent.

This is very important because it shows that all questions of convergence can be considered relative to any norm with the same outcome.

Proof: Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X and let $|\cdot|$ be the norm taken with respect to this basis which was described earlier. Then by Theorem 14.0.6, there are positive constants $\delta_1, \Delta_1, \delta_2, \Delta_2$, all independent of $\mathbf{x} \in X$ such that

$$\delta_2 \, |||\mathbf{x}||| \le |\mathbf{x}| \le \Delta_2 \, |||\mathbf{x}||| \, ,$$

$$\delta_1 \, ||\mathbf{x}|| \le |\mathbf{x}| \le \Delta_1 \, ||\mathbf{x}|| \, .$$

Then

$$\delta_2 \, |||\mathbf{x}||| \le |\mathbf{x}| \le \Delta_1 \, ||\mathbf{x}|| \le \frac{\Delta_1}{\delta_1} \, |\mathbf{x}| \le \frac{\Delta_1 \Delta_2}{\delta_1} \, |||\mathbf{x}|||$$

and so

$$\frac{\delta_2}{\Delta_1} \, |||\mathbf{x}||| \le ||\mathbf{x}|| \le \frac{\Delta_2}{\delta_1} \, |||\mathbf{x}||| \ \blacksquare$$

Definition 14.0.9 *Let X and Y be normed linear spaces with norms $||\cdot||_X$ and $||\cdot||_Y$ respectively. Then $\mathcal{L}(X,Y)$ denotes the space of linear transformations, called bounded linear transformations, mapping X to Y which have the property that*

$$||A|| \equiv \sup\{||Ax||_Y : ||x||_X \leq 1\} < \infty.$$

Then $||A||$ is referred to as the operator norm of the bounded linear transformation A.

It is an easy exercise to verify that $||\cdot||$ is a norm on $\mathcal{L}(X,Y)$ and it is always the case that

$$||Ax||_Y \leq ||A|| \, ||x||_X \, .$$

Furthermore, you should verify that you can replace ≤ 1 with $= 1$ in the definition. Thus

$$||A|| \equiv \sup\{||Ax||_Y : ||x||_X = 1\}.$$

Theorem 14.0.10 *Let X and Y be finite dimensional normed linear spaces of dimension n and m respectively and denote by $||\cdot||$ the norm on either X or Y. Then if A is any linear function mapping X to Y, then $A \in \mathcal{L}(X,Y)$ and $(\mathcal{L}(X,Y), ||\cdot||)$ is a complete normed linear space of dimension nm with*

$$||A\mathbf{x}|| \leq ||A|| \, ||\mathbf{x}|| \, .$$

Proof: It is necessary to show the norm defined on linear transformations really is a norm. Again the first and third properties listed above for norms are obvious. It remains to show the second and verify $||A|| < \infty$. Letting $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis and $|\cdot|$ defined with respect to this basis as above, there exist constants $\delta, \Delta > 0$ such that

$$\delta \, ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta \, ||\mathbf{x}|| \, .$$

Then,

$$\begin{aligned} ||A+B|| &\equiv \sup\{||(A+B)(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\} \\ &\leq \sup\{||A\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} + \sup\{||B\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \\ &\equiv ||A|| + ||B|| \, . \end{aligned}$$

Next consider the claim that $||A|| < \infty$. This follows from

$$||A(\mathbf{x})|| = \left|\left| A\left(\sum_{i=1}^{n} x_i \mathbf{v}_i\right)\right|\right| \leq \sum_{i=1}^{n} |x_i| \, ||A(\mathbf{v}_i)||$$

$$\leq |\mathbf{x}| \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2\right)^{1/2} \leq \Delta \, ||\mathbf{x}|| \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2\right)^{1/2} < \infty.$$

Thus $||A|| \leq \Delta \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2\right)^{1/2}$.

Next consider the assertion about the dimension of $\mathcal{L}(X,Y)$. It follows from Theorem 9.2.3. By Corollary 14.0.7 $(\mathcal{L}(X,Y), ||\cdot||)$ is complete. If $\mathbf{x} \neq \mathbf{0}$,

$$||A\mathbf{x}|| \frac{1}{||\mathbf{x}||} = \left|\left| A \frac{\mathbf{x}}{||\mathbf{x}||} \right|\right| \leq ||A|| \;\; \blacksquare$$

Note by Corollary 14.0.8 you can define a norm any way desired on any finite dimensional linear space which has the field of scalars $\mathbb{R}$ or $\mathbb{C}$ and any other way of defining a norm on this space yields an equivalent norm. Thus, it doesn't much matter as far as notions of convergence are concerned which norm is used for a finite dimensional space. In particular in the space of $m \times n$ matrices, you can use the operator norm defined above, or some other way of giving this space a norm. A popular choice for a norm is the Frobenius norm discussed earlier but reviewed here.

Definition 14.0.11 *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 14.0.12 *Let A be an $m \times n$ matrix. Define the spectral norm of A, written as $||A||_2$ to be*

$$\max\left\{\lambda^{1/2} : \lambda \text{ is an eigenvalue of } A^*A\right\}.$$

That is, the largest singular value of A. (Note the eigenvalues of A^*A are all positive because if $A^*A\mathbf{x} = \lambda\mathbf{x}$, then

$$\lambda\,(\mathbf{x}, \mathbf{x}) = (A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \geq 0.)$$

Actually, this is nothing new. It turns out that $||\cdot||_2$ is nothing more than the operator norm for A taken with respect to the usual Euclidean norm,

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

Proposition 14.0.13 *The following holds.*

$$||A||_2 = \sup\{|A\mathbf{x}| : |\mathbf{x}| = 1\} \equiv ||A||\,.$$

Proof: Note that A^*A is Hermitian and so by Corollary 13.3.5,

$$\begin{aligned} ||A||_2 &= \max\left\{(A^*A\mathbf{x}, \mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &= \max\left\{(A\mathbf{x}, A\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &= \max\{|A\mathbf{x}| : |\mathbf{x}| = 1\} = ||A||\,. \blacksquare \end{aligned}$$

Here is another proof of this proposition. Recall there are unitary matrices of the right size U, V such that $A = U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*$ where the matrix on the inside is as described in the section on the singular value decomposition. Then since unitary matrices preserve norms,

$$\begin{aligned} ||A|| &= \sup_{|\mathbf{x}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*\mathbf{x}\right| = \sup_{|V^*\mathbf{x}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*\mathbf{x}\right| \\ &= \sup_{|\mathbf{y}|\leq 1}\left|U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\mathbf{y}\right| = \sup_{|\mathbf{y}|\leq 1}\left|\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\mathbf{y}\right| = \sigma_1 \equiv ||A||_2 \end{aligned}$$

This completes the alternate proof.

From now on, $||A||_2$ will mean either the operator norm of A taken with respect to the usual Euclidean norm or the largest singular value of A, whichever is most convenient.

An interesting application of the notion of equivalent norms on $\mathbb{R}^n$ is the process of giving a norm on a finite Cartesian product of normed linear spaces.

Definition 14.0.14 *Let X_i, $i = 1, \cdots, n$ be normed linear spaces with norms, $||\cdot||_i$. For*

$$\mathbf{x} \equiv (x_1, \cdots, x_n) \in \prod_{i=1}^{n} X_i$$

define $\theta : \prod_{i=1}^{n} X_i \to \mathbb{R}^n$ by

$$\theta(\mathbf{x}) \equiv (||x_1||_1, \cdots, ||x_n||_n)$$

Then if $||\cdot||$ is any norm on $\mathbb{R}^n$, define a norm on $\prod_{i=1}^{n} X_i$, also denoted by $||\cdot||$ by

$$||\mathbf{x}|| \equiv ||\theta \mathbf{x}||.$$

The following theorem follows immediately from Corollary 14.0.8.

Theorem 14.0.15 *Let X_i and $||\cdot||_i$ be given in the above definition and consider the norms on $\prod_{i=1}^{n} X_i$ described there in terms of norms on $\mathbb{R}^n$. Then any two of these norms on $\prod_{i=1}^{n} X_i$ obtained in this way are equivalent.*

For example, define

$$||\mathbf{x}||_1 \equiv \sum_{i=1}^{n} |x_i|,$$

$$||\mathbf{x}||_\infty \equiv \max\{|x_i|, i = 1, \cdots, n\},$$

or

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

and all three are equivalent norms on $\prod_{i=1}^{n} X_i$.

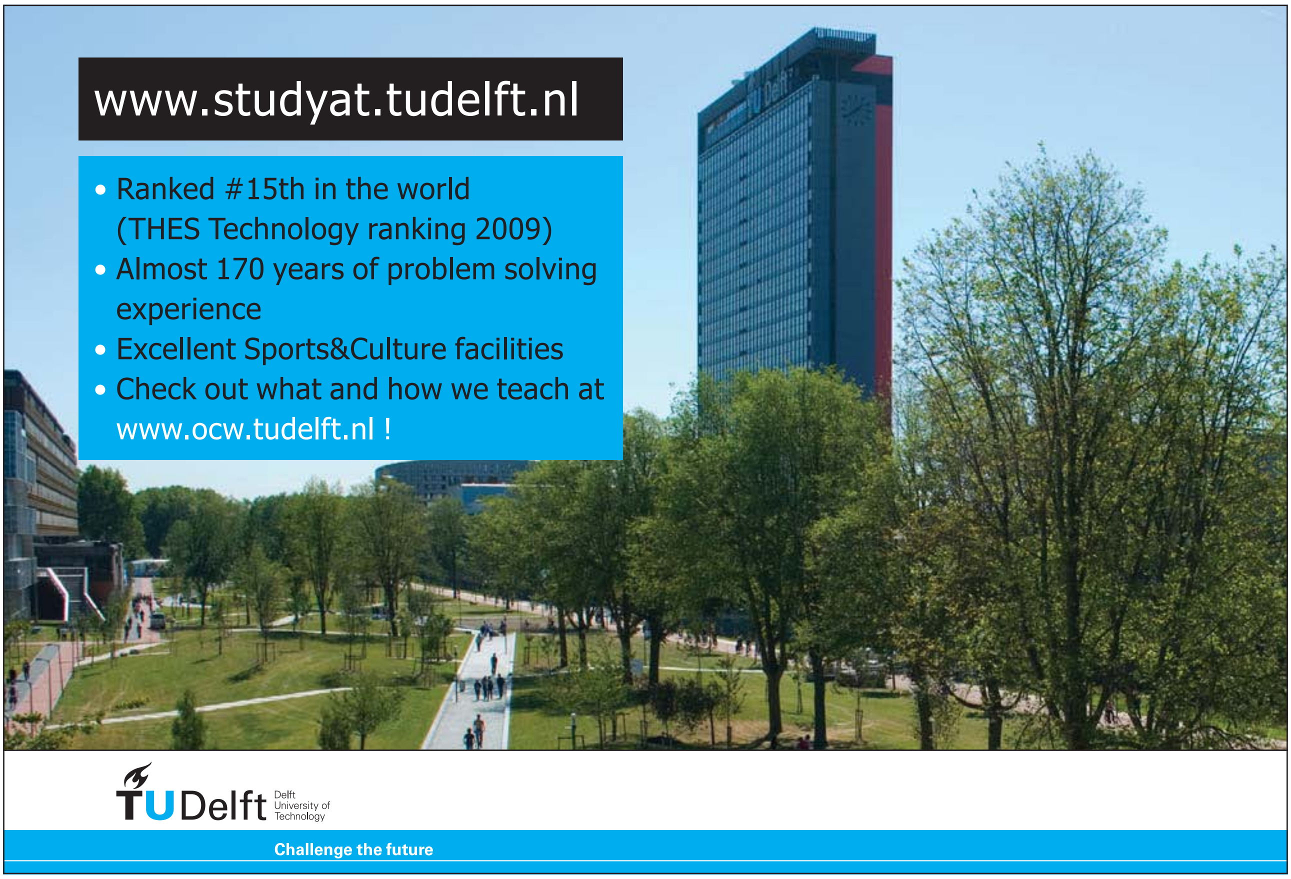
14.1 The p Norms

In addition to $||\cdot||_1$ and $||\cdot||_\infty$ mentioned above, it is common to consider the so called p norms for $\mathbf{x} \in \mathbb{C}^n$.

Definition 14.1.1 *Let* $\mathbf{x} \in \mathbb{C}^n$. *Then define for* $p \geq 1$,

$$||\mathbf{x}||_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

The following inequality is called Holder's inequality.



Proposition 14.1.2 *For* $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,

$$\sum_{i=1}^n |x_i| |y_i| \leq \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \left(\sum_{i=1}^n |y_i|^{p'} \right)^{1/p'}$$

The proof will depend on the following lemma.

Lemma 14.1.3 *If* $a, b \geq 0$ *and* p' *is defined by* $\frac{1}{p} + \frac{1}{p'} = 1$, *then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

Proof of the Proposition: If $\mathbf{x}$ or $\mathbf{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 14.1.3,

$$\begin{aligned} \sum_{i=1}^n \frac{|x_i|}{A} \frac{|y_i|}{B} &\leq \sum_{i=1}^n \left[\frac{1}{p} \left(\frac{|x_i|}{A} \right)^p + \frac{1}{p'} \left(\frac{|y_i|}{B} \right)^{p'} \right] \\ &= \frac{1}{p} \frac{1}{A^p} \sum_{i=1}^n |x_i|^p + \frac{1}{p'} \frac{1}{B^p} \sum_{i=1}^n |y_i|^{p'} \\ &= \frac{1}{p} + \frac{1}{p'} = 1 \end{aligned}$$

and so

$$\sum_{i=1}^n |x_i| |y_i| \leq AB = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \left(\sum_{i=1}^n |y_i|^{p'} \right)^{1/p'}. \blacksquare$$

Theorem 14.1.4 *The* p *norms do indeed satisfy the axioms of a norm.*

Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $\frac{p}{p'} = p - 1$. Then using the Holder inequality,

$$\begin{aligned} ||\mathbf{x} + \mathbf{y}||^p &= \sum_{i=1}^n |x_i + y_i|^p \\ &\leq \sum_{i=1}^n |x_i + y_i|^{p-1} |x_i| + \sum_{i=1}^n |x_i + y_i|^{p-1} |y_i| \\ &= \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}} |x_i| + \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}} |y_i| \\ &\leq \left(\sum_{i=1}^n |x_i + y_i|^p \right)^{1/p'} \left[\left(\sum_{i=1}^n |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p \right)^{1/p} \right] \\ &= ||\mathbf{x} + \mathbf{y}||^{p/p'} \left(||\mathbf{x}||_p + ||\mathbf{y}||_p \right) \end{aligned}$$

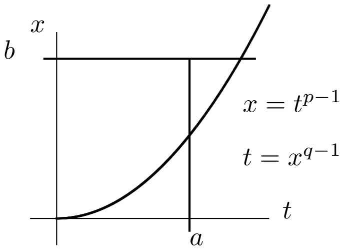
so dividing by $||\mathbf{x} + \mathbf{y}||^{p/p'}$, it follows

$$||\mathbf{x} + \mathbf{y}||^p \, ||\mathbf{x} + \mathbf{y}||^{-p/p'} = ||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$$

$\left(p - \frac{p}{p'} = p\left(1 - \frac{1}{p'}\right) = p\frac{1}{p} = 1.\right)$. ■

It only remains to prove Lemma 14.1.3.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1} dt + \int_0^b x^{q-1} dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$.

Alternate proof of the lemma: Let

$$f(t) \equiv \frac{1}{p}(at)^p + \frac{1}{q}\left(\frac{b}{t}\right)^q, \ t > 0$$

You see right away it is decreasing for a while, having an asymptote at $t = 0$ and then reaches a minimum and increases from then on. Take its derivative.

$$f'(t) = (at)^{p-1} a + \left(\frac{b}{t}\right)^{q-1}\left(\frac{-b}{t^2}\right)$$

Set it equal to 0. This happens when

$$t^{p+q} = \frac{b^q}{a^p}. \tag{14.5}$$

Thus

$$t = \frac{b^{q/(p+q)}}{a^{p/(p+q)}}$$

and so at this value of t,

$$at = (ab)^{q/(p+q)}, \ \left(\frac{b}{t}\right) = (ab)^{p/(p+q)}.$$

Thus the minimum of f is

$$\frac{1}{p}\left((ab)^{q/(p+q)}\right)^p + \frac{1}{q}\left((ab)^{p/(p+q)}\right)^q = (ab)^{pq/(p+q)}$$

but recall $1/p + 1/q = 1$ and so $pq/(p+q) = 1$. Thus the minimum value of f is ab. Letting $t = 1$, this shows

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Note that equality occurs when the minimum value happens for $t = 1$ and this indicates from 14.5 that $a^p = b^q$. ■

Now $||A||_p$ may be considered as the operator norm of A taken with respect to $||\cdot||_p$. In the case when $p = 2$, this is just the spectral norm. There is an easy estimate for $||A||_p$ in terms of the entries of A.

Theorem 14.1.5 *The following holds.*

$$||A||_p \leq \left(\sum_k \left(\sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q}$$

Proof: Let $||\mathbf{x}||_p \leq 1$ and let $A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$ where the $\mathbf{a}_k$ are the columns of A. Then

$$A\mathbf{x} = \left(\sum_k x_k \mathbf{a}_k \right)$$

and so by Holder's inequality,

$$\begin{aligned} ||A\mathbf{x}||_p &\equiv \left|\left| \sum_k x_k \mathbf{a}_k \right|\right|_p \leq \sum_k |x_k| \, ||\mathbf{a}_k||_p \\ &\leq \left(\sum_k |x_k|^p \right)^{1/p} \left(\sum_k ||\mathbf{a}_k||_p^q \right)^{1/q} \\ &\leq \left(\sum_k \left(\sum_j |A_{jk}|^p \right)^{q/p} \right)^{1/q} \quad \blacksquare \end{aligned}$$

14.2 The Condition Number

Let $A \in \mathcal{L}(X, X)$ be a linear transformation where X is a finite dimensional vector space and consider the problem $Ax = b$ where it is assumed there is a unique solution to this problem. How does the solution change if A is changed a little bit and if b is changed a little bit? This is clearly an interesting question because you often do not know A and b exactly. If a small change in these quantities results in a large change in the solution, x, then it seems clear this would be undesirable. In what follows $||\cdot||$ when applied to a linear transformation will always refer to the operator norm.

Lemma 14.2.1 *Let $A, B \in \mathcal{L}(X, X)$ where X is a normed vector space as above. Then for $||\cdot||$ denoting the operator norm,*

$$||AB|| \leq ||A|| \, ||B|| \, .$$

Proof: This follows from the definition. Letting $||x|| \leq 1$, it follows from Theorem 14.0.10

$$||ABx|| \leq ||A|| \, ||Bx|| \leq ||A|| \, ||B|| \, ||x|| \leq ||A|| \, ||B||$$

and so

$$||AB|| \equiv \sup_{||x|| \leq 1} ||ABx|| \leq ||A|| \, ||B|| . \blacksquare$$

Lemma 14.2.2 *Let $A, B \in \mathcal{L}(X,X), A^{-1} \in \mathcal{L}(X,X)$, and suppose $||B|| < 1/\left|\left|A^{-1}\right|\right|$. Then $(A+B)^{-1}$ exists and*

$$\left|\left|(A+B)^{-1}\right|\right| \leq \left|\left|A^{-1}\right|\right| \left|\frac{1}{1-||A^{-1}B||}\right| .$$

The above formula makes sense because $\left|\left|A^{-1}B\right|\right| < 1$.

Proof: By Lemma 14.2.1,

$$\left|\left|A^{-1}B\right|\right| \leq \left|\left|A^{-1}\right|\right| ||B|| < \left|\left|A^{-1}\right|\right| \frac{1}{||A^{-1}||} = 1$$

Suppose $(A+B)x = 0$. Then $0 = A\left(I + A^{-1}B\right)x$ and so since A is one to one, $\left(I + A^{-1}B\right)x = 0$. Therefore,

$$\begin{aligned} 0 &= \left|\left|\left(I + A^{-1}B\right)x\right|\right| \geq ||x|| - \left|\left|A^{-1}Bx\right|\right| \\ &\geq ||x|| - \left|\left|A^{-1}B\right|\right| ||x|| = \left(1 - \left|\left|A^{-1}B\right|\right|\right) ||x|| > 0 \end{aligned}$$

a contradiction. This also shows $\left(I + A^{-1}B\right)$ is one to one. Therefore, both $(A+B)^{-1}$ and $\left(I + A^{-1}B\right)^{-1}$ are in $\mathcal{L}(X,X)$. Hence

$$(A+B)^{-1} = \left(A\left(I + A^{-1}B\right)\right)^{-1} = \left(I + A^{-1}B\right)^{-1} A^{-1}$$

Now if

$$x = \left(I + A^{-1}B\right)^{-1} y$$

for $||y|| \leq 1$, then

$$\left(I + A^{-1}B\right)x = y$$

and so

$$||x|| \left(1 - \left|\left|A^{-1}B\right|\right|\right) \leq \left|\left|x + A^{-1}Bx\right|\right| \leq ||y|| = 1$$

and so

$$||x|| = \left|\left|\left(I + A^{-1}B\right)^{-1} y\right|\right| \leq \frac{1}{1 - ||A^{-1}B||}$$

Since $||y|| \leq 1$ is arbitrary, this shows

$$\left|\left|\left(I + A^{-1}B\right)^{-1}\right|\right| \leq \frac{1}{1 - ||A^{-1}B||}$$

Therefore,

$$\begin{aligned} \left|\left|(A+B)^{-1}\right|\right| &= \left|\left|\left(I + A^{-1}B\right)^{-1} A^{-1}\right|\right| \\ &\leq \left|\left|A^{-1}\right|\right| \left|\left|\left(I + A^{-1}B\right)^{-1}\right|\right| \leq \left|\left|A^{-1}\right|\right| \frac{1}{1 - ||A^{-1}B||} \blacksquare \end{aligned}$$

Proposition 14.2.3 *Suppose A is invertible, $b \neq 0$, $Ax = b$, and $A_1 x_1 = b_1$ where $||A - A_1|| < 1/\left|\left|A^{-1}\right|\right|$. Then*

$$\frac{||x_1 - x||}{||x||} \leq \frac{1}{(1 - ||A^{-1}(A_1 - A)||)} ||A|| \left|\left|A^{-1}\right|\right| \left(\frac{||A_1 - A||}{||A||} + \frac{||b - b_1||}{||b||}\right). \tag{14.6}$$

Proof: It follows from the assumptions that

$$Ax - A_1 x + A_1 x - A_1 x_1 = b - b_1.$$

Hence

$$A_1 (x - x_1) = (A_1 - A) x + b - b_1.$$

Now $A_1 = (A + (A_1 - A))$ and so by the above lemma, A_1^{-1} exists and so

$$(x - x_1) = A_1^{-1} (A_1 - A) x + A_1^{-1} (b - b_1)$$

$$= (A + (A_1 - A))^{-1} (A_1 - A) x + (A + (A_1 - A))^{-1} (b - b_1).$$

By the estimate in Lemma 14.2.2,

$$||x - x_1|| \leq \frac{\left|\left|A^{-1}\right|\right|}{1 - ||A^{-1}(A_1 - A)||} (||A_1 - A|| \, ||x|| + ||b - b_1||).$$

Dividing by $||x||$,

$$\frac{||x - x_1||}{||x||} \leq \frac{\left|\left|A^{-1}\right|\right|}{1 - ||A^{-1}(A_1 - A)||} \left(||A_1 - A|| + \frac{||b - b_1||}{||x||}\right) \tag{14.7}$$

Now $b = Ax = A\left(A^{-1}b\right)$ and so $||b|| \leq ||A|| \left|\left|A^{-1}b\right|\right|$ and so

$$||x|| = \left|\left|A^{-1}b\right|\right| \geq ||b|| \, / \, ||A||.$$

Therefore, from 14.7,

$$\begin{aligned} \frac{||x - x_1||}{||x||} &\leq \frac{\left|\left|A^{-1}\right|\right|}{1 - ||A^{-1}(A_1 - A)||} \left(\frac{||A|| \, ||A_1 - A||}{||A||} + \frac{||A|| \, ||b - b_1||}{||b||}\right) \\ &\leq \frac{\left|\left|A^{-1}\right|\right| ||A||}{1 - ||A^{-1}(A_1 - A)||} \left(\frac{||A_1 - A||}{||A||} + \frac{||b - b_1||}{||b||}\right) \end{aligned}$$

which proves the proposition. ■

This shows that the number, $\left|\left|A^{-1}\right|\right| ||A||$, controls how sensitive the relative change in the solution of $Ax = b$ is to small changes in A and b. This number is called the condition number. It is bad when it is large because a small relative change in b, for example could yield a large relative change in x.

Recall that for A an $n \times n$ matrix, $||A||_2 = \sigma_1$ where σ_1 is the largest singular value. The largest singular value of A^{-1} is therefore, $1/\sigma_n$ where σ_n is the smallest singular value of A. Therefore, the condition number reduces to σ_1/σ_n, the ratio of the largest to the smallest singular value of A.

14.3 The Spectral Radius

Even though it is in general impractical to compute the Jordan form, its existence is all that is needed in order to prove an important theorem about something which is relatively easy to compute. This is the spectral radius of a matrix.

Definition 14.3.1 *Define $\sigma(A)$ to be the eigenvalues of A. Also,*

$$\rho(A) \equiv \max(|\lambda| : \lambda \in \sigma(A))$$

The number, $\rho(A)$ is known as the spectral radius of A.

Recall the following symbols and their meaning.

$$\lim\sup_{n\to\infty} a_n, \ \lim\inf_{n\to\infty} a_n$$

They are respectively the largest and smallest limit points of the sequence $\{a_n\}$ where $\pm\infty$ is allowed in the case where the sequence is unbounded. They are also defined as

$$\begin{aligned}\lim\sup_{n\to\infty} a_n &\equiv \lim_{n\to\infty}\left(\sup\{a_k : k \geq n\}\right),\\ \lim\inf_{n\to\infty} a_n &\equiv \lim_{n\to\infty}\left(\inf\{a_k : k \geq n\}\right).\end{aligned}$$

Thus, the limit of the sequence exists if and only if these are both equal to the same real number.

Lemma 14.3.2 *Let J be a $p \times p$ Jordan matrix*

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{pmatrix}$$

where each J_k is of the form

$$J_k = \lambda_k I + N_k$$

in which N_k is a nilpotent matrix having zeros down the main diagonal and ones down the super diagonal. Then

$$\lim_{n\to\infty} ||J^n||^{1/n} = \rho$$

where $\rho = \max\{|\lambda_k|, k = 1, \ldots, n\}$. Here the norm is defined to equal

$$||B|| = \max\{|B_{ij}|, i, j\}.$$

Proof: Suppose first that $\rho \neq 0$. First note that for this norm, if B, C are $p \times p$ matrices,

$$||BC|| \leq p\,||B||\,||C||$$

which follows from a simple computation. Now

$$||J^n||^{1/n} = \left|\left|\begin{pmatrix} (\lambda_1 I + N_1)^n & & \\ & \ddots & \\ & & (\lambda_s I + N_s)^n \end{pmatrix}\right|\right|^{1/n}$$

$$= \rho \left|\left|\begin{pmatrix} \left(\frac{\lambda_1}{\rho} I + \frac{1}{\rho} N_1\right)^n & & \\ & \ddots & \\ & & \left(\frac{\lambda_2}{\rho} I + \frac{1}{\rho} N_2\right)^n \end{pmatrix}\right|\right|^{1/n} \qquad (14.8)$$

From the definition of ρ, at least one of the λ_k/ρ has absolute value equal to 1. Therefore,

$$\left|\left|\begin{pmatrix} \left(\frac{\lambda_1}{\rho}I+\frac{1}{\rho}N_1\right)^n & & \\ & \ddots & \\ & & \left(\frac{\lambda_2}{\rho}I+\frac{1}{\rho}N_2\right)^n \end{pmatrix}\right|\right|^{1/n} - 1 \equiv e_n \geq 0$$

because each N_k has only zero terms on the main diagonal. Therefore, some term in the matrix has absolute value at least as large as 1. Now also, since $N_k^p = 0$, the norm of the matrix in the above is dominated by an expression of the form Cn^p where C is some constant which does not depend on n. This is because a typical block in the above matrix is of the form

$$\sum_{i=1}^{p} \binom{n}{i} \left(\frac{\lambda_k}{\rho}\right)^{n-i} N_k^i$$

and each $|\lambda_k| \leq \rho$.

It follows that for $n > p+1$,

$$Cn^p \geq (1+e_n)^n \geq \binom{n}{p+1} e_n^{p+1}$$

and so

$$\left(\frac{Cn^p}{\binom{n}{p+1}}\right)^{1/(p+1)} \geq e_n \geq 0$$

Therefore, $\lim_{n\to\infty} e_n = 0$. It follows from 14.8 that the expression in the norms in this equation converges to 1 and so

$$\lim_{n\to\infty} ||J^n||^{1/n} = \rho.$$

In case $\rho = 0$ so that all the eigenvalues equal zero, it follows that $J^n = 0$ for all $n > p$. Therefore, the limit still exists and equals ρ. ■

The following theorem is due to Gelfand around 1941.

Theorem 14.3.3 *(Gelfand) Let A be a complex $p \times p$ matrix. Then if ρ is the absolute value of its largest eigenvalue,*

$$\lim_{n\to\infty} ||A^n||^{1/n} = \rho.$$

Here $||\cdot||$ is any norm on $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$.

Proof: First assume $||\cdot||$ is the special norm of the above lemma. Then letting J denote the Jordan form of $A, S^{-1}AS = J$, it follows from Lemma 14.3.2

$$\begin{aligned} \limsup_{n\to\infty} ||A^n||^{1/n} &= \limsup_{n\to\infty} \left|\left|SJ^nS^{-1}\right|\right|^{1/n} \\ &\leq \limsup_{n\to\infty} \left(\left(p^2\right)||S||\left|\left|S^{-1}\right|\right|\right)^{1/n} ||J^n||^{1/n} = \rho \end{aligned}$$

$$\begin{aligned} &= \liminf_{n\to\infty} ||J^n||^{1/n} = \liminf_{n\to\infty} \left|\left|S^{-1}A^nS\right|\right|^n \\ &= \liminf_{n\to\infty} \left(\left(p^2\right)||S||\left|\left|S^{-1}\right|\right|\right)^{1/n} ||A^n||^{1/n} = \liminf_{n\to\infty} ||A^n||^{1/n} \end{aligned}$$

If follows that $\liminf_{n\to\infty} ||A^n||^{1/n} = \limsup_{n\to\infty} ||A^n||^{1/n} = \lim_{n\to\infty} ||A^n||^{1/n} = \rho$.

Now by equivalence of norms, if $|||\cdot|||$ is any other norm for the set of complex $p \times p$ matrices, there exist constants δ, Δ such that

$$\delta \left|\left|A^n\right|\right| \leq |||A^n||| \leq \Delta \left|\left|A^n\right|\right|$$

Then raising to the $1/n$ power and taking a limit,

$$\rho \leq \lim \inf_{n\to\infty} |||A^n|||^{1/n} \leq \lim \sup_{n\to\infty} |||A^n|||^{1/n} \leq \rho \ \blacksquare$$

Example 14.3.4 *Consider* $\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}$. *Estimate the absolute value of the largest eigenvalue.*

A laborious computation reveals the eigenvalues are 5, and 10. Therefore, the right answer in this case is 10. Consider $\left|\left|A^7\right|\right|^{1/7}$ where the norm is obtained by taking the maximum of all the absolute values of the entries. Thus

$$\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}^7 = \begin{pmatrix} 8015\,625 & -1984\,375 & 3968\,750 \\ -3968\,750 & 6031\,250 & 7937\,500 \\ 1984\,375 & 1984\,375 & 6031\,250 \end{pmatrix}$$

and taking the seventh root of the largest entry gives

$$\rho(A) \approx 8015\,625^{1/7} = 9.\,688\,951\,236\,71.$$

Of course the interest lies primarily in matrices for which the exact roots to the characteristic equation are not known and in the theoretical significance.

14.4 Series And Sequences Of Linear Operators

Before beginning this discussion, it is necessary to define what is meant by convergence in $\mathcal{L}(X,Y)$.

Definition 14.4.1 *Let* $\{A_k\}_{k=1}^{\infty}$ *be a sequence in* $\mathcal{L}(X,Y)$ *where* X, Y *are finite dimensional normed linear spaces. Then* $\lim_{n\to\infty} A_k = A$ *if for every* $\varepsilon > 0$ *there exists* N *such that if* $n > N$, *then*

$$||A - A_n|| < \varepsilon.$$

Here the norm refers to any of the norms defined on $\mathcal{L}(X,Y)$. *By Corollary 14.0.8 and Theorem 9.2.3 it doesn't matter which one is used. Define the symbol for an infinite sum in the usual way. Thus*

$$\sum_{k=1}^{\infty} A_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} A_k$$

Lemma 14.4.2 *Suppose* $\{A_k\}_{k=1}^{\infty}$ *is a sequence in* $\mathcal{L}(X,Y)$ *where* X, Y *are finite dimensional normed linear spaces. Then if*

$$\sum_{k=1}^{\infty} ||A_k|| < \infty,$$

It follows that

$$\sum_{k=1}^{\infty} A_k \tag{14.9}$$

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 14.9 are a Cauchy sequence. Therefore by Corollary 14.0.7 it follows that these partial sums converge. ■

As a special case, suppose $\lambda \in \mathbb{C}$ and consider

$$\sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!}$$

where $t \in \mathbb{R}$. In this case, $A_k = \frac{t^k \lambda^k}{k!}$ and you can think of it as being in $\mathcal{L}(\mathbb{C}, \mathbb{C})$. Then the following corollary is of great interest.

Corollary 14.4.3 *Let*

$$f(t) \equiv \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} \equiv 1 + \sum_{k=1}^{\infty} \frac{t^k \lambda^k}{k!}$$

Then this function is a well defined complex valued function and furthermore, it satisfies the initial value problem,

$$y' = \lambda y, \ y(0) = 1$$

Furthermore, if $\lambda = a + ib$,

$$|f|(t) = e^{at}.$$

Proof: That $f(t)$ makes sense follows right away from Lemma 14.4.2.

$$\sum_{k=0}^{\infty} \left|\frac{t^k \lambda^k}{k!}\right| = \sum_{k=0}^{\infty} \frac{|t|^k |\lambda|^k}{k!} = e^{|t||\lambda|}$$

It only remains to verify f satisfies the differential equation because it is obvious from the series that $f(0) = 1$.

$$\frac{f(t+h) - f(t)}{h} = \frac{1}{h} \sum_{k=1}^{\infty} \frac{\left((t+h)^k - t^k\right) \lambda^k}{k!}$$

and by the mean value theorem this equals an expression of the following form where θ_k is a number between 0 and 1.

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{k (t + \theta_k h)^{k-1} \lambda^k}{k!} &= \sum_{k=1}^{\infty} \frac{(t + \theta_k h)^{k-1} \lambda^k}{(k-1)!} \\ &= \lambda \sum_{k=0}^{\infty} \frac{(t + \theta_k h)^k \lambda^k}{k!} \end{aligned}$$

It only remains to verify this converges to

$$\lambda \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} = \lambda f(t)$$

as $h \to 0$.

$$\left| \sum_{k=0}^{\infty} \frac{(t + \theta_k h)^k \lambda^k}{k!} - \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} \right| = \left| \sum_{k=0}^{\infty} \frac{\left((t + \theta_k h)^k - t^k \right) \lambda^k}{k!} \right|$$

and by the mean value theorem again and the triangle inequality

$$\leq \left| \sum_{k=0}^{\infty} \frac{k \left| (t + \eta_k) \right|^{k-1} |h| \, |\lambda|^k}{k!} \right| \leq |h| \sum_{k=0}^{\infty} \frac{k \left| (t + \eta_k) \right|^{k-1} |\lambda|^k}{k!}$$

where η_k is between 0 and 1. Thus

$$\leq |h| \sum_{k=0}^{\infty} \frac{k \left(|t| + 1 \right)^{k-1} |\lambda|^k}{k!} = |h| \, C(t)$$

It follows $f'(t) = \lambda f(t)$. This proves the first part.

Next note that for $f(t) = u(t) + iv(t)$, both u, v are differentiable. This is because

$$u = \frac{f + \overline{f}}{2}, \; v = \frac{f - \overline{f}}{2i}.$$

Then from the differential equation,

$$(a + ib)(u + iv) = u' + iv'$$

and equating real and imaginary parts,

$$u' = au - bv, \; v' = av + bu.$$

Then a short computation shows

$$\left(u^2 + v^2 \right)' = 2a \left(u^2 + v^2 \right), \; \left(u^2 + v^2 \right)(0) = 1.$$

Now in general, if

$$y' = cy, \; y(0) = 1,$$

with c real it follows $y(t) = e^{ct}$. To see this,

$$y' - cy = 0$$

and so, multiplying both sides by e^{-ct} you get

$$\frac{d}{dt} \left(y e^{-ct} \right) = 0$$

and so ye^{-ct} equals a constant which must be 1 because of the initial condition $y(0) = 1$. Thus

$$\left(u^2 + v^2 \right)(t) = e^{2at}$$

and taking square roots yields the desired conclusion. ■

Definition 14.4.4 *The function in Corollary 14.4.3 given by that power series is denoted as*

$$\exp(\lambda t) \text{ or } e^{\lambda t}.$$

The next lemma is normally discussed in advanced calculus courses but is proved here for the convenience of the reader. It is known as the root test.

Definition 14.4.5 *For $\{a_n\}$ any sequence of real numbers*

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} (\sup\{a_k : k \geq n\})$$

Similarly

$$\liminf_{n\to\infty} a_n \equiv \lim_{n\to\infty} (\inf\{a_k : k \geq n\})$$

In case $A_n$is an increasing (decreasing) sequence which is unbounded above (below) then it is understood that $\lim_{n\to\infty} A_n = \infty(-\infty)$ respectively. Thus either of $\limsup$ *or* $\liminf$ *can equal $+\infty$ or $-\infty$. However, the important thing about these is that unlike the limit, these always exist.*

It is convenient to think of these as the largest point which is the limit of some subsequence of $\{a_n\}$ and the smallest point which is the limit of some subsequence of $\{a_n\}$ respectively. Thus $\lim_{n\to\infty} a_n$ exists and equals some point of $[-\infty, \infty]$ if and only if the two are equal.

Lemma 14.4.6 *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

$$1 < \limsup_{p\to\infty} ||A^p||^{1/p}, \tag{14.10}$$

then $\sum_{k=0}^{\infty} A^k$ fails to converge.

Proof: Suppose $r < 1$. Then there exists N such that if $p > N$,

$$a_p^{1/p} < R$$

where $r < R < 1$. Therefore, for all such p, $a_p < R^p$ and so by comparison with the geometric series, $\sum R^p$, it follows $\sum_{p=1}^{\infty} a_p$ converges.

Next suppose $r > 1$. Then letting $1 < R < r$, it follows there are infinitely many values of p at which

$$R < a_p^{1/p}$$

which implies $R^p < a_p$, showing that a_p cannot converge to 0 and so the series cannot converge either.

To see the last claim, if 14.10 holds, then from the first part of this lemma, $||A^p||$ fails to converge to 0 and so $\left\{\sum_{k=0}^{m} A^k\right\}_{m=0}^{\infty}$ is not a Cauchy sequence. Hence $\sum_{k=0}^{\infty} A^k \equiv \lim_{m\to\infty} \sum_{k=0}^{m} A^k$ cannot exist. ■

Now denote by $\sigma(A)^p$ the collection of all numbers of the form λ^p where $\lambda \in \sigma(A)$.

Lemma 14.4.7 $\sigma(A^p) = \sigma(A)^p$

Proof: In dealing with $\sigma(A^p)$, is suffices to deal with $\sigma(J^p)$ where J is the Jordan form of A because J^p and A^p are similar. Thus if $\lambda \in \sigma(A^p)$, then $\lambda \in \sigma(J^p)$ and so $\lambda = \alpha$ where α is one of the entries on the main diagonal of J^p. These entries are of the form λ^p where $\lambda \in \sigma(A)$. Thus $\lambda \in \sigma(A)^p$ and this shows $\sigma(A^p) \subseteq \sigma(A)^p$.

Now take $\alpha \in \sigma(A)$ and consider α^p.

$$\alpha^p I - A^p = \left(\alpha^{p-1} I + \cdots + \alpha A^{p-2} + A^{p-1}\right)(\alpha I - A)$$

and so $\alpha^p I - A^p$ fails to be one to one which shows that $\alpha^p \in \sigma(A^p)$ which shows that $\sigma(A)^p \subseteq \sigma(A^p)$. ■

14.5 Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$A\mathbf{x} = \mathbf{b} \tag{14.11}$$

where A is an $n \times n$ matrix. In many applications, the matrix A is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. To solve them, it is common to use an iterative technique. I am following the treatment given to this subject by Nobel and Daniel [20].

Definition 14.5.1 *The Jacobi iterative technique, also called the method of simultaneous corrections is defined as follows. Let $\mathbf{x}^1$ be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors, $\mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \cdots$ and hopefully this sequence of vectors will converge to the solution to 14.11. The vectors in this list are called iterates and they are obtained according to the following procedure. Letting $A = (a_{ij})$,*

$$a_{ii} x_i^{r+1} = -\sum_{j \neq i} a_{ij} x_j^r + b_i. \tag{14.12}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{aligned} &\begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} \\ = \quad -&\begin{pmatrix} 0 & * & \cdots & * \\ * & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \end{aligned} \tag{14.13}$$

The matrix on the left in 14.13 is obtained by retaining the main diagonal of A and setting every other entry equal to zero. The matrix on the right in 14.13 is obtained from A by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 14.5.2 *Use the Jacobi method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Of course this is solved most easily using row reductions. The Jacobi method is useful when the matrix is 1000×1000 or larger. This example is just to illustrate how the method works. First lets solve it using row operations. The augmented matrix is

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \frac{6}{29} \\ 0 & 1 & 0 & 0 & \frac{11}{29} \\ 0 & 0 & 1 & 0 & \frac{8}{29} \\ 0 & 0 & 0 & 1 & \frac{25}{29} \end{pmatrix}$$

which in terms of decimals is approximately equal to

$$\begin{pmatrix} 1.0 & 0 & 0 & 0 & .206 \\ 0 & 1.0 & 0 & 0 & .379 \\ 0 & 0 & 1.0 & 0 & .275 \\ 0 & 0 & 0 & 1.0 & .862 \end{pmatrix}.$$

In terms of the matrices, the Jacobi iteration is of the form

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1^{r} \\ x_2^{r} \\ x_3^{r} \\ x_4^{r} \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left, [1]this iteration reduces to

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1^{r} \\ x_2^{r} \\ x_3^{r} \\ x_4^{r} \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}. \tag{14.14}$$

Now iterate this starting with

$$\mathbf{x}^1 \equiv \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus

$$\mathbf{x}^2 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}$$

Then

$$\mathbf{x}^3 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}}^{\mathbf{x}_2} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}$$

[1] You certainly would not compute the invese in solving a large system. This is just to show you how the method works for this simple example. You would use the first description in terms of indices.

Continuing this way one finally gets

$$\mathbf{x}^6 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}}^{\mathbf{x}5} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .216 \\ .386 \\ .295 \\ .871 \end{pmatrix}.$$

You can keep going like this. Recall the solution is approximately equal to

$$\begin{pmatrix} .206 \\ .379 \\ .275 \\ .862 \end{pmatrix}$$

so you see that with no care at all and only 6 iterations, an approximate solution has been obtained which is not too far off from the actual solution.

It is important to realize that a computer would use 14.12 directly. Indeed, writing the problem in terms of matrices as I have done above destroys every benefit of the method. However, it makes it a little easier to see what is happening and so this is why I have presented it in this way.

Definition 14.5.3 *The Gauss Seidel method, also called the method of successive corrections is given as follows. For $A = (a_{ij})$, the iterates for the problem $A\mathbf{x} = \mathbf{b}$ are obtained according to the formula*

$$\sum_{j=1}^{i} a_{ij} x_j^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^r + b_i. \tag{14.15}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix}$$

$$= -\begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{14.16}$$

In words, you set every entry in the original matrix which is strictly above the main diagonal equal to zero to obtain the matrix on the left. To get the matrix on the right, you set every entry of A which is on or below the main diagonal equal to zero. Using the iteration procedure of 14.15 directly, the Gauss Seidel method makes use of the very latest information which is available at that stage of the computation.

The following example is the same as the example used to illustrate the Jacobi method.

Example 14.5.4 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

In terms of matrices, this procedure is

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left[2] this yields

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}$$

[2] As in the case of the Jacobi iteration, the computer would not do this. It would use the iteration procedure in terms of the entries of the matrix directly. Otherwise all benefit to using this method is lost.

As before, I will be totally unoriginal in the choice of $\mathbf{x}^1$. Let it equal the zero vector. Therefore,

$$\mathbf{x}^2 = \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}.$$

Now

$$\mathbf{x}^3 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}}^{\mathbf{x}^2} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix}.$$

It follows

$$\mathbf{x}^4=-\begin{pmatrix}0&\frac{1}{3}&0&0\\0&-\frac{1}{12}&\frac{1}{4}&0\\0&\frac{1}{30}&-\frac{1}{10}&\frac{1}{5}\\0&-\frac{1}{60}&\frac{1}{20}&-\frac{1}{10}\end{pmatrix}\begin{pmatrix}.194\\.343\\.306\\.846\end{pmatrix}+\begin{pmatrix}\frac{1}{3}\\\frac{5}{12}\\\frac{13}{30}\\\frac{47}{60}\end{pmatrix}=\begin{pmatrix}.219\\.368\,75\\.283\,3\\.858\,35\end{pmatrix}$$

and so

$$\mathbf{x}^5=-\begin{pmatrix}0&\frac{1}{3}&0&0\\0&-\frac{1}{12}&\frac{1}{4}&0\\0&\frac{1}{30}&-\frac{1}{10}&\frac{1}{5}\\0&-\frac{1}{60}&\frac{1}{20}&-\frac{1}{10}\end{pmatrix}\begin{pmatrix}.219\\.368\,75\\.283\,3\\.858\,35\end{pmatrix}+\begin{pmatrix}\frac{1}{3}\\\frac{5}{12}\\\frac{13}{30}\\\frac{47}{60}\end{pmatrix}=\begin{pmatrix}.210\,42\\.376\,57\\.277\,7\\.861\,15\end{pmatrix}.$$

Recall the answer is

$$\begin{pmatrix}.206\\.379\\.275\\.862\end{pmatrix}$$

so the iterates are already pretty close to the answer. You could continue doing these iterates and it appears they converge to the solution. Now consider the following example.

Example 14.5.5 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix}1&4&0&0\\1&4&1&0\\0&2&5&1\\0&0&2&4\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}1\\2\\3\\4\end{pmatrix}$$

The exact solution is given by doing row operations on the augmented matrix. When this is done the row echelon form is

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

and so the solution is approximately

$$\begin{pmatrix}6\\-\frac{5}{4}\\1\\\frac{1}{2}\end{pmatrix}=\begin{pmatrix}6.0\\-1.25\\1.0\\.5\end{pmatrix}$$

The Gauss Seidel iterations are of the form

$$\begin{pmatrix}1&0&0&0\\1&4&0&0\\0&2&5&0\\0&0&2&4\end{pmatrix}\begin{pmatrix}x_1^{r+1}\\x_2^{r+1}\\x_3^{r+1}\\x_4^{r+1}\end{pmatrix}=-\begin{pmatrix}0&4&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{pmatrix}\begin{pmatrix}x_1^r\\x_2^r\\x_3^r\\x_4^r\end{pmatrix}+\begin{pmatrix}1\\2\\3\\4\end{pmatrix}$$

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

$$\mathbf{x}^{r+1}=-\begin{pmatrix}0&4&0&0\\0&-1&\frac{1}{4}&0\\0&\frac{2}{5}&-\frac{1}{10}&\frac{1}{5}\\0&-\frac{1}{5}&\frac{1}{20}&-\frac{1}{10}\end{pmatrix}\mathbf{x}^r+\begin{pmatrix}1\\\frac{1}{4}\\\frac{1}{2}\\\frac{3}{4}\end{pmatrix}.$$

This time, I will pick an initial vector close to the answer. Let

$$\mathbf{x}^1 = \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

$$\mathbf{x}^2 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix}$$

You can't expect to be real close after only one iteration. Lets do another.

$$\mathbf{x}^3 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -.975 \\ .88 \\ .56 \end{pmatrix}$$

$$\mathbf{x}^4 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 5.0 \\ -.975 \\ .88 \\ .56 \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 4.9 \\ -.945 \\ .866 \\ .567 \end{pmatrix}$$

The iterates seem to be getting farther from the actual solution. Why is the process which worked so well in the other examples not working here? A better question might be: Why does either process ever work at all?

Both iterative procedures for solving

$$A\mathbf{x} = \mathbf{b} \tag{14.17}$$

are of the form

$$B\mathbf{x}^{r+1} = -C\mathbf{x}^r + \mathbf{b}$$

where $A = B + C$. In the Jacobi procedure, the matrix C was obtained by setting the diagonal of A equal to zero and leaving all other entries the same while the matrix B was obtained by making every entry of A equal to zero other than the diagonal entries which are left unchanged. In the Gauss Seidel procedure, the matrix B was obtained from A by making every entry strictly above the main diagonal equal to zero and leaving the others unchanged and C was obtained from A by making every entry on or below the main diagonal equal to zero and leaving the others unchanged. Thus in the Jacobi procedure, B is a diagonal matrix while in the Gauss Seidel procedure, B is lower triangular. Using matrices to explicitly solve for the iterates, yields

$$\mathbf{x}^{r+1} = -B^{-1}C\mathbf{x}^r + B^{-1}\mathbf{b}. \tag{14.18}$$

This is what you would never have the computer do but this is what will allow the statement of a theorem which gives the condition for convergence of these and all other similar methods. Recall the definition of the spectral radius of $M, \rho(M)$, in Definition 14.3.1 on Page 459.

Theorem 14.5.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 14.18 converge to the unique solution of 14.17.*

I will prove this theorem in the next section. The proof depends on analysis which should not be surprising because it involves a statement about convergence of sequences.

What is an easy to verify sufficient condition which will imply the above holds? It is easy to give one in the case of the Jacobi method. Suppose the matrix A is diagonally dominant. That is $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$. Then B would be the diagonal matrix consisting of the entries $|a_{ii}|$. You can see then that every entry of $B^{-1}C$ has absolute value less than 1. Thus if you let the norm $\left|\left|B^{-1}C\right|\right|_\infty$ be given by the maximum of the absolute values of the entries of the matrix, then $\left|\left|B^{-1}C\right|\right|_\infty = r < 1$. Also, by equivalence of norms it follows there exist positive constants δ, Δ such that

$$\delta \left|\left|\cdot\right|\right| \leq \left|\left|\cdot\right|\right|_\infty \leq \Delta \left|\left|\cdot\right|\right|$$

where here $||\cdot||$ is an operator norm. It follows that if $|\lambda| \geq 1$, then $\left(\lambda I - B^{-1}C\right)^{-1}$ exists. In fact it equals

$$\sum_{k=0}^{\infty} \lambda^{-1} \left(\frac{B^{-1}C}{\lambda}\right)^k,$$

the series converging because

$$\left|\left|\sum_{k=m}^{n} \left(\frac{B^{-1}C}{\lambda}\right)^k\right|\right|_\infty \leq \sum_{k=m}^{\infty} \left|\left|\left(\frac{B^{-1}C}{\lambda}\right)^k\right|\right|_\infty$$

$$\leq \sum_{k=m}^{\infty} \Delta \left|\left|\left(\frac{B^{-1}C}{\lambda}\right)^k\right|\right|_\infty \leq \sum_{k=m}^{\infty} \Delta \left|\left|\left(\frac{B^{-1}C}{\lambda}\right)\right|\right|^k$$

$$\leq \sum_{k=m}^{\infty} \frac{\Delta}{\delta} \left|\left|\left(\frac{B^{-1}C}{\lambda}\right)\right|\right|_\infty^k \leq \frac{\Delta}{\delta} \sum_{k=m}^{\infty} r^k \leq \frac{\Delta}{\delta}\left(\frac{r^m}{1-r}\right)$$

which shows the partial sums form a Cauchy sequence. Therefore, $\rho\left(B^{-1}C\right) < 1$ in this case.

You might try a similar argument in the case of the Gauss Seidel method.

14.6 Theory Of Convergence

Definition 14.6.1 *A normed vector space, E with norm $||\cdot||$ is called a Banach space if it is also complete. This means that every Cauchy sequence converges. Recall that a sequence $\{x_n\}_{n=1}^\infty$ is a Cauchy sequence if for every $\varepsilon > 0$ there exists N such that whenever $m, n > N$,*

$$||x_n - x_m|| < \varepsilon.$$

Thus whenever $\{x_n\}$ is a Cauchy sequence, there exists x such that

$$\lim_{n\to\infty} ||x - x_n|| = 0.$$

Example 14.6.2 *Let Ω be a nonempty subset of a normed linear space, F. Denote by $BC(\Omega; E)$ the set of bounded continuous functions having values in E where E is a Banach space. Then define the norm on $BC(\Omega; E)$ by*

$$||f|| \equiv \sup \{||f(x)||_E : x \in \Omega\} .$$

Lemma 14.6.3 *The space $BC(\Omega; E)$ with the given norm is a Banach space.*

Proof: It is obvious $||\cdot||$ is a norm. It only remains to verify $BC(\Omega; E)$ is complete. Let $\{f_n\}$ be a Cauchy sequence. Then pick $x \in \Omega$.

$$||f_n(x) - f_m(x)||_E \leq ||f_n - f_m|| < \varepsilon$$

whenever m, n are large enough. Thus, for each $x, \{f_n(x)\}$ is a Cauchy sequence in E. Since E is complete, it follows there exists a function, f defined on Ω such that $f(x) = \lim_{n\to\infty} f_n(x)$.

It remains to verify that $f \in BC(\Omega; E)$ and that $||f - f_n|| \to 0$. I will first show that

$$\lim_{n\to\infty} \left(\sup_{x\in\Omega} \{||f(x) - f_n(x)||_E\} \right) = 0. \tag{14.19}$$

From this it will follow that f is bounded. Then I will show that f is continuous and $||f - f_n|| \to 0$. Let $\varepsilon > 0$ be given and let N be such that for $m, n > N$

$$||f_n - f_m|| < \varepsilon/3.$$

Then it follows that for all x,

$$||f(x) - f_m(x)||_E = \lim_{n\to\infty} ||f_n(x) - f_m(x)||_E \leq \varepsilon/3$$

Therefore, for $m > N$,

$$\sup_{x\in\Omega} \{||f(x) - f_m(x)||_E\} \leq \frac{\varepsilon}{3} < \varepsilon.$$

This proves 14.19. Then by the triangle inequality and letting N be as just described, pick $m > N$. Then for any $x \in \Omega$

$$||f(x)||_E \leq ||f_m(x)||_E + \varepsilon \leq ||f_m|| + \varepsilon.$$

Hence f is bounded. Now pick $x \in \Omega$ and let $\varepsilon > 0$ be given and N be as above. Then

$$\begin{aligned} ||f(x) - f(y)||_E &\leq ||f(x) - f_m(x)||_E + ||f_m(x) - f_m(y)||_E + ||f_m(y) - f(y)||_E \\ &\leq \frac{\varepsilon}{3} + ||f_m(x) - f_m(y)||_E + \frac{\varepsilon}{3}. \end{aligned}$$

Now by continuity of f_m, the middle term is less than $\varepsilon/3$ whenever $||x - y||$ is sufficiently small. Therefore, f is also continuous. Finally, from the above,

$$||f - f_n|| \leq \frac{\varepsilon}{3}$$

whenever $n > N$ and so $\lim_{n\to\infty} ||f - f_n|| = 0$ as claimed. ■

The most familiar example of a Banach space is $\mathbb{F}^n$. The following lemma is of great importance so it is stated in general.

Lemma 14.6.4 *Suppose $T : E \to E$ where E is a Banach space with norm $|\cdot|$. Also suppose*

$$|T\mathbf{x} - T\mathbf{y}| \leq r|\mathbf{x} - \mathbf{y}| \tag{14.20}$$

for some $r \in (0, 1)$. Then there exists a unique fixed point, $\mathbf{x} \in E$ such that

$$T\mathbf{x} = \mathbf{x}. \tag{14.21}$$

Letting $\mathbf{x}^1 \in E$, this fixed point, $\mathbf{x}$, is the limit of the sequence of iterates,

$$\mathbf{x}^1, T\mathbf{x}^1, T^2\mathbf{x}^1, \cdots. \tag{14.22}$$

In addition to this, there is a nice estimate which tells how close $\mathbf{x}^1$ is to $\mathbf{x}$ in terms of things which can be computed.

$$\left|\mathbf{x}^1 - \mathbf{x}\right| \leq \frac{1}{1-r} \left|\mathbf{x}^1 - T\mathbf{x}^1\right|. \tag{14.23}$$

Proof: This follows easily when it is shown that the above sequence, $\left\{T^k\mathbf{x}^1\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Note that

$$\left|T^2\mathbf{x}^1 - T\mathbf{x}^1\right| \leq r\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.$$

Suppose

$$\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \leq r^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right|. \tag{14.24}$$

Then

$$\begin{aligned}\left|T^{k+1}\mathbf{x}^1 - T^k\mathbf{x}^1\right| &\leq r\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \\ &\leq rr^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right| = r^k\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.\end{aligned}$$

By induction, this shows that for all $k \geq 2$, 14.24 is valid. Now let $k > l \geq N$.

$$\begin{aligned}\left|T^k\mathbf{x}^1 - T^l\mathbf{x}^1\right| &= \left|\sum_{j=l}^{k-1}\left(T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right)\right| \leq \sum_{j=l}^{k-1}\left|T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right| \\ &\leq \sum_{j=N}^{k-1} r^j\left|T\mathbf{x}^1 - \mathbf{x}^1\right| \leq \left|T\mathbf{x}^1 - \mathbf{x}^1\right|\frac{r^N}{1-r}\end{aligned}$$

which converges to 0 as $N \to \infty$. Therefore, this is a Cauchy sequence so it must converge to $\mathbf{x} \in E$. Then

$$\mathbf{x} = \lim_{k\to\infty} T^k\mathbf{x}^1 = \lim_{k\to\infty} T^{k+1}\mathbf{x}^1 = T\lim_{k\to\infty} T^k\mathbf{x}^1 = T\mathbf{x}.$$

This shows the existence of the fixed point. To show it is unique, suppose there were another one, $\mathbf{y}$. Then

$$|\mathbf{x} - \mathbf{y}| = |T\mathbf{x} - T\mathbf{y}| \leq r|\mathbf{x} - \mathbf{y}|$$

and so $\mathbf{x} = \mathbf{y}$.

It remains to verify the estimate.

$$\begin{aligned}\left|\mathbf{x}^1 - \mathbf{x}\right| &\leq \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + \left|T\mathbf{x}^1 - \mathbf{x}\right| = \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + \left|T\mathbf{x}^1 - T\mathbf{x}\right| \\ &\leq \left|\mathbf{x}^1 - T\mathbf{x}^1\right| + r\left|\mathbf{x}^1 - \mathbf{x}\right|\end{aligned}$$

and solving the inequality for $\left|\mathbf{x}^1 - \mathbf{x}\right|$ gives the estimate desired. ■

The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

Corollary 14.6.5 *Suppose $T : E \to E$, for some constant C*

$$|T\mathbf{x} - T\mathbf{y}| \leq C|\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in E$, and for some $N \in \mathbb{N}$,

$$\left|T^N\mathbf{x} - T^N\mathbf{y}\right| \leq r|\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in E$ where $r \in (0,1)$. Then there exists a unique fixed point for T and it is still the limit of the sequence, $\left\{T^k\mathbf{x}^1\right\}$ for any choice of $\mathbf{x}^1$.

Proof: From Lemma 14.6.4 there exists a unique fixed point for T^N denoted here as $\mathbf{x}$. Therefore, $T^N \mathbf{x} = \mathbf{x}$. Now doing T to both sides,

$$T^N T\mathbf{x} = T\mathbf{x}.$$

By uniqueness, $T\mathbf{x} = \mathbf{x}$ because the above equation shows $T\mathbf{x}$ is a fixed point of T^N and there is only one fixed point of T^N. In fact, there is only one fixed point of T because a fixed point of T is automatically a fixed point of T^N.

It remains to show $T^k \mathbf{x}^1 \to \mathbf{x}$, the unique fixed point of T^N. If this does not happen, there exists $\varepsilon > 0$ and a subsequence, still denoted by T^k such that

$$\left|T^k \mathbf{x}^1 - \mathbf{x}\right| \geq \varepsilon$$

Now $k = j_k N + r_k$ where $r_k \in \{0, \cdots, N-1\}$ and j_k is a positive integer such that $\lim_{k\to\infty} j_k = \infty$. Then there exists a single $r \in \{0, \cdots, N-1\}$ such that for infinitely many $k, r_k = r$. Taking a further subsequence, still denoted by T^k it follows

$$\left|T^{j_k N + r} \mathbf{x}^1 - \mathbf{x}\right| \geq \varepsilon \tag{14.25}$$

However,

$$T^{j_k N + r} \mathbf{x}^1 = T^r T^{j_k N} \mathbf{x}^1 \to T^r \mathbf{x} = \mathbf{x}$$

and this contradicts 14.25. ■

Theorem 14.6.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 14.18 converge to the unique solution of 14.17.*

Proof: Consider the iterates in 14.18. Let $T\mathbf{x} = B^{-1}C\mathbf{x} + \mathbf{b}$. Then

$$\left|T^k \mathbf{x} - T^k \mathbf{y}\right| = \left|\left(B^{-1}C\right)^k \mathbf{x} - \left(B^{-1}C\right)^k \mathbf{y}\right| \leq \left|\left|\left(B^{-1}C\right)^k\right|\right| \left|\mathbf{x} - \mathbf{y}\right|.$$

Here $||\cdot||$ refers to any of the operator norms. It doesn't matter which one you pick because they are all equivalent. I am writing the proof to indicate the operator norm taken with respect to the usual norm on E. Since $\rho\left(B^{-1}C\right) < 1$, it follows from Gelfand's theorem, Theorem 14.3.3 on Page 461, there exists N such that if $k \geq N$, then for some $r^{1/k} < 1$,

$$\left|\left|\left(B^{-1}C\right)^{k}\right|\right|^{1/k} < r^{1/k} < 1.$$

Consequently,

$$\left|T^{N}\mathbf{x} - T^{N}\mathbf{y}\right| \leq r\left|\mathbf{x} - \mathbf{y}\right|.$$

Also $|T\mathbf{x} - T\mathbf{y}| \leq \left|\left|B^{-1}C\right|\right| |\mathbf{x} - \mathbf{y}|$ and so Corollary 14.6.5 applies and gives the conclusion of this theorem. ■

14.7 Exercises

1. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

2. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

3. Solve the system
$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

4. If you are considering a system of the form $A\mathbf{x} = \mathbf{b}$ and A^{-1} does not exist, will either the Gauss Seidel or Jacobi methods work? Explain. What does this indicate about finding eigenvectors for a given eigenvalue?

5. For $||\mathbf{x}||_\infty \equiv \max\{|x_j| : j = 1, 2, \cdots, n\}$, the parallelogram identity does not hold. Explain.

6. A norm $||\cdot||$ is said to be strictly convex if whenever $||x|| = ||y||, x \neq y$, it follows
$$\left|\left|\frac{x+y}{2}\right|\right| < ||x|| = ||y|| \, .$$
Show the norm $|\cdot|$ which comes from an inner product is strictly convex.

7. A norm $||\cdot||$ is said to be uniformly convex if whenever $||x_n||, ||y_n||$ are equal to 1 for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} ||x_n + y_n|| = 2$, it follows $\lim_{n\to\infty} ||x_n - y_n|| = 0$. Show the norm $|\cdot|$ coming from an inner product is always uniformly convex. Also show that uniform convexity implies strict convexity which is defined in Problem 6.

8. Suppose $A : \mathbb{C}^n \to \mathbb{C}^n$ is a one to one and onto matrix. Define
$$||\mathbf{x}|| \equiv |A\mathbf{x}| \, .$$
Show this is a norm.

9. If X is a finite dimensional normed vector space and $A, B \in \mathcal{L}(X, X)$ such that $||B|| < ||A||$, can it be concluded that $\left|\left|A^{-1}B\right|\right| < 1$?

10. Let X be a vector space with a norm $||\cdot||$ and let $V = \text{span}\,(v_1, \cdots, v_m)$ be a finite dimensional subspace of X such that $\{v_1, \cdots, v_m\}$ is a basis for V. Show V is a closed subspace of X. This means that if $w_n \to w$ and each $w_n \in V$, then so is w. Next show that if $w \notin V$,

$$\text{dist}\,(w, V) \equiv \inf \{||w - v|| : v \in V\} > 0$$

is a continuous function of w and

$$|\text{dist}\,(w, V) - \text{dist}\,(w_1, V)| \leq \|w_1 - w\|$$

Next show that if $w \notin V$, there exists z such that $||z|| = 1$ and $\text{dist}\,(z, V) > 1/2$. For those who know some advanced calculus, show that if X is an infinite dimensional vector space having norm $||\cdot||$, then the closed unit ball in X cannot be compact. Thus closed and bounded is never compact in an infinite dimensional normed vector space.

11. Suppose $\rho(A) < 1$ for $A \in \mathcal{L}(V, V)$ where V is a p dimensional vector space having a norm $||\cdot||$. You can use $\mathbb{R}^p$ or $\mathbb{C}^p$ if you like. Show there exists a new norm $|||\cdot|||$ such that with respect to this new norm, $|||A||| < 1$ where $|||A|||$ denotes the operator norm of A taken with respect to this new norm on V,

$$|||A||| \equiv \sup \{|||A\mathbf{x}||| : |||\mathbf{x}||| \leq 1\}$$

Hint: You know from Gelfand's theorem that

$$||A^n||^{1/n} < r < 1$$

provided n is large enough, this operator norm taken with respect to $||\cdot||$. Show there exists $0 < \lambda < 1$ such that

$$\rho\left(\frac{A}{\lambda}\right) < 1.$$

You can do this by arguing the eigenvalues of A/λ are the scalars μ/λ where $\mu \in \sigma(A)$. Now let $\mathbb{Z}_+$ denote the nonnegative integers.

$$|||\mathbf{x}||| \equiv \sup_{n \in \mathbb{Z}_+} \left|\left|\frac{A^n}{\lambda^n}\mathbf{x}\right|\right|$$

First show this is actually a norm. Next explain why

$$|||A\mathbf{x}||| \equiv \lambda \sup_{n \in \mathbb{Z}_+} \left|\left|\frac{A^{n+1}}{\lambda^{n+1}}\mathbf{x}\right|\right| \leq \lambda\,|||\mathbf{x}|||\,.$$

12. Establish a similar result to Problem 11 without using Gelfand's theorem. Use an argument which depends directly on the Jordan form or a modification of it.

13. Using Problem 11 give an easier proof of Theorem 14.6.6 without having to use Corollary 14.6.5. It would suffice to use a different norm of this problem and the contraction mapping principle of Lemma 14.6.4.

14. A matrix A is diagonally dominant if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$. Show that the Gauss Seidel method converges if A is diagonally dominant.

15. Suppose $f(\lambda) = \sum_{k=0}^{\infty} a_n \lambda^n$ converges if $|\lambda| < R$. Show that if $\rho(A) < R$ where A is an $n \times n$ matrix, then

$$f(A) \equiv \sum_{k-0}^{\infty} a_n A^n$$

converges in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$. **Hint:** Use Gelfand's theorem and the root test.

16. Referring to Corollary 14.4.3, for $\lambda = a + ib$ show

$$\exp(\lambda t) = e^{at}(\cos(bt) + i\sin(bt)).$$

Hint: Let $y(t) = \exp(\lambda t)$ and let $z(t) = e^{-at}y(t)$. Show

$$z'' + b^2 z = 0,\ z(0) = 1, z'(0) = ib.$$

Now letting $z = u + iv$ where u, v are real valued, show

$$\begin{aligned} u'' + b^2 u &= 0,\ u(0) = 1, u'(0) = 0 \\ v'' + b^2 v &= 0,\ v(0) = 0, v'(0) = b. \end{aligned}$$

Next show $u(t) = \cos(bt)$ and $v(t) = \sin(bt)$ work in the above and that there is at most one solution to

$$w'' + b^2 w = 0\ w(0) = \alpha, w'(0) = \beta.$$

Thus $z(t) = \cos(bt) + i\sin(bt)$ and so $y(t) = e^{at}(\cos(bt) + i\sin(bt))$. To show there is at most one solution to the above problem, suppose you have two, w_1, w_2. Subtract them. Let $f = w_1 - w_2$. Thus

$$f'' + b^2 f = 0$$

and f is real valued. Multiply both sides by f' and conclude

$$\frac{d}{dt}\left(\frac{(f')^2}{2} + b^2\frac{f^2}{2}\right) = 0$$

Thus the expression in parenthesis is constant. Explain why this constant must equal 0.

17. Let $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. Show the following power series converges in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$.

$$\sum_{k=0}^{\infty} \frac{t^k A^k}{k!}$$

You might want to use Lemma 14.4.2. This is how you can define $\exp(tA)$. Next show using arguments like those of Corollary 14.4.3

$$\frac{d}{dt}\exp(tA) = A\exp(tA)$$

so that this is a matrix valued solution to the differential equation and initial condition

$$\Psi'(t) = A\Psi(t),\ \Psi(0) = I.$$

This $\Psi(t)$ is called a fundamental matrix for the differential equation $\mathbf{y}' = A\mathbf{y}$. Show $t \to \Psi(t)\mathbf{y}_0$ gives a solution to the initial value problem

$$\mathbf{y}' = A\mathbf{y},\ \mathbf{y}(0) = \mathbf{y}_0.$$

18. In Problem 17 $\Psi(t)$ is defined by the given series. Denote by $\exp(t\sigma(A))$ the numbers $\exp(t\lambda)$ where $\lambda \in \sigma(A)$. Show $\exp(t\sigma(A)) = \sigma(\Psi(t))$. This is like Lemma 14.4.7. Letting J be the Jordan canonical form for A, explain why

$$\Psi(t) \equiv \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = S \sum_{k=0}^{\infty} \frac{t^k J^k}{k!} S^{-1}$$

and you note that in J^k, the diagonal entries are of the form λ^k for λ an eigenvalue of A. Also $J = D + N$ where N is nilpotent and commutes with D. Argue then that

$$\sum_{k=0}^{\infty} \frac{t^k J^k}{k!}$$

is an upper triangular matrix which has on the diagonal the expressions $e^{\lambda t}$ where $\lambda \in \sigma(A)$. Thus conclude

$$\sigma(\Psi(t)) \subseteq \exp(t\sigma(A))$$

Next take $e^{t\lambda} \in \exp(t\sigma(A))$ and argue it must be in $\sigma(\Psi(t))$. You can do this as follows:

$$\begin{aligned}\Psi(t) - e^{t\lambda} I &= \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} - \sum_{k=0}^{\infty} \frac{t^k \lambda^k}{k!} I = \sum_{k=0}^{\infty} \frac{t^k}{k!} \left(A^k - \lambda^k I \right) \\ &= \left(\sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j \right) (A - \lambda I)\end{aligned}$$

Now you need to argue

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{j=1}^{k-1} A^{k-j} \lambda^j$$

converges to something in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. To do this, use the ratio test and Lemma 14.4.2 after first using the triangle inequality. Since $\lambda \in \sigma(A)$, $\Psi(t) - e^{t\lambda} I$ is not one to one and so this establishes the other inclusion. You fill in the details. This theorem is a special case of theorems which go by the name "spectral mapping theorem".

19. Suppose $\Psi(t) \in \mathcal{L}(V, W)$ where V, W are finite dimensional inner product spaces and $t \to \Psi(t)$ is continuous for $t \in [a, b]$: For every $\varepsilon > 0$ there there exists $\delta > 0$ such that if $|s - t| < \delta$ then $||\Psi(t) - \Psi(s)|| < \varepsilon$. Show $t \to (\Psi(t) v, w)$ is continuous. Here it is the inner product in W. Also define what it means for $t \to \Psi(t) v$ to be continuous and show this is continuous. Do it all for differentiable in place of continuous. Next show $t \to ||\Psi(t)||$ is continuous.

20. If $z(t) \in W$, a finite dimensional inner product space, what does it mean for $t \to z(t)$ to be continuous or differentiable? If z is continuous, define

$$\int_a^b z(t)\, dt \in W$$

as follows.

$$\left(w, \int_a^b z(t)\, dt \right) \equiv \int_a^b (w, z(t))\, dt.$$

Show that this definition is well defined and furthermore the triangle inequality,

$$\left| \int_a^b z(t)\, dt \right| \leq \int_a^b |z(t)|\, dt,$$

and fundamental theorem of calculus,

$$\frac{d}{dt} \left(\int_a^t z(s)\, ds \right) = z(t)$$

hold along with any other interesting properties of integrals which are true.

21. For V, W two inner product spaces, define

$$\int_a^b \Psi(t)\, dt \in \mathcal{L}(V, W)$$

as follows.

$$\left(w, \int_a^b \Psi(t)\, dt\, (v) \right) \equiv \int_a^b (w, \Psi(t) v)\, dt.$$

Show this is well defined and does indeed give $\int_a^b \Psi(t)\,dt \in \mathcal{L}(V,W)$. Also show the triangle inequality

$$\left|\left|\int_a^b \Psi(t)\,dt\right|\right| \leq \int_a^b ||\Psi(t)||\,dt$$

where $||\cdot||$ is the operator norm and verify the fundamental theorem of calculus holds.

$$\left(\int_a^t \Psi(s)\,ds\right)' = \Psi(t)\,.$$

Also verify the usual properties of integrals continue to hold such as the fact the integral is linear and

$$\int_a^b \Psi(t)\,dt + \int_b^c \Psi(t)\,dt = \int_a^c \Psi(t)\,dt$$

and similar things. **Hint:** On showing the triangle inequality, it will help if you use the fact that

$$|w|_W = \sup_{|v|\leq 1} |(w,v)|\,.$$

You should show this also.

22. Prove Gronwall's inequality. Suppose $u(t) \geq 0$ and for all $t \in [0,T]$,

$$u(t) \leq u_0 + \int_0^t Ku(s)\,ds.$$

where K is some nonnegative constant. Then

$$u(t) \leq u_0 e^{Kt}.$$

Hint: $w(t) = \int_0^t u(s)\,ds$. Then using the fundamental theorem of calculus, $w(t)$ satisfies the following.

$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0,\ w(0) = 0.$$

Now use the usual techniques you saw in an introductory differential equations class. Multiply both sides of the above inequality by e^{-Kt} and note the resulting left side is now a total derivative. Integrate both sides from 0 to t and see what you have got. If you have problems, look ahead in the book. This inequality is proved later in Theorem C.4.3.

23. With Gronwall's inequality and the integral defined in Problem 21 with its properties listed there, prove there is at most one solution to the initial value problem

$$\mathbf{y}' = A\mathbf{y},\ \mathbf{y}(0) = \mathbf{y}_0.$$

Hint: If there are two solutions, subtract them and call the result $\mathbf{z}$. Then

$$\mathbf{z}' = A\mathbf{z},\ \mathbf{z}(0) = \mathbf{0}.$$

It follows

$$\mathbf{z}(t) = \mathbf{0} + \int_0^t A\mathbf{z}(s)\,ds$$

and so

$$||\mathbf{z}(t)|| \leq \int_0^t \|A\| \, ||\mathbf{z}(s)|| \, ds$$

Now consider Gronwall's inequality of Problem 22.

24. Suppose A is a matrix which has the property that whenever $\mu \in \sigma(A)$, $\operatorname{Re} \mu < 0$. Consider the initial value problem

$$\mathbf{y}' = A\mathbf{y}, \mathbf{y}(0) = \mathbf{y}_0.$$

The existence and uniqueness of a solution to this equation has been established above in preceding problems, Problem 17 to 23. Show that in this case where the real parts of the eigenvalues are all negative, the solution to the initial value problem satisfies

$$\lim_{t \to \infty} \mathbf{y}(t) = \mathbf{0}.$$

Hint: A nice way to approach this problem is to show you can reduce it to the consideration of the initial value problem

$$\mathbf{z}' = J_\varepsilon \mathbf{z},\ \mathbf{z}(0) = \mathbf{z}_0$$

where J_ε is the modified Jordan canonical form where instead of ones down the main diagonal, there are ε down the main diagonal (Problem 19). Then

$$\mathbf{z}' = D\mathbf{z} + N_\varepsilon \mathbf{z}$$

where D is the diagonal matrix obtained from the eigenvalues of A and N_ε is a nilpotent matrix commuting with D which is very small provided ε is chosen very small. Now let $\Psi(t)$ be the solution of

$$\Psi' = -D\Psi,\ \Psi(0) = I$$

described earlier as

$$\sum_{k=0}^{\infty} \frac{(-1)^k t^k D^k}{k!}.$$

Thus $\Psi(t)$ commutes with D and N_ε. Tell why. Next argue

$$(\Psi(t)\mathbf{z})' = \Psi(t) N_\varepsilon \mathbf{z}(t)$$

and integrate from 0 to t. Then

$$\Psi(t)\mathbf{z}(t) - \mathbf{z}_0 = \int_0^t \Psi(s) N_\varepsilon \mathbf{z}(s)\, ds.$$

It follows

$$||\Psi(t)\mathbf{z}(t)|| \leq ||z_0|| + \int_0^t ||N_\varepsilon||\, ||\Psi(s)\mathbf{z}(s)||\, ds.$$

It follows from Gronwall's inequality

$$||\Psi(t)\mathbf{z}(t)|| \leq ||z_0||\, e^{||N_\varepsilon||t}$$

Now look closely at the form of $\Psi(t)$ to get an estimate which is interesting. Explain why

$$\Psi(t) = \begin{pmatrix} e^{\mu_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\mu_n t} \end{pmatrix}$$

and now observe that if ε is chosen small enough, $||N_\varepsilon||$ is so small that each component of $\mathbf{z}(t)$ converges to 0.

25. Using Problem 24 show that if A is a matrix having the real parts of all eigenvalues less than 0 then if

$$\Psi'(t) = A\Psi(t),\ \Psi(0) = I$$

it follows

$$\lim_{t\to\infty} \Psi(t) = 0.$$

Hint: Consider the columns of $\Psi(t)$?

26. Let $\Psi(t)$ be a fundamental matrix satisfying

$$\Psi'(t) = A\Psi(t), \ \Psi(0) = I.$$

Show $\Psi(t)^n = \Psi(nt)$. **Hint:** Subtract and show the difference satisfies $\Phi' = A\Phi, \ \Phi(0) = 0$. Use uniqueness.

27. If the real parts of the eigenvalues of A are all negative, show that for every positive t,

$$\lim_{n\to\infty} \Psi(nt) = 0.$$

Hint: Pick $\text{Re}(\sigma(A)) < -\lambda < 0$ and use Problem 18 about the spectrum of $\Psi(t)$ and Gelfand's theorem for the spectral radius along with Problem 26 to argue that $\left|\left|\Psi(nt)/e^{-\lambda nt}\right|\right| < 1$ for all n large enough.

28. Let H be a Hermitian matrix. $(H = H^*)$. Show that $e^{iH} \equiv \sum_{n=0}^{\infty} \frac{(iH)^n}{n!}$ is unitary.

29. Show the converse of the above exercise. If V is unitary, then $V = e^{iH}$ for some H Hermitian.

30. If U is unitary and does not have -1 as an eigenvalue so that $(I+U)^{-1}$ exists, show that

$$H = i(I-U)(I+U)^{-1}$$

is Hermitian. Then, verify that

$$U = (I+iH)(I-iH)^{-1}.$$

31. Suppose that $A \in \mathcal{L}(V,V)$ where V is a normed linear space. Also suppose that $\|A\| < 1$ where this refers to the operator norm on A. Verify that

$$(I - A)^{-1} = \sum_{i=0}^{\infty} A^i$$

This is called the Neumann series. Suppose now that you only know the algebraic condition $\rho(A) < 1$. Is it still the case that the Neumann series converges to $(I - A)^{-1}$?

Numerical Methods, Eigenvalues

15.1 The Power Method For Eigenvalues

This chapter discusses numerical methods for finding eigenvalues. However, to do this correctly, you must include numerical analysis considerations which are distinct from linear algebra. The purpose of this chapter is to give an introduction to some numerical methods without leaving the context of linear algebra. In addition, some examples are given which make use of computer algebra systems. For a more thorough discussion, you should see books on numerical methods in linear algebra like some listed in the references.

Let A be a complex $p \times p$ matrix and suppose that it has distinct eigenvalues

$$\{\lambda_1, \cdots, \lambda_m\}$$

and that $|\lambda_1| > |\lambda_k|$ for all k. Also let the Jordan form of A be

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}$$

with

$$J_k = \lambda_k I_k + N_k$$

where $N_k^{r_k} \neq 0$ but $N_k^{r_k+1} = 0$. Also let

$$P^{-1}AP = J, \ A = PJP^{-1}.$$

Now fix $\mathbf{x} \in \mathbb{F}^p$. Take $A\mathbf{x}$ and let s_1 be the entry of the vector $A\mathbf{x}$ which has largest absolute value. Thus $A\mathbf{x}/s_1$ is a vector $\mathbf{y}_1$ which has a component of 1 and every other entry of this vector has magnitude no larger than 1. If the scalars $\{s_1, \cdots, s_{n-1}\}$ and vectors $\{\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}\}$ have been obtained, let

$$\mathbf{y}_n \equiv \frac{A\mathbf{y}_{n-1}}{s_n}$$

where s_n is the entry of $A\mathbf{y}_{n-1}$ which has largest absolute value. Thus

$$\mathbf{y}_n = \frac{AA\mathbf{y}_{n-2}}{s_n s_{n-1}} \cdots = \frac{A^n \mathbf{x}}{s_n s_{n-1} \cdots s_1} \tag{15.1}$$

Consider one of the blocks in the Jordan form.

$$J_k^n = \lambda_1^n \sum_{i=0}^{r_k} \binom{n}{i} \frac{\lambda_k^{n-i}}{\lambda_1^n} N_k^i \equiv \lambda_1^n K(k,n)$$

Then from the above,

$$\frac{A^n}{s_n s_{n-1} \cdots s_1} = P \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \begin{pmatrix} K(1,n) & & \\ & \ddots & \\ & & K(m,n) \end{pmatrix} P^{-1}$$

Consider one of the terms in the sum for $K(k,n)$ for $k > 1$. Letting the norm of a matrix be the maximum of the absolute values of its entries,

$$\left\| \binom{n}{i} \frac{\lambda_k^{n-i}}{\lambda_1^n} N_k^i \right\| \leq n^{r_k} \left| \frac{\lambda_k}{\lambda_1} \right|^n p^{r_k} C$$

where C depends on the eigenvalues but is independent of n. Then this converges to 0 because the infinite sum of these converges due to the root test. Thus each of the matrices $K(k,n)$ converges to 0 for each $k > 1$ as $n \to \infty$.

Now what about $K(1,n)$? It equals

$$\binom{n}{r_1} \sum_{i=0}^{r_1} \left(\binom{n}{i} / \binom{n}{r_1} \right) \lambda_1^{-i} N_1^i$$

$$= \binom{n}{r_1} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n) \right)$$

where $\lim_{n\to\infty} m(n) = 0$. This follows from

$$\lim_{n\to\infty} \left(\binom{n}{i} / \binom{n}{r_1} \right) = 0, \ i < r_1$$

It follows that 15.1 is of the form

$$\mathbf{y}_n = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \begin{pmatrix} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n) \right) & 0 \\ 0 & E_n \end{pmatrix} P^{-1} \mathbf{x} = \frac{A\mathbf{y}_{n-1}}{s_n}$$

where the entries of E_n converge to 0 as $n \to \infty$. Now denote by $\left(P^{-1}\mathbf{x}\right)_{m_1}$ the first m_1 entries of $P^{-1}\mathbf{x}$ where it is assumed that λ_1 has multiplicity m_1. Assume that

$$\left(P^{-1}\mathbf{x}\right)_{m_1} \notin \ker N_1^{r_1}$$

This will be the case unless you have made an extremely unfortunate choice of $\mathbf{x}$. Then $\mathbf{y}_n$ is of the form

$$\mathbf{y}_n = \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \begin{pmatrix} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n) \right) \left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{z}_n \end{pmatrix} \tag{15.2}$$

where $\binom{n}{r_1}\mathbf{z}_n \to \mathbf{0}$. Also, from the construction, there is a single entry of $\mathbf{y}_n$ equal to 1 and all other entries of the above vector have absolute value no larger than 1. It follows that

$$\frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1}$$

must be bounded independent of n.

Then it follows from this observation, that for large n, the above vector $\mathbf{y}_n$ is approximately equal to

$$\begin{aligned}
&\frac{\lambda_1^n}{s_n s_{n-1}\cdots s_1}\binom{n}{r_1} P\begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1}\left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{0} \end{pmatrix} \\
&= \frac{1}{s_n s_{n-1}\cdots s_1} P\begin{pmatrix} \lambda_1^{n-r_1}\binom{n}{r_1} N_1^{r_1} & 0 \\ 0 & 0 \end{pmatrix} P^{-1}\mathbf{x} \qquad (15.3)
\end{aligned}$$

If $\left(P^{-1}\mathbf{x}\right)_{m_1} \notin \ker\left(N_1^{r_1}\right)$, then the above vector is also not equal to $\mathbf{0}$. What happens when it is multiplied on the left by $A - \lambda_1 I = P\left(J - \lambda_1 I\right) P^{-1}$? This results in

$$\frac{1}{s_n s_{n-1}\cdots s_1} P\begin{pmatrix} \lambda_1^{n-r_1} N_1\binom{n}{r_1} N_1^{r_1} & 0 \\ 0 & 0 \end{pmatrix} P^{-1}\mathbf{x} = \mathbf{0}$$

because $N_1^{r_1+1} = 0$. Therefore, the vector in 15.3 is an eigenvector and $\mathbf{y}_n$ is approximately equal to this eigenvector.

With this preparation, here is a theorem.

Theorem 15.1.1 *Let A be a complex $p \times p$ matrix such that the eigenvalues are*

$$\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$$

with $|\lambda_1| > |\lambda_j|$ for all $j \neq 1$. Then for $\mathbf{x}$ a given vector, let

$$\mathbf{y}_1 = \frac{A\mathbf{x}}{s_1}$$

where s_1 is an entry of $A\mathbf{x}$ which has the largest absolute value. If the scalars $\{s_1, \cdots, s_{n-1}\}$ and vectors $\{\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}\}$ have been obtained, let

$$\mathbf{y}_n \equiv \frac{A\mathbf{y}_{n-1}}{s_n}$$

where s_n is the entry of $A\mathbf{y}_{n-1}$ which has largest absolute value. Then it is probably the case that $\{s_n\}$ will converge to λ_1 and $\{\mathbf{y}_n\}$ will converge to an eigenvector associated with λ_1.

Proof: Consider the claim about s_{n+1}. It was shown above that

$$\mathbf{z} \equiv P\begin{pmatrix} \lambda_1^{-r_1} N_1^{r_1}\left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{0} \end{pmatrix}$$

is an eigenvector for λ_1. Let z_l be the entry of $\mathbf{z}$ which has largest absolute value. Then for large n, it will probably be the case that the entry of $\mathbf{y}_n$ which has largest absolute value will also be in the l^{th} slot. This follows from 15.2 because for large $n, \mathbf{z}_n$ will be very small, smaller than the largest entry of the top part of the vector in that expression. Then, since $m(n)$ is very small, the result follows if $\mathbf{z}$ has a well defined entry which has largest absolute value. Now from the above construction,

$$s_{n+1}\mathbf{y}_{n+1} \equiv A\mathbf{y}_n \approx \frac{\lambda_1^{n+1}}{s_n \cdots s_1}\binom{n}{r_1}\mathbf{z}$$

Applying a similar formula to s_n and the above observation, about the largest entry, it follows that for large n

$$s_{n+1} \approx \frac{\lambda_1^{n+1}}{s_n \cdots s_1}\binom{n}{r_1} z_l,\ s_n \approx \frac{\lambda_1^n}{s_{n-1}\cdots s_1}\binom{n-1}{r_1} z_l$$

Therefore, for large n,

$$\frac{s_{n+1}}{s_n} \approx \frac{\lambda_1}{s_n} \frac{n \cdots (n - r_1 + 1)}{(n-1) \cdots (n - r_1)} \approx \frac{\lambda_1}{s_n}$$

which shows that $s_{n+1} \approx \lambda_1$.

Now from the construction and the formula in 15.2, for large n

$$\begin{aligned}
\mathbf{y}_{n+1} &= \frac{\lambda_1^{n+1}}{s_{n+1} s_{n-1} \cdots s_1} \binom{n+1}{r_1} P \begin{pmatrix} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n)\right) \left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{z}_n \end{pmatrix} \\
&= \frac{\lambda_1}{s_{n+1}} \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n+1}{r_1} P \begin{pmatrix} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n)\right) \left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{z}_n \end{pmatrix} \\
&\approx \frac{\binom{n+1}{r_1}}{\binom{n}{r_1}} \frac{\lambda_1^n}{s_n s_{n-1} \cdots s_1} \binom{n}{r_1} P \begin{pmatrix} \left(\lambda_1^{-r_1} N_1^{r_1} + m(n)\right) \left(P^{-1}\mathbf{x}\right)_{m_1} \\ \mathbf{z}_n \end{pmatrix} \\
&= \frac{\binom{n+1}{r_1}}{\binom{n}{r_1}} \mathbf{y}_n \approx \mathbf{y}_n
\end{aligned}$$

Thus $\{\mathbf{y}_n\}$ is a Cauchy sequence and must converge to a vector $\mathbf{v}$. Now from the construction,

$$\lambda_1 \mathbf{v} = \lim_{n\to\infty} s_{n+1}\mathbf{y}_{n+1} = \lim_{n\to\infty} A\mathbf{y}_n = A\mathbf{v}. \ \blacksquare$$

In summary, here is the procedure.

Finding the largest eigenvalue with its eigenvector.

1. Start with a vector, $\mathbf{u}_1$ which you hope is not unlucky.
2. If $\mathbf{u}_k$ is known,
$$\mathbf{u}_{k+1} = \frac{A\mathbf{u}_k}{s_{k+1}}$$
where s_{k+1} is the entry of $A\mathbf{u}_k$ which has largest absolute value.
3. When the scaling factors s_k are not changing much, s_{k+1} will be close to the eigenvalue and $\mathbf{u}_{k+1}$ will be close to an eigenvector.
4. Check your answer to see if it worked well.

Example 15.1.2 *Find the largest eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$.

You can begin with $\mathbf{u}_1 = (1, \cdots, 1)^T$ and apply the above procedure. However, you can accelerate the process if you begin with $A^n\mathbf{u}_1$ and then divide by the largest entry to get the first approximate eigenvector. Thus

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}^{20} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2.5558 \times 10^{21} \\ -1.2779 \times 10^{21} \\ -3.6562 \times 10^{15} \end{pmatrix}$$

Divide by the largest entry to obtain a good aproximation.

$$\begin{pmatrix} 2.5558 \times 10^{21} \\ -1.2779 \times 10^{21} \\ -3.6562 \times 10^{15} \end{pmatrix} \frac{1}{2.5558 \times 10^{21}} = \begin{pmatrix} 1.0 \\ -0.5 \\ -1.4306 \times 10^{-6} \end{pmatrix}$$

Now begin with this one.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.5 \\ -1.4306 \times 10^{-6} \end{pmatrix} = \begin{pmatrix} 12.000 \\ -6.0000 \\ 4.2918 \times 10^{-6} \end{pmatrix}$$

Divide by 12 to get the next iterate.

$$\begin{pmatrix} 12.000 \\ -6.0000 \\ 4.2918 \times 10^{-6} \end{pmatrix} \frac{1}{12} = \begin{pmatrix} 1.0 \\ -0.5 \\ 3.5765 \times 10^{-7} \end{pmatrix}$$

Another iteration will reveal that the scaling factor is still 12. Thus this is an approximate eigenvalue. In fact, it is **the** largest eigenvalue and the corresponding eigenvector is

$$\begin{pmatrix} 1.0 \\ -0.5 \\ 0 \end{pmatrix}$$

The process has worked very well.

15.1.1 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. One can find complex eigenvalues using this method. The situation is this: You have a number, α which is close to λ, some eigenvalue of an $n \times n$ matrix A. You don't know λ but you know that α is closer to λ than to any other eigenvalue. Your problem is to find both λ and an eigenvector which goes with λ. Another way to look at this is to start with α and seek the eigenvalue λ, which is closest to α along with an eigenvector associated with λ. If α is an eigenvalue of A, then you have what you want. Therefore, I will always assume α is not an eigenvalue of A and so $(A - \alpha I)^{-1}$ exists. The method is based on the following lemma.

Lemma 15.1.3 *Let $\{\lambda_k\}_{k=1}^n$ be the eigenvalues of A. If $\mathbf{x}_k$ is an eigenvector of A for the eigenvalue λ_k, then $\mathbf{x}_k$ is an eigenvector for $(A - \alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_k - \alpha}$. Conversely, if*

$$(A - \alpha I)^{-1} \mathbf{y} = \frac{1}{\lambda - \alpha} \mathbf{y} \tag{15.4}$$

and $\mathbf{y} \neq \mathbf{0}$, then $A\mathbf{y} = \lambda \mathbf{y}$.

Proof: Let λ_k and $\mathbf{x}_k$ be as described in the statement of the lemma. Then

$$(A - \alpha I) \mathbf{x}_k = (\lambda_k - \alpha) \mathbf{x}_k$$

and so

$$\frac{1}{\lambda_k - \alpha} \mathbf{x}_k = (A - \alpha I)^{-1} \mathbf{x}_k.$$

Suppose 15.4. Then $\mathbf{y} = \frac{1}{\lambda - \alpha} [A\mathbf{y} - \alpha \mathbf{y}]$. Solving for $A\mathbf{y}$ leads to $A\mathbf{y} = \lambda \mathbf{y}$. ■

Now assume α is closer to λ than to any other eigenvalue. Then the magnitude of $\frac{1}{\lambda - \alpha}$ is greater than the magnitude of all the other eigenvalues of $(A - \alpha I)^{-1}$. Therefore, the power method applied to $(A - \alpha I)^{-1}$ will yield $\frac{1}{\lambda - \alpha}$. You end up with $s_{n+1} \approx \frac{1}{\lambda - \alpha}$ and solve for λ.

15.1.2 The Explicit Description Of The Method

Here is how you use this method to find the eigenvalue and eigenvector closest to α.

1. Find $(A - \alpha I)^{-1}$.
2. Pick $\mathbf{u}_1$. If you are not phenomenally unlucky, the iterations will converge.
3. If $\mathbf{u}_k$ has been obtained,
$$\mathbf{u}_{k+1} = \frac{(A - \alpha I)^{-1} \mathbf{u}_k}{s_{k+1}}$$
where s_{k+1} is the entry of $(A - \alpha I)^{-1} \mathbf{u}_k$ which has largest absolute value.
4. When the scaling factors, s_k are not changing much and the $\mathbf{u}_k$ are not changing much, find the approximation to the eigenvalue by solving
$$s_{k+1} = \frac{1}{\lambda - \alpha}$$
for λ. The eigenvector is approximated by $\mathbf{u}_{k+1}$.

5. Check your work by multiplying by the original matrix to see how well what you have found works.

Thus this amounts to the power method for the matrix $(A - \alpha I)^{-1}$.

Example 15.1.4 *Find the eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$ *which is closest to* -7. *Also find an eigenvector which goes with this eigenvalue.*

In this case the eigenvalues are $-6, 0$, and 12 so the correct answer is -6 for the eigenvalue. Then from the above procedure, I will start with an initial vector,

$$\mathbf{u}_1 \equiv \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Then I must solve the following equation.

$$\left(\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} + 7 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Simplifying the matrix on the left, I must solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and then divide by the entry which has largest absolute value to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix}$$

and divide by the largest entry, 1.0515 to get

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ .0266 \\ -.97061 \end{pmatrix}$$

Solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .0266 \\ -.97061 \end{pmatrix}$$

and divide by the largest entry, 1. 01 to get

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ 3.8454 \times 10^{-3} \\ -.99604 \end{pmatrix}.$$

These scaling factors are pretty close after these few iterations. Therefore, the predicted eigenvalue is obtained by solving the following for λ.

$$\frac{1}{\lambda + 7} = 1.01$$

which gives $\lambda = -6.01$. You see this is pretty close. In this case the eigenvalue closest to -7 was -6.

How would you know what to start with for an initial guess? You might apply Gerschgorin's theorem.

Example 15.1.5 *Consider the symmetric matrix* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}$. *Find the middle eigenvalue and an eigenvector which goes with it.*

Since A is symmetric, it follows it has three real eigenvalues which are solutions to

$$\begin{aligned} p(\lambda) &= \det\left(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right) \\ &= \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0 \end{aligned}$$

If you use your graphing calculator to graph this polynomial, you find there is an eigenvalue somewhere between $-.9$ and $-.8$ and that this is the middle eigenvalue. Of course you could zoom in and find it very accurately without much trouble but what about the eigenvector which goes with it? If you try to solve

$$\left((-.8)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

there will be only the zero solution because the matrix on the left will be invertible and the same will be true if you replace $-.8$ with a better approximation like $-.86$ or $-.855$. This is because all these are only approximations to the eigenvalue and so the matrix in the above is nonsingular for all of these. Therefore, you will only get the zero solution and

$$\boxed{\underline{\textbf{Eigenvectors}}\ \underline{\textbf{are}}\ \underline{\textbf{never}}\ \underline{\textbf{equal}}\ \underline{\textbf{to}}\ \underline{\textbf{zero}}\textbf{!}}$$

However, there exists such an eigenvector and you can find it using the shifted inverse power method. Pick $\alpha = -.855$. Then you solve

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} + .855\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

or in other words,

$$\begin{pmatrix} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and after finding the solution, divide by the largest entry -67.944, to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ -.58921 \\ -.23044 \end{pmatrix}$$

After a couple more iterations, you obtain

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ -.58777 \\ -.22714 \end{pmatrix} \tag{15.5}$$

Then doing it again, the scaling factor is -513.42 and the next iterate is

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ -.58778 \\ -.22714 \end{pmatrix}$$

Clearly the $\mathbf{u}_k$ are not changing much. This suggests an approximate eigenvector for this eigenvalue which is close to $-.855$ is the above $\mathbf{u}_3$ and an eigenvalue is obtained by solving

$$\frac{1}{\lambda + .855} = -514.01,$$

which yields $\lambda = -.8569$ Lets check this.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.58777 \\ -.22714 \end{pmatrix} = \begin{pmatrix} -.85696 \\ .50367 \\ .19464 \end{pmatrix}.$$

$$-.8569 \begin{pmatrix} 1.0 \\ -.58777 \\ -.22714 \end{pmatrix} = \begin{pmatrix} -.8569 \\ .5037 \\ .1946 \end{pmatrix}$$

Thus the vector of 15.5 is very close to the desired eigenvector, just as $-.8569$ is very close to the desired eigenvalue. For practical purposes, I have found both the eigenvector and the eigenvalue.

Example 15.1.6 *Find the eigenvalues and eigenvectors of the matrix* $A = \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

This is only a 3×3 matrix and so it is not hard to estimate the eigenvalues. Just get the characteristic equation, graph it using a calculator and zoom in to find the eigenvalues. If you do this, you find there is an eigenvalue near -1.2, one near $-.4$, and one near 5.5. (The characteristic equation is $2 + 8\lambda + 4\lambda^2 - \lambda^3 = 0$.) Of course I have no idea what the eigenvectors are.

Lets first try to find the eigenvector and a better approximation for the eigenvalue near -1.2. In this case, let $\alpha = -1.2$. Then

$$(A - \alpha I)^{-1} = \begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}.$$

As before, it helps to get things started if you raise to a power and then go from the approximate eigenvector obtained.

$$\begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}^7 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2.2956 \times 10^{11} \\ 1.1291 \times 10^{11} \\ 2.0865 \times 10^{11} \end{pmatrix}$$

Then the next iterate will be

$$\begin{pmatrix} -2.2956 \times 10^{11} \\ 1.1291 \times 10^{11} \\ 2.0865 \times 10^{11} \end{pmatrix} \frac{1}{-2.2956 \times 10^{11}} = \begin{pmatrix} 1.0 \\ -0.49185 \\ -0.90891 \end{pmatrix}$$

Next iterate:

$$\begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49185 \\ -0.90891 \end{pmatrix} = \begin{pmatrix} -54.115 \\ 26.615 \\ 49.184 \end{pmatrix}$$

Divide by largest entry

$$\begin{pmatrix} -54.115 \\ 26.615 \\ 49.184 \end{pmatrix} \frac{1}{-54.115} = \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix}$$

You can see the vector didn't change much and so the next scaling factor will not be much different than this one. Hence you need to solve for λ

$$\frac{1}{\lambda + 1.2} = -54.115$$

Then $\lambda = -1.2185$ is an approximate eigenvalue and

$$\begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix}$$

is an approximate eigenvector. How well does it work?

$$\begin{aligned} \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix} &= \begin{pmatrix} -1.2185 \\ 0.5993 \\ 1.1075 \end{pmatrix} \\ (-1.2185) \begin{pmatrix} 1.0 \\ -0.49182 \\ -0.90888 \end{pmatrix} &= \begin{pmatrix} -1.2185 \\ 0.59928 \\ 1.1075 \end{pmatrix} \end{aligned}$$

You can see that for practical purposes, this has found the eigenvalue closest to -1.2185 and the corresponding eigenvector.

The other eigenvectors and eigenvalues can be found similarly. In the case of $-.4$, you could let $\alpha = -.4$ and then

$$(A - \alpha I)^{-1} = \begin{pmatrix} 8.0645161 \times 10^{-2} & -9.2741935 & 6.4516129 \\ -.40322581 & 11.370968 & -7.2580645 \\ .40322581 & 3.6290323 & -2.7419355 \end{pmatrix}.$$

Following the procedure of the power method, you find that after about 5 iterations, the scaling factor is 9.7573139, they are not changing much, and

$$\mathbf{u}_5 = \begin{pmatrix} -.7812248 \\ 1.0 \\ .26493688 \end{pmatrix}.$$

Thus the approximate eigenvalue is

$$\frac{1}{\lambda + .4} = 9.7573139$$

which shows $\lambda = -.\,297\,512\,78$ is an approximation to the eigenvalue near .4. How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} -.\,781\,224\,8 \\ 1.\,0 \\ .\,264\,936\,88 \end{pmatrix} = \begin{pmatrix} .\,232\,361\,04 \\ -.\,297\,512\,72 \\ -.0\,787\,375\,2 \end{pmatrix}.$$

$$-.\,297\,512\,78 \begin{pmatrix} -.\,781\,224\,8 \\ 1.\,0 \\ .\,264\,936\,88 \end{pmatrix} = \begin{pmatrix} .\,232\,424\,36 \\ -.\,297\,512\,78 \\ -7.\,882\,210\,8 \times 10^{-2} \end{pmatrix}.$$

It works pretty well. For practical purposes, the eigenvalue and eigenvector have now been found. If you want better accuracy, you could just continue iterating.

Next I will find the eigenvalue and eigenvector for the eigenvalue near 5.5. In this case,

$$(A - \alpha I)^{-1} = \begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix}.$$

As before, I have no idea what the eigenvector is but I am tired of always using $(1,1,1)^T$ and I don't want to give the impression that you always need to start with this vector. Therefore, I shall let $\mathbf{u}_1 = (1,2,3)^T$. Also, I will begin by raising the matrix to a power.

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix}^9 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3.\,009 \times 10^{16} \\ 1.\,968\,2 \times 10^{16} \\ 2.\,870\,6 \times 10^{16} \end{pmatrix}.$$

Divide by largest entry to get the next iterate.

$$\begin{pmatrix} 3.\,009 \times 10^{16} \\ 1.\,968\,2 \times 10^{16} \\ 2.\,870\,6 \times 10^{16} \end{pmatrix} \frac{1}{3.\,009 \times 10^{16}} = \begin{pmatrix} 1.0 \\ 0.654\,1 \\ 0.954 \end{pmatrix}$$

Now

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.654\,1 \\ 0.954 \end{pmatrix} = \begin{pmatrix} 62.\,322 \\ 40.\,765 \\ 59.\,454 \end{pmatrix}$$

Then the next iterate is

$$\begin{pmatrix} 62.\,322 \\ 40.\,765 \\ 59.\,454 \end{pmatrix} \frac{1}{62.322} = \begin{pmatrix} 1.0 \\ 0.654\,1 \\ 0.953\,98 \end{pmatrix}$$

This is very close to the eigenvector given above and so the next scaling factor will also be close to 62.322. Thus the approximate eigenvalue is obtained by solving

$$\frac{1}{\lambda - 5.5} = 62.322$$

An approximate eigenvalue is $\lambda = 5.\,516$ and an approximate eigenvector is the above vector. How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.654\,1 \\ 0.953\,98 \end{pmatrix} = \begin{pmatrix} 5.\,516 \\ 3.\,608\,1 \\ 5.\,262\,2 \end{pmatrix}$$

$$5.\,516 \begin{pmatrix} 1.0 \\ 0.654\,1 \\ 0.953\,98 \end{pmatrix} = \begin{pmatrix} 5.\,516 \\ 3.\,608 \\ 5.\,262\,2 \end{pmatrix}$$

It appears this is very close.

15.1.3 Complex Eigenvalues

What about complex eigenvalues? If your matrix is real, you won't see these by graphing the characteristic equation on your calculator. Will the shifted inverse power method find these eigenvalues and their associated eigenvectors? The answer is yes. However, for a real matrix, you must pick α to be complex. This is because the eigenvalues occur in conjugate pairs so if you don't pick it complex, it will be the same distance between any conjugate pair of complex numbers and so nothing in the above argument for convergence implies you will get convergence to a complex number. Also, the process of iteration will yield only real vectors and scalars.

Example 15.1.7 *Find the complex eigenvalues and corresponding eigenvectors for the matrix*

$$\begin{pmatrix} 5 & -8 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Here the characteristic equation is $\lambda^3 - 5\lambda^2 + 8\lambda - 6 = 0$. One solution is $\lambda = 3$. The other two are $1+i$ and $1-i$. I will apply the process to $\alpha = i$ to find the eigenvalue closest to i.

$$(A - \alpha I)^{-1} = \begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix}$$

Then let $\mathbf{u}_1 = (1,1,1)^T$ for lack of any insight into anything better.

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} .38+.66i \\ .66+.62i \\ .62+.34i \end{pmatrix}$$

$s_2 = .66 + .62i$.

$$\mathbf{u}_2 = \begin{pmatrix} .80487805+.24390244i \\ 1.0 \\ .75609756-.19512195i \end{pmatrix}$$

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \cdot$$
$$\begin{pmatrix} .80487805+.24390244i \\ 1.0 \\ .75609756-.19512195i \end{pmatrix}$$
$$= \begin{pmatrix} .64634146+.81707317i \\ .81707317+.35365854i \\ .54878049-6.0975609\times 10^{-2}i \end{pmatrix}$$

$s_3 = .64634146 + .81707317i$. After more iterations, of this sort, you find $s_9 = 1.0027485 + 2.1376217 \times 10^{-4} i$ and

$$\mathbf{u}_9 = \begin{pmatrix} 1.0 \\ .50151417-.49980733i \\ 1.5620881\times 10^{-3}-.49977855i \end{pmatrix}.$$

Then

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \cdot$$
$$\begin{pmatrix} 1.0 \\ .501\,514\,17-.499\,807\,33i \\ 1.562\,088\,1\times 10^{-3}-.499\,778\,55i \end{pmatrix}$$
$$= \begin{pmatrix} 1.000\,407\,8+1.269\,979\times 10^{-3}i \\ .501\,077\,31-.498\,893\,66i \\ 8.848\,928\times 10^{-4}-.499\,515\,22i \end{pmatrix}$$

$s_{10} = 1.000\,407\,8+1.269\,979\times 10^{-3}i.$

$$\mathbf{u}_{10} = \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$

The scaling factors are not changing much at this point. Thus you would solve the following for λ.

$$1.000\,407\,8+1.269\,979\times 10^{-3}i = \frac{1}{\lambda - i}$$

The approximate eigenvalue is then $\lambda = .999\,590\,76+.998\,731\,06i$. This is pretty close to $1+i$. How well does the eigenvector work?

$$\begin{pmatrix} 5 & -8 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$
$$= \begin{pmatrix} .999\,590\,61+.998\,731\,12i \\ 1.0 \\ .500\,239\,18-.499\,325\,33i \end{pmatrix}$$

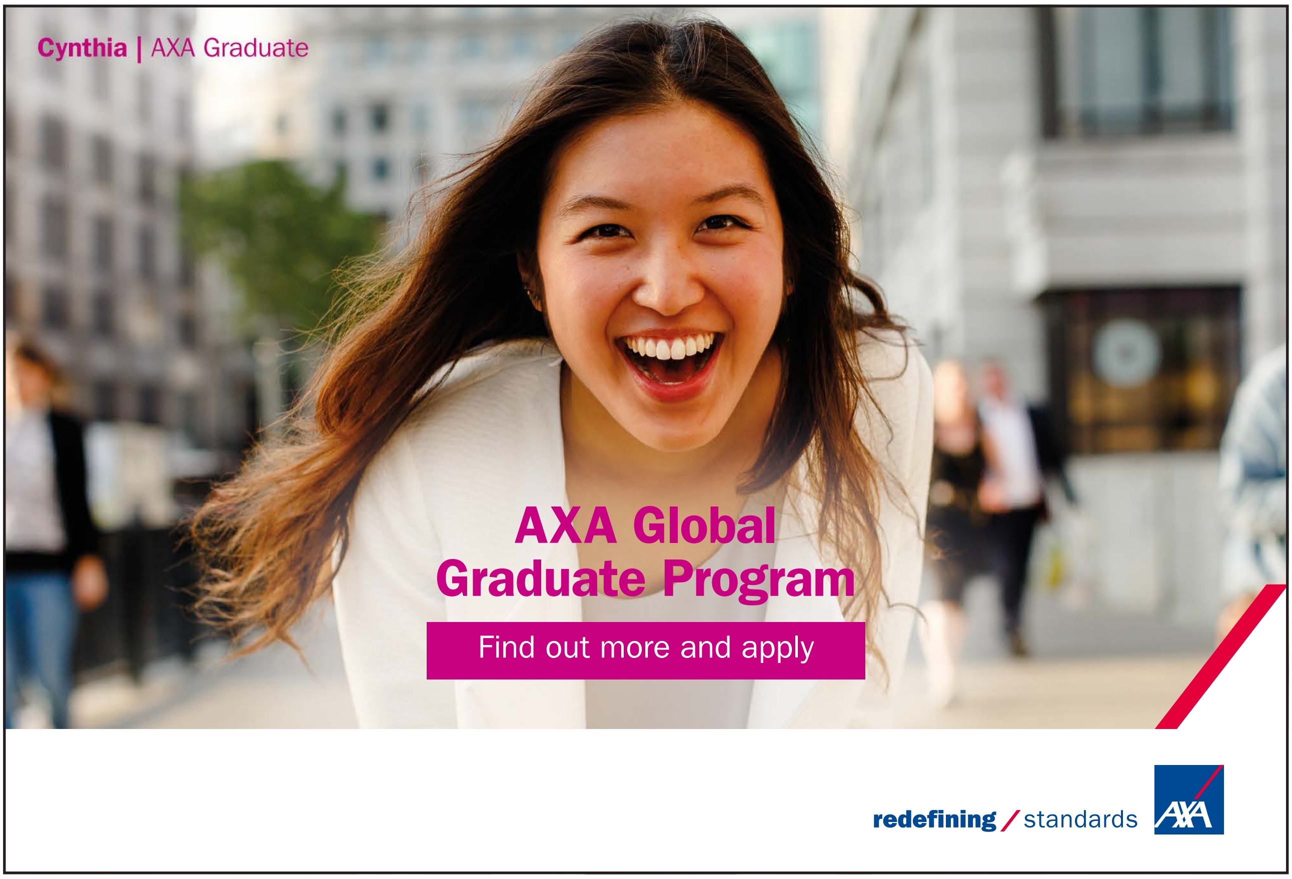
$$(.999\,590\,76+.998\,731\,06i) \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$
$$= \begin{pmatrix} .999\,590\,76+.998\,731\,06i \\ .998\,726\,18+4.834\,203\,9\times 10^{-4}i \\ .498\,928\,9-.498\,857\,22i \end{pmatrix}$$

It took more iterations than before because α was not very close to $1+i$.

This illustrates an interesting topic which leads to many related topics. If you have a polynomial, $x^4+ax^3+bx^2+cx+d$, you can consider it as the characteristic polynomial of a certain matrix, called a **companion matrix**. In this case,

$$\begin{pmatrix} -a & -b & -c & -d \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The above example was just a companion matrix for $\lambda^3-5\lambda^2+8\lambda-6$. You can see the pattern which will enable you to obtain a companion matrix for any polynomial of the form $\lambda^n+a_1\lambda^{n-1}+\cdots+a_{n-1}\lambda+a_n$. This illustrates that one way to find the complex zeros of a polynomial is to use the shifted inverse power method on a companion matrix for the polynomial. Doubtless there are better ways but this does illustrate how impressive this procedure is. Do you have a better way?



Note that the shifted inverse power method is a way you can begin with something close but not equal to an eigenvalue and end up with something close to an eigenvector.

15.1.4 Rayleigh Quotients And Estimates for Eigenvalues

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. Recall a matrix A is Hermitian if $A = A^*$ where A^* means to take the transpose of the conjugate of A. In the case of a real matrix, Hermitian reduces to symmetric. Recall also that for $\mathbf{x} \in \mathbb{F}^n$,

$$|\mathbf{x}|^2 = \mathbf{x}^*\mathbf{x} = \sum_{j=1}^{n} |x_j|^2 .$$

Recall the following corollary found on Page 239 which is stated here for convenience.

Corollary 15.1.8 *If A is Hermitian, then all the eigenvalues of A are real and there exists an orthonormal basis of eigenvectors.*

Thus for $\{\mathbf{x}_k\}_{k=1}^n$ this orthonormal basis,

$$\mathbf{x}_i^*\mathbf{x}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

For $\mathbf{x} \in \mathbb{F}^n$, $\mathbf{x} \neq \mathbf{0}$, the Rayleigh quotient is defined by

$$\frac{\mathbf{x}^* A\mathbf{x}}{|\mathbf{x}|^2}.$$

Now let the eigenvalues of A be $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $A\mathbf{x}_k = \lambda_k \mathbf{x}_k$ where $\{\mathbf{x}_k\}_{k=1}^n$ is the above orthonormal basis of eigenvectors mentioned in the corollary. Then if $\mathbf{x}$ is an arbitrary vector, there exist constants, a_i such that

$$\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{x}_i.$$

Also,

$$\begin{aligned} |\mathbf{x}|^2 &= \sum_{i=1}^{n} \overline{a}_i \mathbf{x}_i^* \sum_{j=1}^{n} a_j \mathbf{x}_j \\ &= \sum_{ij} \overline{a}_i a_j \mathbf{x}_i^* \mathbf{x}_j = \sum_{ij} \overline{a}_i a_j \delta_{ij} = \sum_{i=1}^{n} |a_i|^2 . \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\mathbf{x}^* A\mathbf{x}}{|\mathbf{x}|^2} &= \frac{\left(\sum_{i=1}^n \overline{a}_i \mathbf{x}_i^*\right)\left(\sum_{j=1}^n a_j \lambda_j \mathbf{x}_j\right)}{\sum_{i=1}^n |a_i|^2} \\ &= \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \mathbf{x}_i^* \mathbf{x}_j}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \delta_{ij}}{\sum_{i=1}^n |a_i|^2} \\ &= \frac{\sum_{i=1}^n |a_i|^2 \lambda_i}{\sum_{i=1}^n |a_i|^2} \in [\lambda_1, \lambda_n] . \end{aligned}$$

In other words, the Rayleigh quotient is always between the largest and the smallest eigenvalues of A. When $\mathbf{x} = \mathbf{x}_n$, the Rayleigh quotient equals the largest eigenvalue and when $\mathbf{x} = \mathbf{x}_1$ the Rayleigh quotient equals the smallest eigenvalue. Suppose you calculate a Rayleigh quotient. How close is it to some eigenvalue?

Theorem 15.1.9 *Let* $\mathbf{x} \neq \mathbf{0}$ *and form the Rayleigh quotient,*

$$\frac{\mathbf{x}^* A\mathbf{x}}{|\mathbf{x}|^2} \equiv q.$$

Then there exists an eigenvalue of A*, denoted here by* λ_q *such that*

$$|\lambda_q - q| \leq \frac{|A\mathbf{x} - q\mathbf{x}|}{|\mathbf{x}|}. \tag{15.6}$$

Proof: Let $\mathbf{x} = \sum_{k=1}^{n} a_k \mathbf{x}_k$ where $\{\mathbf{x}_k\}_{k=1}^{n}$ is the orthonormal basis of eigenvectors.

$$\begin{aligned}
|A\mathbf{x} - q\mathbf{x}|^2 &= (A\mathbf{x} - q\mathbf{x})^* (A\mathbf{x} - q\mathbf{x}) \\
&= \left(\sum_{k=1}^{n} a_k \lambda_k \mathbf{x}_k - q a_k \mathbf{x}_k\right)^* \left(\sum_{k=1}^{n} a_k \lambda_k \mathbf{x}_k - q a_k \mathbf{x}_k\right) \\
&= \left(\sum_{j=1}^{n} (\lambda_j - q)\, \overline{a}_j \mathbf{x}_j^*\right) \left(\sum_{k=1}^{n} (\lambda_k - q)\, a_k \mathbf{x}_k\right) \\
&= \sum_{j,k} (\lambda_j - q)\, \overline{a}_j\, (\lambda_k - q)\, a_k \mathbf{x}_j^* \mathbf{x}_k \\
&= \sum_{k=1}^{n} |a_k|^2 (\lambda_k - q)^2
\end{aligned}$$

Now pick the eigenvalue λ_q which is closest to q. Then

$$|A\mathbf{x} - q\mathbf{x}|^2 = \sum_{k=1}^{n} |a_k|^2 (\lambda_k - q)^2 \geq (\lambda_q - q)^2 \sum_{k=1}^{n} |a_k|^2 = (\lambda_q - q)^2 |\mathbf{x}|^2$$

which implies 15.6. ■

Example 15.1.10 *Consider the symmetric matrix* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix}$. *Let* $\mathbf{x} = (1,1,1)^T$. *How close is the Rayleigh quotient to some eigenvalue of* A*? Find the eigenvector and eigenvalue to several decimal places.*

Everything is real and so there is no need to worry about taking conjugates. Therefore, the Rayleigh quotient is

$$\frac{\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}{3} = \frac{19}{3}$$

According to the above theorem, there is some eigenvalue of this matrix λ_q such that

$$\begin{aligned}\left|\lambda_q - \frac{19}{3}\right| &\leq \frac{\left|\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{19}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right|}{\sqrt{3}} \\ &= \frac{1}{\sqrt{3}}\begin{pmatrix} -\frac{1}{3} \\ -\frac{4}{3} \\ \frac{5}{3} \end{pmatrix} \\ &= \frac{\sqrt{\frac{1}{9}+\left(\frac{4}{3}\right)^2+\left(\frac{5}{3}\right)^2}}{\sqrt{3}} = 1.\,247\,2\end{aligned}$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the shifted inverse power method is all about.

Solve

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} - \frac{19}{3}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

In other words solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and divide by the entry which is largest, $3.\,870\,7$, to get

$$\mathbf{u}_2 = \begin{pmatrix} .\,699\,25 \\ .\,493\,89 \\ 1.0 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .\,699\,25 \\ .\,493\,89 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, 2.9979 to get

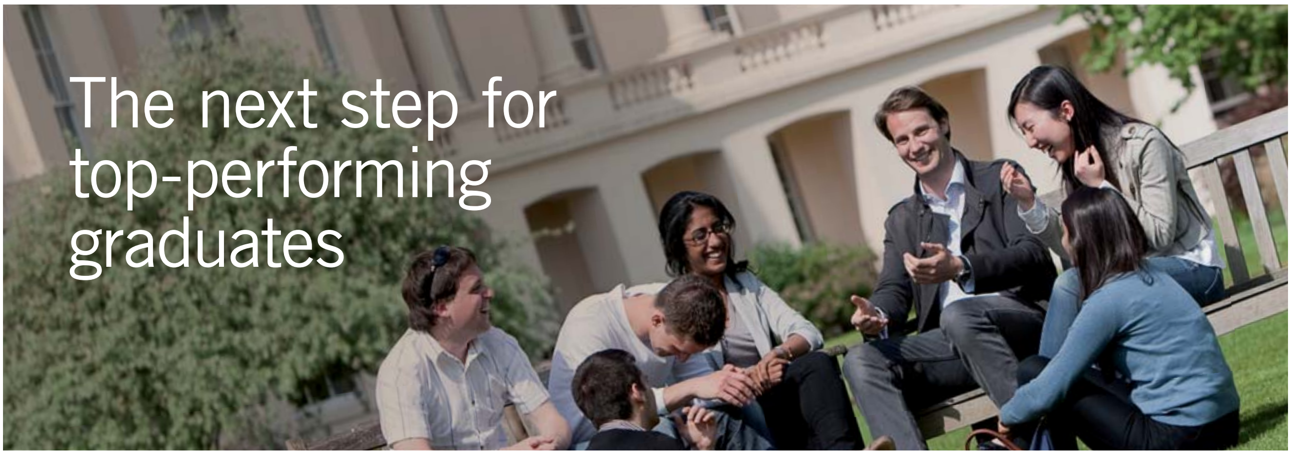
$$\mathbf{u}_3 = \begin{pmatrix} .71473 \\ .52263 \\ 1.0 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .71473 \\ .52263 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, 3.0454, to get

$$\mathbf{u}_4 = \begin{pmatrix} .7137 \\ .52056 \\ 1.0 \end{pmatrix}$$



Solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .713\,7 \\ .520\,56 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, $3.\,042\,1$ to get

$$\mathbf{u}_5 = \begin{pmatrix} .\,713\,78 \\ .\,520\,73 \\ 1.0 \end{pmatrix}$$

You can see these scaling factors are not changing much. The predicted eigenvalue is then about

$$\frac{1}{3.\,042\,1} + \frac{19}{3} = 6.\,662\,1.$$

How close is this?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} .\,713\,78 \\ .\,520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.\,755\,2 \\ 3.\,469 \\ 6.\,662\,1 \end{pmatrix}$$

while

$$6.\,662\,1 \begin{pmatrix} .\,713\,78 \\ .\,520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.\,755\,3 \\ 3.\,469\,2 \\ 6.\,662\,1 \end{pmatrix}.$$

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

15.2 The QR Algorithm

15.2.1 Basic Properties And Definition

Recall the theorem about the QR factorization in Theorem 5.7.5. It says that given an $n \times n$ real matrix A, there exists a real orthogonal matrix Q and an upper triangular matrix R such that $A = QR$ and that this factorization can be accomplished by a systematic procedure. One such procedure was given in proving this theorem.

There is also a way to generalize the QR factorization to the case where A is just a complex $n \times n$ matrix and Q is unitary while R is upper triangular with nonnegative entries on the main diagonal. Letting $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$ be the matrix with the $\mathbf{a}_j$ the columns, each a vector in $\mathbb{C}^n$, let Q_1 be a unitary matrix which maps $\mathbf{a}_1$ to $|\mathbf{a}_1|\,\mathbf{e}_1$ in the case that $\mathbf{a}_1 \neq \mathbf{0}$. If $\mathbf{a}_1 = \mathbf{0}$, let $Q_1 = I$. Why does such a unitary matrix exist? Let

$$\{\mathbf{a}_1 / |\mathbf{a}_1|, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$$

be an orthonormal basis and let $Q_1\left(\frac{\mathbf{a}_1}{|\mathbf{a}_1|}\right) = \mathbf{e}_1, Q_1(\mathbf{u}_2) = \mathbf{e}_2$ etc. Extend Q_1 linearly. Then Q_1 preserves lengths so it is unitary by Lemma 13.6.1. Now

$$\begin{aligned} Q_1 A &= \begin{pmatrix} Q_1\mathbf{a}_1 & Q_1\mathbf{a}_2 & \cdots & Q_1\mathbf{a}_n \end{pmatrix} \\ &= \begin{pmatrix} |\mathbf{a}_1|\,\mathbf{e}_1 & Q_1\mathbf{a}_2 & \cdots & Q_1\mathbf{a}_n \end{pmatrix} \end{aligned}$$

which is a matrix of the form

$$\begin{pmatrix} |\mathbf{a}_1| & \mathbf{b} \\ \mathbf{0} & A_1 \end{pmatrix}$$

Now do the same thing for A_1 obtaining an $n-1 \times n-1$ unitary matrix Q_2' which when multiplied on the left of A_1 yields something of the form

$$\begin{pmatrix} a & \mathbf{b}_1 \\ \mathbf{0} & A_2 \end{pmatrix}$$

Then multiplying A on the left by the product

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2' \end{pmatrix} Q_1 \equiv Q_2 Q_1$$

yields a matrix which is upper triangular with respect to the first two columns. Continuing this way

$$Q_n Q_{n-1} \cdots Q_1 A = R$$

where R is upper triangular having all positive entries on the main diagonal. Then the desired unitary matrix is

$$Q = (Q_n Q_{n-1} \cdots Q_1)^*$$

►►

The QR algorithm is described in the following definition.

Definition 15.2.1 *The QR algorithm is the following. In the description of this algorithm, Q is unitary and R is upper triangular having nonnegative entries on the main diagonal. Starting with A an $n \times n$ matrix, form*

$$A_0 \equiv A = Q_1 R_1 \tag{15.7}$$

Then

$$A_1 \equiv R_1 Q_1. \tag{15.8}$$

In general given

$$A_k = R_k Q_k, \tag{15.9}$$

obtain A_{k+1} by

$$A_k = Q_{k+1} R_{k+1},\ A_{k+1} = R_{k+1} Q_{k+1} \tag{15.10}$$

This algorithm was proposed by Francis in 1961. The sequence $\{A_k\}$ is the desired sequence of iterates. Now with the above definition of the algorithm, here are its properties. The next lemma shows each of the A_k is unitarily similar to A and the amazing thing about this algorithm is that often it becomes increasingly easy to find the eigenvalues of the A_k.

Lemma 15.2.2 *Let A be an $n \times n$ matrix and let the Q_k and R_k be as described in the algorithm. Then each A_k is unitarily similar to A and denoting by $Q^{(k)}$ the product $Q_1 Q_2 \cdots Q_k$ and $R^{(k)}$ the product $R_k R_{k-1} \cdots R_1$, it follows that*

$$A^k = Q^{(k)} R^{(k)}$$

(The matrix on the left is A raised to the k^{th} power.)

$$A = Q^{(k)} A_k Q^{(k)*}, \; A_k = Q^{(k)*} A Q^{(k)}.$$

Proof: From the algorithm, $R_{k+1} = A_{k+1} Q_{k+1}^*$ and so

$$A_k = Q_{k+1} R_{k+1} = Q_{k+1} A_{k+1} Q_{k+1}^*$$

Now iterating this, it follows

$$A_{k-1} = Q_k A_k Q_k^* = Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^*$$

$$A_{k-2} = Q_{k-1}A_{k-1}Q_{k-1}^* = Q_{k-1}Q_kQ_{k+1}A_{k+1}Q_{k+1}^*Q_k^*Q_{k-1}^*$$

etc. Thus, after $k-2$ more iterations,

$$A = Q^{(k+1)}A_{k+1}Q^{(k+1)*}$$

The product of unitary matrices is unitary and so this proves the first claim of the lemma.

Now consider the part about A^k. From the algorithm, this is clearly true for $k = 1$. ($A^1 = QR$) Suppose then that

$$A^k = Q_1Q_2\cdots Q_kR_kR_{k-1}\cdots R_1$$

What was just shown indicated

$$A = Q_1Q_2\cdots Q_{k+1}A_{k+1}Q_{k+1}^*Q_k^*\cdots Q_1^*$$

and now from the algorithm, $A_{k+1} = R_{k+1}Q_{k+1}$ and so

$$A = Q_1Q_2\cdots Q_{k+1}R_{k+1}Q_{k+1}Q_{k+1}^*Q_k^*\cdots Q_1^*$$

Then

$$A^{k+1} = AA^k =$$

$$\overbrace{Q_1Q_2\cdots Q_{k+1}R_{k+1}Q_{k+1}Q_{k+1}^*Q_k^*\cdots Q_1^*}^{A}Q_1\cdots Q_kR_kR_{k-1}\cdots R_1$$
$$= Q_1Q_2\cdots Q_{k+1}R_{k+1}R_kR_{k-1}\cdots R_1 \equiv Q^{(k+1)}R^{(k+1)} \blacksquare$$

Here is another very interesting lemma.

Lemma 15.2.3 *Suppose $Q^{(k)}, Q$ are unitary and R_k is upper triangular such that the diagonal entries on R_k are all positive and*

$$Q = \lim_{k\to\infty} Q^{(k)}R_k$$

Then

$$\lim_{k\to\infty} Q^{(k)} = Q, \ \lim_{k\to\infty} R_k = I.$$

Also the QR factorization of A is unique whenever A^{-1} exists.

Proof: Let

$$Q = (\mathbf{q}_1, \cdots, \mathbf{q}_n), \ Q^{(k)} = \left(\mathbf{q}_1^k, \cdots, \mathbf{q}_n^k\right)$$

where the $\mathbf{q}$ are the columns. Also denote by r_{ij}^k the ij^{th} entry of R_k. Thus

$$Q^{(k)}R_k = \left(\mathbf{q}_1^k, \cdots, \mathbf{q}_n^k\right)\begin{pmatrix} r_{11}^k & & * \\ & \ddots & \\ 0 & & r_{nn}^k \end{pmatrix}$$

It follows

$$r_{11}^k\mathbf{q}_1^k \to \mathbf{q}_1$$

and so

$$r_{11}^k = \left|r_{11}^k\mathbf{q}_1^k\right| \to 1$$

Therefore,

$$\mathbf{q}_1^k \to \mathbf{q}_1.$$

Next consider the second column.

$$r_{12}^k \mathbf{q}_1^k + r_{22}^k \mathbf{q}_2^k \to \mathbf{q}_2$$

Taking the inner product of both sides with $\mathbf{q}_1^k$ it follows

$$\lim_{k\to\infty} r_{12}^k = \lim_{k\to\infty} \left(\mathbf{q}_2 \cdot \mathbf{q}_1^k\right) = (\mathbf{q}_2 \cdot \mathbf{q}_1) = 0.$$

Therefore,

$$\lim_{k\to\infty} r_{22}^k \mathbf{q}_2^k = \mathbf{q}_2$$

and since $r_{22}^k > 0$, it follows as in the first part that $r_{22}^k \to 1$. Hence

$$\lim_{k\to\infty} \mathbf{q}_2^k = \mathbf{q}_2.$$

Continuing this way, it follows

$$\lim_{k\to\infty} r_{ij}^k = 0$$

for all $i \neq j$ and

$$\lim_{k\to\infty} r_{jj}^k = 1, \ \lim_{k\to\infty} \mathbf{q}_j^k = \mathbf{q}_j.$$

Thus $R_k \to I$ and $Q^{(k)} \to Q$. This proves the first part of the lemma.

The second part follows immediately. If $QR = Q'R' = A$ where A^{-1} exists, then

$$Q^*Q' = R\left(R'\right)^{-1}$$

and I need to show both sides of the above are equal to I. The left side of the above is unitary and the right side is upper triangular having positive entries on the diagonal. This is because the inverse of such an upper triangular matrix having positive entries on the main diagonal is still upper triangular having positive entries on the main diagonal and the product of two such upper triangular matrices gives another of the same form having positive entries on the main diagonal. Suppose then that $Q = R$ where Q is unitary and R is upper triangular having positive entries on the main diagonal. Let $Q_k = Q$ and $R_k = R$. It follows

$$IR_k \to R = Q$$

and so from the first part, $R_k \to I$ but $R_k = R$ and so $R = I$. Thus applying this to $Q^*Q' = R\left(R'\right)^{-1}$ yields both sides equal I. ■

A case of all this is of great interest. Suppose A has a largest eigenvalue λ which is real. Then A^n is of the form $\left(A^{n-1}\mathbf{a}_1, \cdots, A^{n-1}\mathbf{a}_n\right)$ and so likely each of these columns will be pointing roughly in the direction of an eigenvector of A which corresponds to this eigenvalue. Then when you do the QR factorization of this, it follows from the fact that R is upper triangular, that the first column of Q will be a multiple of $A^{n-1}\mathbf{a}_1$ and so will end up being roughly parallel to the eigenvector desired. Also this will require the entries below the top in the first column of $A_n = Q^TAQ$ will all be small because they will be of the form $\mathbf{q}_i^TA\mathbf{q}_1 \approx \lambda\mathbf{q}_i^T\mathbf{q}_1 = 0$. Therefore, A_n will be of the form

$$\begin{pmatrix} \lambda' & \mathbf{a} \\ \mathbf{e} & B \end{pmatrix}$$

where $\mathbf{e}$ is small. It follows that λ' will be close to λ and $\mathbf{q}_1$ will be close to an eigenvector for λ. Then if you like, you could do the same thing with the matrix B to obtain approximations for the other eigenvalues. Finally, you could use the shifted inverse power method to get more exact solutions.

15.2.2 The Case Of Real Eigenvalues

With these lemmas, it is possible to prove that for the QR algorithm and certain conditions, the sequence A_k converges pointwise to an upper triangular matrix having the eigenvalues of A down the diagonal. I will assume all the matrices are real here.

This convergence won't always happen. Consider for example the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. You can verify quickly that the algorithm will return this matrix for each k. The problem here is that, although the matrix has the two eigenvalues $-1, 1$, they have the same absolute value. The QR algorithm works in somewhat the same way as the power method, exploiting differences in the size of the eigenvalues.

If A has all real eigenvalues and you are interested in finding these eigenvalues along with the corresponding eigenvectors, you could always consider $A + \lambda I$ instead where λ is sufficiently large and positive that $A + \lambda I$ has all positive eigenvalues. (Recall Gerschgorin's theorem.) Then if μ is an eigenvalue of $A + \lambda I$ with

$$(A + \lambda I)\,\mathbf{x} = \mu \mathbf{x}$$

then

$$A\mathbf{x} = (\mu - \lambda)\,\mathbf{x}$$

so to find the eigenvalues of A you just subtract λ from the eigenvalues of $A + \lambda I$. Thus there is no loss of generality in assuming at the outset that the eigenvalues of A are all positive. Here is the theorem. It involves a technical condition which will often hold. The proof presented here follows [26] and is a special case of that presented in this reference.

Before giving the proof, note that the product of upper triangular matrices is upper triangular. If they both have positive entries on the main diagonal so will the product. Furthermore, the inverse of an upper triangular matrix is upper triangular. I will use these simple facts without much comment whenever convenient.

Theorem 15.2.4 *Let A be a real matrix having eigenvalues*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$$

and let

$$A = SDS^{-1} \tag{15.11}$$

where

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

and suppose S^{-1} has an LU factorization. Then the matrices A_k in the QR algorithm described above converge to an upper triangular matrix T' having the eigenvalues of A, $\lambda_1, \cdots, \lambda_n$ descending on the main diagonal. The matrices $Q^{(k)}$ converge to Q', an orthogonal matrix which equals Q except for possibly having some columns multiplied by -1 for Q the unitary part of the QR factorization of S,

$$S = QR,$$

and

$$\lim_{k\to\infty} A_k = T' = Q'^T A Q'$$

Proof: From Lemma 15.2.2

$$A^k = Q^{(k)} R^{(k)} = SD^k S^{-1} \tag{15.12}$$

Let $S = QR$ where this is just a QR factorization which is known to exist and let $S^{-1} = LU$ which is assumed to exist. Thus

$$Q^{(k)} R^{(k)} = QRD^k LU \tag{15.13}$$

and so

$$Q^{(k)} R^{(k)} = QRD^k LU = QRD^k LD^{-k} D^k U$$

That matrix in the middle, $D^k LD^{-k}$ satisfies

$$\left(D^k LD^{-k}\right)_{ij} = \lambda_i^k L_{ij} \lambda_j^{-k} \text{ for } j \leq i, \ 0 \text{ if } j > i.$$

Thus for $j < i$ the expression converges to 0 because $\lambda_j > \lambda_i$ when this happens. When $i = j$ it reduces to 1. Thus the matrix in the middle is of the form

$$I + E_k$$

where $E_k \to 0$. Then it follows

$$A^k = Q^{(k)}R^{(k)} = QR\left(I + E_k\right)D^kU$$
$$= Q\left(I + RE_kR^{-1}\right)RD^kU \equiv Q\left(I + F_k\right)RD^kU$$

where $F_k \to 0$. Then let $I + F_k = Q_kR_k$ where this is another QR factorization. Then it reduces to

$$Q^{(k)}R^{(k)} = QQ_kR_kRD^kU$$

This looks really interesting because by Lemma 15.2.3 $Q_k \to I$ and $R_k \to I$ because $Q_kR_k = (I + F_k) \to I$. So it follows QQ_k is an orthogonal matrix converging to Q while

$$R_kRD^kU\left(R^{(k)}\right)^{-1}$$

is upper triangular, being the product of upper triangular matrices. Unfortunately, it is not known that the diagonal entries of this matrix are nonnegative because of the U. Let Λ be just like the identity matrix but having some of the ones replaced with -1 in such a way that ΛU is an upper triangular matrix having positive diagonal entries. Note $\Lambda^2 = I$ and also Λ commutes with a diagonal matrix. Thus

$$Q^{(k)}R^{(k)} = QQ_kR_kRD^k\Lambda^2U = QQ_kR_kR\Lambda D^k\left(\Lambda U\right)$$

At this point, one does some inspired massaging to write the above in the form

$$\begin{aligned} & QQ_k\left(\Lambda D^k\right)\left[\left(\Lambda D^k\right)^{-1}R_kR\Lambda D^k\right]\left(\Lambda U\right) \\ = \quad & Q\left(Q_k\Lambda\right)D^k\left[\left(\Lambda D^k\right)^{-1}R_kR\Lambda D^k\right]\left(\Lambda U\right) \\ = \quad & Q\left(Q_k\Lambda\right)\overbrace{D^k\left[\left(\Lambda D^k\right)^{-1}R_kR\Lambda D^k\right]\left(\Lambda U\right)}^{\equiv G_k} \end{aligned}$$

Now I claim the middle matrix in $[\cdot]$ is upper triangular and has all positive entries on the diagonal. This is because it is an upper triangular matrix which is similar to the upper triangular matrix R_kR and so it has the same eigenvalues (diagonal entries) as R_kR. Thus the matrix $G_k \equiv D^k\left[\left(\Lambda D^k\right)^{-1}R_kR\Lambda D^k\right]\left(\Lambda U\right)$ is upper triangular and has all positive entries on the diagonal. Multiply on the right by G_k^{-1} to get

$$Q^{(k)}R^{(k)}G_k^{-1} = QQ_k\Lambda \to Q'$$

where Q' is essentially equal to Q but might have some of the columns multiplied by -1. This is because $Q_k \to I$ and so $Q_k\Lambda \to \Lambda$. Now by Lemma 15.2.3, it follows

$$Q^{(k)} \to Q',\ R^{(k)}G_k^{-1} \to I.$$

It remains to verify A_k converges to an upper triangular matrix. Recall that from 15.12 and the definition below this $(S = QR)$

$$A = SDS^{-1} = (QR)\,D\,(QR)^{-1} = QRDR^{-1}Q^T = QTQ^T$$

Where T is an upper triangular matrix. This is because it is the product of upper triangular matrices R, D, R^{-1}. Thus

$$Q^TAQ = T.$$

If you replace Q with Q' in the above, it still results in an upper triangular matrix T' having the same diagonal entries as T. This is because

$$T = Q^T AQ = (Q'\Lambda)^T A (Q'\Lambda) = \Lambda Q'^T AQ'\Lambda$$

and considering the ii^{th} entry yields

$$\left(Q^T AQ\right)_{ii} \equiv \sum_{j,k} \Lambda_{ij} \left(Q'^T AQ'\right)_{jk} \Lambda_{ki} = \Lambda_{ii}\Lambda_{ii} \left(Q'^T AQ'\right)_{ii} = \left(Q'^T AQ'\right)_{ii}$$

Recall from Lemma 15.2.2,

$$A_k = Q^{(k)T} AQ^{(k)}$$

Thus taking a limit and using the first part,

$$A_k = Q^{(k)T} AQ^{(k)} \to Q'^T AQ' = T'. \blacksquare$$

An easy case is for A symmetric. Recall Corollary 7.4.13. By this corollary, there exists an orthogonal (real unitary) matrix Q such that

$$Q^T AQ = D$$

where D is diagonal having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right.

Corollary 15.2.5 *Let A be a real symmetric $n \times n$ matrix having eigenvalues*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$$

and let Q be defined by

$$QDQ^T = A, \ D = Q^T AQ, \tag{15.14}$$

where Q is orthogonal and D is a diagonal matrix having the eigenvalues on the main diagonal decreasing in size from the upper left corner to the lower right. Let Q^T have an LU factorization. Then in the QR algorithm, the matrices $Q^{(k)}$ converge to Q' where Q' is the same as Q except having some columns multiplied by (-1). Thus the columns of Q' are eigenvectors of A. The matrices A_k converge to D.

Proof: This follows from Theorem 15.2.4. Here $S = Q, S^{-1} = Q^T$. Thus

$$Q = S = QR$$

and $R = I$. By Theorem 15.2.4 and Lemma 15.2.2,

$$A_k = Q^{(k)T} AQ^{(k)} \to Q'^T AQ' = Q^T AQ = D.$$

because formula 15.14 is unaffected by replacing Q with Q'. $\blacksquare$

When using the QR algorithm, it is not necessary to check technical condition about S^{-1} having an LU factorization. The algorithm delivers a sequence of matrices which are similar to the original one. If that sequence converges to an upper triangular matrix, then the algorithm worked. Furthermore, the technical condition is sufficient but not necessary. The algorithm will work even without the technical condition.

Example 15.2.6 *Find the eigenvalues and eigenvectors of the matrix*

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

It is a symmetric matrix but other than that, I just pulled it out of the air. By Lemma 15.2.2 it follows $A_k = Q^{(k)T}AQ^{(k)}$. And so to get to the answer quickly I could have the computer raise A to a power and then take the QR factorization of what results to get the k^{th} iteration using the above formula. Lets pick $k = 10$.

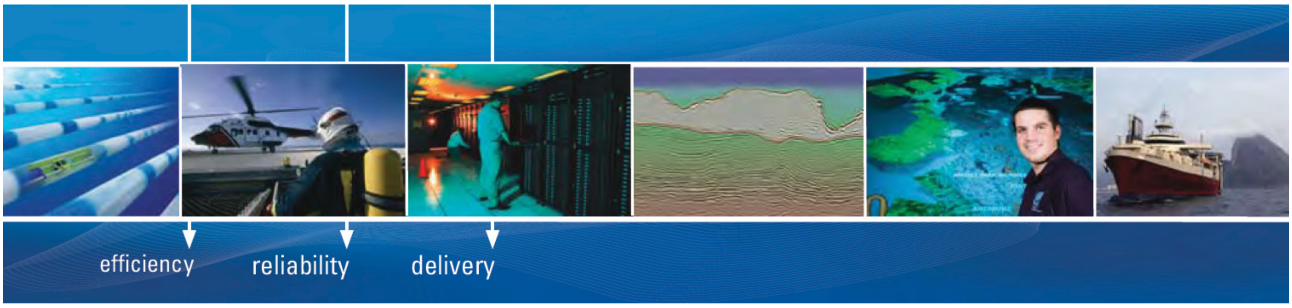
$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}^{10} = \begin{pmatrix} 4.2273 \times 10^{7} & 2.5959 \times 10^{7} & 1.8611 \times 10^{7} \\ 2.5959 \times 10^{7} & 1.6072 \times 10^{7} & 1.1506 \times 10^{7} \\ 1.8611 \times 10^{7} & 1.1506 \times 10^{7} & 8.2396 \times 10^{6} \end{pmatrix}$$

Now take QR factorization of this. The computer will do that also.
This yields

$$\begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix} \cdot \begin{pmatrix} 5.2983 \times 10^{7} & 3.2627 \times 10^{7} & 2.338 \times 10^{7} \\ 0 & 1.2172 \times 10^{5} & 71946. \\ 0 & 0 & 277.03 \end{pmatrix}$$

Next it follows

$$A_{10} = \begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix}^{T} \cdot \begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} .79785 & -.59912 & -6.6943 \times 10^{-2} \\ .48995 & .70912 & -.50706 \\ .35126 & .37176 & .85931 \end{pmatrix}$$



and this equals

$$\begin{pmatrix} 6.0571 & 3.698 \times 10^{-3} & 3.4346 \times 10^{-5} \\ 3.698 \times 10^{-3} & 3.2008 & -4.0643 \times 10^{-4} \\ 3.4346 \times 10^{-5} & -4.0643 \times 10^{-4} & -.2579 \end{pmatrix}$$

By Gerschgorin's theorem, the eigenvalues are pretty close to the diagonal entries of the above matrix. Note I didn't use the theorem, just Lemma 15.2.2 and Gerschgorin's theorem to verify the eigenvalues are close to the above numbers. The eigenvectors are close to

$$\begin{pmatrix} .79785 \\ .48995 \\ .35126 \end{pmatrix}, \begin{pmatrix} -.59912 \\ .70912 \\ .37176 \end{pmatrix}, \begin{pmatrix} -6.6943 \times 10^{-2} \\ -.50706 \\ .85931 \end{pmatrix}$$

Lets check one of these.

$$\left(\left(\begin{array}{ccc} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{array}\right) - 6.0571\left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)\right)\left(\begin{array}{c} .79785 \\ .48995 \\ .35126 \end{array}\right)$$

$$= \left(\begin{array}{c} -2.1972\times 10^{-3} \\ 2.5439\times 10^{-3} \\ 1.3931\times 10^{-3} \end{array}\right) \approx \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right)$$

Now lets see how well the smallest approximate eigenvalue and eigenvector works.

$$\left(\left(\begin{array}{ccc} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{array}\right) - (-.2579)\left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)\right)\left(\begin{array}{c} -6.6943\times 10^{-2} \\ -.50706 \\ .85931 \end{array}\right)$$

$$= \left(\begin{array}{c} 2.704\times 10^{-4} \\ -2.7377\times 10^{-4} \\ -1.3695\times 10^{-4} \end{array}\right) \approx \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array}\right)$$

For practical purposes, this has found the eigenvalues and eigenvectors.

15.2.3 The QR Algorithm In The General Case

In the case where A has distinct positive eigenvalues it was shown above that under reasonable conditions related to a certain matrix having an LU factorization the QR algorithm produces a sequence of matrices $\{A_k\}$ which converges to an upper triangular matrix. What if A is just an $n\times n$ matrix having possibly complex eigenvalues but A is nondefective? What happens with the QR algorithm in this case? The short answer to this question is that the A_k of the algorithm **typically cannot converge**. However, this does not mean the algorithm is not useful in finding eigenvalues. It turns out the sequence of matrices $\{A_k\}$ have the appearance of a block upper triangular matrix for large k in the sense that the entries below the blocks on the main diagonal are small. Then looking at these blocks gives a way to approximate the eigenvalues. An important example of the concept of a block triangular matrix is the real Schur form for a matrix discussed in Theorem 7.4.6 but the concept as described here allows for any size block centered on the diagonal.

First it is important to note a simple fact about unitary diagonal matrices. In what follows Λ will denote a unitary matrix which is also a diagonal matrix. These matrices are just the identity matrix with some of the ones replaced with a number of the form $e^{i\theta}$ for some θ. The important property of multiplication of any matrix by Λ on either side is that it leaves all the zero entries the same and also preserves the absolute values of the other entries. Thus a block triangular matrix multiplied by Λ on either side is still block triangular. If the matrix is close to being block triangular this property of being close to a block triangular matrix is also preserved by multiplying on either side by Λ. Other patterns depending only on the size of the absolute value occurring in the matrix are also preserved by multiplying on either side by Λ. In other words, in looking for a pattern in a matrix, multiplication by Λ is irrelevant.

Now let A be an $n\times n$ matrix having real or complex entries. By Lemma 15.2.2 and the assumption that A is nondefective, there exists an invertible S,

$$A^k = Q^{(k)}R^{(k)} = SD^kS^{-1} \tag{15.15}$$

where

$$D = \left(\begin{array}{ccc} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{array}\right)$$

and by rearranging the columns of S, D can be made such that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| .$$

Assume S^{-1} has an LU factorization. Then

$$A^k = SD^k LU = SD^k LD^{-k} D^k U.$$

Consider the matrix in the middle, $D^k LD^{-k}$. The ij^{th} entry is of the form

$$\left(D^k LD^{-k}\right)_{ij} = \begin{cases} \lambda_i^k L_{ij} \lambda_j^{-k} \text{ if } j < i \\ 1 \text{ if } i = j \\ 0 \text{ if } j > i \end{cases}$$

and these all converge to 0 whenever $|\lambda_i| < |\lambda_j|$. Thus

$$D^k LD^{-k} = (L_k + E_k)$$

where L_k is a lower triangular matrix which has all ones down the diagonal and some subdiagonal terms of the form

$$\lambda_i^k L_{ij} \lambda_j^{-k} \tag{15.16}$$

for which $|\lambda_i| = |\lambda_j|$ while $E_k \to 0$. (Note the entries of L_k are all bounded independent of k but some may fail to converge.) Then

$$Q^{(k)} R^{(k)} = S (L_k + E_k) D^k U$$

Let

$$SL_k = Q_k R_k \tag{15.17}$$

where this is the QR factorization of SL_k. Then

$$\begin{aligned} Q^{(k)} R^{(k)} &= (Q_k R_k + SE_k) D^k U \\ &= Q_k \left(I + Q_k^* SE_k R_k^{-1}\right) R_k D^k U \\ &= Q_k (I + F_k) R_k D^k U \end{aligned}$$

where $F_k \to 0$. Let $I + F_k = Q_k' R_k'$. Then

$$Q^{(k)} R^{(k)} = Q_k Q_k' R_k' R_k D^k U$$

By Lemma 15.2.3

$$Q_k' \to I \text{ and } R_k' \to I. \tag{15.18}$$

Now let Λ_k be a diagonal unitary matrix which has the property that

$$\Lambda_k^* D^k U$$

is an upper triangular matrix which has all the diagonal entries positive. Then

$$Q^{(k)} R^{(k)} = Q_k Q_k' \Lambda_k (\Lambda_k^* R_k' R_k \Lambda_k) \Lambda_k^* D^k U$$

That matrix in the middle has all positive diagonal entries because it is itself an upper triangular matrix, being the product of such, and is similar to the matrix $R_k' R_k$ which is upper triangular with positive diagonal entries. By Lemma 15.2.3 again, this time using the uniqueness assertion,

$$Q^{(k)} = Q_k Q_k' \Lambda_k ,\ R^{(k)} = (\Lambda_k^* R_k' R_k \Lambda_k) \Lambda_k^* D^k U$$

Note the term $Q_k Q_k' \Lambda_k$ must be real because the algorithm gives all $Q^{(k)}$ as real matrices. By 15.18 it follows that for k large enough

$$Q^{(k)} \approx Q_k \Lambda_k$$

where $\approx$ means the two matrices are close. Recall

$$A_k = Q^{(k)T} A Q^{(k)}$$

and so for large k,

$$A_k \approx (Q_k \Lambda_k)^* A (Q_k \Lambda_k) = \Lambda_k^* Q_k^* A Q_k \Lambda_k$$

As noted above, the form of $\Lambda_k^* Q_k^* A Q_k \Lambda_k$ in terms of which entries are large and small is not affected by the presence of Λ_k and Λ_k^*. Thus, in considering what form this is in, it suffices to consider $Q_k^* A Q_k$.

This could get pretty complicated but I will consider the case where

$$\text{if } |\lambda_i| = |\lambda_{i+1}|, \text{ then } |\lambda_{i+2}| < |\lambda_{i+1}|. \tag{15.19}$$

This is typical of the situation where the eigenvalues are all distinct and the matrix A is real so the eigenvalues occur as conjugate pairs. Then in this case, L_k above is lower triangular with some nonzero terms on the diagonal right below the main diagonal but zeros everywhere else. Thus maybe

$$(L_k)_{s+1,s} \neq 0$$

Recall 15.17 which implies

$$Q_k = S L_k R_k^{-1} \tag{15.20}$$

where R_k^{-1} is upper triangular. Also recall that from the definition of S in 15.15,

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

and so the columns of S are eigenvectors of A, the i^{th} being an eigenvector for λ_i. Now from the form of L_k, it follows $L_k R_k^{-1}$ is a block upper triangular matrix denoted by T_B and so $Q_k = S T_B$. It follows from the above construction in 15.16 and the given assumption on the sizes of the eigenvalues, there are finitely many 2×2 blocks centered on the main diagonal along with possibly some diagonal entries. Therefore, for large k the matrix

$$A_k = Q^{(k)T} A Q^{(k)}$$

is approximately of the same form as that of

$$Q_k^* A Q_k = T_B^{-1} S^{-1} A S T_B = T_B^{-1} D T_B$$

which is a block upper triangular matrix. As explained above, multiplication by the various diagonal unitary matrices does not affect this form. Therefore, for large k, A_k is approximately a block upper triangular matrix.

How would this change if the above assumption on the size of the eigenvalues were relaxed but the matrix was still nondefective with appropriate matrices having an LU factorization as above? It would mean the blocks on the diagonal would be larger. This immediately makes the problem more cumbersome to deal with. However, in the case that the eigenvalues of A are distinct, the above situation really is typical of what occurs and in any case can be quickly reduced to this case.

To see this, suppose condition 15.19 is violated and $\lambda_j, \cdots, \lambda_{j+p}$ are complex eigenvalues having nonzero imaginary parts such that each has the same absolute value but they are all distinct. Then let $\mu > 0$ and consider the matrix $A+\mu I$. Thus the corresponding eigenvalues of $A+\mu I$ are $\lambda_j+\mu, \cdots, \lambda_{j+p}+\mu$. A short computation shows shows $|\lambda_j + \mu|, \cdots, |\lambda_{j+p} + \mu|$

are all distinct and so the above situation of 15.19 is obtained. Of course, if there are repeated eigenvalues, it may not be possible to reduce to the case above and you would end up with large blocks on the main diagonal which could be difficult to deal with.

So how do you identify the eigenvalues? You know A_k and behold that it is close to a block upper triangular matrix T'_B. You know A_k is also similar to A. Therefore, T'_B has eigenvalues which are close to the eigenvalues of A_k and hence those of A provided k is sufficiently large. See Theorem 7.9.2 which depends on complex analysis or the exercise on Page 264 which gives another way to see this. Thus you find the eigenvalues of this block triangular matrix T'_B and assert that these are good approximations of the eigenvalues of A_k and hence to those of A. How do you find the eigenvalues of a block triangular matrix? This is easy from Lemma 7.4.5. Say

$$T'_B = \begin{pmatrix} B_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & B_m \end{pmatrix}$$

Then forming $\lambda I - T'_B$ and taking the determinant, it follows from Lemma 7.4.5 this equals

$$\prod_{j=1}^{m} \det(\lambda I_j - B_j)$$

and so all you have to do is take the union of the eigenvalues for each B_j. In the case emphasized here this is very easy because these blocks are just 2 × 2 matrices.

How do you identify approximate eigenvectors from this? First try to find the approximate eigenvectors for A_k. Pick an approximate eigenvalue λ, an exact eigenvalue for T'_B. Then find $\mathbf{v}$ solving $T'_B\mathbf{v} = \lambda\mathbf{v}$. It follows since T'_B is close to A_k that

$$A_k\mathbf{v} \approx \lambda\mathbf{v}$$

and so

$$Q^{(k)}AQ^{(k)T}\mathbf{v} = A_k\mathbf{v} \approx \lambda\mathbf{v}$$

Hence

$$AQ^{(k)T}\mathbf{v} \approx \lambda Q^{(k)T}\mathbf{v}$$

and so $Q^{(k)T}\mathbf{v}$ is an approximation to the eigenvector which goes with the eigenvalue of A which is close to λ.

Example 15.2.7 *Here is a matrix.*

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

It happens that the eigenvalues of this matrix are $1, 1+i, 1-i$. *Lets apply the* QR *algorithm as if the eigenvalues were not known.*

Applying the QR algorithm to this matrix yields the following sequence of matrices.

$$A_1 = \begin{pmatrix} 1.2353 & 1.9412 & 4.3657 \\ -.39215 & 1.5425 & 5.3886 \times 10^{-2} \\ -.16169 & -.18864 & .22222 \end{pmatrix}$$

$$\vdots$$

$$A_{12} = \begin{pmatrix} 9.1772 \times 10^{-2} & .63089 & -2.0398 \\ -2.8556 & 1.9082 & -3.1043 \\ 1.0786 \times 10^{-2} & 3.4614 \times 10^{-4} & 1.0 \end{pmatrix}$$

At this point the bottom two terms on the left part of the bottom row are both very small so it appears the real eigenvalue is near 1.0. The complex eigenvalues are obtained from solving

$$\det \left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 9.1772 \times 10^{-2} & .63089 \\ -2.8556 & 1.9082 \end{pmatrix} \right) = 0$$

This yields

$$\lambda = 1.0 - .98828i, \ 1.0 + .98828i$$

Example 15.2.8 *The equation $x^4 + x^3 + 4x^2 + x - 2 = 0$ has exactly two real solutions. You can see this by graphing it. However, the rational root theorem from algebra shows neither of these solutions are rational. Also, graphing it does not yield any information about the complex solutions. Lets use the QR algorithm to approximate all the solutions, real and complex.*

A matrix whose characteristic polynomial is the given polynomial is

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

Using the QR algorithm yields the following sequence of iterates for A_k

$$A_1 = \begin{pmatrix} .99999 & -2.5927 & -1.7588 & -1.2978 \\ 2.1213 & -1.7778 & -1.6042 & -.99415 \\ 0 & .34246 & -.32749 & -.91799 \\ 0 & 0 & -.44659 & .10526 \end{pmatrix}$$

$$\vdots$$

$$A_9 = \begin{pmatrix} -.83412 & -4.1682 & -1.939 & -.7783 \\ 1.05 & .14514 & .2171 & 2.5474 \times 10^{-2} \\ 0 & 4.0264 \times 10^{-4} & -.85029 & -.61608 \\ 0 & 0 & -1.8263 \times 10^{-2} & .53939 \end{pmatrix}$$

Now this is similar to A and the eigenvalues are close to the eigenvalues obtained from the two blocks on the diagonal. Of course the lower left corner of the bottom block is vanishing but it is still fairly large so the eigenvalues are approximated by the solution to

$$\det \left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -.85029 & -.61608 \\ -1.8263 \times 10^{-2} & .53939 \end{pmatrix} \right) = 0$$

The solution to this is

$$\lambda = -.85834, \ .54744$$

and for the complex eigenvalues,

$$\det \left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -.83412 & -4.1682 \\ 1.05 & .14514 \end{pmatrix} \right) = 0$$

The solution is

$$\lambda = -.\,344\,49 - 2.\,033\,9i,\ -.\,344\,49 + 2.\,033\,9i$$

How close are the complex eigenvalues just obtained to giving a solution to the original equation? Try $-.\,344\,49 + 2.\,033\,9i$. When this is plugged in it yields

$$-.00\,12 + 2.\,006\,8 \times 10^{-4}i$$

which is pretty close to 0. The real eigenvalues are also very close to the corresponding real solutions to the original equation.

It seems like most of the attention to the QR algorithm has to do with finding ways to get it to "converge" faster. Great and marvelous are the clever tricks which have been proposed to do this but my intent is to present the basic ideas, not to go in to the numerous refinements of this algorithm. However, there is one thing which is usually done. It involves reducing to the case of an upper Hessenberg matrix which is one which is zero below the main sub diagonal. To see that every matrix is unitarily similar to an upper Hessenberg matrix , see Problem 1 on Page 360. What follows is a construction which also proves this.

Let A be an invertible $n \times n$ matrix. Let Q_1' be a unitary matrix

$$Q_1' \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} \sqrt{\sum_{j=2}^{n} |a_{j1}|^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv \begin{pmatrix} a \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The vector Q_1' is multiplying is just the bottom $n-1$ entries of the first column of A. Then let Q_1 be

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1' \end{pmatrix}$$

It follows

$$Q_1 A Q_1^* = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1' \end{pmatrix} A Q_1^* = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a & & & \\ \vdots & & A_1' & \\ 0 & & & \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1'^* \end{pmatrix}$$

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

Now let Q_2' be the $n-2 \times n-2$ matrix which does to the first column of A_1 the same sort of thing that the $n-1 \times n-1$ matrix Q_1' did to the first column of A. Let

$$Q_2 \equiv \begin{pmatrix} I & 0 \\ 0 & Q_2' \end{pmatrix}$$

where I is the 2×2 identity. Then applying block multiplication,

$$Q_2 Q_1 A Q_1^* Q_2^* = \begin{pmatrix} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$$

where A_2 is now an $n-2 \times n-2$ matrix. Continuing this way you eventually get a unitary

matrix Q which is a product of those discussed above such that

$$QAQ^T = \begin{pmatrix} * & * & \cdots & * & * \\ * & * & \cdots & * & * \\ 0 & * & * & & \vdots \\ \vdots & \vdots & \ddots & \ddots & * \\ 0 & 0 & & * & * \end{pmatrix}$$

This matrix equals zero below the subdiagonal. It is called an upper Hessenberg matrix.

It happens that in the QR algorithm, if A_k is upper Hessenberg, so is A_{k+1}. To see this, note that the matrix is upper Hessenberg means that $A_{ij} = 0$ whenever $i - j \geq 2$.

$$A_{k+1} = R_k Q_k$$

where $A_k = Q_k R_k$. Therefore as shown before,

$$A_{k+1} = R_k A_k R_k^{-1}$$

Let the ij^{th} entry of A_k be a_{ij}^k. Then if $i - j \geq 2$

$$a_{ij}^{k+1} = \sum_{p=i}^{n} \sum_{q=1}^{j} r_{ip} a_{pq}^k r_{qj}^{-1}$$

It is given that $a_{pq}^k = 0$ whenever $p - q \geq 2$. However, from the above sum,

$$p - q \geq i - j \geq 2$$

and so the sum equals 0.

Since upper Hessenberg matrices stay that way in the algorithm and it is closer to being upper triangular, it is reasonable to suppose the QR algorithm will yield good results more quickly for this upper Hessenberg matrix than for the original matrix. This would be especially true if the matrix is good sized. The other important thing to observe is that, starting with an upper Hessenberg matrix, the algorithm will restrict the size of the blocks which occur to being 2×2 blocks which are easy to deal with. These blocks allow you to identify the complex roots.

15.3 Exercises

In these exercises which call for a computation, don't waste time on them unless you use a computer or calculator which can raise matrices to powers and take QR factorizations.

1. In Example 15.1.10 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

2. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   In this case the exact eigenvalues are $\pm\sqrt{3}, 6$. Compare with the exact answers.

3. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   The exact eigenvalues are $2, 4 + \sqrt{15}, 4 - \sqrt{15}$. Compare your numerical results with the exact values. Is it much fun to compute the exact eigenvectors?

4. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

5. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.

   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

6. Consider the matrix $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

7. Consider the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

8. Consider the matrix $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

9. Using Gerschgorin's theorem, find upper and lower bounds for the eigenvalues of $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$.

10. Tell how to find a matrix whose characteristic polynomial is a given monic polynomial. This is called a companion matrix. Find the roots of the polynomial $x^3 + 7x^2 + 3x + 7$.

11. Find the roots to $x^4 + 3x^3 + 4x^2 + x + 1$. It has two complex roots.

12. Suppose A is a real symmetric matrix and the technique of reducing to an upper Hessenberg matrix is followed. Show the resulting upper Hessenberg matrix is actually equal to 0 on the top as well as the bottom.

Positive Matrices

Earlier theorems about Markov matrices were presented. These were matrices in which all the entries were nonnegative and either the columns or the rows added to 1. It turns out that many of the theorems presented can be generalized to positive matrices. When this is done, the resulting theory is mainly due to Perron and Frobenius. I will give an introduction to this theory here following Karlin and Taylor [18].

Definition A.0.1 *For A a matrix or vector, the notation, $A >> 0$ will mean every entry of A is positive. By $A > 0$ is meant that every entry is nonnegative and at least one is positive. By $A \geq 0$ is meant that every entry is nonnegative. Thus the matrix or vector consisting only of zeros is ≥ 0. An expression like $A >> B$ will mean $A - B >> 0$ with similar modifications for $>$ and $\geq$.*

For the sake of this section only, define the following for $\mathbf{x} = (x_1, \cdots, x_n)^T$, a vector.

$$|\mathbf{x}| \equiv (|x_1|, \cdots, |x_n|)^T .$$

Thus $|\mathbf{x}|$ is the vector which results by replacing each entry of $\mathbf{x}$ with its absolute value[1]*. Also define for $\mathbf{x} \in \mathbb{C}^n$,*

$$||\mathbf{x}||_1 \equiv \sum_k |x_k| .$$

Lemma A.0.2 *Let $A >> 0$ and let $\mathbf{x} > \mathbf{0}$. Then $A\mathbf{x} >> \mathbf{0}$.*

Proof: $(A\mathbf{x})_i = \sum_j A_{ij} x_j > 0$ because all the $A_{ij} > 0$ and at least one $x_j > 0$.

Lemma A.0.3 *Let $A >> 0$. Define*

$$S \equiv \{\lambda : A\mathbf{x} > \lambda\mathbf{x} \text{ for some } \mathbf{x} >> \mathbf{0}\} ,$$

and let

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\} .$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\} .$$

Then

$$\sup(S) = \sup(S_1) .$$

[1] This notation is just about the most abominable thing imaginable. However, it saves space in the presentation of this theory of positive matrices and avoids the use of new symbols. Please forget about it when you leave this section.

Proof: Let $\lambda \in S$. Then there exists $\mathbf{x} >> \mathbf{0}$ such that $A\mathbf{x} > \lambda\mathbf{x}$. Consider $\mathbf{y} \equiv \mathbf{x} / \left|\left|\mathbf{x}\right|\right|_1$. Then $\left|\left|\mathbf{y}\right|\right|_1 = 1$ and $A\mathbf{y} > \lambda\mathbf{y}$. Therefore, $\lambda \in S_1$ and so $S \subseteq S_1$. Therefore, $\sup(S) \leq \sup(S_1)$.

Now let $\lambda \in S_1$. Then there exists $\mathbf{x} \geq \mathbf{0}$ such that $\left|\left|\mathbf{x}\right|\right|_1 = 1$ so $\mathbf{x} > \mathbf{0}$ and $A\mathbf{x} > \lambda\mathbf{x}$. Letting $\mathbf{y} \equiv A\mathbf{x}$, it follows from Lemma A.0.2 that $A\mathbf{y} >> \lambda\mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. Thus $\lambda \in S$ and so $S_1 \subseteq S$ which shows that $\sup(S_1) \leq \sup(S)$. ■

This lemma is significant because the set, $\{\mathbf{x} \geq \mathbf{0}$ such that $\left|\left|\mathbf{x}\right|\right|_1 = 1\} \equiv K$ is a compact set in $\mathbb{R}^n$. Define

$$\lambda_0 \equiv \sup(S) = \sup(S_1). \tag{1.1}$$

The following theorem is due to Perron.

Theorem A.0.4 *Let $A >> 0$ be an $n \times n$ matrix and let λ_0 be given in 1.1. Then*

1. *$\lambda_0 > 0$ and there exists $\mathbf{x}_0 >> \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$ so λ_0 is an eigenvalue for A.*
2. *If $A\mathbf{x} = \mu\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq \lambda_0$. Then $|\mu| < \lambda_0$.*
3. *The eigenspace for λ_0 has dimension* 1.

Proof: To see $\lambda_0 > 0$, consider the vector, $\mathbf{e} \equiv (1, \cdots, 1)^T$. Then

$$(A\mathbf{e})_i = \sum_j A_{ij} > 0$$

and so λ_0 is at least as large as

$$\min_i \sum_j A_{ij}.$$

Let $\{\lambda_k\}$ be an increasing sequence of numbers from S_1 converging to λ_0. Letting $\mathbf{x}_k$ be the vector from K which occurs in the definition of S_1, these vectors are in a compact set. Therefore, there exists a subsequence, still denoted by $\mathbf{x}_k$ such that $\mathbf{x}_k \to \mathbf{x}_0 \in K$ and $\lambda_k \to \lambda_0$. Then passing to the limit,

$$A\mathbf{x}_0 \geq \lambda_0\mathbf{x}_0,\ \mathbf{x}_0 > \mathbf{0}.$$

If $A\mathbf{x}_0 > \lambda_0\mathbf{x}_0$, then letting $\mathbf{y} \equiv A\mathbf{x}_0$, it follows from Lemma A.0.2 that $A\mathbf{y} >> \lambda_0\mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. But this contradicts the definition of λ_0 as the supremum of the elements of S because since $A\mathbf{y} >> \lambda_0\mathbf{y}$, it follows $A\mathbf{y} >> (\lambda_0 + \varepsilon)\mathbf{y}$ for ε a small positive number. Therefore, $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$. It remains to verify that $\mathbf{x}_0 >> \mathbf{0}$. But this follows immediately from

$$0 < \sum_j A_{ij}x_{0j} = (A\mathbf{x}_0)_i = \lambda_0 x_{0i}.$$

This proves 1.

Next suppose $A\mathbf{x} = \mu\mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Then $|A\mathbf{x}| = |\mu|\,|\mathbf{x}|$. But this implies $A|\mathbf{x}| \geq |\mu|\,|\mathbf{x}|$. (See the above abominable definition of $|\mathbf{x}|$.)

Case 1: $|\mathbf{x}| \neq \mathbf{x}$ and $|\mathbf{x}| \neq -\mathbf{x}$.

In this case, $A|\mathbf{x}| > |A\mathbf{x}| = |\mu|\,|\mathbf{x}|$ and letting $\mathbf{y} = A|\mathbf{x}|$, it follows $\mathbf{y} >> \mathbf{0}$ and $A\mathbf{y} >> |\mu|\,\mathbf{y}$ which shows $A\mathbf{y} >> (|\mu| + \varepsilon)\mathbf{y}$ for sufficiently small positive ε and verifies $|\mu| < \lambda_0$.

Case 2: $|\mathbf{x}| = \mathbf{x}$ or $|\mathbf{x}| = -\mathbf{x}$

In this case, the entries of $\mathbf{x}$ are all real and have the same sign. Therefore, $A|\mathbf{x}| = |A\mathbf{x}| = |\mu|\,|\mathbf{x}|$. Now let $\mathbf{y} \equiv |\mathbf{x}| / \left|\left|\mathbf{x}\right|\right|_1$. Then $A\mathbf{y} = |\mu|\,\mathbf{y}$ and so $|\mu| \in S_1$ showing that

$|\mu| \leq \lambda_0$. But also, the fact the entries of $\mathbf{x}$ all have the same sign shows $\mu = |\mu|$ and so $\mu \in S_1$. Since $\mu \neq \lambda_0$, it must be that $\mu = |\mu| < \lambda_0$. This proves 2.

It remains to verify 3. Suppose then that $A\mathbf{y} = \lambda_0 \mathbf{y}$ and for all scalars $\alpha, \alpha \mathbf{x}_0 \neq \mathbf{y}$. Then

$$A \operatorname{Re} \mathbf{y} = \lambda_0 \operatorname{Re} \mathbf{y},\, A \operatorname{Im} \mathbf{y} = \lambda_0 \operatorname{Im} \mathbf{y}.$$

If $\operatorname{Re} \mathbf{y} = \alpha_1 \mathbf{x}_0$ and $\operatorname{Im} \mathbf{y} = \alpha_2 \mathbf{x}_0$ for real numbers, α_i,then $\mathbf{y} = (\alpha_1 + i\alpha_2)\, \mathbf{x}_0$ and it is assumed this does not happen. Therefore, either

$$t \operatorname{Re} \mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}$$

or

$$t \operatorname{Im} \mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}.$$

Assume the first holds. Then varying $t \in \mathbb{R}$, there exists a value of t such that $\mathbf{x}_0 + t \operatorname{Re} \mathbf{y} > \mathbf{0}$ but it is not the case that $\mathbf{x}_0 + t \operatorname{Re} \mathbf{y} >> 0$. Then $A\,(\mathbf{x}_0 + t \operatorname{Re} \mathbf{y}) >> 0$ by Lemma A.0.2. But this implies $\lambda_0\,(\mathbf{x}_0 + t \operatorname{Re} \mathbf{y}) >> 0$ which is a contradiction. Hence there exist real numbers, α_1 and α_2 such that $\operatorname{Re} \mathbf{y} = \alpha_1 \mathbf{x}_0$ and $\operatorname{Im} \mathbf{y} = \alpha_2 \mathbf{x}_0$ showing that $\mathbf{y} = (\alpha_1 + i\alpha_2)\, \mathbf{x}_0$. This proves 3.

It is possible to obtain a simple corollary to the above theorem.

Corollary A.0.5 *If $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$, then all the conclusions of the above theorem hold.*

Proof: There exists $\mu_0 > 0$ such that $A^m \mathbf{y}_0 = \mu_0 \mathbf{y}_0$ for $\mathbf{y}_0 >> 0$ by Theorem A.0.4 and

$$\mu_0 = \sup \{\mu : A^m \mathbf{x} \geq \mu \mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Let $\lambda_0^m = \mu_0$. Then

$$(A - \lambda_0 I)\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0 = (A^m - \lambda_0^m I)\,\mathbf{y}_0 = \mathbf{0}$$

and so letting $\mathbf{x}_0 \equiv \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0$, it follows $\mathbf{x}_0 >> 0$ and $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$.

Suppose now that $A\mathbf{x} = \mu \mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Suppose $|\mu| \geq \lambda_0$. Multiplying both sides by A, it follows $A^m \mathbf{x} = \mu^m \mathbf{x}$ and $|\mu^m| = |\mu|^m \geq \lambda_0^m = \mu_0$ and so from Theorem A.0.4, since $|\mu^m| \geq \mu_0$, and μ^m is an eigenvalue of A^m, it follows that $\mu^m = \mu_0$. But by Theorem A.0.4 again, this implies $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c and hence $A\mathbf{y}_0 = \mu \mathbf{y}_0$. Since $\mathbf{y}_0 >> \mathbf{0}$, it follows $\mu \geq 0$ and so $\mu = \lambda_0$, a contradiction. Therefore, $|\mu| < \lambda_0$.

Finally, if $A\mathbf{x} = \lambda_0 \mathbf{x}$, then $A^m \mathbf{x} = \lambda_0^m \mathbf{x}$ and so $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c. Consequently,

$$\begin{aligned}\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{x} &= c\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0 \\ &= c\mathbf{x}_0.\end{aligned}$$

Hence

$$m\lambda_0^{m-1}\mathbf{x} = c\mathbf{x}_0$$

which shows the dimension of the eigenspace for λ_0 is one. ■

The following corollary is an extremely interesting convergence result involving the powers of positive matrices.

Corollary A.0.6 *Let $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$. Then for λ_0 given in 1.1, there exists a rank one matrix P such that* $\lim_{m\to\infty} \left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = 0.$

Proof: Considering A^T, and the fact that A and A^T have the same eigenvalues, Corollary A.0.5 implies the existence of a vector, $\mathbf{v} >> \mathbf{0}$ such that

$$A^T\mathbf{v} = \lambda_0\mathbf{v}.$$

Also let $\mathbf{x}_0$ denote the vector such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$ with $\mathbf{x}_0 >> \mathbf{0}$. First note that $\mathbf{x}_0^T\mathbf{v} > 0$ because both these vectors have all entries positive. Therefore, $\mathbf{v}$ may be scaled such that

$$\mathbf{v}^T\mathbf{x}_0 = \mathbf{x}_0^T\mathbf{v} = 1. \tag{1.2}$$

Define

$$P \equiv \mathbf{x}_0\mathbf{v}^T.$$

Thanks to 1.2,

$$\frac{A}{\lambda_0}P = \mathbf{x}_0\mathbf{v}^T = P,\ P\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T = P, \tag{1.3}$$

and

$$P^2 = \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0\mathbf{v}^T = \mathbf{v}^T\mathbf{x}_0 = P. \tag{1.4}$$

Therefore,

$$\begin{aligned}\left(\frac{A}{\lambda_0} - P\right)^2 &= \left(\frac{A}{\lambda_0}\right)^2 - 2\left(\frac{A}{\lambda_0}\right)P + P^2 \\ &= \left(\frac{A}{\lambda_0}\right)^2 - P.\end{aligned}$$

Continuing this way, using 1.3 repeatedly, it follows

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m = \left(\frac{A}{\lambda_0}\right)^m - P. \tag{1.5}$$

The eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are of interest because it is powers of this matrix which determine the convergence of $\left(\frac{A}{\lambda_0}\right)^m$ to P. Therefore, let μ be a nonzero eigenvalue of this matrix. Thus

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x} = \mu\mathbf{x} \tag{1.6}$$

for $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq 0$. Applying P to both sides and using the second formula of 1.3 yields

$$\mathbf{0} = (P - P)\,\mathbf{x} = \left(P\left(\frac{A}{\lambda_0}\right) - P^2\right)\mathbf{x} = \mu P\mathbf{x}.$$

But since $P\mathbf{x} = \mathbf{0}$, it follows from 1.6 that

$$A\mathbf{x} = \lambda_0\mu\mathbf{x}$$

which implies $\lambda_0\mu$ is an eigenvalue of A. Therefore, by Corollary A.0.5 it follows that either $\lambda_0\mu = \lambda_0$ in which case $\mu = 1$, or $\lambda_0\,|\mu| < \lambda_0$ which implies $|\mu| < 1$. But if $\mu = 1$, then $\mathbf{x}$ is a multiple of $\mathbf{x}_0$ and 1.6 would yield

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x}_0 = \mathbf{x}_0$$

which says $\mathbf{x}_0 - \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0 = \mathbf{x}_0$ and so by 1.2, $\mathbf{x}_0 = \mathbf{0}$ contrary to the property that $\mathbf{x}_0 >> \mathbf{0}$. Therefore, $|\mu| < 1$ and so this has shown that the absolute values of all eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are less than 1. By Gelfand's theorem, Theorem 14.3.3, it follows

$$\left|\left|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right|\right|^{1/m} < r < 1$$

whenever m is large enough. Now by 1.5 this yields

$$\left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = \left|\left|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right|\right| \leq r^m$$

whenever m is large enough. It follows

$$\lim_{m\to\infty}\left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = 0$$

as claimed.

What about the case when $A > 0$ but maybe it is not the case that $A >> 0$? As before,

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\}.$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and

$$\lambda_0 \equiv \sup(S_1) \tag{1.7}$$

Theorem A.0.7 *Let $A > 0$ and let λ_0 be defined in 1.7. Then there exists $\mathbf{x}_0 > \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$.*

Proof: Let E consist of the matrix which has a one in every entry. Then from Theorem A.0.4 it follows there exists $\mathbf{x}_\delta >> \mathbf{0}$, $||\mathbf{x}_\delta||_1 = 1$, such that $(A + \delta E)\mathbf{x}_\delta = \lambda_{0\delta}\mathbf{x}_\delta$ where

$$\lambda_{0\delta} \equiv \sup\{\lambda : (A + \delta E)\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Now if $\alpha < \delta$

$$\{\lambda : (A + \alpha E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\} \subseteq$$
$$\{\lambda : (A + \delta E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and so $\lambda_{0\delta} \geq \lambda_{0\alpha}$ because $\lambda_{0\delta}$ is the sup of the second set and $\lambda_{0\alpha}$ is the sup of the first. It follows the limit, $\lambda_1 \equiv \lim_{\delta\to 0+} \lambda_{0\delta}$ exists. Taking a subsequence and using the compactness of K, there exists a subsequence, still denoted by δ such that as $\delta \to 0$, $\mathbf{x}_\delta \to \mathbf{x} \in K$. Therefore,

$$A\mathbf{x} = \lambda_1\mathbf{x}$$

and so, in particular, $A\mathbf{x} \geq \lambda_1\mathbf{x}$ and so $\lambda_1 \leq \lambda_0$. But also, if $\lambda \leq \lambda_0$,

$$\lambda\mathbf{x} \leq A\mathbf{x} < (A + \delta E)\mathbf{x}$$

showing that $\lambda_{0\delta} \geq \lambda$ for all such λ. But then $\lambda_{0\delta} \geq \lambda_0$ also. Hence $\lambda_1 \geq \lambda_0$, showing these two numbers are the same. Hence $A\mathbf{x} = \lambda_0\mathbf{x}$. ■

If $A^m >> 0$ for some m and $A > 0$, it follows that the dimension of the eigenspace for λ_0 is one and that the absolute value of every other eigenvalue of A is less than λ_0. If it is only assumed that $A > 0$, not necessarily $>> 0$, this is no longer true. However, there is something which is very interesting which can be said. First here is an interesting lemma.

Lemma A.0.8 *Let M be a matrix of the form*

$$M = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix}$$

or

$$M = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$

where A is an $r \times r$ matrix and C is an $(n - r) \times (n - r)$ matrix. Then $\det(M) = \det(A)\det(B)$ and $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Proof: To verify the claim about the determinants, note

$$\begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}$$

Therefore,

$$\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}.$$

But it is clear from the method of Laplace expansion that

$$\det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} = \det A$$

and from the multilinear properties of the determinant and row operations that

$$\det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det C.$$

The case where M is upper block triangular is similar.

This immediately implies $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Theorem A.0.9 *Let $A > 0$ and let λ_0 be given in 1.7. If λ is an eigenvalue for A such that $|\lambda| = \lambda_0$, then λ/λ_0 is a root of unity. Thus $(\lambda/\lambda_0)^m = 1$ for some $m \in \mathbb{N}$.*

Proof: Applying Theorem A.0.7 to A^T, there exists $\mathbf{v} > \mathbf{0}$ such that $A^T\mathbf{v} = \lambda_0\mathbf{v}$. In the first part of the argument it is assumed $\mathbf{v} >> \mathbf{0}$. Now suppose $A\mathbf{x} = \lambda\mathbf{x}, \mathbf{x} \neq \mathbf{0}$ and that $|\lambda| = \lambda_0$. Then

$$A|\mathbf{x}| \geq |\lambda||\mathbf{x}| = \lambda_0|\mathbf{x}|$$

and it follows that if $A|\mathbf{x}| > |\lambda||\mathbf{x}|$, then since $\mathbf{v} >> \mathbf{0}$,

$$\lambda_0(\mathbf{v}, |\mathbf{x}|) < (\mathbf{v}, A|\mathbf{x}|) = \left(A^T\mathbf{v}, |\mathbf{x}|\right) = \lambda_0(\mathbf{v}, |\mathbf{x}|),$$

a contradiction. Therefore,

$$A|\mathbf{x}| = \lambda_0|\mathbf{x}|. \tag{1.8}$$

It follows that

$$\left|\sum_j A_{ij}x_j\right| = \lambda_0|\mathbf{x}_i| = \sum_j A_{ij}|x_j|$$

and so the complex numbers,

$$A_{ij}x_j,\ A_{ik}x_k$$

must have the same argument for every k, j because equality holds in the triangle inequality. Therefore, there exists a complex number, μ_i such that

$$A_{ij}x_j = \mu_i A_{ij} |x_j| \tag{1.9}$$

and so, letting $r \in \mathbb{N}$,

$$A_{ij}x_j \mu_j^r = \mu_i A_{ij} |x_j| \mu_j^r.$$

Summing on j yields

$$\sum_j A_{ij}x_j \mu_j^r = \mu_i \sum_j A_{ij} |x_j| \mu_j^r. \tag{1.10}$$

Also, summing 1.9 on j and using that λ is an eigenvalue for $\mathbf{x}$, it follows from 1.8 that

$$\lambda x_i = \sum_j A_{ij}x_j = \mu_i \sum_j A_{ij} |x_j| = \mu_i \lambda_0 |x_i| . \tag{1.11}$$

From 1.10 and 1.11,

$$\begin{aligned}\sum_j A_{ij} x_j \mu_j^r &= \mu_i \sum_j A_{ij} \left|x_j\right| \mu_j^r \\ &= \mu_i \sum_j A_{ij} \overbrace{\mu_j \left|x_j\right|}^{\text{see 1.11}} \mu_j^{r-1} \\ &= \mu_i \sum_j A_{ij} \left(\frac{\lambda}{\lambda_0}\right) x_j \mu_j^{r-1} \\ &= \mu_i \left(\frac{\lambda}{\lambda_0}\right) \sum_j A_{ij} x_j \mu_j^{r-1}\end{aligned}$$

Now from 1.10 with r replaced by $r-1$, this equals

$$\begin{aligned}\mu_i^2 \left(\frac{\lambda}{\lambda_0}\right) \sum_j A_{ij} \left|x_j\right| \mu_j^{r-1} &= \mu_i^2 \left(\frac{\lambda}{\lambda_0}\right) \sum_j A_{ij} \mu_j \left|x_j\right| \mu_j^{r-2} \\ &= \mu_i^2 \left(\frac{\lambda}{\lambda_0}\right)^2 \sum_j A_{ij} x_j \mu_j^{r-2}.\end{aligned}$$

Continuing this way,

$$\sum_j A_{ij} x_j \mu_j^r = \mu_i^k \left(\frac{\lambda}{\lambda_0}\right)^k \sum_j A_{ij} x_j \mu_j^{r-k}$$

and eventually, this shows

$$\begin{aligned}\sum_j A_{ij} x_j \mu_j^r &= \mu_i^r \left(\frac{\lambda}{\lambda_0}\right)^r \sum_j A_{ij} x_j \\ &= \left(\frac{\lambda}{\lambda_0}\right)^r \lambda \left(x_i \mu_i^r\right)\end{aligned}$$

and this says $\left(\frac{\lambda}{\lambda_0}\right)^{r+1}$ is an eigenvalue for $\left(\frac{A}{\lambda_0}\right)$ with the eigenvector being

$$\left(x_1 \mu_1^r, \cdots, x_n \mu_n^r\right)^T.$$

Now recall that $r \in \mathbb{N}$ was arbitrary and so this has shown that $\left(\frac{\lambda}{\lambda_0}\right)^2, \left(\frac{\lambda}{\lambda_0}\right)^3, \left(\frac{\lambda}{\lambda_0}\right)^4, \cdots$ are each eigenvalues of $\left(\frac{A}{\lambda_0}\right)$ which has only finitely many and hence this sequence must repeat. Therefore, $\left(\frac{\lambda}{\lambda_0}\right)$ is a root of unity as claimed. This proves the theorem in the case that $\mathbf{v} >> \mathbf{0}$.

Now it is necessary to consider the case where $\mathbf{v} > \mathbf{0}$ but it is not the case that $\mathbf{v} >> \mathbf{0}$. Then in this case, there exists a permutation matrix P such that

$$P\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix} \equiv \mathbf{v}_1$$

Then

$$\lambda_0 \mathbf{v} = A^T \mathbf{v} = A^T P \mathbf{v}_1.$$

Therefore,

$$\lambda_0 \mathbf{v}_1 = P A^T P \mathbf{v}_1 = G \mathbf{v}_1$$

Now $P^2 = I$ because it is a permutation matrix. Therefore, the matrix $G \equiv PA^T P$ and A are similar. Consequently, they have the same eigenvalues and it suffices from now on to consider the matrix G rather than A. Then

$$\lambda_0 \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix}$$

where M_1 is $r \times r$ and M_4 is $(n-r) \times (n-r)$. It follows from block multiplication and the assumption that A and hence G are > 0 that

$$G = \begin{pmatrix} A' & B \\ 0 & C \end{pmatrix}.$$

Now let λ be an eigenvalue of G such that $|\lambda| = \lambda_0$. Then from Lemma A.0.8, either $\lambda \in \sigma(A')$ or $\lambda \in \sigma(C)$. Suppose without loss of generality that $\lambda \in \sigma(A')$. Since $A' > 0$ it has a largest positive eigenvalue λ_0' which is obtained from 1.7. Thus $\lambda_0' \leq \lambda_0$ but λ being an eigenvalue of A', has its absolute value bounded by λ_0' and so $\lambda_0 = |\lambda| \leq \lambda_0' \leq \lambda_0$ showing that $\lambda_0 \in \sigma(A')$. Now if there exists $\mathbf{v} >> \mathbf{0}$ such that $A'^T \mathbf{v} = \lambda_0 \mathbf{v}$, then the first part of this proof applies to the matrix A and so (λ/λ_0) is a root of unity. If such a vector, $\mathbf{v}$ does not exist, then let A' play the role of A in the above argument and reduce to the consideration of

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

where G' is similar to A' and $\lambda, \lambda_0 \in \sigma(A'')$. Stop if $A''^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Otherwise, decompose A'' similar to the above and add another prime. Continuing this way you must eventually obtain the situation where $(A'^{\cdots\prime})^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Indeed, this happens no later than when $A'^{\cdots\prime}$ is a 1×1 matrix. ■

Functions Of Matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \tag{2.1}$$

for all $|\lambda| < R$. There is a formula for $f(A) \equiv \sum_{n=0}^{\infty} a_n A^n$ which makes sense whenever $\rho(A) < R$. Thus you can speak of $\sin(A)$ or e^A for A an $n \times n$ matrix. To begin with, define

$$f_P(\lambda) \equiv \sum_{n=0}^{P} a_n \lambda^n$$

so for $k < P$

$$\begin{aligned} f_P^{(k)}(\lambda) &= \sum_{n=k}^{P} a_n n \cdots (n-k+1) \lambda^{n-k} \\ &= \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}. \end{aligned} \tag{2.2}$$

Thus

$$\frac{f_P^{(k)}(\lambda)}{k!} = \sum_{n=k}^{P} a_n \binom{n}{k} \lambda^{n-k} \tag{2.3}$$

To begin with consider $f(J_m(\lambda))$ where $J_m(\lambda)$ is an $m \times m$ Jordan block. Thus $J_m(\lambda) = D + N$ where $N^m = 0$ and N commutes with D. Therefore, letting $P > m$

$$\begin{aligned} \sum_{n=0}^{P} a_n J_m(\lambda)^n &= \sum_{n=0}^{P} a_n \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{P} \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{m-1} N^k \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k}. \end{aligned} \tag{2.4}$$

From 2.3 this equals

$$\sum_{k=0}^{m-1} N^k \operatorname{diag}\left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right) \tag{2.5}$$

where for $k = 0, \cdots, m-1$, define $\operatorname{diag}_k(a_1, \cdots, a_{m-k})$ the $m \times m$ matrix which equals zero everywhere except on the k^{th} super diagonal where this diagonal is filled with the numbers, $\{a_1, \cdots, a_{m-k}\}$ from the upper left to the lower right. With no subscript, it is just the diagonal matrices having the indicated entries. Thus in 4×4 matrices, $\operatorname{diag}_2(1, 2)$ would be the matrix

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

Then from 2.5 and 2.2,

$$\sum_{n=0}^{P} a_n J_m(\lambda)^n = \sum_{k=0}^{m-1} \operatorname{diag}_k \left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right).$$

Therefore, $\sum_{n=0}^{P} a_n J_m(\lambda)^n =$

$$\begin{pmatrix} f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \frac{f_P^{(2)}(\lambda)}{2!} & \cdots & \frac{f_P^{(m-1)}(\lambda)}{(m-1)!} \\ & f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \ddots & \vdots \\ & & f_P(\lambda) & \ddots & \frac{f_P^{(2)}(\lambda)}{2!} \\ & & & \ddots & \frac{f_P'(\lambda)}{1!} \\ 0 & & & & f_P(\lambda) \end{pmatrix} \tag{2.6}$$

Now let A be an $n \times n$ matrix with $\rho(A) < R$ where R is given above. Then the Jordan form of A is of the form

$$J = \begin{pmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_r \end{pmatrix} \tag{2.7}$$

where $J_k = J_{m_k}(\lambda_k)$ is an $m_k \times m_k$ Jordan block and $A = S^{-1}JS$. Then, letting $P > m_k$ for all k,

$$\sum_{n=0}^{P} a_n A^n = S^{-1} \sum_{n=0}^{P} a_n J^n S,$$

and because of block multiplication of matrices,

$$\sum_{n=0}^{P} a_n J^n = \begin{pmatrix} \sum_{n=0}^{P} a_n J_1^n & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \sum_{n=0}^{P} a_n J_r^n \end{pmatrix}$$

and from 2.6 $\sum_{n=0}^{P} a_n J_k^n$ converges as $P \to \infty$ to the $m_k \times m_k$ matrix

$$\begin{pmatrix} f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \frac{f^{(2)}(\lambda_k)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \ddots & \vdots \\ 0 & 0 & f(\lambda_k) & \ddots & \frac{f^{(2)}(\lambda_k)}{2!} \\ \vdots & & \ddots & \ddots & \frac{f'(\lambda_k)}{1!} \\ 0 & 0 & \cdots & 0 & f(\lambda_k) \end{pmatrix} \tag{2.8}$$

There is no convergence problem because $|\lambda| < R$ for all $\lambda \in \sigma(A)$. This has proved the following theorem.

Theorem B.0.10 *Let f be given by 2.1 and suppose $\rho(A) < R$ where R is the radius of convergence of the power series in 2.1. Then the series,*

$$\sum_{k=0}^{\infty} a_n A^n \tag{2.9}$$

converges in the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ with respect to any of the norms on this space and furthermore,

$$\sum_{k=0}^{\infty} a_n A^n = S^{-1} \begin{pmatrix} \sum_{n=0}^{\infty} a_n J_1^n & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \sum_{n=0}^{\infty} a_n J_r^n \end{pmatrix} S$$

where $\sum_{n=0}^{\infty} a_n J_k^n$ is an $m_k \times m_k$ matrix of the form given in 2.8 where $A = S^{-1}JS$ and the Jordan form of A, J is given by 2.7. Therefore, you can define $f(A)$ by the series in 2.9.

Here is a simple example.

Example B.0.11 *Find* $\sin(A)$ *where* $A = \begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4 \end{pmatrix}$.

In this case, the Jordan canonical form of the matrix is not too hard to find.

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

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

Then from the above theorem $\sin(J)$ is given by

$$\sin \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix}.$$

Therefore, $\sin(A) =$

$$\begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix} \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = M$$

where the columns of M are as follows from left to right,

$$\begin{pmatrix} \sin 4 \\ \frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 \\ 0 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 \end{pmatrix}, \begin{pmatrix} \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 + \frac{3}{2}\sin 2 - 2\cos 2 \\ -\cos 2 \\ -\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}, \begin{pmatrix} -\cos 2 \\ \sin 2 \\ \sin 2 - \cos 2 \\ \cos 2 - \sin 2 \end{pmatrix}$$

$$\begin{pmatrix} \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 - 2\cos 2 \\ -\cos 2 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}.$$

Perhaps this isn't the first thing you would think of. Of course the ability to get this nice closed form description of $\sin(A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

The following corollary is known as the spectral mapping theorem.

Corollary B.0.12 *Let A be an $n \times n$ matrix and let $\rho(A) < R$ where for $|\lambda| < R$,*

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n.$$

Then $f(A)$ is also an $n \times n$ matrix and furthermore, $\sigma(f(A)) = f(\sigma(A))$. Thus the eigenvalues of $f(A)$ are exactly the numbers $f(\lambda)$ where λ is an eigenvalue of A. Furthermore, the algebraic multiplicity of $f(\lambda)$ coincides with the algebraic multiplicity of λ.

All of these things can be generalized to linear transformations defined on infinite dimensional spaces and when this is done the main tool is the Dunford integral along with the methods of complex analysis. It is good to see it done for finite dimensional situations first because it gives an idea of what is possible. Actually, some of the most interesting functions in applications do not come in the above form as a power series expanded about 0. One example of this situation has already been encountered in the proof of the right polar decomposition with the square root of an Hermitian transformation which had all nonnegative eigenvalues. Another example is that of taking the positive part of an Hermitian matrix. This is important in some physical models where something may depend on the positive part of the strain which is a symmetric real matrix. Obviously there is no way to consider this as a power series expanded about 0 because the function $f(r) = r^+$ is not even differentiable at 0. Therefore, a totally different approach must be considered. First the notion of a positive part is defined.

Definition B.0.13 *Let A be an Hermitian matrix. Thus it suffices to consider A as an element of $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ according to the usual notion of matrix multiplication. Then there exists an orthonormal basis of eigenvectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ such that*

$$A = \sum_{j=1}^{n} \lambda_j \mathbf{u}_j \otimes \mathbf{u}_j,$$

for λ_j the eigenvalues of A, all real. Define

$$A^+ \equiv \sum_{j=1}^{n} \lambda_j^+ \mathbf{u}_j \otimes \mathbf{u}_j$$

where $\lambda^+ \equiv \frac{|\lambda|+\lambda}{2}$.

This gives us a nice definition of what is meant but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix A. The following addresses this question.

Theorem B.0.14 *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.

$$\left|A^{+} - B^{+}\right|^2 = \operatorname{trace}\left(\sum_i \lambda_i^{+} \mathbf{v}_i \otimes \mathbf{v}_i - \sum_j \mu_j^{+} \mathbf{w}_j \otimes \mathbf{w}_j\right)^2 =$$

$$\operatorname{trace}\left[\sum_i \left(\lambda_i^{+}\right)^2 \mathbf{v}_i \otimes \mathbf{v}_i + \sum_j \left(\mu_j^{+}\right)^2 \mathbf{w}_j \otimes \mathbf{w}_j\right.$$

$$\left.-\sum_{i,j} \lambda_i^{+} \mu_j^{+} \left(\mathbf{w}_j, \mathbf{v}_i\right) \mathbf{v}_i \otimes \mathbf{w}_j - \sum_{i,j} \lambda_i^{+} \mu_j^{+} \left(\mathbf{v}_i, \mathbf{w}_j\right) \mathbf{w}_j \otimes \mathbf{v}_i\right]$$

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,

$$= \sum_i \left(\lambda_i^+\right)^2 + \sum_j \left(\mu_j^+\right)^2 - 2\sum_{i,j} \lambda_i^+ \mu_j^+ \left|(\mathbf{v}_i, \mathbf{w}_j)\right|^2 . \tag{2.10}$$

Since these are orthonormal bases,

$$\sum_i \left|(\mathbf{v}_i, \mathbf{w}_j)\right|^2 = 1 = \sum_j \left|(\mathbf{v}_i, \mathbf{w}_j)\right|^2$$

and so 2.10 equals

$$= \sum_i \sum_j \left(\left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+ \mu_j^+\right) \left|(\mathbf{v}_i, \mathbf{w}_j)\right|^2 .$$

Similarly,

$$|A - B|^2 = \sum_i \sum_j \left((\lambda_i)^2 + \left(\mu_j\right)^2 - 2\lambda_i \mu_j\right) \left|(\mathbf{v}_i, \mathbf{w}_j)\right|^2 .$$

Now it is easy to check that $(\lambda_i)^2 + \left(\mu_j\right)^2 - 2\lambda_i \mu_j \geq \left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+ \mu_j^+$. ■

Differential Equations

C.1 Theory Of Ordinary Differential Equations

Here I will present fundamental existence and uniqueness theorems for initial value problems for the differential equation,

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

Suppose that $\mathbf{f} : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies the following two conditions.

$$|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{x}_1)| \leq K |\mathbf{x} - \mathbf{x}_1|, \tag{3.1}$$

$$\mathbf{f} \text{ is continuous.} \tag{3.2}$$

The first of these conditions is known as a Lipschitz condition.

Lemma C.1.1 *Suppose $\mathbf{x} : [a, b] \to \mathbb{R}^n$ is a continuous function and $c \in [a, b]$. Then $\mathbf{x}$ is a solution to the initial value problem,*

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), \ \mathbf{x}(c) = \mathbf{x}_0 \tag{3.3}$$

if and only if $\mathbf{x}$ is a solution to the integral equation,

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_c^t \mathbf{f}(s, \mathbf{x}(s))\, ds. \tag{3.4}$$

Proof: If $\mathbf{x}$ solves 3.4, then since $\mathbf{f}$ is continuous, we may apply the fundamental theorem of calculus to differentiate both sides and obtain $\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t))$. Also, letting $t = c$ on both sides, gives $\mathbf{x}(c) = \mathbf{x}_0$. Conversely, if $\mathbf{x}$ is a solution of the initial value problem, we may integrate both sides from c to t to see that $\mathbf{x}$ solves 3.4. ■

Theorem C.1.2 *Let* $\mathbf{f}$ *satisfy 3.1 and 3.2. Then there exists a unique solution to the initial value problem, 3.3 on the interval* $[a, b]$.

Proof: Let $||\mathbf{x}||_\lambda \equiv \sup\left\{e^{\lambda t}|\mathbf{x}(t)| : t \in [a, b]\right\}$. Then this norm is equivalent to the usual norm on $BC([a, b], \mathbb{F}^n)$ described in Example 14.6.2. This means that for $||\cdot||$ the norm given there, there exist constants δ and Δ such that

$$||\mathbf{x}||_\lambda \delta \leq ||\mathbf{x}|| \leq \Delta ||\mathbf{x}||$$

for all $\mathbf{x} \in BC([a,b], \mathbb{F}^n)$. In fact, you can take $\delta \equiv e^{\lambda a}$ and $\Delta \equiv e^{\lambda b}$ in case $\lambda > 0$ with the two reversed in case $\lambda < 0$. Thus $BC([a,b], \mathbb{F}^n)$ is a Banach space with this norm, $||\cdot||_\lambda$. Then let $F : BC([a,b], \mathbb{F}^n) \to BC([a,b], \mathbb{F}^n)$ be defined by

$$F\mathbf{x}(t) \equiv \mathbf{x}_0 + \int_c^t \mathbf{f}(s, \mathbf{x}(s))\, ds.$$

Let $\lambda < 0$. It follows

$$\begin{aligned} e^{\lambda t} |F\mathbf{x}(t) - F\mathbf{y}(t)| &\leq \left| e^{\lambda t} \int_c^t |\mathbf{f}(s, \mathbf{x}(s)) - \mathbf{f}(s, \mathbf{y}(s))|\, ds \right| \\ &\leq \left| \int_c^t K e^{\lambda(t-s)} |\mathbf{x}(s) - \mathbf{y}(s)| e^{\lambda s} ds \right| \end{aligned}$$

$$\leq ||\mathbf{x} - \mathbf{y}||_\lambda \int_a^t K e^{\lambda(t-s)} ds \leq ||\mathbf{x} - \mathbf{y}||_\lambda \frac{K}{|\lambda|}$$

and therefore,

$$||F\mathbf{x} - F\mathbf{y}||_\lambda \leq ||\mathbf{x} - \mathbf{y}|| \frac{K}{|\lambda|}.$$

If $|\lambda|$ is chosen larger than K, this implies F is a contraction mapping on $BC([a, b], \mathbb{F}^n)$. Therefore, there exists a unique fixed point. With Lemma C.1.1 this proves the theorem. ■

C.2 Linear Systems

As an example of the above theorem, consider for $t \in [a, b]$ the system

$$\mathbf{x}' = A(t)\,\mathbf{x}(t) + \mathbf{g}(t), \ \mathbf{x}(c) = \mathbf{x}_0 \tag{3.5}$$

where $A(t)$ is an $n \times n$ matrix whose entries are continuous functions of $t, (a_{ij}(t))$ and $\mathbf{g}(t)$ is a vector whose components are continuous functions of t satisfies the conditions of Theorem C.1.2 with $\mathbf{f}(t, \mathbf{x}) = A(t)\,\mathbf{x} + \mathbf{g}(t)$. To see this, let $\mathbf{x} = (x_1, \cdots, x_n)^T$ and $\mathbf{x}_1 = (x_{11}, \cdots x_{1n})^T$. Then letting $M = \max\{|a_{ij}(t)| : t \in [a, b], i, j \leq n\}$,

$$|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{x}_1)| = |A(t)(\mathbf{x} - \mathbf{x}_1)|$$

$$= \left| \left(\sum_{i=1}^n \left| \sum_{j=1}^n a_{ij}(t)(x_j - x_{1j}) \right|^2 \right)^{1/2} \right| \leq M \left| \left(\sum_{i=1}^n \left(\sum_{j=1}^n |x_j - x_{1j}| \right)^2 \right)^{1/2} \right|$$

$$\leq M \left| \left(\sum_{i=1}^n n \sum_{j=1}^n |x_j - x_{1j}|^2 \right)^{1/2} \right| = Mn \left(\sum_{j=1}^n |x_j - x_{1j}|^2 \right)^{1/2} = Mn\,|\mathbf{x} - \mathbf{x}_1|.$$

Therefore, let $K = Mn$. This proves

Theorem C.2.1 *Let $A(t)$ be a continuous $n \times n$ matrix and let $\mathbf{g}(t)$ be a continuous vector for $t \in [a, b]$ and let $c \in [a, b]$ and $\mathbf{x}_0 \in \mathbb{F}^n$. Then there exists a unique solution to 3.5 valid for $t \in [a, b]$.*

This includes more examples of linear equations than are typically encountered in an entire differential equations course.

C.3 Local Solutions

Lemma C.3.1 *Let $D(\mathbf{x}_0, r) \equiv \{\mathbf{x} \in \mathbb{F}^n : |\mathbf{x} - \mathbf{x}_0| \leq r\}$ and suppose U is an open set containing $D(\mathbf{x}_0, r)$ such that $\mathbf{f} : U \to \mathbb{F}^n$ is $C^1(U)$. (Recall this means all partial derivatives of $\mathbf{f}$ exist and are continuous.) Then for $K = Mn$, where M denotes the maximum of $\left|\frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{z})\right|$ for $\mathbf{z} \in D(\mathbf{x}_0, r)$, it follows that for all $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$,*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K\,|\mathbf{x} - \mathbf{y}|.$$

Proof: Let $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$ and consider the line segment joining these two points, $\mathbf{x}+t(\mathbf{y} - \mathbf{x})$ for $t \in [0, 1]$. Letting $\mathbf{h}(t) = \mathbf{f}(\mathbf{x}+t(\mathbf{y} - \mathbf{x}))$ for $t \in [0, 1]$, then

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) = \mathbf{h}(1) - \mathbf{h}(0) = \int_0^1 \mathbf{h}'(t)\,dt.$$

Also, by the chain rule,

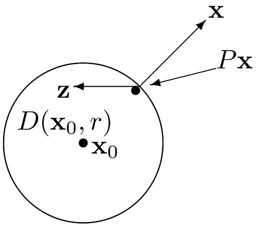
$$\mathbf{h}'(t) = \sum_{i=1}^{n} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(y_i - x_i).$$

Therefore,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| =$$

$$\begin{aligned} & \left|\int_0^1 \sum_{i=1}^{n} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(y_i - x_i)\,dt\right| \\ \leq\ & \int_0^1 \sum_{i=1}^{n} \left|\frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\right| |y_i - x_i|\,dt \\ \leq\ & M\sum_{i=1}^{n} |y_i - x_i| \leq Mn\,|\mathbf{x}-\mathbf{y}|. \ \blacksquare \end{aligned}$$

Now consider the map, P which maps all of $\mathbb{R}^n$ to $D(\mathbf{x}_0, r)$ given as follows. For $\mathbf{x} \in D(\mathbf{x}_0, r)$, $P\mathbf{x} = \mathbf{x}$. For $\mathbf{x} \notin D(\mathbf{x}_0, r)$, $P\mathbf{x}$ will be the closest point in $D(\mathbf{x}_0, r)$ to $\mathbf{x}$. Such a closest point exists because $D(\mathbf{x}_0, r)$ is a closed and bounded set. Taking $f(\mathbf{y}) \equiv |\mathbf{y}-\mathbf{x}|$, it follows f is a continuous function defined on $D(\mathbf{x}_0, r)$ which must achieve its minimum value by the extreme value theorem from calculus.



Lemma C.3.2 *For any pair of points,* $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, $|P\mathbf{x} - P\mathbf{y}| \leq |\mathbf{x}-\mathbf{y}|$.

Proof: The above picture suggests the geometry of what is going on. Letting $\mathbf{z} \in D(\mathbf{x}_0, r)$, it follows that for all $t \in [0, 1]$,

$$|\mathbf{x} - P\mathbf{x}|^2 \leq |\mathbf{x}-(P\mathbf{x} + t(\mathbf{z}-P\mathbf{x}))|^2$$

$$= |\mathbf{x}-P\mathbf{x}|^2 + 2t\operatorname{Re}((\mathbf{x}-P\mathbf{x})\cdot(P\mathbf{x}-\mathbf{z})) + t^2|\mathbf{z}-P\mathbf{x}|^2$$

Hence

$$2t\operatorname{Re}((\mathbf{x}-P\mathbf{x})\cdot(P\mathbf{x}-\mathbf{z})) + t^2|\mathbf{z}-P\mathbf{x}|^2 \geq 0$$

and this can only happen if

$$\operatorname{Re}((\mathbf{x}-P\mathbf{x})\cdot(P\mathbf{x}-\mathbf{z})) \geq 0.$$

Therefore,

$$\begin{aligned} \operatorname{Re}((\mathbf{x}-P\mathbf{x})\cdot(P\mathbf{x}-P\mathbf{y})) &\geq 0 \\ \operatorname{Re}((\mathbf{y}-P\mathbf{y})\cdot(P\mathbf{y}-P\mathbf{x})) &\geq 0 \end{aligned}$$

and so

$$\operatorname{Re}(\mathbf{x}-P\mathbf{x}-(\mathbf{y}-P\mathbf{y}))\cdot(P\mathbf{x}-P\mathbf{y}) \geq 0$$

which implies

$$\operatorname{Re}(\mathbf{x}-\mathbf{y})\cdot(P\mathbf{x}-P\mathbf{y})\geq|P\mathbf{x}-P\mathbf{y}|^2$$

Then using the Cauchy Schwarz inequality it follows

$$|\mathbf{x}-\mathbf{y}|\geq|P\mathbf{x}-P\mathbf{y}|\,.$$

■

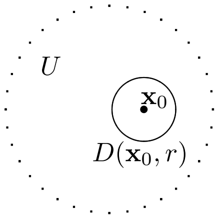
With this here is the local existence and uniqueness theorem.

Theorem C.3.3 *Let $[a,b]$ be a closed interval and let U be an open subset of $\mathbb{F}^n$. Let $\mathbf{f}:[a,b]\times U\rightarrow\mathbb{F}^n$ be continuous and suppose that for each $t\in[a,b]$, the map $\mathbf{x}\rightarrow\frac{\partial\mathbf{f}}{\partial x_i}(t,\mathbf{x})$ is continuous. Also let $\mathbf{x}_0\in U$ and $c\in[a,b]$. Then there exists an interval, $I\subseteq[a,b]$ such that $c\in I$ and there exists a unique solution to the initial value problem,*

$$\mathbf{x}'=\mathbf{f}(t,\mathbf{x})\,,\ \mathbf{x}(c)=\mathbf{x}_0 \tag{3.6}$$

valid for $t\in I$.

Proof: Consider the following picture.



The large dotted circle represents U and the little solid circle represents $D(\mathbf{x}_0, r)$ as indicated. Here r is so small that $D(\mathbf{x}_0, r)$ is contained in U as shown. Now let P denote the projection map defined above. Consider the initial value problem

$$\mathbf{x}' = \mathbf{f}(t, P\mathbf{x}),\ \mathbf{x}(c) = \mathbf{x}_0. \tag{3.7}$$

From Lemma C.3.1 and the continuity of $\mathbf{x} \to \frac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x})$, there exists a constant, K such that if $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$, then $|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})| \leq K|\mathbf{x} - \mathbf{y}|$ for all $t \in [a, b]$. Therefore, by Lemma C.3.2

$$|\mathbf{f}(t, P\mathbf{x}) - \mathbf{f}(t, P\mathbf{y})| \leq K|P\mathbf{x}-P\mathbf{y}| \leq K|\mathbf{x} - \mathbf{y}|.$$

It follows from Theorem C.1.2 that 3.7 has a unique solution valid for $t \in [a, b]$. Since $\mathbf{x}$ is continuous, it follows that there exists an interval, I containing c such that for $t \in I$, $\mathbf{x}(t) \in D(\mathbf{x}_0, r)$. Therefore, for these values of t, $\mathbf{f}(t, P\mathbf{x}) = \mathbf{f}(t, \mathbf{x})$ and so there is a unique solution to 3.6 on I. ■

Now suppose $\mathbf{f}$ has the property that for every $R > 0$ there exists a constant, K_R such that for all $\mathbf{x}, \mathbf{x}_1 \in \overline{B(0, R)}$,

$$|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{x}_1)| \leq K_R|\mathbf{x} - \mathbf{x}_1|. \tag{3.8}$$

Corollary C.3.4 *Let* $\mathbf{f}$ *satisfy 3.8 and suppose also that* $(t, \mathbf{x}) \to \mathbf{f}(t, \mathbf{x})$ *is continuous. Suppose now that* $\mathbf{x}_0$ *is given and there exists an estimate of the form* $|\mathbf{x}(t)| < R$ *for all* $t \in [0, T)$ *where* $T \leq \infty$ *on the local solution to*

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0. \tag{3.9}$$

Then there exists a unique solution to the initial value problem, 3.9 valid on $[0, T)$.

Proof: Replace $\mathbf{f}(t, \mathbf{x})$ with $\mathbf{f}(t, P\mathbf{x})$ where P is the projection onto $\overline{B(0, R)}$. Then by Theorem C.1.2 there exists a unique solution to the system

$$\mathbf{x}' = \mathbf{f}(t, P\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0$$

valid on $[0, T_1]$ for every $T_1 < T$. Therefore, the above system has a unique solution on $[0, T)$ and from the estimate, $P\mathbf{x} = \mathbf{x}$. ■

C.4 First Order Linear Systems

Here is a discussion of linear systems of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t)$$

where A is a constant $n \times n$ matrix and $\mathbf{f}$ is a vector valued function having all entries continuous. Of course the existence theory is a very special case of the general considerations above but I will give a self contained presentation based on elementary first order scalar differential equations and linear algebra.

Definition C.4.1 *Suppose $t \to M(t)$ is a matrix valued function of t. Thus $M(t) = (m_{ij}(t))$. Then define*

$$M'(t) \equiv \left(m_{ij}'(t)\right).$$

In words, the derivative of $M(t)$ is the matrix whose entries consist of the derivatives of the entries of $M(t)$. Integrals of matrices are defined the same way. Thus

$$\int_a^b M(t)\,di \equiv \left(\int_a^b m_{ij}(t)\,dt\right).$$

In words, the integral of $M(t)$ is the matrix obtained by replacing each entry of $M(t)$ by the integral of that entry.

With this definition, it is easy to prove the following theorem.

Theorem C.4.2 *Suppose $M(t)$ and $N(t)$ are matrices for which $M(t)N(t)$ makes sense. Then if $M'(t)$ and $N'(t)$ both exist, it follows that*

$$(M(t)N(t))' = M'(t)N(t) + M(t)N'(t).$$

Proof:

$$\begin{aligned}
\left((M(t)N(t))'\right)_{ij} &\equiv \left((M(t)N(t))_{ij}\right)' = \left(\sum_k M(t)_{ik} N(t)_{kj}\right)' \\
&= \sum_k \left(M(t)_{ik}\right)' N(t)_{kj} + M(t)_{ik}\left(N(t)_{kj}\right)' \\
&\equiv \sum_k \left(M(t)'\right)_{ik} N(t)_{kj} + M(t)_{ik}\left(N(t)'\right)_{kj} \\
&\equiv \left(M'(t)N(t) + M(t)N'(t)\right)_{ij} \quad \blacksquare
\end{aligned}$$

In the study of differential equations, one of the most important theorems is Gronwall's inequality which is next.

Theorem C.4.3 *Suppose $u(t) \geq 0$ and for all $t \in [0,T]$,*

$$u(t) \leq u_0 + \int_0^t K u(s)\,ds. \tag{3.10}$$

where K is some nonnegative constant. Then

$$u(t) \leq u_0 e^{Kt}. \tag{3.11}$$

Proof: Let $w(t) = \int_0^t u(s)\,ds$. Then using the fundamental theorem of calculus, 3.10 $w(t)$ satisfies the following.

$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0,\ w(0) = 0. \tag{3.12}$$

Multiply both sides of this inequality by e^{-Kt} and using the product rule and the chain rule,

$$e^{-Kt}\left(w'(t) - Kw(t)\right) = \frac{d}{dt}\left(e^{-Kt}w(t)\right) \leq u_0 e^{-Kt}.$$

Integrating this from 0 to t,

$$e^{-Kt}w(t) \leq u_0 \int_0^t e^{-Ks}ds = u_0\left(-\frac{e^{-tK}-1}{K}\right).$$

Now multiply through by e^{Kt} to obtain

$$w(t) \leq u_0 \left(-\frac{e^{-tK}-1}{K}\right) e^{Kt} = -\frac{u_0}{K} + \frac{u_0}{K} e^{tK}.$$

Therefore, 3.12 implies

$$u(t) \leq u_0 + K\left(-\frac{u_0}{K} + \frac{u_0}{K} e^{tK}\right) = u_0 e^{Kt}.$$

■

With Gronwall's inequality, here is a theorem on uniqueness of solutions to the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t), \ \mathbf{x}(a) = \mathbf{x}_a, \tag{3.13}$$

in which A is an $n \times n$ matrix and $\mathbf{f}$ is a continuous function having values in $\mathbb{C}^n$.

Theorem C.4.4 *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy 3.13. Then* $\mathbf{x}(t) = \mathbf{y}(t)$ *for all* t.

Proof: Let $\mathbf{z}(t) = \mathbf{x}(t+a) - \mathbf{y}(t+a)$. Then for $t \geq 0$,

$$\mathbf{z}' = A\mathbf{z}, \ \mathbf{z}(0) = \mathbf{0}. \tag{3.14}$$

Note that for $K = \max\{|a_{ij}|\}$, where $A = (a_{ij})$,

$$|(A\mathbf{z},\mathbf{z})| = \left|\sum_{ij} a_{ij} z_j \overline{z_i}\right| \leq K \sum_{ij} |z_i| |z_j| \leq K \sum_{ij} \left(\frac{|z_i|^2}{2} + \frac{|z_j|^2}{2}\right) = nK |\mathbf{z}|^2.$$

(For x and y real numbers, $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ because this is equivalent to saying $(x-y)^2 \geq 0$.) Similarly, $|(\mathbf{z},A\mathbf{z})| \leq nK|\mathbf{z}|^2$ .Thus,

$$|(\mathbf{z},A\mathbf{z})|, |(A\mathbf{z},\mathbf{z})| \leq nK |\mathbf{z}|^2. \tag{3.15}$$

Now multiplying 3.14 by $\mathbf{z}$ and observing that

$$\frac{d}{dt}\left(|\mathbf{z}|^2\right) = (\mathbf{z}',\mathbf{z}) + (\mathbf{z},\mathbf{z}') = (A\mathbf{z},\mathbf{z}) + (\mathbf{z},A\mathbf{z}),$$

it follows from 3.15 and the observation that $\mathbf{z}(0) = 0$,

$$|\mathbf{z}(t)|^2 \leq \int_0^t 2nK |\mathbf{z}(s)|^2 \, ds$$

and so by Gronwall's inequality, $|\mathbf{z}(t)|^2 = 0$ for all $t \geq 0$. Thus,

$$\mathbf{x}(t) = \mathbf{y}(t)$$

for all $t \geq a$.

Now let $\mathbf{w}(t) = \mathbf{x}(a-t) - \mathbf{y}(a-t)$ for $t \geq 0$. Then $\mathbf{w}'(t) = (-A)\mathbf{w}(t)$ and you can repeat the argument which was just given to conclude that $\mathbf{x}(t) = \mathbf{y}(t)$ for all $t \leq a$. ■

Definition C.4.5 *Let A be an $n \times n$ matrix. We say $\Phi(t)$ is a fundamental matrix for A if*

$$\Phi'(t) = A\Phi(t), \ \Phi(0) = I, \tag{3.16}$$

and $\Phi(t)^{-1}$ exists for all $t \in \mathbb{R}$.

Why should anyone care about a fundamental matrix? The reason is that such a matrix valued function makes possible a convenient description of the solution of the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t), \ \mathbf{x}(0) = \mathbf{x}_0, \tag{3.17}$$

on the interval, $[0,T]$. First consider the special case where $n=1$. This is the first order linear differential equation,

$$r' = \lambda r + g,\ r(0) = r_0, \tag{3.18}$$

where g is a continuous scalar valued function. First consider the case where $g=0$.

Lemma C.4.6 *There exists a unique solution to the initial value problem,*

$$r' = \lambda r,\ r(0) = 1, \tag{3.19}$$

and the solution for $\lambda = a+ib$ is given by

$$r(t) = e^{at}(\cos bt + i \sin bt). \tag{3.20}$$

This solution to the initial value problem is denoted as $e^{\lambda t}$. (If λ is real, $e^{\lambda t}$ as defined here reduces to the usual exponential function so there is no contradiction between this and earlier notation seen in Calculus.)

Proof: From the uniqueness theorem presented above, Theorem C.4.4, applied to the case where $n=1$, there can be no more than one solution to the initial value problem, 3.19. Therefore, it only remains to verify 3.20 is a solution to 3.19. However, this is an easy calculus exercise. ■

Note the differential equation in 3.19 says

$$\frac{d}{dt}\left(e^{\lambda t}\right) = \lambda e^{\lambda t}. \tag{3.21}$$

With this lemma, it becomes possible to easily solve the case in which $g \neq 0$.

Theorem C.4.7 *There exists a unique solution to 3.18 and this solution is given by the formula,*

$$r(t) = e^{\lambda t} r_0 + e^{\lambda t} \int_0^t e^{-\lambda s} g(s)\, ds. \tag{3.22}$$

Proof: By the uniqueness theorem, Theorem C.4.4, there is no more than one solution. It only remains to verify that 3.22 is a solution. But $r(0) = e^{\lambda 0} r_0 + \int_0^0 e^{-\lambda s} g(s)\, ds = r_0$ and so the initial condition is satisfied. Next differentiate this expression to verify the differential equation is also satisfied. Using 3.21, the product rule and the fundamental theorem of calculus,

$$r'(t) = \lambda e^{\lambda t} r_0 + \lambda e^{\lambda t} \int_0^t e^{-\lambda s} g(s)\, ds + e^{\lambda t} e^{-\lambda t} g(t) = \lambda r(t) + g(t).\ ■$$

Now consider the question of finding a fundamental matrix for A. When this is done, it will be easy to give a formula for the general solution to 3.17 known as the variation of constants formula, arguably the most important result in differential equations.

The next theorem gives a formula for the fundamental matrix 3.16. It is known as Putzer's method [1],[21].

Theorem C.4.8 *Let A be an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Define*

$$P_k(A) \equiv \prod_{m=1}^{k} (A - \lambda_m I),\ P_0(A) \equiv I,$$

and let the scalar valued functions, $r_k(t)$ be defined as the solutions to the following initial value problem

$$\begin{pmatrix} r_0'(t) \\ r_1'(t) \\ r_2'(t) \\ \vdots \\ r_n'(t) \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_1 r_1(t) + r_0(t) \\ \lambda_2 r_2(t) + r_1(t) \\ \vdots \\ \lambda_n r_n(t) + r_{n-1}(t) \end{pmatrix}, \begin{pmatrix} r_0(0) \\ r_1(0) \\ r_2(0) \\ \vdots \\ r_n(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Note the system amounts to a list of single first order linear differential equations. Now define

$$\Phi(t) \equiv \sum_{k=0}^{n-1} r_{k+1}(t) P_k(A).$$

Then

$$\Phi'(t) = A\Phi(t), \ \Phi(0) = I. \tag{3.23}$$

Furthermore, if $\Phi(t)$ is a solution to 3.23 for all t, then it follows $\Phi(t)^{-1}$ exists for all t and $\Phi(t)$ is the unique fundamental matrix for A.

Proof: The first part of this follows from a computation. First note that by the Cayley Hamilton theorem, $P_n(A) = 0$. Now for the computation:

$$\Phi'(t) = \sum_{k=0}^{n-1} r_{k+1}'(t) P_k(A) = \sum_{k=0}^{n-1} (\lambda_{k+1} r_{k+1}(t) + r_k(t)) P_k(A) =$$

$$\sum_{k=0}^{n-1} \lambda_{k+1} r_{k+1}(t) P_k(A) + \sum_{k=0}^{n-1} r_k(t) P_k(A) = \sum_{k=0}^{n-1} (\lambda_{k+1} I - A) r_{k+1}(t) P_k(A) +$$

$$\sum_{k=0}^{n-1} r_k(t) P_k(A) + \sum_{k=0}^{n-1} A r_{k+1}(t) P_k(A)$$

$$= -\sum_{k=0}^{n-1} r_{k+1}(t) P_{k+1}(A) + \sum_{k=0}^{n-1} r_k(t) P_k(A) + A \sum_{k=0}^{n-1} r_{k+1}(t) P_k(A). \tag{3.24}$$

Now using $r_0(t) = 0$, the first term equals

$$-\sum_{k=1}^{n} r_k(t) P_k(A) = -\sum_{k=1}^{n-1} r_k(t) P_k(A) = -\sum_{k=0}^{n-1} r_k(t) P_k(A)$$

and so 3.24 reduces to

$$A \sum_{k=0}^{n-1} r_{k+1}(t) P_k(A) = A\Phi(t).$$

This shows $\Phi'(t) = A\Phi(t)$. That $\Phi(0) = 0$ follows from

$$\Phi(0) = \sum_{k=0}^{n-1} r_{k+1}(0) P_k(A) = r_1(0) P_0 = I.$$

It remains to verify that if 3.23 holds, then $\Phi(t)^{-1}$ exists for all t. To do so, consider $\mathbf{v} \neq \mathbf{0}$ and suppose for some t_0, $\Phi(t_0)\mathbf{v} = \mathbf{0}$. Let $\mathbf{x}(t) \equiv \Phi(t_0 + t)\mathbf{v}$. Then

$$\mathbf{x}'(t) = A\Phi(t_0 + t)\mathbf{v} = A\mathbf{x}(t),\ \mathbf{x}(0) = \Phi(t_0)\mathbf{v} = \mathbf{0}.$$

But also $\mathbf{z}(t) \equiv \mathbf{0}$ also satisfies

$$\mathbf{z}'(t) = A\mathbf{z}(t),\ \mathbf{z}(0) = \mathbf{0},$$

and so by the theorem on uniqueness, it must be the case that $\mathbf{z}(t) = \mathbf{x}(t)$ for all t, showing that $\Phi(t + t_0)\mathbf{v} = \mathbf{0}$ for all t, and in particular for $t = -t_0$. Therefore,

$$\Phi(-t_0 + t_0)\mathbf{v} = I\mathbf{v} = \mathbf{0}$$

and so $\mathbf{v} = \mathbf{0}$, a contradiction. It follows that $\Phi(t)$ must be one to one for all t and so, $\Phi(t)^{-1}$ exists for all t.

It only remains to verify the solution to 3.23 is unique. Suppose Ψ is another fundamental matrix solving 3.23. Then letting $\mathbf{v}$ be an arbitrary vector,

$$\mathbf{z}(t) \equiv \Phi(t)\mathbf{v},\ \mathbf{y}(t) \equiv \Psi(t)\mathbf{v}$$

both solve the initial value problem,

$$\mathbf{x}' = A\mathbf{x},\ \mathbf{x}(0) = \mathbf{v},$$

and so by the uniqueness theorem, $\mathbf{z}(t) = \mathbf{y}(t)$ for all t showing that $\Phi(t)\mathbf{v} = \Psi(t)\mathbf{v}$ for all t. Since $\mathbf{v}$ is arbitrary, this shows that $\Phi(t) = \Psi(t)$ for every t. ■

It is useful to consider the differential equations for the r_k for $k \geq 1$. As noted above, $r_0(t) = 0$ and $r_1(t) = e^{\lambda_1 t}$.

$$r_{k+1}' = \lambda_{k+1} r_{k+1} + r_k,\ r_{k+1}(0) = 0.$$

Thus

$$r_{k+1}(t) = \int_0^t e^{\lambda_{k+1}(t-s)} r_k(s)\, ds.$$

Therefore,

$$r_2(t) = \int_0^t e^{\lambda_2(t-s)} e^{\lambda_1 s} ds = \frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{-\lambda_2 + \lambda_1}$$

assuming $\lambda_1 \neq \lambda_2$.

Sometimes people define a fundamental matrix to be a matrix $\Phi(t)$ such that $\Phi'(t) = A\Phi(t)$ and $\det(\Phi(t)) \neq 0$ for all t. Thus this avoids the initial condition, $\Phi(0) = I$. The next proposition has to do with this situation.

Proposition C.4.9 *Suppose A is an $n \times n$ matrix and suppose $\Phi(t)$ is an $n \times n$ matrix for each $t \in \mathbb{R}$ with the property that*

$$\Phi'(t) = A\Phi(t). \tag{3.25}$$

Then either $\Phi(t)^{-1}$ exists for all $t \in \mathbb{R}$ or $\Phi(t)^{-1}$ fails to exist for all $t \in \mathbb{R}$.

Proof: Suppose $\Phi(0)^{-1}$ exists and 3.25 holds. Let $\Psi(t) \equiv \Phi(t)\Phi(0)^{-1}$. Then $\Psi(0) = I$ and

$$\Psi'(t) = \Phi'(t)\Phi(0)^{-1} = A\Phi(t)\Phi(0)^{-1} = A\Psi(t)$$

so by Theorem C.4.8, $\Psi(t)^{-1}$ exists for all t. Therefore, $\Phi(t)^{-1}$ also exists for all t.

Next suppose $\Phi(0)^{-1}$ does not exist. I need to show $\Phi(t)^{-1}$ does not exist for any t. Suppose then that $\Phi(t_0)^{-1}$ does exist. Then let$\Psi(t) \equiv \Phi(t_0+t)\Phi(t_0)^{-1}$. Then $\Psi(0) = I$ and $\Psi' = A\Psi$ so by Theorem C.4.8 it follows $\Psi(t)^{-1}$ exists for all t and so for all $t, \Phi(t+t_0)^{-1}$ must also exist, even for $t = -t_0$ which implies $\Phi(0)^{-1}$ exists after all. ■

The conclusion of this proposition is usually referred to as the Wronskian alternative and another way to say it is that if 3.25 holds, then either $\det(\Phi(t)) = 0$ for all t or $\det(\Phi(t))$ is never equal to 0. The Wronskian is the usual name of the function, $t \to \det(\Phi(t))$.

The following theorem gives the variation of constants formula,.

Theorem C.4.10 *Let* $\mathbf{f}$ *be continuous on* $[0,T]$ *and let* A *be an* $n\times n$ *matrix and* $\mathbf{x}_0$ *a vector in* $\mathbb{C}^n$. *Then there exists a unique solution to 3.17,* $\mathbf{x}$, *given by the variation of constants formula,*

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t)\int_0^t \Phi(s)^{-1}\mathbf{f}(s)\,ds \tag{3.26}$$

for $\Phi(t)$ *the fundamental matrix for* A. *Also,* $\Phi(t)^{-1} = \Phi(-t)$ *and* $\Phi(t+s) = \Phi(t)\Phi(s)$ *for all* t, s *and the above variation of constants formula can also be written as*

$$\begin{aligned}\mathbf{x}(t) &= \Phi(t)\mathbf{x}_0 + \int_0^t \Phi(t-s)\mathbf{f}(s)\,ds && (3.27)\\ &= \Phi(t)\mathbf{x}_0 + \int_0^t \Phi(s)\mathbf{f}(t-s)\,ds && (3.28)\end{aligned}$$

Proof: From the uniqueness theorem there is at most one solution to 3.17. Therefore, if 3.26 solves 3.17, the theorem is proved. The verification that the given formula works is identical with the verification that the scalar formula given in Theorem C.4.7 solves the initial value problem given there. $\Phi(s)^{-1}$ is continuous because of the formula for the inverse of a matrix in terms of the transpose of the cofactor matrix. Therefore, the integrand in 3.26 is continuous and the fundamental theorem of calculus applies. To verify the formula for the inverse, fix s and consider $\mathbf{x}(t) = \Phi(s+t)\mathbf{v}$, and $\mathbf{y}(t) = \Phi(t)\Phi(s)\mathbf{v}$. Then

$$\mathbf{x}'(t) = A\Phi(t+s)\mathbf{v} = A\mathbf{x}(t),\ \mathbf{x}(0) = \Phi(s)\mathbf{v}$$

$$\mathbf{y}'(t) = A\Phi(t)\Phi(s)\mathbf{v} = A\mathbf{y}(t),\ \mathbf{y}(0) = \Phi(s)\mathbf{v}.$$

By the uniqueness theorem, $\mathbf{x}(t) = \mathbf{y}(t)$ for all t. Since s and $\mathbf{v}$ are arbitrary, this shows $\Phi(t+s) = \Phi(t)\Phi(s)$ for all t, s. Letting $s = -t$ and using $\Phi(0) = I$ verifies $\Phi(t)^{-1} = \Phi(-t)$.

Next, note that this also implies $\Phi(t-s)\Phi(s) = \Phi(t)$ and so $\Phi(t-s) = \Phi(t)\Phi(s)^{-1}$. Therefore, this yields 3.27 and then 3.28follows from changing the variable. ■

If $\Phi' = A\Phi$ and $\Phi(t)^{-1}$ exists for all t, you should verify that the solution to the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f},\ \mathbf{x}(t_0) = \mathbf{x}_0$$

is given by

$$\mathbf{x}(t) = \Phi(t-t_0)\mathbf{x}_0 + \int_{t_0}^t \Phi(t-s)\mathbf{f}(s)\,ds.$$

Theorem C.4.10 is general enough to include all constant coefficient linear differential equations or any order. Thus it includes as a special case the main topics of an entire elementary differential equations class. This is illustrated in the following example. One can reduce an arbitrary linear differential equation to a first order system and then apply the above theory to solve the problem. The next example is a differential equation of damped vibration.

Example C.4.11 *The differential equation is $y'' + 2y' + 2y = \cos t$ and initial conditions, $y(0) = 1$ and $y'(0) = 0$.*

To solve this equation, let $x_1 = y$ and $x_2 = x_1' = y'$. Then, writing this in terms of these new variables, yields the following system.

$$\begin{gathered} x_2' + 2x_2 + 2x_1 = \cos t \\ x_1' = x_2 \end{gathered}$$

This system can be written in the above form as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{pmatrix} x_2 \\ -2x_2 - 2x_1 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix}.$$

and the initial condition is of the form

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Now $P_0(A) \equiv I$. The eigenvalues are $-1+i, -1-i$ and so

$$P_1(A) = \left(\begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} - (-1+i) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1-i & 1 \\ -2 & -1-i \end{pmatrix}.$$

Recall $r_0(t) \equiv 0$ and $r_1(t) = e^{(-1+i)t}$. Then

$$r_2' = (-1-i)\, r_2 + e^{(-1+i)t},\ r_2(0) = 0$$

and so

$$r_2(t) = \frac{e^{(-1+i)t} - e^{(-1-i)t}}{2i} = e^{-t} \sin(t)$$

Putzer's method yields the fundamental matrix as

$$\begin{aligned} \Phi(t) &= e^{(-1+i)t} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + e^{-t} \sin(t) \begin{pmatrix} 1-i & 1 \\ -2 & -1-i \end{pmatrix} \\ &= \begin{pmatrix} e^{-t}(\cos(t) + \sin(t)) & e^{-t} \sin t \\ -2e^{-t} \sin t & e^{-t}(\cos(t) - \sin(t)) \end{pmatrix} \end{aligned}$$

From variation of constants formula the desired solution is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}(t) = \begin{pmatrix} e^{-t}(\cos(t) + \sin(t)) & e^{-t} \sin t \\ -2e^{-t} \sin t & e^{-t}(\cos(t) - \sin(t)) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$+ \int_0^t \begin{pmatrix} e^{-s}(\cos(s) + \sin(s)) & e^{-s} \sin s \\ -2e^{-s} \sin s & e^{-s}(\cos(s) - \sin(s)) \end{pmatrix} \begin{pmatrix} 0 \\ \cos(t-s) \end{pmatrix}$$

$$= \begin{pmatrix} e^{-t}(\cos(t) + \sin(t)) \\ -2e^{-t} \sin t \end{pmatrix} + \int_0^t \begin{pmatrix} e^{-s} \sin(s) \cos(t-s) \\ e^{-s}(\cos s - \sin s) \cos(t-s) \end{pmatrix} ds$$

$$= \begin{pmatrix} e^{-t}(\cos(t) + \sin(t)) \\ -2e^{-t} \sin t \end{pmatrix} + \begin{pmatrix} -\frac{1}{5}(\cos t)\, e^{-t} - \frac{3}{5} e^{-t} \sin t + \frac{1}{5} \cos t + \frac{2}{5} \sin t \\ -\frac{2}{5}(\cos t)\, e^{-t} + \frac{4}{5} e^{-t} \sin t + \frac{2}{5} \cos t - \frac{1}{5} \sin t \end{pmatrix}$$

$$= \begin{pmatrix} \frac{4}{5}(\cos t)\, e^{-t} + \frac{2}{5} e^{-t} \sin t + \frac{1}{5} \cos t + \frac{2}{5} \sin t \\ -\frac{6}{5} e^{-t} \sin t - \frac{2}{5}(\cos t)\, e^{-t} + \frac{2}{5} \cos t - \frac{1}{5} \sin t \end{pmatrix}$$

Thus $y(t) = x_1(t) = \frac{4}{5}(\cos t)\, e^{-t} + \frac{2}{5} e^{-t} \sin t + \frac{1}{5} \cos t + \frac{2}{5} \sin t$.

C.5 Geometric Theory Of Autonomous Systems

Here a sufficient condition is given for stability of a first order system. First of all, here is a fundamental estimate for the entries of a fundamental matrix.

Lemma C.5.1 *Let the functions, r_k be given in the statement of Theorem C.4.8 and suppose that A is an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Suppose that these eigenvalues are ordered such that*

$$\operatorname{Re}(\lambda_1) \leq \operatorname{Re}(\lambda_2) \leq \cdots \leq \operatorname{Re}(\lambda_n) < 0.$$

Then if $0 > -\delta > \operatorname{Re}(\lambda_n)$ is given, there exists a constant, C such that for each $k = 0, 1, \cdots, n$,

$$|r_k(t)| \leq Ce^{-\delta t} \tag{3.29}$$

for all $t > 0$.

Proof: This is obvious for $r_0(t)$ because it is identically equal to 0. From the definition of the $r_k, r_1' = \lambda_1 r_1, r_1(0) = 1$ and so $r_1(t) = e^{\lambda_1 t}$ which implies

$$|r_1(t)| \leq e^{\operatorname{Re}(\lambda_1)t}.$$

Suppose for some $m \geq 1$ there exists a constant, C_m such that

$$|r_k(t)| \leq C_m t^m e^{\operatorname{Re}(\lambda_m)t}$$

for all $k \leq m$ for all $t > 0$. Then

$$r_{m+1}'(t) = \lambda_{m+1} r_{m+1}(t) + r_m(t),\ r_{m+1}(0) = 0$$

and so

$$r_{m+1}(t) = e^{\lambda_{m+1}t} \int_0^t e^{-\lambda_{m+1}s} r_m(s)\, ds.$$

Then by the induction hypothesis,

$$\begin{aligned} |r_{m+1}(t)| &\leq e^{\operatorname{Re}(\lambda_{m+1})t} \int_0^t \left|e^{-\lambda_{m+1}s}\right| C_m s^m e^{\operatorname{Re}(\lambda_m)s} ds \\ &\leq e^{\operatorname{Re}(\lambda_{m+1})t} \int_0^t s^m C_m e^{-\operatorname{Re}(\lambda_{m+1})s} e^{\operatorname{Re}(\lambda_m)s} ds \\ &\leq e^{\operatorname{Re}(\lambda_{m+1})t} \int_0^t s^m C_m ds = \frac{C_m}{m+1} t^{m+1} e^{\operatorname{Re}(\lambda_{m+1})t} \end{aligned}$$

It follows by induction there exists a constant, C such that for all $k \leq n$,

$$|r_k(t)| \leq Ct^n e^{\operatorname{Re}(\lambda_n)t}$$

and this obviously implies the conclusion of the lemma.

The proof of the above lemma yields the following corollary.

Corollary C.5.2 *Let the functions, r_k be given in the statement of Theorem C.4.8 and suppose that A is an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Suppose that these eigenvalues are ordered such that*

$$\operatorname{Re}(\lambda_1) \leq \operatorname{Re}(\lambda_2) \leq \cdots \leq \operatorname{Re}(\lambda_n).$$

Then there exists a constant C such that for all $k \leq m$

$$|r_k(t)| \leq Ct^m e^{\operatorname{Re}(\lambda_m)t}.$$

With the lemma, the following sloppy estimate is available for a fundamental matrix.

Theorem C.5.3 *Let A be an $n \times n$ matrix and let $\Phi(t)$ be the fundamental matrix for A. That is,*

$$\Phi'(t) = A\Phi(t), \ \Phi(0) = I.$$

Suppose also the eigenvalues of A are $\{\lambda_1, \cdots, \lambda_n\}$ where these eigenvalues are ordered such that

$$\operatorname{Re}(\lambda_1) \leq \operatorname{Re}(\lambda_2) \leq \cdots \leq \operatorname{Re}(\lambda_n) < 0.$$

Then if $0 > -\delta > \operatorname{Re}(\lambda_n)$, is given, there exists a constant, C such that $\left|\Phi(t)_{ij}\right| \leq Ce^{-\delta t}$ for all $t > 0$. Also

$$|\Phi(t)\mathbf{x}| \leq Cn^{3/2}e^{-\delta t}|\mathbf{x}|. \tag{3.30}$$

Proof: Let

$$M \equiv \max\left\{\left|P_k(A)_{ij}\right| \text{ for all } i, j, k\right\}.$$

Then from Putzer's formula for $\Phi(t)$ and Lemma C.5.1, there exists a constant, C such that

$$\left|\Phi(t)_{ij}\right| \leq \sum_{k=0}^{n-1} Ce^{-\delta t}M.$$

Let the new C be given by nCM. ■

Next,

$$|\Phi(t)\mathbf{x}|^2 \equiv \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\Phi_{ij}(t)x_j\right)^2 \leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}|\Phi_{ij}(t)||x_j|\right)^2$$

$$\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}Ce^{-\delta t}|\mathbf{x}|\right)^2 = C^2e^{-2\delta t}\sum_{i=1}^{n}(n|\mathbf{x}|)^2 = C^2e^{-2\delta t}n^3|\mathbf{x}|^2$$

This proves 3.30 and completes the proof.

Definition C.5.4 *Let $\mathbf{f} : U \to \mathbb{R}^n$ where U is an open subset of $\mathbb{R}^n$ such that $\mathbf{a} \in U$ and $\mathbf{f}(\mathbf{a}) = \mathbf{0}$. A point, $\mathbf{a}$ where $\mathbf{f}(\mathbf{a}) = \mathbf{0}$ is called an equilibrium point. Then $\mathbf{a}$ is asymptotically stable if for any $\varepsilon > 0$ there exists $r > 0$ such that whenever $|\mathbf{x}_0 - \mathbf{a}| < r$ and $\mathbf{x}(t)$ the solution to the initial value problem,*

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}), \ \mathbf{x}(0) = \mathbf{x}_0,$$

it follows

$$\lim_{t\to\infty}\mathbf{x}(t) = \mathbf{a}, \ |\mathbf{x}(t) - \mathbf{a}| < \varepsilon$$

A differential equation of the form $\mathbf{x}' = \mathbf{f}(\mathbf{x})$ is called autonomous as opposed to a nonautonomous equation of the form $\mathbf{x}' = \mathbf{f}(t, \mathbf{x})$. The equilibrium point $\mathbf{a}$ is stable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\mathbf{x}_0 - \mathbf{a}| < \delta$, then if $\mathbf{x}$ is the solution of

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}), \ \mathbf{x}(0) = \mathbf{x}_0, \tag{3.31}$$

then $|\mathbf{x}(t) - \mathbf{a}| < \varepsilon$ for all $t > 0$.

Obviously asymptotic stability implies stability.

An ordinary differential equation is called almost linear if it is of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x})$$

where A is an $n \times n$ matrix and

$$\lim_{\mathbf{x} \to \mathbf{0}} \frac{\mathbf{g}(\mathbf{x})}{|\mathbf{x}|} = \mathbf{0}.$$

Now the stability of an equilibrium point of an autonomous system, $\mathbf{x}' = \mathbf{f}(\mathbf{x})$ can always be reduced to the consideration of the stability of $\mathbf{0}$ for an almost linear system. Here is why. If you are considering the equilibrium point, $\mathbf{a}$ for $\mathbf{x}' = \mathbf{f}(\mathbf{x})$, you could define a new variable, $\mathbf{y}$ by $\mathbf{a} + \mathbf{y} = \mathbf{x}$. Then asymptotic stability would involve $|\mathbf{y}(t)| < \varepsilon$ and $\lim_{t\to\infty} \mathbf{y}(t) = \mathbf{0}$ while stability would only require $|\mathbf{y}(t)| < \varepsilon$. Then since $\mathbf{a}$ is an equilibrium point, $\mathbf{y}$ solves the following initial value problem.

$$\mathbf{y}' = \mathbf{f}(\mathbf{a} + \mathbf{y}) - \mathbf{f}(\mathbf{a}),\ \mathbf{y}(0) = \mathbf{y}_0,$$

where $\mathbf{y}_0 = \mathbf{x}_0 - \mathbf{a}$.

Let $A = D\mathbf{f}(\mathbf{a})$. Then from the definition of the derivative of a function,

$$\mathbf{y}' = A\mathbf{y} + \mathbf{g}(\mathbf{y}),\ \mathbf{y}(0) = \mathbf{y}_0 \tag{3.32}$$

where

$$\lim_{\mathbf{y} \to \mathbf{0}} \frac{\mathbf{g}(\mathbf{y})}{|\mathbf{y}|} = \mathbf{0}.$$

Thus there is never any loss of generality in considering only the equilibrium point $\mathbf{0}$ for an almost linear system.[1] Therefore, from now on I will only consider the case of almost linear systems and the equilibrium point $\mathbf{0}$.

Theorem C.5.5 *Consider the almost linear system of equations,*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{3.33}$$

where

$$\lim_{\mathbf{x} \to \mathbf{0}} \frac{\mathbf{g}(\mathbf{x})}{|\mathbf{x}|} = \mathbf{0}$$

and $\mathbf{g}$ *is a* C^1 *function. Suppose that for all* λ *an eigenvalue of* A, $\operatorname{Re}\lambda < 0$. *Then* $\mathbf{0}$ *is asymptotically stable.*

Proof: By Theorem C.5.3 there exist constants $\delta > 0$ and K such that for $\Phi(t)$ the fundamental matrix for A,

$$|\Phi(t)\mathbf{x}| \leq Ke^{-\delta t}|\mathbf{x}|.$$

Let $\varepsilon > 0$ be given and let r be small enough that $Kr < \varepsilon$ and for $|\mathbf{x}| < (K+1)r$, $|\mathbf{g}(\mathbf{x})| < \eta|\mathbf{x}|$ where η is so small that $K\eta < \delta$, and let $|\mathbf{y}_0| < r$. Then by the variation of constants formula, the solution to 3.33, at least for small t satisfies

$$\mathbf{y}(t) = \Phi(t)\mathbf{y}_0 + \int_0^t \Phi(t-s)\,\mathbf{g}(\mathbf{y}(s))\,ds.$$

[1] This is no longer true when you study partial differential equations as ordinary differential equations in infinite dimensional spaces.

The following estimate holds.

$$|\mathbf{y}(t)| \leq Ke^{-\delta t}|\mathbf{y}_0| + \int_0^t Ke^{-\delta(t-s)}\eta|\mathbf{y}(s)|\,ds < Ke^{-\delta t}r + \int_0^t Ke^{-\delta(t-s)}\eta|\mathbf{y}(s)|\,ds.$$

Therefore,

$$e^{\delta t}|\mathbf{y}(t)| < Kr + \int_0^t K\eta e^{\delta s}|\mathbf{y}(s)|\,ds.$$

By Gronwall's inequality,

$$e^{\delta t}|\mathbf{y}(t)| < Kre^{K\eta t}$$

and so

$$|\mathbf{y}(t)| < Kre^{(K\eta-\delta)t} < \varepsilon e^{(K\eta-\delta)t}$$

Therefore, $|\mathbf{y}(t)| < Kr < \varepsilon$ for all t and so from Corollary C.3.4, the solution to 3.33 exists for all $t \geq 0$ and since $K\eta - \delta < 0$,

$$\lim_{t\to\infty}|\mathbf{y}(t)| = 0.\blacksquare$$

C.6 General Geometric Theory

Here I will consider the case where the matrix A has both positive and negative eigenvalues. First here is a useful lemma.

Lemma C.6.1 *Suppose A is an $n \times n$ matrix and there exists $\delta > 0$ such that*

$$0 < \delta < \operatorname{Re}(\lambda_1) \leq \cdots \leq \operatorname{Re}(\lambda_n)$$

where $\{\lambda_1, \cdots, \lambda_n\}$ are the eigenvalues of A, with possibly some repeated. Then there exists a constant, C such that for all $t < 0$,

$$|\Phi(t)\mathbf{x}| \leq Ce^{\delta t}|\mathbf{x}|$$

Proof: I want an estimate on the solutions to the system

$$\Phi'(t) = A\Phi(t), \ \Phi(0) = I.$$

for $t < 0$. Let $s = -t$ and let $\Psi(s) = \Phi(t)$. Then writing this in terms of Ψ,

$$\Psi'(s) = -A\Psi(s), \ \Psi(0) = I.$$

Now the eigenvalues of $-A$ have real parts less than $-\delta$ because these eigenvalues are obtained from the eigenvalues of A by multiplying by -1. Then by Theorem C.5.3 there exists a constant, C such that for any $\mathbf{x}$,

$$|\Psi(s)\mathbf{x}| \leq Ce^{-\delta s}|\mathbf{x}|.$$

Therefore, from the definition of Ψ,

$$|\Phi(t)\mathbf{x}| \leq Ce^{\delta t}|\mathbf{x}|.\blacksquare$$

Here is another essential lemma which is found in Coddington and Levinson [6]

Lemma C.6.2 *Let $p_j(t)$ be polynomials with complex coefficients and let*

$$f(t) = \sum_{j=1}^{m} p_j(t) e^{\lambda_j t}$$

where $m \geq 1, \lambda_j \neq \lambda_k$ for $j \neq k$, and none of the $p_j(t)$ vanish identically. Let

$$\sigma = \max(\operatorname{Re}(\lambda_1), \cdots, \operatorname{Re}(\lambda_m)).$$

Then there exists a positive number, r and arbitrarily large positive values of t such that

$$e^{-\sigma t}|f(t)| > r.$$

In particular, $|f(t)|$ is unbounded.

Proof: Suppose the largest exponent of any of the p_j is M and let $\lambda_j = a_j + ib_j$. First assume each $a_j = 0$. This is convenient because $\sigma = 0$ in this case and the largest of the $\operatorname{Re}(\lambda_j)$ occurs in every λ_j.

Then arranging the above sum as a sum of decreasing powers of t,

$$f(t) = t^M f_M(t) + \cdots + tf_1(t) + f_0(t).$$

Then

$$t^{-M} f(t) = f_M(t) + O\left(\frac{1}{t}\right)$$

where the last term means that $tO\left(\frac{1}{t}\right)$ is bounded. Then

$$f_M(t) = \sum_{j=1}^{m} c_j e^{ib_j t}$$

It can't be the case that all the c_j are equal to 0 because then M would not be the highest power exponent. Suppose $c_k \neq 0$. Then

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T t^{-M} f(t) e^{-ib_k t} dt = \sum_{j=1}^{m} c_j \frac{1}{T}\int_0^T e^{i(b_j - b_k)t} dt = c_k \neq 0.$$

Letting $r = |c_k/2|$, it follows $\left|t^{-M} f(t) e^{-ib_k t}\right| > r$ for arbitrarily large values of t. Thus it is also true that $|f(t)| > r$ for arbitrarily large values of t.

Next consider the general case in which σ is given above. Thus

$$e^{-\sigma t} f(t) = \sum_{j:a_j=\sigma} p_j(t) e^{b_j t} + g(t)$$

where $\lim_{t\to\infty} g(t) = 0$, $g(t)$ being of the form $\sum_s p_s(t) e^{(a_s - \sigma + ib_s)t}$ where $a_s - \sigma < 0$. Then this reduces to the case above in which $\sigma = 0$. Therefore, there exists $r > 0$ such that

$$\left|e^{-\sigma t} f(t)\right| > r$$

for arbitrarily large values of t. ■

Next here is a Banach space which will be useful.

Lemma C.6.3 *For $\gamma > 0$, let*

$$E_\gamma = \left\{\mathbf{x} \in BC([0,\infty), \mathbb{F}^n) : t \to e^{\gamma t}\mathbf{x}(t) \text{ is also in } BC([0,\infty), \mathbb{F}^n)\right\}$$

and let the norm be given by

$$||\mathbf{x}||_\gamma \equiv \sup\left\{\left|e^{\gamma t}\mathbf{x}(t)\right| : t \in [0,\infty)\right\}$$

Then E_γ is a Banach space.

Proof: Let $\{\mathbf{x}_k\}$ be a Cauchy sequence in E_γ. Then since $BC([0,\infty), \mathbb{F}^n)$ is a Banach space, there exists $\mathbf{y} \in BC([0,\infty), \mathbb{F}^n)$ such that $e^{\gamma t}\mathbf{x}_k(t)$ converges uniformly on $[0,\infty)$ to $\mathbf{y}(t)$. Therefore $e^{-\gamma t}e^{\gamma t}\mathbf{x}_k(t) = \mathbf{x}_k(t)$ converges uniformly to $e^{-\gamma t}\mathbf{y}(t)$ on $[0,\infty)$. Define $\mathbf{x}(t) \equiv e^{-\gamma t}\mathbf{y}(t)$. Then $\mathbf{y}(t) = e^{\gamma t}\mathbf{x}(t)$ and by definition,

$$||\mathbf{x}_k - \mathbf{x}||_\gamma \to 0.$$

■

C.7 The Stable Manifold

Here assume

$$A = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix} \tag{3.34}$$

where A_- and A_+ are square matrices of size $k \times k$ and $(n-k) \times (n-k)$ respectively. Also assume A_- has eigenvalues whose real parts are all less than $-\alpha$ while A_+ has eigenvalues whose real parts are all larger than α. Assume also that each of A_- and A_+ is upper triangular.

Also, I will use the following convention. For $\mathbf{v} \in \mathbb{F}^n$,

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_- \\ \mathbf{v}_+ \end{pmatrix}$$

where $\mathbf{v}_-$ consists of the first k entries of $\mathbf{v}$.

Then from Theorem C.5.3 and Lemma C.6.1 the following lemma is obtained.

Lemma C.7.1 *Let A be of the form given in 3.34 as explained above and let $\Phi_+(t)$ and $\Phi_-(t)$ be the fundamental matrices corresponding to A_+ and A_- respectively. Then there exist positive constants, α and γ such that*

$$|\Phi_+(t)\,\mathbf{y}| \leq Ce^{\alpha t} \textit{ for all } t < 0 \tag{3.35}$$

$$|\Phi_-(t)\,\mathbf{y}| \leq Ce^{-(\alpha+\gamma)t} \textit{ for all } t > 0. \tag{3.36}$$

Also for any nonzero $\mathbf{x} \in \mathbb{C}^{n-k}$,

$$|\Phi_+(t)\,\mathbf{x}| \textit{ is unbounded.} \tag{3.37}$$

Proof: The first two claims have been established already. It suffices to pick α and γ such that $-(\alpha+\gamma)$ is larger than all eigenvalues of A_- and α is smaller than all eigenvalues of A_+. It remains to verify 3.37. From the Putzer formula for $\Phi_+(t)$,

$$\Phi_+(t)\,\mathbf{x} = \sum_{k=0}^{n-1} r_{k+1}(t)\,P_k(A)\,\mathbf{x}$$

where $P_0(A) \equiv I$. Now each r_k is a polynomial (possibly a constant) times an exponential. This follows easily from the definition of the r_k as solutions of the differential equations

$$r'_{k+1} = \lambda_{k+1} r_{k+1} + r_k.$$

Now by assumption the eigenvalues have positive real parts so

$$\sigma \equiv \max\left(\operatorname{Re}(\lambda_1), \cdots, \operatorname{Re}(\lambda_{n-k})\right) > 0.$$

It can also be assumed

$$\operatorname{Re}(\lambda_1) \geq \cdots \geq \operatorname{Re}(\lambda_{n-k})$$

By Lemma C.6.2 it follows $|\Phi_+(t)\mathbf{x}|$ is unbounded. This follows because

$$\Phi_+(t)\mathbf{x} = r_1(t)\mathbf{x} + \sum_{k=1}^{n-1} r_{k+1}(t)\mathbf{y}_k, \; r_1(t) = e^{\lambda_1 t}.$$

Since $\mathbf{x} \neq \mathbf{0}$, it has a nonzero entry, say $x_m \neq 0$. Consider the m^{th} entry of the vector $\Phi_+(t)\mathbf{x}$. By this Lemma the m^{th} entry is unbounded and this is all it takes for $\mathbf{x}(t)$ to be unbounded. ■

Lemma C.7.2 *Consider the initial value problem for the almost linear system*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0,$$

where $\mathbf{g}$ *is* C^1 *and* A *is of the special form*

$$A = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix}$$

in which A_- *is a* $k \times k$ *matrix which has eigenvalues for which the real parts are all negative and* A_+ *is a* $(n-k) \times (n-k)$ *matrix for which the real parts of all the eigenvalues are positive. Then* $\mathbf{0}$ *is not stable. More precisely, there exists a set of points* $(\mathbf{a}_-, \boldsymbol{\psi}(\mathbf{a}_-))$ *for* $\mathbf{a}_-$ *small such that for* $\mathbf{x}_0$ *on this set,*

$$\lim_{t\to\infty} \mathbf{x}(t, \mathbf{x}_0) = \mathbf{0}$$

and for $\mathbf{x}_0$ *not on this set, there exists a* $\delta > 0$ *such that* $|\mathbf{x}(t, \mathbf{x}_0)|$ *cannot remain less than* δ *for all positive* t.

Proof: Consider the initial value problem for the almost linear equation,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{a} = \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix}.$$

Then by the variation of constants formula, a local solution has the form

$$\begin{aligned} \mathbf{x}(t, \mathbf{a}) &= \begin{pmatrix} \Phi_-(t) & 0 \\ 0 & \Phi_+(t) \end{pmatrix} \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix} \\ &\quad + \int_0^t \begin{pmatrix} \Phi_-(t-s) & 0 \\ 0 & \Phi_+(t-s) \end{pmatrix} \mathbf{g}(\mathbf{x}(s, \mathbf{a}))\, ds \end{aligned} \tag{3.38}$$

Write $\mathbf{x}(t)$ for $\mathbf{x}(t, \mathbf{a})$ for short. Let $\varepsilon > 0$ be given and suppose δ is such that if $|\mathbf{x}| < \delta$, then $|\mathbf{g}_\pm(\mathbf{x})| < \varepsilon |\mathbf{x}|$. Assume from now on that $|\mathbf{a}| < \delta$. Then suppose $|\mathbf{x}(t)| < \delta$ for all $t > 0$. Writing 3.38 differently yields

$$\begin{aligned} \mathbf{x}(t, \mathbf{a}) &= \begin{pmatrix} \Phi_-(t) & 0 \\ 0 & \Phi_+(t) \end{pmatrix} \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix} + \begin{pmatrix} \int_0^t \Phi_-(t-s)\, \mathbf{g}_-(\mathbf{x}(s, \mathbf{a}))\, ds \\ 0 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 \\ \int_0^t \Phi_+(t-s)\, \mathbf{g}_+(\mathbf{x}(s, \mathbf{a}))\, ds \end{pmatrix} \end{aligned}$$

$$\begin{aligned} &= \begin{pmatrix} \Phi_-(t) & 0 \\ 0 & \Phi_+(t) \end{pmatrix} \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix} + \begin{pmatrix} \int_0^t \Phi_-(t-s)\, \mathbf{g}_-(\mathbf{x}(s, \mathbf{a}))\, ds \\ 0 \end{pmatrix} \\ &+ \begin{pmatrix} 0 \\ \int_0^\infty \Phi_+(t-s)\, \mathbf{g}_+(\mathbf{x}(s, \mathbf{a}))\, ds - \int_t^\infty \Phi_+(t-s)\, \mathbf{g}_+(\mathbf{x}(s, \mathbf{a}))\, ds \end{pmatrix}. \end{aligned}$$

These improper integrals converge thanks to the assumption that $\mathbf{x}$ is bounded and the estimates 3.35 and 3.36. Continuing the rewriting,

$$\begin{aligned} \begin{pmatrix} \mathbf{x}_-(t) \\ \mathbf{x}_+(t) \end{pmatrix} &= \begin{pmatrix} \left(\Phi_-(t)\, \mathbf{a}_- + \int_0^t \Phi_-(t-s)\, \mathbf{g}_-(\mathbf{x}(s, \mathbf{a}))\, ds\right) \\ \Phi_+(t) \left(\mathbf{a}_+ + \int_0^\infty \Phi_+(-s)\, \mathbf{g}_+(\mathbf{x}(s, \mathbf{a}))\, ds\right) \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 \\ -\int_t^\infty \Phi_+(t-s)\, \mathbf{g}_+(\mathbf{x}(s, \mathbf{a}))\, ds \end{pmatrix}. \end{aligned}$$

It follows from Lemma C.7.1 that if $|\mathbf{x}(t,\mathbf{a})|$ is bounded by δ as asserted, then it must be the case that $\mathbf{a}_+ + \int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds = \mathbf{0}$. Consequently, it must be the case that

$$\mathbf{x}(t) = \Phi(t)\begin{pmatrix} \mathbf{a}_- \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s,\mathbf{a}))\,ds \\ -\int_t^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds \end{pmatrix} \tag{3.39}$$

Letting $t \to 0$, this requires that for a solution to the initial value problem to exist and also satisfy $|\mathbf{x}(t)| < \delta$ for all $t > 0$ it must be the case that

$$\mathbf{x}(0) = \begin{pmatrix} \mathbf{a}_- \\ -\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds \end{pmatrix}$$

where $\mathbf{x}(t,\mathbf{a})$ is the solution of

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \begin{pmatrix} \mathbf{a}_- \\ -\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds \end{pmatrix}$$

This is because in 3.39, if $\mathbf{x}$ is bounded by δ then the reverse steps show $\mathbf{x}$ is a solution of the above differential equation and initial condition.

It follows if I can show that for all $\mathbf{a}_-$ sufficiently small and $\mathbf{a} = (\mathbf{a}_-, \mathbf{0})^T$, there exists a solution to 3.39 $\mathbf{x}(s,\mathbf{a})$ on $(0,\infty)$ for which $|\mathbf{x}(s,\mathbf{a})| < \delta$, then I can define

$$\boldsymbol{\psi}(\mathbf{a}) \equiv -\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds$$

and conclude that $|\mathbf{x}(t,\mathbf{x}_0)| < \delta$ for all $t > 0$ if and only if $\mathbf{x}_0 = (\mathbf{a}_-, \boldsymbol{\psi}(\mathbf{a}_-))^T$ for some sufficiently small $\mathbf{a}_-$.

Let C, α, γ be the constants of Lemma C.7.1. Let η be a small positive number such that

$$\frac{C\eta}{\alpha} < \frac{1}{6}$$

Note that $\frac{\partial \mathbf{g}}{\partial x_i}(0) = \mathbf{0}$. Therefore, by Lemma C.3.1, there exists $\delta > 0$ such that if $|\mathbf{x}|, |\mathbf{y}| \le \delta$, then

$$|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \eta\,|\mathbf{x} - \mathbf{y}|$$

and in particular,

$$|\mathbf{g}_\pm(\mathbf{x}) - \mathbf{g}_\pm(\mathbf{y})| < \eta\,|\mathbf{x} - \mathbf{y}| \tag{3.40}$$

because each $\frac{\partial \mathbf{g}}{\partial x_i}(\mathbf{x})$ is very small. In particular, this implies

$$|\mathbf{g}_-(\mathbf{x})| < \eta\,|\mathbf{x}|,\ |\mathbf{g}_+(\mathbf{x})| < \eta\,|\mathbf{x}|.$$

For $\mathbf{x} \in E_\gamma$ defined in Lemma C.6.3 and $|\mathbf{a}_-| < \frac{\delta}{2C}$,

$$F\mathbf{x}(t) \equiv \begin{pmatrix} \Phi_-(t)\,\mathbf{a}_- + \int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s))\,ds \\ -\int_t^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s))\,ds \end{pmatrix}.$$

I need to find a fixed point of F. Letting $||\mathbf{x}||_\gamma < \delta$, and using the estimates of Lemma C.7.1,

$$\begin{aligned} e^{\gamma t}\,|F\mathbf{x}(t)| \quad &\le \quad e^{\gamma t}\,|\Phi_-(t)\,\mathbf{a}_-| + e^{\gamma t}\int_0^t Ce^{-(\alpha+\gamma)(t-s)}\eta\,|\mathbf{x}(s)|\,ds \\ &\quad +e^{\gamma t}\int_t^\infty Ce^{\alpha(t-s)}\eta\,|\mathbf{x}(s)|\,ds \end{aligned}$$

$$\begin{aligned}
&\leq e^{\gamma t}C\frac{\delta}{2C}e^{-(\alpha+\gamma)t}+e^{\gamma t}\left|\left|\mathbf{x}\right|\right|_{\gamma}C\eta\int_{0}^{t}e^{-(\alpha+\gamma)(t-s)}e^{-\gamma s}ds \\
&\quad +e^{\gamma t}C\eta\int_{t}^{\infty}e^{\alpha(t-s)}e^{-\gamma s}ds\left|\left|\mathbf{x}\right|\right|_{\gamma} \\
&< \frac{\delta}{2}+\delta C\eta\int_{0}^{t}e^{-\alpha(t-s)}ds+C\eta\delta\int_{t}^{\infty}e^{(\alpha+\gamma)(t-s)}ds \\
&< \frac{\delta}{2}+\delta C\eta\frac{1}{\alpha}+\frac{\delta C\eta}{\alpha+\gamma}\leq\delta\left(\frac{1}{2}+\frac{C\eta}{\alpha}\right)<\frac{2\delta}{3}.
\end{aligned}$$

Thus F maps every $\mathbf{x}\in E_{\gamma}$ having $\left|\left|\mathbf{x}\right|\right|_{\gamma}<\delta$ to $F\mathbf{x}$ where $\left|\left|F\mathbf{x}\right|\right|_{\gamma}\leq\frac{2\delta}{3}$.

Now let $\mathbf{x},\mathbf{y}\in E_{\gamma}$ where $\left|\left|\mathbf{x}\right|\right|_{\gamma},\left|\left|\mathbf{y}\right|\right|_{\gamma}<\delta$. Then

$$\begin{aligned}
e^{\gamma t}\left|F\mathbf{x}(t)-F\mathbf{y}(t)\right| &\leq e^{\gamma t}\int_{0}^{t}\left|\Phi_{-}(t-s)\right|\eta e^{-\gamma s}e^{\gamma s}\left|\mathbf{x}(s)-\mathbf{y}(s)\right|ds \\
&\quad +e^{\gamma t}\int_{t}^{\infty}\left|\Phi_{+}(t-s)\right|e^{-\gamma s}e^{\gamma s}\eta\left|\mathbf{x}(s)-\mathbf{y}(s)\right|ds
\end{aligned}$$

$$\leq C\eta\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_{\gamma}\left(\int_{0}^{t}e^{-\alpha(t-s)}ds\right)+\int_{t}^{\infty}e^{(\alpha+\gamma)(t-s)}ds$$

$$\leq C\eta\left(\frac{1}{\alpha}+\frac{1}{\alpha+\gamma}\right)\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_{\gamma}<\frac{2C\eta}{\alpha}\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_{\gamma}<\frac{1}{3}\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_{\gamma}.$$

It follows from Lemma 14.6.4, for each $\mathbf{a}_{-}$ such that $\left|\mathbf{a}_{-}\right|<\frac{\delta}{2C}$, there exists a unique solution to 3.39 in E_{γ}.

As pointed out earlier, if

$$\boldsymbol{\psi}(\mathbf{a})\equiv-\int_{0}^{\infty}\Phi_{+}(-s)\,\mathbf{g}_{+}(\mathbf{x}(s,\mathbf{a}))\,ds$$

then for $\mathbf{x}(t,\mathbf{x}_0)$ the solution to the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}), \ \mathbf{x}(0) = \mathbf{x}_0$$

has the property that if $\mathbf{x}_0$ is not of the form $\begin{pmatrix} \mathbf{a}_- \\ \boldsymbol{\psi}(\mathbf{a}_-) \end{pmatrix}$, then $|\mathbf{x}(t,\mathbf{x}_0)|$ cannot be less than δ for all $t > 0$.

On the other hand, if $\mathbf{x}_0 = \begin{pmatrix} \mathbf{a}_- \\ \boldsymbol{\psi}(\mathbf{a}_-) \end{pmatrix}$ for $|\mathbf{a}_-| < \frac{\delta}{2C}$, then $\mathbf{x}(t,\mathbf{x}_0)$,the solution to 3.39 is the unique solution to the initial value problem

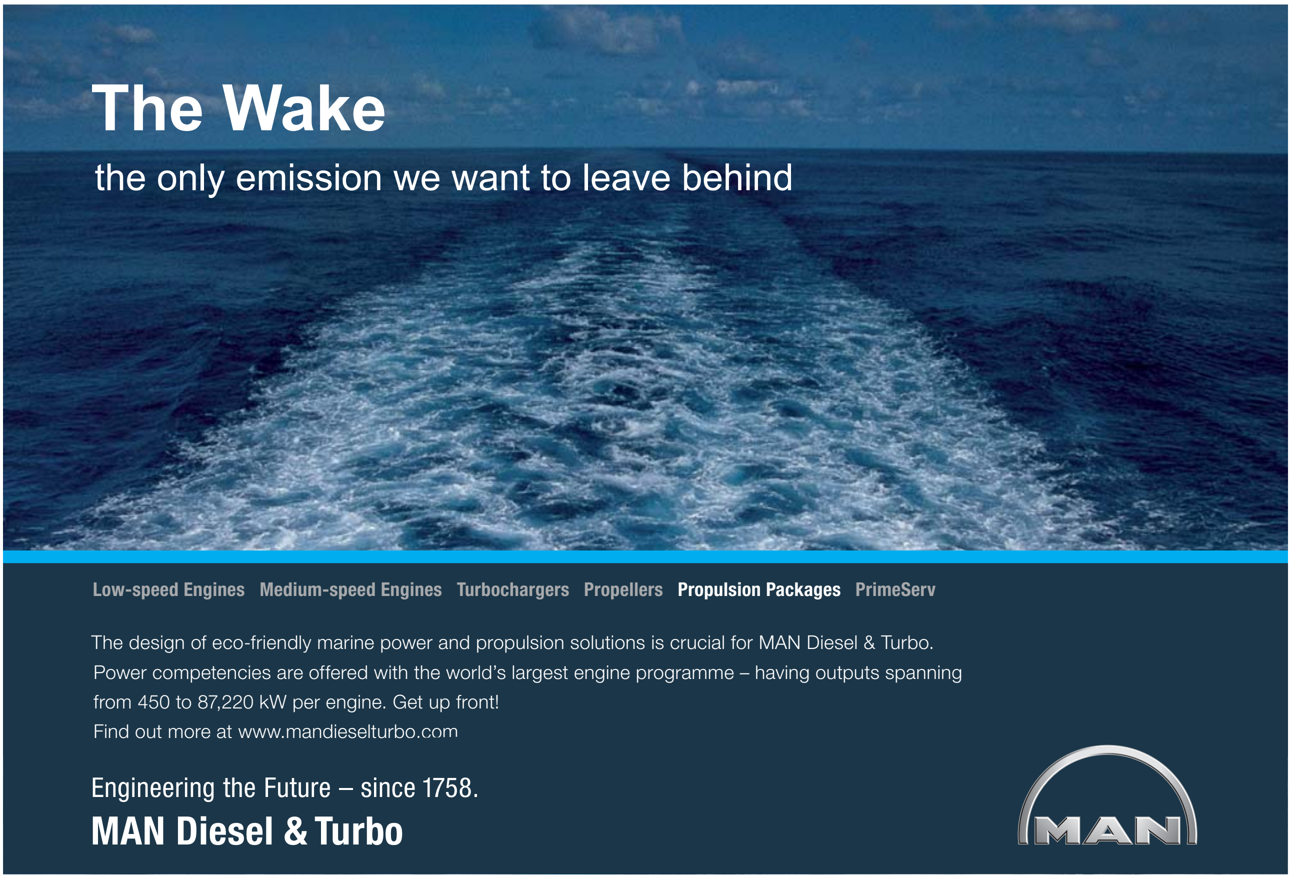
$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}), \ \mathbf{x}(0) = \mathbf{x}_0.$$

and it was shown that $||\mathbf{x}(\cdot,\mathbf{x}_0)||_\gamma < \delta$ and so in fact,

$$|\mathbf{x}(t,\mathbf{x}_0)| \leq \delta e^{-\gamma t}$$

showing that

$$\lim_{t\to\infty} \mathbf{x}(t,\mathbf{x}_0) = \mathbf{0}.$$



■

The following theorem is the main result. It involves a use of linear algebra and the above lemma.

Theorem C.7.3 *Consider the initial value problem for the almost linear system*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}), \ \mathbf{x}(0) = \mathbf{x}_0$$

in which $\mathbf{g}$ *is* C^1 *and where at there are* $k < n$ *eigenvalues of* A *which have negative real parts and* $n - k$ *eigenvalues of* A *which have positive real parts. Then* $\mathbf{0}$ *is not stable. More precisely, there exists a set of points* $(\mathbf{a}, \boldsymbol{\psi}(\mathbf{a}))$ *for* $\mathbf{a}$ *small and in a* k *dimensional subspace such that for* $\mathbf{x}_0$ *on this set,*

$$\lim_{t\to\infty} \mathbf{x}(t, \mathbf{x}_0) = \mathbf{0}$$

and for $\mathbf{x}_0$ *not on this set, there exists a* $\delta > 0$ *such that* $|\mathbf{x}(t, \mathbf{x}_0)|$ *cannot remain less than* δ *for all positive* t.

Proof: This involves nothing more than a reduction to the situation of Lemma C.7.2. From Theorem 10.5.2 on Page 10.5.2 A is similar to a matrix of the form described in Lemma C.7.2. Thus $A = S^{-1}\begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix} S$. Letting $\mathbf{y} = S\mathbf{x}$, it follows

$$\mathbf{y}' = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix} \mathbf{y} + \mathbf{g}\left(S^{-1}\mathbf{y}\right)$$

Now $|\mathbf{x}| = \left|S^{-1}S\mathbf{x}\right| \leq \left|\left|S^{-1}\right|\right| |\mathbf{y}|$ and $|\mathbf{y}| = \left|SS^{-1}\mathbf{y}\right| \leq ||S|| \, |\mathbf{x}|$. Therefore,

$$\frac{1}{||S||} |\mathbf{y}| \leq |\mathbf{x}| \leq \left|\left|S^{-1}\right|\right| |\mathbf{y}| .$$

It follows all conclusions of Lemma C.7.2 are valid for this theorem. ■

The set of points $(\mathbf{a}, \boldsymbol{\psi}(\mathbf{a}))$ for $\mathbf{a}$ small is called the stable manifold. Much more can be said about the stable manifold and you should look at a good differential equations book for this.

Compactness And Completeness

D.0.1 The Nested Interval Lemma

First, here is the one dimensional nested interval lemma.

Lemma D.0.4 *Let $I_k = [a_k, b_k]$ be closed intervals, $a_k \leq b_k$, such that $I_k \supseteq I_{k+1}$ for all k. Then there exists a point c which is contained in all these intervals. If $\lim_{k\to\infty} (b_k - a_k) = 0$, then there is exactly one such point.*

Proof: Note that the $\{a_k\}$ are an increasing sequence and that $\{b_k\}$ is a decreasing sequence. Now note that if $m < n$, then

$$a_m \leq a_n \leq b_n$$

while if $m > n$,

$$b_n \geq b_m \geq a_m.$$

It follows that $a_m \leq b_n$ for any pair m, n. Therefore, each b_n is an upper bound for all the a_m and so if $c \equiv \sup \{a_k\}$, then for each n, it follows that $c \leq b_n$ and so for all, $a_n \leq c \leq b_n$ which shows that c is in all of these intervals.

If the condition on the lengths of the intervals holds, then if c, c' are in all the intervals, then if they are not equal, then eventually, for large enough k, they cannot both be contained in $[a_k, b_k]$ since eventually $b_k - a_k < |c - c'|$. This would be a contradiction. Hence $c = c'$. ■

Definition D.0.5 *The **diameter** of a set S, is defined as*

$$\operatorname{diam}(S) \equiv \sup \{|\mathbf{x} - \mathbf{y}| : \mathbf{x}, \mathbf{y} \in S\}.$$

Thus $\operatorname{diam}(S)$ is just a careful description of what you would think of as the diameter. It measures how stretched out the set is.

Here is a multidimensional version of the nested interval lemma.

Lemma D.0.6 *Let $I_k = \prod_{i=1}^{p} \left[a_i^k, b_i^k\right] \equiv \left\{\mathbf{x} \in \mathbb{R}^p : x_i \in \left[a_i^k, b_i^k\right]\right\}$ and suppose that for all $k = 1, 2, \cdots$,*

$$I_k \supseteq I_{k+1}.$$

Then there exists a point $\mathbf{c} \in \mathbb{R}^p$ which is an element of every I_k. If $\lim_{k\to\infty} \operatorname{diam}(I_k) = 0$, then the point $\mathbf{c}$ is unique.

Proof: For each $i = 1, \cdots, p$, $\left[a_i^k, b_i^k\right] \supseteq \left[a_i^{k+1}, b_i^{k+1}\right]$ and so, by Lemma D.0.4, there exists a point $c_i \in \left[a_i^k, b_i^k\right]$ for all k. Then letting $\mathbf{c} \equiv (c_1, \cdots, c_p)$ it follows $\mathbf{c} \in I_k$ for all k. If the condition on the diameters holds, then the lengths of the intervals $\lim_{k\to\infty} \left[a_i^k, b_i^k\right] = 0$ and so by the same lemma, each c_i is unique. Hence $\mathbf{c}$ is unique. ■

D.0.2 Convergent Sequences, Sequential Compactness

A mapping $\mathbf{f} : \{k, k+1, k+2, \cdots\} \to \mathbb{R}^p$ is called a sequence. We usually write it in the form $\{\mathbf{a}_j\}$ where it is understood that $\mathbf{a}_j \equiv \mathbf{f}(j)$.

Definition D.0.7 *A sequence, $\{\mathbf{a}_k\}$ is said to* ***converge*** *to $\mathbf{a}$ if for every $\varepsilon > 0$ there exists n_ε such that if $n > n_\varepsilon$, then $|\mathbf{a} - \mathbf{a}_n| < \varepsilon$. The usual notation for this is $\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a}$ although it is often written as $\mathbf{a}_n \to \mathbf{a}$. A closed set $K \subseteq \mathbb{R}^n$ is one which has the property that if $\{\mathbf{k}_j\}_{j=1}^\infty$ is a sequence of points of K which converges to $\mathbf{x}$, then $\mathbf{x} \in K$.*

One can also define a subsequence.

Definition D.0.8 *$\{\mathbf{a}_{n_k}\}$ is a* ***subsequence*** *of $\{\mathbf{a}_n\}$ if $n_1 < n_2 < \cdots$.*

The following theorem says the limit, if it exists, is unique.

Theorem D.0.9 *If a sequence, $\{\mathbf{a}_n\}$ converges to $\mathbf{a}$ and to $\mathbf{b}$ then $\mathbf{a} = \mathbf{b}$.*

Proof: There exists n_ε such that if $n > n_\varepsilon$ then $|\mathbf{a}_n - \mathbf{a}| < \frac{\varepsilon}{2}$ and if $n > n_\varepsilon$, then $|\mathbf{a}_n - \mathbf{b}| < \frac{\varepsilon}{2}$. Then pick such an n.

$$|\mathbf{a} - \mathbf{b}| < |\mathbf{a} - \mathbf{a}_n| + |\mathbf{a}_n - \mathbf{b}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since ε is arbitrary, this proves the theorem. ■

The following is the definition of a Cauchy sequence in $\mathbb{R}^p$.

Definition D.0.10 *$\{\mathbf{a}_n\}$ is a Cauchy sequence if for all $\varepsilon > 0$, there exists n_ε such that whenever $n, m \geq n_\varepsilon$, if follows that $|\mathbf{a}_n - \mathbf{a}_m| < \varepsilon$.*

A sequence is Cauchy, means the terms are “bunching up to each other” as m, n get large.

Theorem D.0.11 *The set of terms in a Cauchy sequence in $\mathbb{R}^p$ is bounded in the sense that for all n, $|\mathbf{a}_n| < M$ for some $M < \infty$.*

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition, $|\mathbf{a}_n - \mathbf{a}_{n_1}| < 1$.It follows that for all $n > n_1$, $|\mathbf{a}_n| < 1 + |\mathbf{a}_{n_1}|$ .Therefore, for all n,

$$|\mathbf{a}_n| \leq 1 + |\mathbf{a}_{n_1}| + \sum_{k=1}^{n_1} |\mathbf{a}_k| . \blacksquare$$

Theorem D.0.12 *If a sequence $\{\mathbf{a}_n\}$ in $\mathbb{R}^p$ converges, then the sequence is a Cauchy sequence. Also, if some subsequence of a Cauchy sequence converges, then the original sequence converges.*

Proof: Let $\varepsilon > 0$ be given and suppose $\mathbf{a}_n \to \mathbf{a}$. Then from the definition of convergence, there exists n_ε such that if $n > n_\varepsilon$, it follows that $|\mathbf{a}_n - \mathbf{a}| < \frac{\varepsilon}{2}$. Therefore, if $m, n \geq n_\varepsilon + 1$, it follows that

$$|\mathbf{a}_n - \mathbf{a}_m| \leq |\mathbf{a}_n - \mathbf{a}| + |\mathbf{a} - \mathbf{a}_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

showing that, since $\varepsilon > 0$ is arbitrary, $\{\mathbf{a}_n\}$ is a Cauchy sequence. It remains to that the last claim.

Suppose then that $\{\mathbf{a}_n\}$ is a Cauchy sequence and $\mathbf{a} = \lim_{k\to\infty} \mathbf{a}_{n_k}$ where $\{\mathbf{a}_{n_k}\}_{k=1}^\infty$ is a subsequence. Let $\varepsilon > 0$ be given. Then there exists K such that if $k, l \geq K$, then

$|\mathbf{a}_k - \mathbf{a}_l| < \frac{\varepsilon}{2}$. Then if $k > K$, it follows $n_k > K$ because $n_1, n_2, n_3, \cdots$ is strictly increasing as the subscript increases. Also, there exists K_1 such that if $k > K_1, |\mathbf{a}_{n_k} - \mathbf{a}| < \frac{\varepsilon}{2}$. Then letting $n > \max(K, K_1)$, pick $k > \max(K, K_1)$. Then

$$|\mathbf{a} - \mathbf{a}_n| \leq |\mathbf{a} - \mathbf{a}_{n_k}| + |\mathbf{a}_{n_k} - \mathbf{a}_n| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Therefore, the sequence converges. ■

Definition D.0.13 *A set K in $\mathbb{R}^p$ is said to be **sequentially compact** if every sequence in K has a subsequence which converges to a point of K.*

Theorem D.0.14 *If $I_0 = \prod_{i=1}^p [a_i, b_i]$ where $a_i \leq b_i$, then I_0 is sequentially compact.*

Proof: Let $\{\mathbf{a}_k\}_{k=1}^\infty \subseteq I_0$ and consider all sets of the form $\prod_{i=1}^p [c_i, d_i]$ where $[c_i, d_i]$ equals either $\left[a_i, \frac{a_i+b_i}{2}\right]$ or $[c_i, d_i] = \left[\frac{a_i+b_i}{2}, b_i\right]$. Thus there are 2^p of these sets because there are two choices for the i^{th} slot for $i = 1, \cdots, p$. Also, if $\mathbf{x}$ and $\mathbf{y}$ are two points in one of these sets, $|x_i - y_i| \leq 2^{-1} |b_i - a_i|$ where $\operatorname{diam}(I_0) = \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2}$,

$$|\mathbf{x} - \mathbf{y}| = \left(\sum_{i=1}^p |x_i - y_i|^2\right)^{1/2} \leq 2^{-1} \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2} \equiv 2^{-1} \operatorname{diam}(I_0).$$

In particular, since $\mathbf{d} \equiv (d_1, \cdots, d_p)$ and $\mathbf{c} \equiv (c_1, \cdots, c_p)$ are two such points,

$$D_1 \equiv \left(\sum_{i=1}^p |d_i - c_i|^2\right)^{1/2} \leq 2^{-1} \operatorname{diam}(I_0)$$

Denote by $\{J_1, \cdots, J_{2^p}\}$ these sets determined above. Since the union of these sets equals all of $I_0 \equiv I$, it follows that for some J_k, the sequence, $\{\mathbf{a}_i\}$ is contained in J_k for infinitely many k. Let that one be called I_1. Next do for I_1 what was done for I_0 to get $I_2 \subseteq I_1$ such that the diameter is half that of I_1 and I_2 contains $\{\mathbf{a}_k\}$ for infinitely many values of k. Continue in this way obtaining a nested sequence $\{I_k\}$ such that $I_k \supseteq I_{k+1}$, and if $\mathbf{x}, \mathbf{y} \in I_k$, then $|\mathbf{x} - \mathbf{y}| \leq 2^{-k} \operatorname{diam}(I_0)$, and I_n contains $\{\mathbf{a}_k\}$ for infinitely many values of k for each n. Then by the nested interval lemma, there exists $\mathbf{c}$ such that $\mathbf{c}$ is contained in each I_k. Pick $\mathbf{a}_{n_1} \in I_1$. Next pick $n_2 > n_1$ such that $\mathbf{a}_{n_2} \in I_2$. If $\mathbf{a}_{n_1}, \cdots, \mathbf{a}_{n_k}$ have been chosen, let $\mathbf{a}_{n_{k+1}} \in I_{k+1}$ and $n_{k+1} > n_k$. This can be done because in the construction, I_n contains $\{\mathbf{a}_k\}$ for infinitely many k. Thus the distance between $\mathbf{a}_{n_k}$ and $\mathbf{c}$ is no larger than $2^{-k} \operatorname{diam}(I_0)$, and so $\lim_{k\to\infty} \mathbf{a}_{n_k} = \mathbf{c} \in I_0$. ■

Corollary D.0.15 *Let K be a closed and bounded set of points in $\mathbb{R}^p$. Then K is sequentially compact.*

Proof: Since K is closed and bounded, there exists a closed rectangle, $\prod_{k=1}^p [a_k, b_k]$ which contains K. Now let $\{\mathbf{x}_k\}$ be a sequence of points in K. By Theorem D.0.14, there exists a subsequence $\{\mathbf{x}_{n_k}\}$ such that $\mathbf{x}_{n_k} \to \mathbf{x} \in \prod_{k=1}^p [a_k, b_k]$. However, K is closed and each $\mathbf{x}_{n_k}$ is in K so $\mathbf{x} \in K$. ■

Theorem D.0.16 *Every Cauchy sequence in $\mathbb{R}^p$ converges.*

Proof: Let $\{\mathbf{a}_k\}$ be a Cauchy sequence. By Theorem D.0.11, there is some box $\prod_{i=1}^p [a_i, b_i]$ containing all the terms of $\{\mathbf{a}_k\}$. Therefore, by Theorem D.0.14, a subsequence converges to a point of $\prod_{i=1}^p [a_i, b_i]$. By Theorem D.0.12, the original sequence converges. ■

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 proved by Gauss in 1797 then proved with no loose ends by Argand in 1806 although others also worked on it. The proof given here follows Rudin [22]. See also Hardy [12] for another proof, more discussion and references. Recall De Moivre's theorem on Page 19 which is listed below for convenience.

Theorem E.0.17 *Let $r>0$ be given. Then if n is a positive integer,*

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

Now from this theorem, the following corollary on Page 1.5.5 is obtained.

Corollary E.0.18 *Let z be a non zero complex number and let k be a positive integer. Then there are always exactly k k^{th} roots of z in $\mathbb{C}$.*

Lemma E.0.19 *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 $\varepsilon > 0$ is given, let

$$\delta < \min \left(1, \frac{\varepsilon}{|a| \, n \left(1 + |z|\right)^n} \right) .$$

It follows from the above inequality that for $|z-w| < \delta$, $|az^n - aw^n| < \varepsilon$. The function of the lemma is just the sum of functions of this sort and so it follows that it is also continuous.

Theorem E.0.20 *(Fundamental theorem of Algebra) Let $p(z)$ be a nonconstant polynomial. Then there exists $z \in \mathbb{C}$ such that $p(z) = 0$.*

Proof: Suppose not. Then

$$p(z)=\sum_{k=0}^{n}a_kz^k$$

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

$$\lim_{|z|\to\infty}|p(z)|=\infty. \tag{5.1}$$

Now let

$$\lambda\equiv\inf\{|p(z)|:z\in\mathbb{C}\}.$$

By 5.1, there exists an $R>0$ such that if $|z|>R$, it follows that $|p(z)|>\lambda+1$. Therefore,

$$\lambda\equiv\inf\{|p(z)|:z\in\mathbb{C}\}=\inf\{|p(z)|:|z|\leq R\}.$$

The set $\{z : |z| \leq R\}$ is a closed and bounded set and so this infimum is achieved at some point w with $|w| \leq R$. A contradiction is obtained if $|p(w)| = 0$ so assume $|p(w)| > 0$. Then consider

$$q(z) \equiv \frac{p(z+w)}{p(w)}.$$

It follows $q(z)$ is of the form

$$q(z) = 1 + c_k z^k + \cdots + c_n z^n$$

where $c_k \neq 0$, because $q(0) = 1$. It is also true that $|q(z)| \geq 1$ by the assumption that $|p(w)|$ is the smallest value of $|p(z)|$. Now let $\theta \in \mathbb{C}$ be a complex number with $|\theta| = 1$ and

$$\theta c_k w^k = -|w|^k |c_k|.$$

If

$$w \neq 0, \theta = \frac{-|w^k| |c_k|}{w^k c_k}$$

and if $w = 0$, $\theta = 1$ will work. Now let $\eta^k = \theta$ and let t be a small positive number.

$$q(t\eta w) \equiv 1 - t^k |w|^k |c_k| + \cdots + c_n t^n (\eta w)^n$$

which is of the form

$$1 - t^k |w|^k |c_k| + t^k (g(t, w))$$

where $\lim_{t\to 0} g(t, w) = 0$. Letting t be small enough,

$$|g(t, w)| < |w|^k |c_k| / 2$$

and so for such t,

$$|q(t\eta w)| < 1 - t^k |w|^k |c_k| + t^k |w|^k |c_k| / 2 < 1,$$

a contradiction to $|q(z)| \geq 1$. ■

Fields And Field Extensions

F.1 The Symmetric Polynomial Theorem

First here is a definition of polynomials in many variables which have coefficients in a commutative ring. A commutative ring would be a field except you don't know that every nonzero element has a multiplicative inverse. If you like, let these coefficients be in a field it is still interesting. A good example of a commutative ring is the integers. In particular, every field is a commutative ring.

Definition F.1.1 *Let* $\mathbf{k} \equiv (k_1, k_2, \cdots, k_n)$ *where each* k_i *is a nonnegative integer. Let*

$$|\mathbf{k}| \equiv \sum_i k_i$$

Polynomials of degree p *in the variables* $x_1, x_2, \cdots, x_n$ *are expressions of the form*

$$g(x_1, x_2, \cdots, x_n) = \sum_{|\mathbf{k}| \leq p} a_{\mathbf{k}} x_1^{k_1} \cdots x_n^{k_n}$$

where each $a_{\mathbf{k}}$ *is in a commutative ring. If all* $a_{\mathbf{k}} = 0$*, the polynomial has no degree. Such a polynomial is said to be symmetric if whenever* σ *is a permutation of* $\{1, 2, \cdots, n\}$*,*

$$g\left(x_{\sigma(1)}, x_{\sigma(2)}, \cdots, x_{\sigma(n)}\right) = g(x_1, x_2, \cdots, x_n)$$

An example of a symmetric polynomial is

$$s_1(x_1, x_2, \cdots, x_n) \equiv \sum_{i=1}^{n} x_i$$

Another one is

$$s_n(x_1, x_2, \cdots, x_n) \equiv x_1 x_2 \cdots x_n$$

Definition F.1.2 *The elementary symmetric polynomial* $s_k(x_1, x_2, \cdots, x_n), k = 1, \cdots, n$ *is the coefficient of* $(-1)^k x^{n-k}$ *in the following polynomial.*

$$\begin{gathered}(x - x_1)(x - x_2) \cdots (x - x_n) \\ = x^n - s_1 x^{n-1} + s_2 x^{n-2} - \cdots \pm s_n\end{gathered}$$

Thus

$$\begin{gathered}s_1 = x_1 + x_2 + \cdots + x_n \\ s_2 = \sum_{i<j} x_i x_j,\ s_3 = \sum_{i<j<k} x_i x_j x_k, \ldots,\ s_n = x_1 x_2 \cdots x_n\end{gathered}$$

Then the following result is the fundamental theorem in the subject. It is the symmetric polynomial theorem. It says that these elementary symmetric polynomials are a lot like a basis for the symmetric polynomials.

Theorem F.1.3 *Let $g(x_1, x_2, \cdots, x_n)$ be a symmetric polynomial. Then $g(x_1, x_2, \cdots, x_n)$ equals a polynomial in the elementary symmetric functions.*

$$g(x_1, x_2, \cdots, x_n) = \sum_{\mathbf{k}} a_{\mathbf{k}} s_1^{k_1} \cdots s_n^{k_n}$$

and the $a_{\mathbf{k}}$ are unique.

Proof: If $n = 1$, it is obviously true because $s_1 = x_1$. Suppose the theorem is true for $n - 1$ and $g(x_1, x_2, \cdots, x_n)$ has degree d. Let

$$g'(x_1, x_2, \cdots, x_{n-1}) \equiv g(x_1, x_2, \cdots, x_{n-1}, 0)$$

By induction, there are unique $a_{\mathbf{k}}$ such that

$$g'(x_1, x_2, \cdots, x_{n-1}) = \sum_{\mathbf{k}} a_{\mathbf{k}} s_1'^{k_1} \cdots s_{n-1}'^{k_{n-1}}$$

where s_i' is the corresponding symmetric polynomial which pertains to $x_1, x_2, \cdots, x_{n-1}$. Note that

$$s_k(x_1, x_2, \cdots, x_{n-1}, 0) = s_k'(x_1, x_2, \cdots, x_{n-1})$$

Now consider

$$g(x_1, x_2, \cdots, x_n) - \sum_{\mathbf{k}} a_{\mathbf{k}} s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} \equiv q(x_1, x_2, \cdots, x_n)$$

is a symmetric polynomial and it equals 0 when x_n equals 0. Since it is symmetric, it is also 0 whenever $x_i = 0$. Therefore,

$$q(x_1, x_2, \cdots, x_n) = s_n h(x_1, x_2, \cdots, x_n)$$

and it follows that $h(x_1, x_2, \cdots, x_n)$ is symmetric of degree no more than $d - n$ and is uniquely determined. Thus, if $g(x_1, x_2, \cdots, x_n)$ is symmetric of degree d,

$$g(x_1, x_2, \cdots, x_n) = \sum_{\mathbf{k}} a_{\mathbf{k}} s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} + s_n h(x_1, x_2, \cdots, x_n)$$

where h has degree no more than $d - n$. Now apply the same argument to $h(x_1, x_2, \cdots, x_n)$ and continue, repeatedly obtaining a sequence of symmetric polynomials h_i, of strictly decreasing degree, obtaining expressions of the form

$$g(x_1, x_2, \cdots, x_n) = \sum_{\mathbf{k}} b_{\mathbf{k}} s_1^{k_1} \cdots s_{n-1}^{k_{n-1}} s_n^{k_n} + s_n h_m(x_1, x_2, \cdots, x_n)$$

Eventually h_m must be a constant or zero. By induction, each step in the argument yields uniqueness and so, the final sum of combinations of elementary symmetric functions is uniquely determined. ■

Here is a very interesting result which I saw claimed in a paper by Steinberg and Redheffer on Lindemannn's theorem which follows from the above corollary.

Theorem F.1.4 *Let $\alpha_1, \cdots, \alpha_n$ be roots of the polynomial equation*

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$

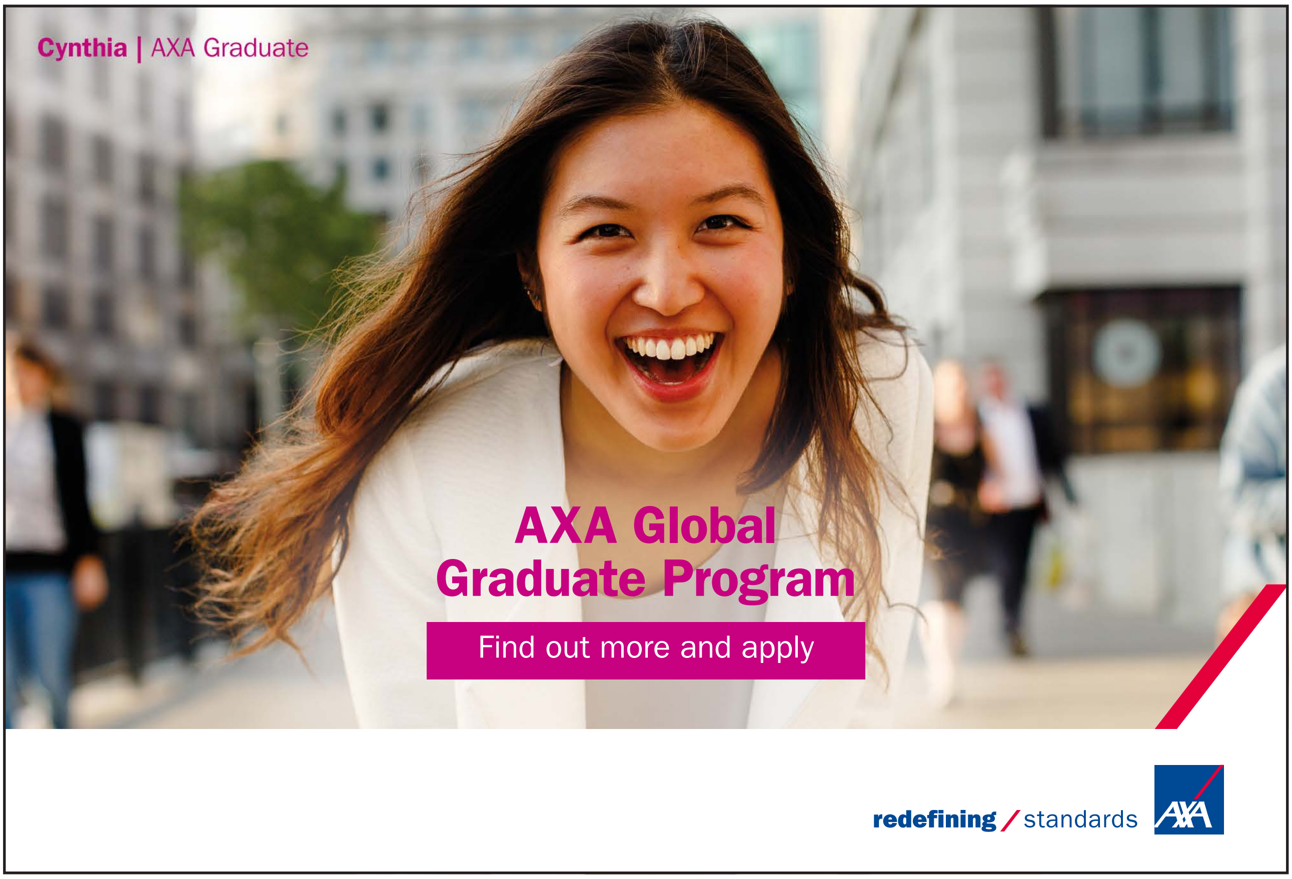
where each a_i is an integer. Then any symmetric polynomial in the quantities $a_n\alpha_1, \cdots, a_n\alpha_n$ having integer coefficients is also an integer. Also any symmetric polynomial in the quantities $\alpha_1, \cdots, \alpha_n$ having rational coefficients is a rational number.

Proof: Let $f(x_1, \cdots, x_n)$ be the symmetric polynomial. Thus

$$f(x_1, \cdots, x_n) \in \mathbb{Z}[x_1 \cdots x_n]$$

From Corollary F.1.3 it follows there are integers $a_{k_1 \cdots k_n}$ such that

$$f(x_1, \cdots, x_n) = \sum_{k_1 + \cdots + k_n \leq m} a_{k_1 \cdots k_n} p_1^{k_1} \cdots p_n^{k_n}$$



where the p_i are the elementary symmetric polynomials defined as the coefficients of

$$\prod_{j=1}^{n}(x-x_j)$$

Thus

$$\begin{aligned}&f(a_n\alpha_1,\cdots,a_n\alpha_n)\\=&\sum_{k_1+\cdots+k_n}a_{k_1\cdots k_n}p_1^{k_1}(a_n\alpha_1,\cdots,a_n\alpha_n)\cdots p_n^{k_n}(a_n\alpha_1,\cdots,a_n\alpha_n)\end{aligned}$$

Now the given polynomial is of the form

$$a_n\prod_{j=1}^{n}(x-\alpha_j)$$

and so the coefficient of x^{n-k} is $p_k(\alpha_1,\cdots,\alpha_n)a_n=a_{n-k}$. Also

$$p_k(a_n\alpha_1,\cdots,a_n\alpha_n)=a_n^k p_k(\alpha_1,\cdots,\alpha_n)=a_n^k\frac{a_{n-k}}{a_n}$$

It follows

$$f(a_n\alpha_1,\cdots,a_n\alpha_n)=\sum_{k_1+\cdots+k_n}a_{k_1\cdots k_n}\left(a_n^1\frac{a_{n-1}}{a_n}\right)^{k_1}\left(a_n^2\frac{a_{n-2}}{a_n}\right)^{k_2}\cdots\left(a_n^n\frac{a_0}{a_n}\right)^{k_n}$$

which is an integer. To see the last claim follows from this, take the symmetric polynomial in $\alpha_1,\cdots,\alpha_n$ and multiply by the product of the denominators of the rational coefficients to get one which has integer coefficients. Then by the first part, each homogeneous term is just an integer divided by a_n raised to some power. ■

F.2 The Fundamental Theorem Of Algebra

This is devoted to a mostly algebraic proof of the fundamental theorem of algebra. It depends on the interesting results about symmetric polynomials which are presented above. I found it on the Wikipedia article about the fundamental theorem of algebra. You google "fundamental theorem of algebra" and go to the Wikipedia article. It gives several other proofs in addition to this one. According to this article, the first completely correct proof of this major theorem is due to Argand in 1806. Gauss and others did it earlier but their arguments had gaps in them.

You can't completely escape analysis when you prove this theorem. The necessary analysis is in the following lemma.

Lemma F.2.1 *Suppose $p(x)=x^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ where n is odd and the coefficients are real. Then $p(x)$ has a real root.*

Proof: This follows from the intermediate value theorem from calculus.

Next is an algebraic consideration. First recall some notation.

$$\prod_{i=1}^{m}a_i\equiv a_1a_2\cdots a_m$$

Recall a polynomial in $\{z_1, \cdots, z_n\}$ is symmetric only if it can be written as a sum of elementary symmetric polynomials raised to various powers multiplied by constants. This follows from Proposition F.1.3 or Theorem F.1.3 both of which are the theorem on symmetric polynomials.

The following is the main part of the theorem. In fact this is one version of the fundamental theorem of algebra which people studied earlier in the 1700's.

Lemma F.2.2 *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial with real coefficients. Then it has a complex root.*

Proof: It is possible to write

$$n = 2^k m$$

where m is odd. If n is odd, $k = 0$. If n is even, keep dividing by 2 until you are left with an odd number. If $k = 0$ so that n is odd, it follows from Lemma F.2.1 that $p(x)$ has a real, hence complex root. The proof will be by induction on k, the case $k = 0$ being done. Suppose then that it works for $n = 2^l m$ where m is odd and $l \leq k - 1$ and let $n = 2^k m$ where m is odd. Let $\{z_1, \cdots, z_n\}$ be the roots of the polynomial in a splitting field, the existence of this field being given by the above proposition. Then

$$p(x) = \prod_{j=1}^{n} (x - z_j) = \sum_{k=0}^{n} (-1)^k p_k x^k \tag{6.1}$$

where p_k is the k^{th} elementary symmetric polynomial. Note this shows

$$a_{n-k} = p_k (-1)^k . \tag{6.2}$$

There is another polynomial which has coefficients which are sums of real numbers times the p_k raised to various powers and it is

$$q_t(x) \equiv \prod_{1 \leq i < j \leq n} (x - (z_i + z_j + t z_i z_j)), \ t \in \mathbb{R}$$

I need to verify this is really the case for $q_t(x)$. When you switch any two of the z_i in $q_t(x)$ the polynomial does not change. For example, let $n = 3$ when $q_t(x)$ is

$$(x - (z_1 + z_2 + t z_1 z_2))(x - (z_1 + z_3 + t z_1 z_3))(x - (z_2 + z_3 + t z_2 z_3))$$

and you can observe the assertion about the polynomial is true when you switch two different z_i. Thus the coefficients of $q_t(x)$ must be symmetric polynomials in the z_i with real coefficients. Hence by Proposition F.1.3 these coefficients are real polynomials in terms of the elementary symmetric polynomials p_k. Thus by 6.2 the coefficients of $q_t(x)$ are real polynomials in terms of the a_k of the original polynomial. Recall these were all real. It follows, and this is what was wanted, that $q_t(x)$ has all real coefficients.

Note that the degree of $q_t(x)$ is $\begin{pmatrix} n \\ 2 \end{pmatrix}$ because there are this number of ways to pick $i < j$ out of $\{1, \cdots, n\}$. Now

$$\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n(n-1)}{2} = 2^{k-1} m \left(2^k m - 1\right)$$

$$= 2^{k-1} (\text{odd})$$

and so by induction, for each $t \in \mathbb{R}, q_t(x)$ has a complex root.

There must exist $s \neq t$ such that for a single pair of indices i, j, with $i < j$,

$$(z_i + z_j + tz_iz_j), (z_i + z_j + sz_iz_j)$$

are both complex. Here is why. Let $A(i,j)$ denote those $t \in \mathbb{R}$ such that $(z_i + z_j + tz_iz_j)$ is complex. It was just shown that every $t \in \mathbb{R}$ must be in some $A(i,j)$. There are infinitely many $t \in \mathbb{R}$ and so some $A(i,j)$ contains two of them.

Now for that t, s,

$$\begin{aligned} z_i + z_j + tz_iz_j &= a \\ z_i + z_j + sz_iz_j &= b \end{aligned}$$

where $t \neq s$ and so by Cramer's rule,

$$z_i + z_j = \frac{\begin{vmatrix} a & t \\ b & s \end{vmatrix}}{\begin{vmatrix} 1 & t \\ 1 & s \end{vmatrix}} \in \mathbb{C}$$

and also

$$z_iz_j = \frac{\begin{vmatrix} 1 & a \\ 1 & b \end{vmatrix}}{\begin{vmatrix} 1 & t \\ 1 & s \end{vmatrix}} \in \mathbb{C}$$

At this point, note that z_i, z_j are both solutions to the equation

$$x^2 - (z_1 + z_2)x + z_1z_2 = 0,$$

which from the above has complex coefficients. By the quadratic formula the z_i, z_j are both complex. Thus the original polynomial has a complex root. ■

With this lemma, it is easy to prove the fundamental theorem of algebra. The difference between the lemma and this theorem is that in the theorem, the coefficients are only assumed to be complex. What this means is that if you have any polynomial with complex coefficients it has a complex root and so it is not irreducible. Hence the field extension is the same field. Another way to say this is that for **every** complex polynomial there exists a factorization into linear factors or in other words a splitting field for a complex polynomial is the field of complex numbers.

Theorem F.2.3 *Let* $p(x) \equiv a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ *be any complex polynomial,* $n \geq 1, a_n \neq 0$. *Then it has a complex root. Furthermore, there exist complex numbers* $z_1, \cdots, z_n$ *such that*

$$p(x) = a_n \prod_{k=1}^{n} (x - z_k)$$

Proof: First suppose $a_n = 1$. Consider the polynomial

$$q(x) \equiv p(x) \overline{p(\overline{x})}$$

this is a polynomial and it has real coefficients. This is because it equals

$$\begin{gathered}\left(x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\right) \cdot \\ \left(x^n + \overline{a_{n-1}}x^{n-1} + \cdots + \overline{a_1}x + \overline{a_0}\right)\end{gathered}$$

The x^{j+k} term of the above product is of the form

$$a_k x^k \overline{a_j} x^j + \overline{a_k} x^k a_j x^j = x^{k+j} \left(a_k \overline{a_j} + \overline{a_k} a_j \right)$$

and

$$a_k \overline{a_j} + \overline{a_k} a_j = a_k \overline{a_j} + \overline{a_k \overline{a_j}}$$

so it is of the form of a complex number added to its conjugate. Hence $q(x)$ has real coefficients as claimed. Therefore, by by Lemma F.2.2 it has a complex root z. Hence either $p(z) = 0$ or $p(\overline{z}) = 0$. Thus $p(x)$ has a complex root.

Next suppose $a_n \neq 0$. Then simply divide by it and get a polynomial in which $a_n = 1$. Denote this modified polynomial as $q(x)$. Then by what was just shown and the Euclidean algorithm, there exists $z_1 \in \mathbb{C}$ such that

$$q(x) = (x - z_1) q_1(x)$$

where $q_1(x)$ has complex coefficients. Now do the same thing for $q_1(x)$ to obtain

$$q(x) = (x - z_1)(x - z_2) q_2(x)$$

and continue this way. Thus

$$\frac{p(x)}{a_n} = \prod_{j=1}^{n} (x - z_j) \quad \blacksquare$$

Obviously this is a harder proof than the other proof of the fundamental theorem of algebra presented earlier. However, this is a better proof. Consider the algebraic numbers $\mathbb{A}$ consisting of the real numbers which are roots of some polynomial having rational coefficients. By Theorem 8.3.32 they are a field. Now consider the field $\mathbb{A} + i\mathbb{A}$ with the usual conventions for complex arithmetic. You could repeat the above argument with small changes and conclude that every polynomial having coefficients in $\mathbb{A} + i\mathbb{A}$ has a root in $\mathbb{A} + i\mathbb{A}$. Recall from Problem 41 on Page 298 that $\mathbb{A}$ is countable and so this is also the case for $\mathbb{A} + i\mathbb{A}$. Thus this gives an algebraically complete field which is countable and so very different than $\mathbb{C}$. Of course there are other situations in which the above harder proof will work and yield interesting results.

F.3 Transcendental Numbers

Most numbers are like this. Here the algebraic numbers are those which are roots of a polynomial equation having rational numbers as coefficients. By the fundamental theorem of calculus, all these numbers are in $\mathbb{C}$. There are only countably many of these algebraic numbers, (Problem 41 on Page 298). Therefore, most numbers are transcendental. Nevertheless, it is very hard to prove that this or that number is transcendental. Probably the most famous theorem about this is the Lindemannn Weierstrass theorem.

Theorem F.3.1 *Let the α_i be distinct nonzero algebraic numbers and let the a_i be nonzero algebraic numbers. Then*

$$\sum_{i=1}^{n} a_i e^{a_i} \neq 0$$

I am following the interesting Wikepedia article on this subject. You can also look at the book by Baker [4], Transcendental Number Theory, Cambridge University Press. There are also many other treatments which you can find on the web including an interesting article by Steinberg and Redheffer which appeared in about 1950.

The proof makes use of the following identity. For $f(x)$ a polynomial,

$$I(s) \equiv \int_0^s e^{s-x} f(x)\, dx = e^s \sum_{j=0}^{\deg(f)} f^{(j)}(0) - \sum_{j=0}^{\deg(f)} f^{(j)}(s). \tag{6.3}$$

where $f^{(j)}$ denotes the j^{th} derivative. In this formula, $s \in \mathbb{C}$ and the integral is defined in the natural way as

$$\int_0^1 sf(ts)\, e^{s-ts} dt \tag{6.4}$$

The identity follows from integration by parts.

$$\begin{aligned}
\int_0^1 sf(ts)\, e^{s-ts} dt &= se^s \int_0^1 f(ts)\, e^{-ts} dt \\
&= se^s \left[-\frac{e^{-ts}}{s} f(ts) |_0^1 + \int_0^1 \frac{e^{-ts}}{s} sf'(st)\, dt \right] \\
&= se^s \left[\frac{1}{s} f(s) - \frac{e^{-s}}{s} f(0) + \int_0^1 e^{-ts} f'(st)\, dt \right] \\
&= f(0) - e^s f(s) + \int_0^1 se^{s-ts} f'(st)\, dt \\
&\equiv f(0) - f(s)\, e^s + \int_0^s e^{s-x} f'(x)\, dx
\end{aligned}$$

Continuing this way establishes the identity.

Lemma F.3.2 *If K and c are nonzero integers, and $\beta_1, \cdots, \beta_m$ are the roots of a single polynomial with integer coefficients,*

$$Q(x) = vx^m + \cdots + u$$

where $v, u \neq 0$, then

$$K + c\left(e^{\beta_1} + \cdots + e^{\beta_m}\right) \neq 0.$$

Letting

$$f(x) = \frac{v^{(m-1)p} Q^p(x)\, x^{p-1}}{(p-1)!}$$

and $I(s)$ be defined in terms of $f(x)$ as above, it follows,

$$\lim_{p\to\infty} \sum_{i=1}^m I(\beta_i) = 0$$

and

$$\sum_{j=0}^n f^{(j)}(0) = v^{p(m-1)} u^p + m_1(p)\, p$$

$$\sum_{i=1}^m \sum_{j=0}^n f^{(j)}(\beta_i) = m_2(p)\, p$$

where $m_i(p)$ is some integer.

Proof: Let p be a prime number. Then consider the polynomial $f(x)$ of degree $n \equiv pm + p - 1$,

$$f(x) = \frac{v^{(m-1)p} Q^p(x) x^{p-1}}{(p-1)!}$$

From 6.3

$$c \sum_{i=1}^{m} I(\beta_i) = c \sum_{i=1}^{m} \left(e^{\beta_i} \sum_{j=0}^{n} f^{(j)}(0) - \sum_{j=0}^{n} f^{(j)}(\beta_i) \right)$$

$$= \left(K + c \sum_{i=1}^{m} e^{\beta_i} \right) \sum_{j=0}^{n} f^{(j)}(0) - K \sum_{j=0}^{n} f^{(j)}(0) - c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}(\beta_i) \tag{6.5}$$

Claim 1: $\lim_{p\to\infty} c \sum_{i=1}^{m} I(\beta_i) = 0.$

Proof: This follows right away from the definition of $I(\beta_j)$ and the definition of $f(x)$.

$$\left| I(\beta_j) \right| \leq \int_0^1 \left| \beta_j f(t\beta_j) e^{\beta_j - t\beta_j} \right| dt$$

$$\leq \int_0^1 \left| \frac{|v|^{(m-1)p} \left| Q(t\beta_j) \right|^p t^{p-1} \left| \beta_j \right|^{p-1}}{(p-1)!} dt \right|$$

which clearly converges to 0. This proves the claim.

The next thing to consider is the term on the end in 6.5,

$$K \sum_{j=0}^{n} f^{(j)}(0) + c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}(\beta_i) \tag{6.6}$$

The idea is to show that for large enough p it is always an integer. When this is done, it can't happen that $K + c\sum_{i=1}^{m} e^{\beta_i} = 0$ because if this were so, you would have a very small number equal to an integer. Now

$$\begin{aligned} f(x) &= \frac{v^{(m-1)p} \left(\overbrace{v(x-\beta_1)(x-\beta_2)\cdots(x-\beta_m)}^{Q(x)} \right)^p x^{p-1}}{(p-1)!} \\ &= \frac{v^{mp} \left((x-\beta_1)(x-\beta_2)\cdots(x-\beta_m)\right)^p x^{p-1}}{(p-1)!} \end{aligned} \tag{6.7}$$

It follows that for $j < p-1$, $f^{(j)}(0) = 0$. This is because of that term x^{p-1}. If $j \geq p$, $f^{(j)}(0)$ is an integer multiple of p. Here is why. The terms in this derivative which are nonzero involve taking $p-1$ derivatives of x^{p-1} and this introduces a $(p-1)!$ which cancels out the denominator. Then there are some other derivatives of the product of the $(x - \beta_i)$ raised to the power p. By the chain rule, these all involve a multiple of p. Thus this j^{th} derivative is of the form

$$pg(x, v\beta_1, \cdots, v\beta_m), \tag{6.8}$$

where $g(x, v\beta_1, \cdots, v\beta_m)$ is a polynomial in x with coefficients which are symmetric polynomials in $\{v\beta_1, \cdots, v\beta_m\}$ having integer coefficients. Then derivatives of g with respect to x also yield polynomials in x which have coefficients which are symmetric polynomials in $\{v\beta_1, \cdots, v\beta_m\}$ having integer coefficients. Evaluating g at $x = 0$ must therefore yield a polynomial which is symmetric in the $\{v\beta_1, \cdots, v\beta_m\}$ with integer coefficients. Since the $\{\beta_1, \cdots, \beta_m\}$ are the roots of a polynomial having integer coefficients with leading coefficient v, it follows from Theorem F.1.4 that this last polynomial is an integer and so the j^{th} derivative of f given by 6.8 when evaluated at $x = 0$ yields an integer times p. Now consider the case of the $(p-1)$ derivative of f. The only nonzero term of $f^{(j)}(0)$ is the one which

comes from taking $p-1$ derivatives of x^{p-1} and so it reduces to

$$v^{mp}(-1)^{mp}(\beta_1\beta_2\cdots\beta_m)^p$$

Now $Q(0)=v(-1)^m(\beta_1\beta_2\cdots\beta_m)=u$ and so $v^p(-1)^{mp}(\beta_1\beta_2\cdots\beta_m)^p=u^p$ which yields

$$f^{(p-1)}(0)=v^{mp}u^pv^{-p}=v^{p(m-1)}u^p$$

Note this is not necessarily a multiple of p and in fact will not be so if $p>u,v$ because p is a prime number. It follows

$$\sum_{j=0}^{n}f^{(j)}(0)=v^{p(m-1)}u^p+m(p)p$$

where $m(p)$ is some integer.

Now consider the other sum in 6.6,

$$c\sum_{i=1}^{m}\sum_{j=0}^{n}f^{(j)}(\beta_i)$$

Using the formula in 6.7 it follows that for $j<p,f^{(j)}(\beta_i)=0$. This is because for such derivatives, each term will have that product of the $(x-\beta_i)$ in it. Next consider the case where $j\geq p$. In this case, the nonzero terms must involve at least p derivatives of the expression

$$((x-\beta_1)(x-\beta_2)\cdots(x-\beta_m))^p$$

since otherwise, when evaluated at any β_k the result would be 0. Hence the $(p-1)!$ will vanish from the denominator and so all coefficients of the polynomials in the β_j and x will be integers and in fact, there will be an extra factor of p left over. Thus the j^{th} derivatives for $j\geq p$ involve taking the k^{th} derivative, $k\geq 0$ with respect to x of

$$pv^{mp}g(x,\beta_1,\cdots,\beta_m)$$

where $g(x,\beta_1,\cdots,\beta_m)$ is a polynomial in x having coefficients which are integers times symmetric polynomials in the $\{\beta_1,\cdots,\beta_m\}$. It follows that the k^{th} derivative for $k\geq 0$ is also a polynomial in x having the same properties. Therefore, taking the k^{th} derivative where k corresponds to $j\geq p$ and adding, yields

$$\sum_{i=1}^{m}pv^{mp}g_{,k}(\beta_i,\beta_1,\cdots,\beta_m)=\sum_{i=1}^{m}f^{(j)}(\beta_i) \tag{6.9}$$

where $g_{,k}$ denotes the k^{th} derivative of g taken with respect to x. Now

$$\sum_{i=1}^{m}g_{,k}(\beta_i,\beta_1,\cdots,\beta_m)$$

is a symmetric polynomial in the $\{\beta_1,\cdots,\beta_m\}$ with no term having degree more than mp and[1] so by Corollary F.1.3 this is of the form

$$\sum_{i=1}^{m}g_{,k}(\beta_i,\beta_1,\cdots,\beta_m)=\sum_{k_1,\cdots,k_m}a_{k_1\cdots k_m}p_1^{k_1}\cdots p_m^{k_m}$$

[1] Note the claim about this being a symmetric polynomial is about the sum, not an individual term.

where the $a_{k_1 \cdots k_m}$ are integers and the p_k are the elementary symmetric polynomials in $\{\beta_1, \cdots, \beta_m\}$. Recall these were roots of $vx^m + \cdots + u$ and so from the definition of the elementary symmetric polynomials given in Definition F.1.2, these p_k are each an integer divided by v, the integers being the coefficients of $Q(x)$. Therefore, from 6.9

$$\sum_{i=1}^{m} f^{(j)}(\beta_i) = pv^{mp} \sum_{i=1}^{m} g_{,k}(\beta_i, \beta_1, \cdots, \beta_m)$$
$$= pv^{mp} \sum_{k_1, \cdots, k_m} a_{k_1 \cdots k_m} p_1^{k_1} \cdots p_m^{k_m}$$

which is pv^{mp} times an expression which consists of integers times products of coefficients of $Q(x)$ divided by v raised to various powers, the sum of which is always no more than mp. Therefore, it reduces to an integer multiple of p and so the same is true of

$$c \sum_{i=1}^{m} \sum_{j=0}^{n} f^{(j)}(\beta_i)$$

which just involves adding up these integer multiples of p. Therefore, 6.6 is of the form

$$Kv^{p(m-1)}u^p + M(p)p$$

for some integer $M(p)$. Summarizing, it follows

$$c \sum_{i=1}^{m} I(\beta_i) = \left(K + c \sum_{i=1}^{m} e^{\beta_i}\right) \sum_{j=0}^{n} f^{(j)}(0) + Kv^{p(m-1)}u^p + M(p)p$$

where the left side is very small whenever p is large enough. Let p be larger than $\max(K, v, u)$. Since p is prime, it follows it cannot divide $Kv^{p(m-1)}u^p$ and so the last two terms must sum to a nonzero integer and so the equation cannot hold unless

$$K + c \sum_{i=1}^{m} e^{\beta_i} \neq 0 \quad \blacksquare$$

Note this shows π is irrational. If $\pi = k/m$ where k, m are integers, then both $i\pi$ and $-i\pi$ are roots of the polynomial with integer coefficients,

$$m^2x^2 + k^2$$

which would require from what was just shown that

$$0 \neq 2 + e^{i\pi} + e^{-i\pi}$$

which is not the case since the sum on the right equals 0.

The following corollary follows from this.

Corollary F.3.3 *Let K and c_i for $i = 1, \cdots, n$ be nonzero integers. For each k between 1 and n let $\{\beta(k)_i\}_{i=1}^{m(k)}$ be the roots of a polynomial with integer coefficients,*

$$Q_k(x) \equiv v_k x^{m_k} + \cdots + u_k$$

where $v_k, u_k \neq 0$. Then

$$K + c_1 \left(\sum_{j=1}^{m_1} e^{\beta(1)_j}\right) + c_2 \left(\sum_{j=1}^{m_2} e^{\beta(2)_j}\right) + \cdots + c_n \left(\sum_{j=1}^{m_n} e^{\beta(n)_j}\right) \neq 0.$$

Proof: Defining $f_k(x)$ and $I_k(s)$ as in Lemma F.3.2, it follows from Lemma F.3.2 that for each $k = 1, \cdots, n$,

$$\begin{aligned} c_k \sum_{i=1}^{m_k} I_k\left(\beta(k)_i\right) &= \left(K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right) \sum_{j=0}^{\deg(f_k)} f_k^{(j)}(0) \\ &\quad -K_k \sum_{j=0}^{\deg(f_k)} f_k^{(j)}(0) - c_k \sum_{i=1}^{m_k} \sum_{j=0}^{\deg(f_k)} f_k^{(j)}\left(\beta(k)_i\right) \end{aligned}$$

This is exactly the same computation as in the beginning of that lemma except one adds and subtracts $K_k \sum_{j=0}^{\deg(f_k)} f_k^{(j)}(0)$ rather than $K \sum_{j=0}^{\deg(f_k)} f_k^{(j)}(0)$ where the K_k are chosen such that their sum equals K. By Lemma F.3.2,

$$c_k \sum_{i=1}^{m_k} I_k\left(\beta(k)_i\right) = \left(K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right)\left(v_k^{(m_k-1)p} u_k^p + N_k p\right)$$

$$-K_k\left(v_k^{(m_k-1)p} u_k^p + N_k p\right) - c_k N_k' p$$

and so

$$c_k \sum_{i=1}^{m_k} I_k\left(\beta(k)_i\right) = \left(K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right)\left(v_k^{(m_k-1)p} u_k^p + N_k p\right)$$

$$-K_k v_k^{(m_k-1)p} u_k^p + M_k p$$

for some integer M_k. By multiplying each $Q_k(x)$ by a suitable constant, it can be assumed without loss of generality that all the $v_k^{m_k-1} u_k$ are equal to a constant integer U. Then the above equals

$$c_k \sum_{i=1}^{m_k} I_k\left(\beta(k)_i\right) = \left(K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right)\left(U^p + N_k p\right)$$

$$-K_k U^p + M_k p$$

Adding these for all k gives

$$\sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} I_k \left(\beta\left(k\right)_i\right) = U^p \left(K + \sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right) - KU^p + Mp$$

$$+ \sum_{k=1}^{n} N_k p \left(K_k + c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}\right) \tag{6.10}$$

For large p it follows from Lemma F.3.2 that the left side is very small. If

$$K + \sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i} = 0$$

then $\sum_{k=1}^{n} c_k \sum_{i=1}^{m_k} e^{\beta(k)_i}$ is an integer and so the last term in 6.10 is an integer times p. Thus for large p it reduces to

$$\text{small number} = -KU^p + Ip$$

where I is an integer. Picking prime $p > \max(U, K)$ it follows $-KU^p + Ip$ is a nonzero integer and this contradicts the left side being a small number less than 1 in absolute value. ■

Next is an even more interesting Lemma which follows from the above corollary.

Lemma F.3.4 *If $b_0, b_1, \cdots, b_n$ are non zero integers, and $\gamma_1, \cdots, \gamma_n$ are distinct algebraic numbers, then*

$$b_0 e^{\gamma_0} + b_1 e^{\gamma_1} + \cdots + b_n e^{\gamma_n} \neq 0$$

Proof: Assume

$$b_0 e^{\gamma_0} + b_1 e^{\gamma_1} + \cdots + b_n e^{\gamma_n} = 0 \tag{6.11}$$

Divide by e^{γ_0} and letting $K = b_0$,

$$K + b_1 e^{\alpha(1)} + \cdots + b_n e^{\alpha(n)} = 0 \tag{6.12}$$

where $\alpha(k) = \gamma_k - \gamma_0$. These are still distinct algebraic numbers none of which is 0 thanks to Theorem 8.3.32. Therefore, $\alpha(k)$ is a root of a polynomial

$$v_k x^{m_k} + \cdots + u_k \tag{6.13}$$

having integer coefficients, $v_k, u_k \neq 0$. Recall algebraic numbers were defined as roots of polynomial equations having rational coefficients. Just multiply by the denominators to get one with integer coefficients. Let the roots of this polynomial equation be

$$\left\{\alpha(k)_1, \cdots, \alpha(k)_{m_k}\right\}$$

and suppose they are listed in such a way that $\alpha(k)_1 = \alpha(k)$. Letting i_k be an integer in $\{1, \cdots, m_k\}$ it follows from the assumption 6.11 that

$$\prod_{\substack{(i_1, \cdots, i_n) \\ i_k \in \{1, \cdots, m_k\}}} \left(K + b_1 e^{\alpha(1)_{i_1}} + b_2 e^{\alpha(2)_{i_2}} + \cdots + b_n e^{\alpha(n)_{i_n}}\right) = 0 \tag{6.14}$$

This is because one of the factors is the one occurring in 6.12 when $i_k = 1$ for every k. The product is taken over all distinct ordered lists $(i_1, \cdots, i_n)$ where i_k is as indicated. Expand this possibly huge product. This will yield something like the following.

$$\begin{aligned} K' + c_1 \left(e^{\beta(1)_1} + \cdots + e^{\beta(1)_{\mu(1)}}\right) + c_2 \left(e^{\beta(2)_1} + \cdots + e^{\beta(2)_{\mu(2)}}\right) + \cdots + \\ c_N \left(e^{\beta(N)_1} + \cdots + e^{\beta(N)_{\mu(N)}}\right) = 0 \end{aligned} \tag{6.15}$$

These integers c_j come from products of the b_i and K. The $\beta(i)_j$ are the distinct exponents which result. Note that a typical term in this product 6.14 would be something like

$$\overbrace{K^{p+1} b_{k_1} \cdots b_{k_{n-p}}}^{\text{integer}} e^{\overbrace{\alpha(k_1)_{i_1} + \alpha(k_2)_{i_2} \cdots + \alpha(k_{n-p})_{i_{n-p}}}^{\beta(j)_r}}$$

the k_j possibly not distinct and each $i_k \in \{1, \cdots, m_{i_k}\}$. Other terms of this sort are

$$K^{p+1} b_{k_1} \cdots b_{k_{n-p}} e^{\alpha(k_1)_{i_1'} + \alpha(k_2)_{i_2'} \cdots + \alpha(k_{n-p})_{i'_{n-p}}},$$
$$K^{p+1} b_{k_1} \cdots b_{k_{n-p}} e^{\alpha(k_1)_1 + \alpha(k_2)_1 \cdots + \alpha(k_{n-p})_1}$$

where each i_k' is another index in $\{1, \cdots, m_{i_k}\}$ and so forth. A given j in the sum of 6.15 corresponds to such a choice of $\left\{b_{k_1}, \cdots, b_{k_{n-p}}\right\}$ which leads to $K^{p+1} b_{k_1} \cdots b_{k_{n-p}}$ times a sum of exponentials like those just described. Since the product in 6.14 is taken over all choices $i_k \in \{1, \cdots, m_k\}$, it follows that if you switch $\alpha(r)_i$ and $\alpha(r)_j$, two of the roots of the polynomial

$$v_r x^{m_r} + \cdots + u_r$$

mentioned above, the result in 6.15 would be the same except for permuting the

$$\beta(s)_1, \beta(s)_2, \cdots, \beta(s)_{\mu(s)}.$$

Thus a symmetric polynomial in

$$\beta(s)_1, \beta(s)_2, \cdots, \beta(s)_{\mu(s)}$$

is also a symmetric polynomial in the $\alpha(k)_1, \alpha(k)_2, \cdots, \alpha(k)_{m_k}$ for each k. Thus for a given $r, \beta(r)_1, \cdots, \beta(r)_{\mu(r)}$ are roots of the polynomial

$$(x - \beta(r)_1)(x - \beta(r)_2) \cdots \left(x - \beta(r)_{\mu(r)}\right)$$

whose coefficients are symmetric polynomials in the $\beta(r)_j$ which is a symmetric polynomial in the $\alpha(k)_j, j = 1, \cdots, m_k$ for each k. Letting g be one of these symmetric polynomials and writing it in terms of the $\alpha(k)_i$ you would have

$$\sum_{l_1, \cdots, l_n} A_{l_1 \cdots l_n} \alpha(n)_1^{l_1} \alpha(n)_2^{l_2} \cdots \alpha(n)_{m_n}^{l_n}$$

where $A_{l_1 \cdots l_n}$ is a symmetric polynomial in $\alpha(k)_j, j = 1, \cdots, m_k$ for each $k \leq n-1$. These coefficients are in the field (Proposition 8.3.31) $\mathbb{Q}[A(1), \cdots, A(n-1)]$ where $A(k)$ denotes

$$\left\{\alpha(k)_1, \cdots, \alpha(k)_{m_k}\right\}$$

and so from Proposition F.1.3, the above symmetric polynomial is of the form

$$\sum_{(k_1 \cdots k_{m_n})} B_{k_1 \cdots k_{m_n}} p_1^{k_1}\left(\alpha(n)_1, \cdots, \alpha(n)_{m_n}\right) \cdots p_{m_n}^{k_{m_n}}\left(\alpha(n)_1, \cdots, \alpha(n)_{m_n}\right)$$

where $B_{k_1 \cdots k_{m_n}}$ is a symmetric polynomial in $\alpha(k)_j, j = 1, \cdots, m_k$ for each $k \leq n-1$. Now do for each $B_{k_1 \cdots k_{m_n}}$ what was just done for g featuring this time

$$\left\{\alpha(n-1)_1, \cdots, \alpha(n-1)_{m_{n-1}}\right\}$$

and continuing this way, it must be the case that eventually you have a sum of integer multiples of products of elementary symmetric polynomials in $\alpha(k)_j, j = 1, \cdots, m_k$ for each $k \leq n$. By Theorem F.1.4, these are each rational numbers. Therefore, each such g is a rational number and so the $\beta(r)_j$ are algebraic. Now 6.15 contradicts Corollary F.3.3. ■

Note this lemma is sufficient to prove Lindemann's theorem that π is transcendental. Here is why. If π is algebraic, then so is $i\pi$ and so from this lemma, $e^0 + e^{i\pi} \neq 0$ but this is not the case because $e^{i\pi} = -1$.

The next theorem is the main result, the Lindemannn Weierstrass theorem.

Theorem F.3.5 *Suppose $a(1), \cdots, a(n)$ are nonzero algebraic numbers and suppose*

$$\alpha(1), \cdots, \alpha(n)$$

are distinct algebraic numbers. Then

$$a(1) e^{\alpha(1)} + a(2) e^{\alpha(2)} + \cdots + a(n) e^{\alpha(n)} \neq 0$$

Proof: Suppose $a(j) \equiv a(j)_1$ is a root of the polynomial

$$v_j x^{m_j} + \cdots + u_j$$

where $v_j, u_j \neq 0$. Let the roots of this polynomial be $a(j)_1, \cdots, a(j)_{m_j}$. Suppose to the contrary that

$$a(1)_1 e^{\alpha(1)} + a(2)_1 e^{\alpha(2)} + \cdots + a(n)_1 e^{\alpha(n)} = 0$$

Then consider the big product

$$\prod_{\substack{(i_1,\cdots,i_n) \\ i_k \in \{1,\cdots,m_k\}}} \left(a(1)_{i_1} e^{\alpha(1)} + a(2)_{i_2} e^{\alpha(2)} + \cdots + a(n)_{i_n} e^{\alpha(n)} \right) \tag{6.16}$$

the product taken over all ordered lists $(i_1, \cdots, i_n)$. This product equals

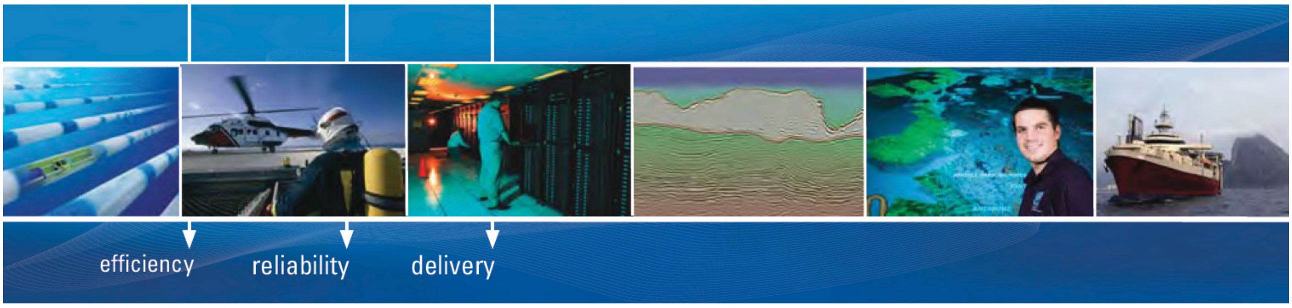
$$0 = b_1 e^{\beta(1)} + b_2 e^{\beta(2)} + \cdots + b_N e^{\beta(N)} \tag{6.17}$$

where the $\beta(j)$ are the distinct exponents which result. The $\beta(i)$ are clearly algebraic because they are the sum of the $\alpha(i)$. Since the product in 6.16 is taken for all ordered lists as described above, it follows that for a given k,if $\alpha(k)_i$ is switched with $\alpha(k)_j$, that is, two of the roots of $v_k x^{m_k} + \cdots + u_k$ are switched, then the product is unchanged and so 6.17 is also unchanged. Thus each b_k is a symmetric polynomial in the $a(k)_j, j = 1, \cdots, m_k$ for each k. It follows

$$b_k = \sum_{(j_1,\cdots,j_{m_n})} A_{j_1,\cdots,j_{m_n}} a(n)_1^{j_1} \cdots a(n)_{m_n}^{j_{m_n}}$$

and this is symmetric in the $\{a(n)_1, \cdots, a(n)_{m_n}\}$ the coefficients $A_{j_1,\cdots,j_{m_n}}$ being in the field (Proposition 8.3.31) $\mathbb{Q}[A(1), \cdots, A(n-1)]$ where $A(k)$ denotes

$$a(k)_1, \cdots, a(k)_{m_k}$$



and so from Proposition F.1.3,

$$b_k = \sum_{(j_1,\cdots,j_{m_n})} B_{j_1,\cdots,j_{m_n}} p_1^{j_1}\left(a(n)_1 \cdots a(n)_{m_n}\right) \cdots p_{m_n}^{j_{m_n}}\left(a(n)_1 \cdots a(n)_{m_n}\right)$$

where the $B_{j_1,\cdots,j_{m_n}}$ are symmetric in $\left\{a(k)_j\right\}_{j=1}^{m_k}$ for each $k \leq n-1$. Now doing to $B_{j_1,\cdots,j_{m_n}}$ what was just done to b_k and continuing this way, it follows b_k is a finite sum of integers times elementary polynomials in the various $\left\{a(k)_j\right\}_{j=1}^{m_k}$ for $k \leq n$. By Theorem F.1.4 this is a rational number. Thus b_k is a rational number. Multiplying by the product of all the denominators, it follows there exist integers c_i such that

$$0 = c_1 e^{\beta(1)} + c_2 e^{\beta(2)} + \cdots + c_N e^{\beta(N)}$$

which contradicts Lemma F.3.4. ■

This theorem is sufficient to show e is transcendental. If it were algebraic, then

$$ee^{-1} + (-1)e^0 \neq 0$$

but this is not the case. If $a \neq 1$ is algebraic, then $\ln(a)$ is transcendental. To see this, note that

$$1e^{\ln(a)} + (-1)\, ae^0 = 0$$

which cannot happen according to the above theorem. If a is algebraic and $\sin(a) \neq 0$, then $\sin(a)$ is transcendental because

$$\frac{1}{2i}e^{ia} - \frac{1}{2i}e^{-ia} + (-1)\sin(a)\, e^0 = 0$$

which cannot occur if $\sin(a)$ is algebraic. There are doubtless other examples of numbers which are transcendental by this amazing theorem.

F.4 More On Algebraic Field Extensions

The next few sections have to do with fields and field extensions. There are many linear algebra techniques which are used in this discussion and it seems to me to be very interesting. However, this is definitely far removed from my own expertise so there may be some parts of this which are not too good. I am following various algebra books in putting this together.

Consider the notion of splitting fields. It is desired to show that any two are isomorphic, meaning that there exists a one to one and onto mapping from one to the other which preserves all the algebraic structure. To begin with, here is a theorem about extending homomorphisms. [17]

Definition F.4.1 *Suppose $\mathbb{F}, \bar{\mathbb{F}}$ are two fields and that $f : \mathbb{F} \to \bar{\mathbb{F}}$ is a homomorphism. This means that*

$$f(xy) = f(x)\, f(y)\,,\ f(x+y) = f(x) + f(y)$$

An isomorphism is a homomorphism which is one to one and onto. A monomorphism is a homomorphism which is one to one. An automorphism is an isomorphism of a single field. Sometimes people use the symbol $\simeq$ to indicate something is an isomorphism. Then if $p(x) \in \mathbb{F}[x]$, say

$$p(x) = \sum_{k=0}^{n} a_k x^k,$$

$\bar{p}(x)$ will be the polynomial in $\bar{\mathbb{F}}[x]$ defined as

$$\bar{p}(x) \equiv \sum_{k=0}^{n} f(a_k)\, x^k.$$

Also consider f as a homomorphism of $\mathbb{F}[x]$ and $\bar{\mathbb{F}}[x]$ in the obvious way.

$$f(p(x)) = \bar{p}(x)$$

The following is a nice theorem which will be useful.

Theorem F.4.2 *Let $\mathbb{F}$ be a field and let r be algebraic over $\mathbb{F}$. Let $p(x)$ be the minimal polynomial of r. Thus $p(r) = 0$ and $p(x)$ is monic and no nonzero polynomial having coefficients in $\mathbb{F}$ of smaller degree has r as a root. In particular, $p(x)$ is irreducible over $\mathbb{F}$. Then define $f : \mathbb{F}[x] \to \mathbb{F}[r]$, the polynomials in r by*

$$f\left(\sum_{i=0}^{m} a_i x^i\right) \equiv \sum_{i=0}^{m} a_i r^i$$

Then f is a homomorphism. Also, defining $g : \mathbb{F}[x] / (p(x))$ by

$$g([q(x)]) \equiv f(q(x)) \equiv q(r)$$

it follows that g is an isomorphism from the field $\mathbb{F}[x] / (p(x))$ to $\mathbb{F}[r]$.

Proof: First of all, consider why f is a homomorphism. The preservation of sums is obvious. Consider products.

$$\begin{aligned} f\left(\sum_i a_i x^i \sum_j b_j x^j\right) &= f\left(\sum_{i,j} a_i b_j x^{i+j}\right) = \sum_{ij} a_i b_j r^{i+j} \\ &= \sum_i a_i r^i \sum_j b_j r^j = f\left(\sum_i a_i x^i\right) f\left(\sum_j b_j x^j\right) \end{aligned}$$

Thus it is clear that f is a homomorphism.

First consider why g is even well defined. If $[q(x)] = [q_1(x)]$, this means that

$$q_1(x) - q(x) = p(x) l(x)$$

for some $l(x) \in \mathbb{F}[x]$. Therefore,

$$\begin{aligned} f(q_1(x)) &= f(q(x)) + f(p(x) l(x)) \\ &= f(q(x)) + f(p(x)) f(l(x)) \\ &\equiv q(r) + p(r) l(r) = q(r) = f(q(x)) \end{aligned}$$

Now from this, it is obvious that g is a homomorphism.

$$\begin{aligned} g([q(x)][q_1(x)]) &= g([q(x) q_1(x)]) = f(q(x) q_1(x)) = q(r) q_1(r) \\ g([q(x)]) g([q_1(x)]) &\equiv q(r) q_1(r) \end{aligned}$$

Similarly, g preserves sums. Now why is g one to one? It suffices to show that if $g([q(x)]) = 0$, then $[q(x)] = 0$. Suppose then that

$$g([q(x)]) \equiv q(r) = 0$$

Then

$$q(x) = p(x) l(x) + \rho(x)$$

where the degree of $\rho(x)$ is less than the degree of $p(x)$ or else $\rho(x) = 0$. If $\rho(x) \neq 0$, then it follows that

$$\rho(r) = 0$$

and $\rho(x)$ has smaller degree than that of $p(x)$ which contradicts the definition of $p(x)$ as the minimal polynomial of r. Since $p(x)$ is irreducible, $\mathbb{F}[x] / (p(x))$ is a field. It is clear that g is onto. Therefore, $\mathbb{F}[r]$ is a field also. (This was shown earlier by different reasoning.) ■

Here is a diagram of what the following theorem says.

Extending f to g

$$\begin{tikzcd}
\mathbb{F} \arrow[r, "f"] \arrow[r, "\simeq"'] & \bar{\mathbb{F}} \\
p(x) \in \mathbb{F}[x] \arrow[r, "f"] & \bar{p}(x) \in \bar{\mathbb{F}}[x] \\
p(x) = \sum_{k=0}^{n} a_k x^k \arrow[r] & \sum_{k=0}^{n} f(a_k) x^k = \bar{p}(x) \\
p(r) = 0 & \bar{p}(\bar{r}) = 0 \\
\mathbb{F}[r] \arrow[r, "g"] \arrow[r, "\simeq"'] & \bar{\mathbb{F}}[\bar{r}] \\
r \arrow[r, "g"] & \bar{r}
\end{tikzcd}$$

One such g for each $\bar{r}$

Theorem F.4.3 *Let $f : \mathbb{F} \to \bar{\mathbb{F}}$ be an isomorphism of the two fields. Let r be algebraic over $\mathbb{F}$ with minimal polynomial $p(x)$ and suppose there exists $\bar{r}$ algebraic over $\bar{\mathbb{F}}$ such that $\bar{p}(\bar{r}) = 0$. Then there exists an isomorphism $g : \mathbb{F}[r] \to \bar{\mathbb{F}}[\bar{r}]$ which agrees with f on $\mathbb{F}$. If $g : \mathbb{F}[r] \to \bar{\mathbb{F}}[\bar{r}]$ is an isomorphism which agrees with f on $\mathbb{F}$ and if $\alpha([k(x)]) \equiv k(r)$ is the homomorphism mapping $\mathbb{F}[x]/(p(x))$ to $\mathbb{F}[r]$, then there must exist $\bar{r}$ such that $\bar{p}(\bar{r}) = 0$ and $g = \beta\alpha^{-1}$ where β*

$$\beta : \mathbb{F}[x]/(p(x)) \to \bar{\mathbb{F}}[\bar{r}]$$

is given by $\beta([k(x)]) = \bar{k}(\bar{r})$. In particular, $g(r) = \bar{r}$.

Proof: From Theorem F.4.2, there exists α, an isomorphism in the following picture, $\alpha([k(x)]) = k(r)$.

$$\mathbb{F}[r] \overset{\alpha}{\leftarrow} \mathbb{F}[x]/(p(x)) \overset{\beta}{\to} \bar{\mathbb{F}}[\bar{r}]$$

where $\beta([k(x)]) \equiv \bar{k}(\bar{r})$. ($\bar{k}(x)$ comes from f as described in the above definition.) This β is a well defined monomorphism because of the assumption that $\bar{p}(\bar{r}) = 0$. This needs to be verified. Assume then that it is so. Then just let $g = \beta\alpha^{-1}$.

Why is β well defined? Suppose $[k(x)] = [k'(x)]$ so that $k(x) - k'(x) = l(x)p(x)$. Then since f is a homomorphism,

$$\bar{k}(x) - \bar{k}'(x) = \bar{l}(x)\bar{p}(x), \ \bar{k}(\bar{r}) - \bar{g}\bar{k}'(\bar{r}) = \bar{l}(\bar{r})\bar{p}(\bar{r}) = 0$$

so β is indeed well defined. It is clear from the definition that β is a homomorphism. Suppose $\beta([k(x)]) = 0$. Does it follow that $[k(x)] = 0$? By assumption, $\bar{g}(\bar{r}) = 0$ and also,

$$\bar{k}(x) = \bar{p}(x)\bar{l}(x) + \bar{\rho}(x)$$

where the degree of $\bar{\rho}(x)$ is less than the degree of $\bar{p}(x)$ or else it equals 0. But then, since f is an isomorphism,

$$k(x) = p(x)l(x) + \rho(x)$$

where the degree of $\rho(x)$ is less than the degree of $p(x)$. However, the above shows that $\rho(r) = 0$ contrary to $p(x)$ being the minimal polynomial. Hence $\rho(x) = 0$ and this implies that $[k(x)] = 0$. Thus β is one to one and a homomorphism. Hence $g = \beta\alpha^{-1}$ works if it is also onto. However, it is clear that α^{-1} is onto and that β is onto. Hence the desired extension exists.

Now suppose such an isomorphism g exists. Then $\bar{r}$ must equal $g(r)$ and

$$0 = g(p(r)) = \bar{p}(g(r)) = \bar{p}(\bar{r})$$

Hence, β can be defined as above as $\beta([k(x)]) \equiv \bar{k}(\bar{r})$ relative to this $\bar{r} \equiv g(r)$ and

$$\beta\alpha^{-1}(k(r)) \equiv \beta([k(x)]) \equiv \bar{k}(g(r)) = g(k(r))$$

■

What is the meaning of the above in simple terms? It says that the monomorphisms from $\mathbb{F}[r]$ to a field $\bar{\mathbb{K}}$ containing $\bar{\mathbb{F}}$ correspond to the roots of $\bar{p}(x)$ in $\bar{\mathbb{K}}$. That is, for each root of $\bar{p}(x)$, there is a monomorphism and for each monomorphism, there is a root. Also, for each root $\bar{r}$ of $\bar{p}(x)$ in $\bar{\mathbb{K}}$, there is an isomorphism from $\mathbb{F}[r]$ to $\bar{\mathbb{F}}[\bar{r}]$.

Note that if $p(x)$ is a monic irreducible polynomial, then it is the minimal polynomial for each of its roots. This is the situation which is about to be considered. It involves the splitting fields $\mathbb{K}, \bar{\mathbb{K}}$ of $p(x), \bar{p}(x)$ where η is an isomorphism of $\mathbb{F}$ and $\bar{\mathbb{F}}$ as described above. See [17]. Here is a little diagram which describes what this theorem says.

Definition F.4.4 *The symbol $[\mathbb{K} : \mathbb{F}]$ where $\mathbb{K}$ is a field extension of $\mathbb{F}$ means the dimension of the vector space $\mathbb{K}$ with field of scalars $\mathbb{F}$.*

$$\begin{array}{ccc}
\mathbb{F} & \overset{\eta}{\underset{\simeq}{\to}} & \bar{\mathbb{F}} \\
p(x) & \text{“}\eta p(x) = \bar{p}(x)\text{”} & \bar{p}(x) \\
\mathbb{F}[r_1, \cdots, r_n] & \overset{\zeta_i}{\underset{\simeq}{\to}} & \bar{\mathbb{F}}[r_1, \cdots, r_n] \\
 & i = 1, \cdots, m, \left\{ \begin{array}{l} m \leq [\mathbb{K} : \mathbb{F}] \\ m = [\mathbb{K} : \mathbb{F}], \bar{r}_i \neq \bar{r}_j \end{array} \right. &
\end{array}$$

Theorem F.4.5 *Let η be an isomorphism from $\mathbb{F}$ to $\bar{\mathbb{F}}$ and let $\mathbb{K} = \mathbb{F}[r_1, \cdots, r_n], \bar{\mathbb{K}} = \bar{\mathbb{F}}[\bar{r}_1, \cdots, \bar{r}_n]$ be splitting fields of $p(x)$ and $\bar{p}(x)$ respectively. Then there exist at most $[\mathbb{K} : \mathbb{F}]$ isomorphisms $\zeta_i : \mathbb{K} \to \bar{\mathbb{K}}$ which extend η. If $\{\bar{r}_1, \cdots, \bar{r}_n\}$ are distinct, then there exist exactly $[\mathbb{K} : \mathbb{F}]$ isomorphisms of the above sort. In either case, the two splitting fields are isomorphic with any of these ζ_i serving as an isomorphism.*

Proof: Suppose $[\mathbb{K} : \mathbb{F}] = 1$. Say a basis for $\mathbb{K}$ is $\{r\}$. Then $\{1, r\}$ is dependent and so there exist $a, b \in \mathbb{F}$, not both zero such that $a + br = 0$. Then it follows that $r \in \mathbb{F}$ and so in this case $\mathbb{F} = \mathbb{K}$. Then the isomorphism which extends η is just η itself and there is exactly 1 isomorphism.

Next suppose $[\mathbb{K} : \mathbb{F}] > 1$. Then $p(x)$ has an irreducible factor over $\mathbb{F}$ of degree larger than 1, $q(x)$. If not, you would have

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_n$$

and it would factor as

$$= (x - r_1) \cdots (x - r_n)$$

with each $r_j \in \mathbb{F}$, so $\mathbb{F} = \mathbb{K}$ contrary to $[\mathbb{K} : \mathbb{F}] > 1$.Without loss of generality, let the roots of $q(x)$ in $\mathbb{K}$ be $\{r_1, \cdots, r_m\}$. Thus

$$q(x) = \prod_{i=1}^{m} (x - r_i), \quad p(x) = \prod_{i=1}^{n} (x - r_i)$$

Now $\bar{q}(x)$ defined analogously to $p(x)$, also has degree at least 2. Furthermore, it divides $\bar{p}(x)$ all of whose roots are in $\bar{\mathbb{K}}$. Denote the roots of $\bar{q}(x)$ in $\mathbb{K}$ as $\{\bar{r}_1, \cdots, \bar{r}_m\}$ where they are counted according to multiplicity.

Then from Theorem F.4.3, there exist $k \leq m$ one to one homomorphisms ζ_i mapping $\mathbb{F}[r_1]$ to $\bar{\mathbb{K}}$, one for each distinct root of $\bar{q}(x)$ in $\bar{\mathbb{K}}$. If the roots of $\bar{p}(x)$ are distinct, then this is sufficient to imply that the roots of $\bar{q}(x)$ are also distinct, and $k = m$. Otherwise, maybe $k < m$. (It is conceivable that $\bar{q}(x)$ might have repeated roots in $\bar{\mathbb{K}}$.) Then

$$[\mathbb{K} : \mathbb{F}] = [\mathbb{K} : \mathbb{F}[r_1]] [\mathbb{F}[r_1] : \mathbb{F}]$$

and since the degree of $q(x) > 1$ and $q(x)$ is irreducible, this shows that $[\mathbb{F}[r_1] : \mathbb{F}] = m > 1$ and so

$$[\mathbb{K} : \mathbb{F}[r_1]] < [\mathbb{K} : \mathbb{F}]$$

Therefore, by induction, each of these $k \leq m = [\mathbb{F}[r_1] : \mathbb{F}]$ one to one homomorphisms extends to an isomorphism from $\mathbb{K}$ to $\bar{\mathbb{K}}$ and for each of these ζ_i, there are no more than

$[\mathbb{K} : \mathbb{F}[r_1]]$ of these isomorphisms extending $\mathbb{F}$. If the roots of $\bar{p}(x)$ are distinct, then there are exactly m of these ζ_i and for each, there are $[\mathbb{K} : \mathbb{F}[r_1]]$ extensions. Therefore, if the roots of $\bar{p}(x)$ are distinct, this has identified

$$[\mathbb{K} : \mathbb{F}[r_1]]\, m = [\mathbb{K} : \mathbb{F}[r_1]]\, [\mathbb{F}[r_1] : \mathbb{F}] = [\mathbb{K} : \mathbb{F}]$$

isomorphisms of $\mathbb{K}$ to $\bar{\mathbb{K}}$ which agree with η on $\mathbb{F}$. If the roots of $\bar{p}(x)$ are not distinct, then maybe there are fewer than $[\mathbb{K} : \mathbb{F}]$ extensions of η.

Is this all of them? Suppose ζ is such an isomorphism of $\mathbb{K}$ and $\bar{\mathbb{K}}$. Then consider its restriction to $\mathbb{F}[r_1]$. By Theorem F.4.3, this restriction must coincide with one of the ζ_i chosen earlier. Then by induction, ζ is one of the extensions of the ζ_i just mentioned. ■

Definition F.4.6 *Let $\mathbb{K}$ be a finite dimensional extension of a field $\mathbb{F}$ such that every element of $\mathbb{K}$ is algebraic over $\mathbb{F}$, that is, each element of $\mathbb{K}$ is a root of some polynomial in $\mathbb{F}[x]$. Then $\mathbb{K}$ is called a normal extension if for every $k \in \mathbb{K}$ all roots of the minimal polynomial of k are contained in $\mathbb{K}$.*

So what are some ways to tell a field is a normal extension? It turns out that if $\mathbb{K}$ is a splitting field of $f(x) \in \mathbb{F}[x]$, then $\mathbb{K}$ is a normal extension. I found this in [17]. This is an amazing result.

Proposition F.4.7 *Let $\mathbb{K}$ be a splitting field of $f(x) \in \mathbb{F}[x]$. Then $\mathbb{K}$ is a normal extension. In fact, if $\mathbb{L}$ is an intermediate field between $\mathbb{F}$ and $\mathbb{K}$, then $\mathbb{L}$ is also a normal extension of $\mathbb{F}$.*

Proof: Let $r \in \mathbb{K}$ be a root of $g(x)$, an irreducible monic polynomial in $\mathbb{F}[x]$. It is required to show that every other root of $g(x)$ is in $\mathbb{K}$. Let the roots of $g(x)$ in a splitting field be $\{r_1 = r, r_2, \cdots, r_m\}$. Now $g(x)$ is the minimal polynomial of r_j over $\mathbb{F}$ because $g(x)$ is irreducible. Recall why this was. If $p(x)$ is the minimal polynomial of r_j,

$$g(x) = p(x)\, l(x) + r(x)$$

where $r(x)$ either is 0 or it has degree less than the degree of $p(x)$. However, $r(r_j) = 0$ and this is impossible if $p(x)$ is the minimal polynomial. Hence $r(x) = 0$ and now it follows that $g(x)$ was not irreducible unless $l(x) = 1$.

By Theorem F.4.3, there exists an isomorphism η of $\mathbb{F}[r_1]$ and $\mathbb{F}[r_j]$ which fixes $\mathbb{F}$ and maps r_1 to r_j. Now $\mathbb{K}[r_1]$ and $\mathbb{K}[r_j]$ are splitting fields of $f(x)$ over $\mathbb{F}[r_1]$ and $\mathbb{F}[r_j]$ respectively. By Theorem F.4.5, the two fields $\mathbb{K}[r_1]$ and $\mathbb{K}[r_j]$ are isomorphic, the isomorphism, ζ extending η. Hence

$$[\mathbb{K}[r_1] : \mathbb{K}] = [\mathbb{K}[r_j] : \mathbb{K}]$$

But $r_1 \in \mathbb{K}$ and so $\mathbb{K}[r_1] = \mathbb{K}$. Therefore, $\mathbb{K} = \mathbb{K}[r_j]$ and so r_j is also in $\mathbb{K}$. Thus all the roots of $g(x)$ are actually in $\mathbb{K}$. Consider the last assertion.

Suppose $r = r_1 \in \mathbb{L}$ where the minimal polynomial for r is denoted by $q(x)$. Then letting the roots of $q(x)$ in $\mathbb{K}$ be $\{r_1, \cdots, r_m\}$. By Theorem F.4.3 applied to the identity map on $\mathbb{L}$, there exists an isomorphism $\theta : \mathbb{L}[r_1] \to \mathbb{L}[r_j]$ which fixes $\mathbb{L}$ and takes r_1 to r_j. But this implies that

$$1 = [\mathbb{L}[r_1] : \mathbb{L}] = [\mathbb{L}[r_j] : \mathbb{L}]$$

Hence $r_j \in \mathbb{L}$ also. Since r was an arbitrary element of $\mathbb{L}$, this shows that $\mathbb{L}$ is normal. ■

Definition F.4.8 *When you have $\mathbb{F}[a_1, \cdots, a_m]$ with each a_i algebraic so that $\mathbb{F}[a_1, \cdots, a_m]$ is a field, you could consider*

$$f(x) \equiv \prod_{i=1}^{m} f_i(x)$$

where $f_i(x)$ is the minimal polynomial of a_i. Then if $\mathbb{K}$ is a splitting field for $f(x)$, this $\mathbb{K}$ is called the normal closure. It is at least as large as $\mathbb{F}[a_1, \cdots, a_m]$ and it has the advantage of being a normal extension.

F.5 The Galois Group

In the case where $\mathbb{F} = \bar{\mathbb{F}}$, the above suggests the following definition.

Definition F.5.1 *When $\mathbb{K}$ is a splitting field for a polynomial $p(x)$ having coefficients in $\mathbb{F}$, we say that $\mathbb{K}$ is a splitting field of $p(x)$ over the field $\mathbb{F}$. Let $\mathbb{K}$ be a splitting field of $p(x)$ over the field $\mathbb{F}$. Then $G(\mathbb{K}, \mathbb{F})$ denotes the group of automorphisms of $\mathbb{K}$ which leave $\mathbb{F}$ fixed. For a finite set S, denote by $|S|$ as the number of elements of S. More generally, when $\mathbb{K}$ is a finite extension of $\mathbb{L}$, denote by $G(\mathbb{K}, \mathbb{L})$ the group of automorphisms of $\mathbb{K}$ which leave $\mathbb{L}$ fixed.*

It is shown later that $G(\mathbb{K}, \mathbb{F})$ really is a group according to the strict definition of a group. For right now, just regard it as a set of automorphisms which keeps $\mathbb{F}$ fixed. Theorem F.4.5 implies the following important result.

Theorem F.5.2 *Let $\mathbb{K}$ be a splitting field of $p(x)$ over the field $\mathbb{F}$. Then*

$$|G(\mathbb{K}, \mathbb{F})| \leq [\mathbb{K} : \mathbb{F}]$$

When the roots of $p(x)$ are distinct, equality holds in the above.

So how large is $|G(\mathbb{K}, \mathbb{F})|$ in case $p(x)$ is a polynomial of degree n which has n distinct roots? Let $p(x)$ be a monic polynomial with roots in $\mathbb{K}$, $\{r_1, \cdots, r_n\}$ and suppose that none of the r_i is in $\mathbb{F}$. Thus

$$\begin{aligned} p(x) &= x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n \\ &= \prod_{k=1}^{n} (x - r_k), \ a_i \in \mathbb{F} \end{aligned}$$

Thus $\mathbb{K}$ consists of all rational functions in the $r_1, \cdots, r_n$. Let σ be a mapping from $\{r_1, \cdots, r_n\}$ to $\{r_1, \cdots, r_n\}$, say $r_j \to r_{i_j}$. In other words σ produces a permutation of these roots. Consider the following way of obtaining something in $G(\mathbb{K}, \mathbb{F})$ from σ. If you have a typical thing in $\mathbb{K}$, you can obtain another thing in $\mathbb{K}$ by replacing each r_j with r_{i_j} in a rational function, a quotient of two polynomials which have coefficients in $\mathbb{F}$. Furthermore, if you do this, you can see right away that the resulting map form $\mathbb{K}$ to $\mathbb{K}$ is obviously an automorphism, preserving the operations of multiplication and addition. Does it keep $\mathbb{F}$ fixed? Of course. You don't change the coefficients of the polynomials in the rational function which are always in $\mathbb{F}$. Thus every permutation of the roots determines an automorphism of $\mathbb{K}$. Now suppose σ is an automorphism of $\mathbb{K}$. Does it determine a permutation of the roots?

$$0 = \sigma(p(r_i)) = \sigma(p(\sigma(r_i)))$$

and so $\sigma(r_i)$ is also a root, say r_{i_j}. Thus it is clear that each $\sigma \in G(\mathbb{K}, \mathbb{F})$ determines a permutation of the roots. Since the roots are distinct, it follows that $|G(\mathbb{K}, \mathbb{F})|$ equals the number of permutations of $\{r_1, \cdots, r_n\}$ which is $n!$ and that there is a one to one correspondence between the permutations of the roots and $G(\mathbb{K}, \mathbb{F})$. More will be done on

this later after discussing permutation groups.

This is a good time to make a very important observation about irreducible polynomials.

Lemma F.5.3 *Suppose $q(x) \neq p(x)$ are both irreducible polynomials over a field $\mathbb{F}$. Then for $\mathbb{K}$ a field which contains all the roots of both polynomials, there is no root common to both $p(x)$ and $q(x)$.*

Proof: If $l(x)$ is a monic polynomial which divides them both, then $l(x)$ must equal 1. Otherwise, it would equal $p(x)$ and $q(x)$ which would require these two to be equal. Thus $p(x)$ and $q(x)$ are relatively prime and there exist polynomials $a(x), b(x)$ having coefficients in $\mathbb{F}$ such that

$$a(x) p(x) + b(x) q(x) = 1$$

Now if $p(x)$ and $q(x)$ share a root r, then $(x - r)$ divides both sides of the above in $\mathbb{K}[x]$, but this is impossible. ■

Now here is an important definition of a class of polynomials which yield equality in the inequality of Theorem F.5.2.

Definition F.5.4 *Let $p(x)$ be a polynomial having coefficients in a field $\mathbb{F}$. Also let $\mathbb{K}$ be a splitting field. Then $p(x)$ is separable if it is of the form*

$$p(x) = \prod_{i=1}^{m} q_i(x)^{k_i}$$

where each $q_i(x)$ is irreducible over $\mathbb{F}$ and each $q_i(x)$ has distinct roots in $\mathbb{K}$. From the above lemma, no two $q_i(x)$ share a root. Thus

$$p_1(x) \equiv \prod_{i=1}^{m} q_i(x)$$

has distinct roots in $\mathbb{K}$.

For example, consider the case where $\mathbb{F} = \mathbb{Q}$ and the polynomial is of the form

$$\left(x^2 + 1\right)^2 \left(x^2 - 2\right)^2 = x^8 - 2x^6 - 3x^4 + 4x^2 + 4$$

Then let $\mathbb{K}$ be the splitting field over $\mathbb{Q}$, $\mathbb{Q}\left[i, \sqrt{2}\right]$ .The polynomials $x^2 + 1$ and $x^2 - 2$ are irreducible over $\mathbb{Q}$ and each has distinct roots in $\mathbb{K}$.

This is also a convenient time to show that $G(\mathbb{K}, \mathbb{F})$ for $\mathbb{K}$ a finite extension of $\mathbb{F}$ really is a group. First, here is the definition.

Definition F.5.5 *A group G is a nonempty set with an operation, denoted here as $\cdot$ such that the following axioms hold.*

1. *For $\alpha, \beta, \gamma \in G, (\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$. We usually don't bother to write the $\cdot$.*
2. *There exists $\iota \in G$ such that $\alpha\iota = \iota\alpha = \alpha$*
3. *For every $\alpha \in G$, there exists $\alpha^{-1} \in G$ such that $\alpha\alpha^{-1} = \alpha^{-1}\alpha = \iota$.*

Then why is $G \equiv G(\mathbb{K}, \mathbb{F})$, where $\mathbb{K}$ is a finite extension of $\mathbb{F}$, a group? If you simply look at the automorphisms of $\mathbb{K}$ then it is obvious that this is a group with the operation being composition. Also, from Theorem F.4.5 $|G(\mathbb{K}, \mathbb{F})|$ is finite. Clearly $\iota \in G$. It is just the automorphism which takes everything to itself. The operation in this case is just composition. Thus the associative law is obvious. What about the existence of the inverse? Clearly, you can define the inverse of α, but does it fix $\mathbb{F}$? If $\alpha = \iota$, then the inverse is clearly ι. Otherwise, consider $\alpha, \alpha^2, \cdots$. Since $|G(\mathbb{K}, \mathbb{F})|$ is finite, eventually there is a repeat. Thus $\alpha^m = \alpha^n$, $n > m$. Simply multiply on the left by $\left(\alpha^{-1}\right)^m$ to get $\iota = \alpha\alpha^{n-m}$. Hence α^{-1} is a suitable power of α and so α^{-1} obviously leaves $\mathbb{F}$ fixed. Thus $G(\mathbb{K}, \mathbb{F})$ which has been called a group all along, really is a group.

Then the following corollary is the reason why separable polynomials are so important. Also, one can show that if $\mathbb{F}$ contains a field which is isomorphic to $\mathbb{Q}$ then every polynomial with coefficients in $\mathbb{F}$ is separable. This will be done later after presenting the big results. This is equivalent to saying that the field has characteristic zero. In addition, the property of being separable holds in other situations which are described later.

Corollary F.5.6 *Let $\mathbb{K}$ be a splitting field of $p(x)$ over the field $\mathbb{F}$. Assume $p(x)$ is separable. Then*

$$|G(\mathbb{K}, \mathbb{F})| = [\mathbb{K} : \mathbb{F}]$$

Proof: Just note that $\mathbb{K}$ is also the splitting field of $p_1(x)$, the product of the distinct irreducible factors and that from Lemma F.5.3, $p_1(x)$ has distinct roots. Thus the conclusion follows from Theorem F.4.5. ■

What if $\mathbb{L}$ is an intermediate field between $\mathbb{F}$ and $\mathbb{K}$? Then $p_1(x)$ still has coefficients in $\mathbb{L}$ and distinct roots in $\mathbb{K}$ and so it also follows that

$$|G(\mathbb{K}, \mathbb{L})| = [\mathbb{K} : \mathbb{L}]$$

Definition F.5.7 *Let G be a group of automorphisms of a field $\mathbb{K}$. Then denote by $\mathbb{K}_G$ the fixed field of G. Thus*

$$\mathbb{K}_G \equiv \{x \in \mathbb{K} : \sigma(x) = x \textit{ for all } \sigma \in G\}$$

Thus there are two new things, the fixed field of a group of automorphisms H denoted by $\mathbb{K}_H$ and the Gallois group $G(\mathbb{K}, \mathbb{L})$. How are these related? First here is a simple lemma which comes from the definitions.

Lemma F.5.8 *Let $\mathbb{K}$ be an algebraic extension of $\mathbb{L}$ (each element of $\mathbb{L}$ is a root of some polynomial in $\mathbb{L}$) for $\mathbb{L}, \mathbb{K}$ fields. Then*

$$G(\mathbb{K}, \mathbb{L}) = G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K},\mathbb{L})}\right)$$

Proof: It is clear that $\mathbb{L} \subseteq \mathbb{K}_{G(\mathbb{K},\mathbb{L})}$ because if $r \in \mathbb{L}$ then by definition, everything in $G(\mathbb{K}, \mathbb{L})$ fixes r and so r is in $\mathbb{K}_{G(\mathbb{K},\mathbb{L})}$. Therefore,

$$G(\mathbb{K}, \mathbb{L}) \supseteq G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K},\mathbb{L})}\right).$$

Now let $\sigma \in G(\mathbb{K}, \mathbb{L})$ then it is one of the automorphisms of $\mathbb{K}$ which fixes everything in the fixed field of $G(\mathbb{K}, \mathbb{L})$. Thus, by definition, $\sigma \in G\left(\mathbb{K}, \mathbb{K}_{G(\mathbb{K},\mathbb{L})}\right)$ and so the two are the same. ■

Now the following says that you can start with $\mathbb{L}$, go to the group $G(\mathbb{K}, \mathbb{L})$ and then to the fixed field of this group and end up back where you started. More precisely,

Proposition F.5.9 *If $\mathbb{K}$ is a splitting field of $p(x)$ over the field $\mathbb{F}$ for separable $p(x)$, and if $\mathbb{L}$ is a field between $\mathbb{K}$ and $\mathbb{F}$, then $\mathbb{K}$ is also a splitting field of $p(x)$ over $\mathbb{L}$ and also*

$$\mathbb{L} = \mathbb{K}_{G(\mathbb{K},\mathbb{L})}$$

Proof: By the above lemma, and Corollary F.5.6,

$$\begin{aligned}|G(\mathbb{K},\mathbb{L})| &= [\mathbb{K}:\mathbb{L}] = \left[\mathbb{K}:\mathbb{K}_{G(\mathbb{K},\mathbb{L})}\right]\left[\mathbb{K}_{G(\mathbb{K},\mathbb{L})}:\mathbb{L}\right] \\ &= \left|G\left(\mathbb{K},\mathbb{K}_{G(\mathbb{K},\mathbb{L})}\right)\right|\left[\mathbb{K}_{G(\mathbb{K},\mathbb{L})}:\mathbb{L}\right] = |G(\mathbb{K},\mathbb{L})|\left[\mathbb{K}_{G(\mathbb{K},\mathbb{L})}:\mathbb{L}\right]\end{aligned}$$

which shows that $\left[\mathbb{K}_{G(\mathbb{K},\mathbb{L})}:\mathbb{L}\right]=1$ and so, since $\mathbb{L}\subseteq\mathbb{K}_{G(\mathbb{K},\mathbb{L})}$, it follows that $\mathbb{L}=\mathbb{K}_{G(\mathbb{K},\mathbb{L})}$. ■

This has shown the following diagram in the context of $\mathbb{K}$ being a splitting field of a separable polynomial over $\mathbb{F}$ and $\mathbb{L}$ being an intermediate field.

$$\mathbb{L}\to G(\mathbb{K},\mathbb{L})\to\mathbb{K}_{G(\mathbb{K},\mathbb{L})}=\mathbb{L}$$

In particular, every intermediate field is a fixed field of a subgroup of $G(\mathbb{K},\mathbb{F})$. Is every subgroup of $G(\mathbb{K},\mathbb{F})$ obtained in the form $G(\mathbb{K},\mathbb{L})$ for some intermediate field? This involves another estimate which is apparently due to Artin. I also found this in [17]. There is more there about some of these things than what I am including.

Theorem F.5.10 *Let $\mathbb{K}$ be a field and let G be a finite group of automorphisms of $\mathbb{K}$. Then*

$$[\mathbb{K}:\mathbb{K}_G]\leq|G|$$

Proof: Let $G=\{\sigma_1,\cdots,\sigma_n\},\sigma_1=\iota$ the identity map and suppose $\{u_1,\cdots,u_m\}$ is a linearly independent set in $\mathbb{K}$ with respect to the field $\mathbb{K}_G$. Suppose $m>n$. Then consider the system of equations

$$\begin{array}{c}\sigma_1(u_1)x_1+\sigma_1(u_2)x_2+\cdots+\sigma_1(u_m)x_m=0\\ \sigma_2(u_1)x_1+\sigma_2(u_2)x_2+\cdots+\sigma_2(u_m)x_m=0\\ \vdots\\ \sigma_n(u_1)x_1+\sigma_n(u_2)x_2+\cdots+\sigma_n(u_m)x_m=0\end{array}\tag{6.18}$$

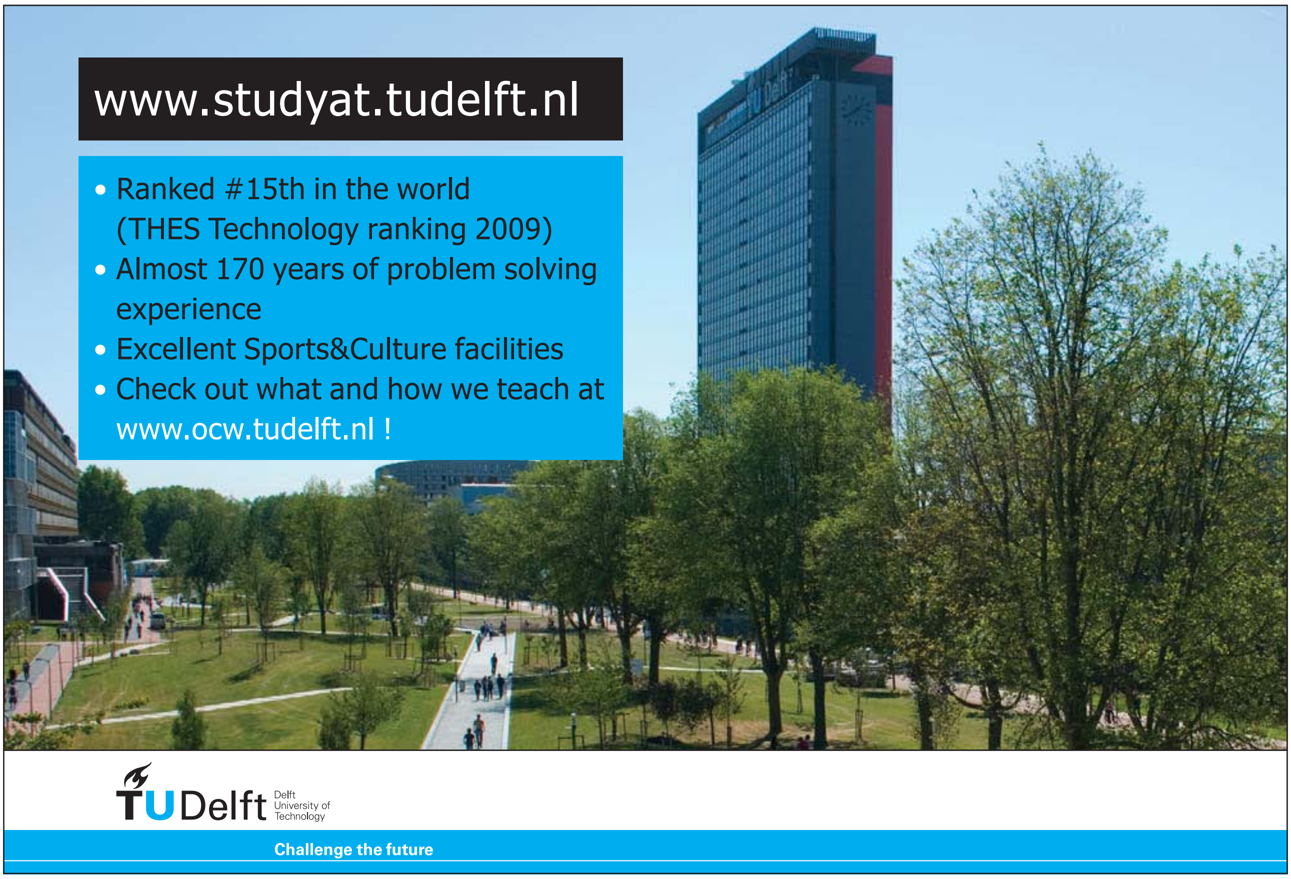
which is of the form $M\mathbf{x}=\mathbf{0}$ for $\mathbf{x}\in\mathbb{K}^m$. Since M has more columns than rows, there exists a nonzero solution $\mathbf{x}\in\mathbb{K}^m$ to the above system. Note that this could not happen if $\mathbf{x}\in\mathbb{K}_G^m$ because of independence of $\{u_1,\cdots,u_m\}$ and the fact that $\sigma_1=\iota$. Let the solution $\mathbf{x}$ be one which has the least possible number of nonzero entries. Without loss of generality, some $x_k=1$ for some k. If $\sigma_r(x_k)=x_k$ for all x_k and for each r, then the x_k are each in $\mathbb{K}_G$ and so the first equation above would be impossible as just noted. Therefore, there exists $l\neq k$ and σ_r such that $\sigma_r(x_l)\neq x_l$. For purposes of illustration, say $l>k$. Now do σ_r to both sides of all the above equations. This yields, after re ordering the resulting equations a list of equations of the form

$$\begin{aligned}
\sigma_1(u_1)\,\sigma_r(x_1)+\cdots+\sigma_1(u_k)\,1+\cdots+\sigma_1(u_l)\,\sigma_r(x_l)+\cdots+\sigma_1(u_m)\,\sigma_r(x_m)&=0\\
\sigma_2(u_1)\,\sigma_r(x_1)+\cdots+\sigma_2(u_k)\,1+\cdots+\sigma_2(u_l)\,\sigma_r(x_l)+\cdots+\sigma_2(u_m)\,\sigma_r(x_m)&=0\\
&\vdots\\
\sigma_n(u_1)\,\sigma_r(x_1)+\cdots+\sigma_n(u_k)\,1+\cdots+\sigma_n(u_l)\,\sigma_r(x_l)+\cdots+\sigma_n(u_m)\,\sigma_r(x_m)&=0
\end{aligned}$$

This is because $\sigma(1)=1$ if σ is an automorphism. The original system in 6.18 is of the form

$$\begin{aligned}
\sigma_1(u_1)\,x_1+\cdots+\sigma_1(u_k)\,1+\cdots+\sigma_1(u_l)\,x_l+\cdots+\sigma_1(u_m)\,x_m&=0\\
\sigma_2(u_1)\,x_1+\cdots+\sigma_2(u_k)\,1+\cdots+\sigma_1(u_l)\,x_l+\cdots+\sigma_2(u_m)\,x_m&=0\\
&\vdots\\
\sigma_n(u_1)\,x_1+\cdots+\sigma_n(u_k)\,1+\cdots+\sigma_1(u_l)\,x_l+\cdots+\sigma_n(u_m)\,x_m&=0
\end{aligned}$$

Now replace the k^{th} equation with the difference of the k^{th} equations in the original system and the one in which σ_r was done to both sides of the equations. Since $\sigma_r(x_l)\neq x_l$ the



result will be a linear system of the form $M\mathbf{y} = \mathbf{0}$ where $\mathbf{y} \neq \mathbf{0}$ has fewer nonzero entries than $\mathbf{x}$, contradicting the choice of $\mathbf{x}$. ■

With the above estimate, here is another relation between the fixed fields and subgroups of automorphisms. It doesn't seem to depend on anything being a splitting field of a separable polynomial.

Proposition F.5.11 *Let H be a finite group of automorphisms defined on a field $\mathbb{K}$. Then for $\mathbb{K}_H$ the fixed field,*

$$G(\mathbb{K}, \mathbb{K}_H) = H$$

Proof: If $\sigma \in H$, then by definition, $\sigma \in G(\mathbb{K}, \mathbb{K}_H)$. It is clear that $H \subseteq G(\mathbb{K}, \mathbb{K}_H)$. Then by Proposition F.5.10 and Theorem F.5.2,

$$|H| \geq [\mathbb{K} : \mathbb{K}_H] \geq |G(\mathbb{K}, \mathbb{K}_H)| \geq |H|$$

and so $H = G(\mathbb{K}, \mathbb{K}_H)$. ■

This leads to the following interesting correspondence in the case where $\mathbb{K}$ is a splitting field of a separable polynomial over a field $\mathbb{F}$.

$$\text{Fixed fields} \quad \begin{array}{ccc} \mathbb{L} & \overset{\beta}{\to} & G(\mathbb{K}, \mathbb{L}) \\ \mathbb{K}_H & \overset{\alpha}{\leftarrow} & H \end{array} \quad \text{Subgroups of } G(\mathbb{K}, \mathbb{F}) \tag{6.19}$$

Then $\alpha\beta\mathbb{L} = \mathbb{L}$ and $\beta\alpha H = H$. Thus there exists a one to one correspondence between the fixed fields and the subgroups of $G(\mathbb{K}, \mathbb{F})$. The following theorem summarizes the above result.

Theorem F.5.12 *Let $\mathbb{K}$ be a splitting field of a separable polynomial over a field $\mathbb{F}$. Then there exists a one to one correspondence between the fixed fields $\mathbb{K}_H$ for H a subgroup of $G(\mathbb{K}, \mathbb{F})$ and the intermediate fields as described in the above. $H_1 \subseteq H_2$ if and only if $\mathbb{K}_{H_1} \supseteq \mathbb{K}_{H_2}$. Also*

$$|H| = [\mathbb{K} : \mathbb{K}_H]$$

Proof: The one to one correspondence is established above. The claim about the fixed fields is obvious because if the group is larger, then the fixed field must get harder because it is more difficult to fix everything using more automorphisms than with fewer automorphisms. Consider the estimate. From Theorem F.5.10, $|H| \geq [\mathbb{K} : \mathbb{K}_H]$. But also, $H = G(\mathbb{K}, \mathbb{K}_H)$ from Proposition F.5.11 $G(\mathbb{K}, \mathbb{K}_H) = H$ and from Theorem F.5.2,

$$|H| = |G(\mathbb{K}, \mathbb{K}_H)| \leq [\mathbb{K} : \mathbb{K}_H].$$

■

Note that from the above discussion, when $\mathbb{K}$ is a splitting field of $p(x) \in \mathbb{F}[x]$, this implies that if $\mathbb{L}$ is an intermediate field, then it is also a fixed field of a subgroup of $G(\mathbb{K}, \mathbb{F})$. In fact, from the above,

$$\mathbb{L} = \mathbb{K}_{G(\mathbb{K}, \mathbb{L})}$$

If H is a subgroup, then it is also the Galois group

$$H = G(\mathbb{K}, \mathbb{K}_H).$$

By Proposition F.4.7, each of these intermediate fields $\mathbb{L}$ is also a normal extension of $\mathbb{F}$. Now there is also something called a normal subgroup which will end up corresponding with these normal field extensions consisting of the intermediate fields between $\mathbb{F}$ and $\mathbb{K}$.

F.6 Normal Subgroups

When you look at groups, one of the first things to consider is the notion of a normal subgroup.

Definition F.6.1 *Let G be a group. Then a subgroup N is said to be a normal subgroup if whenever $\alpha \in G$,*

$$\alpha^{-1} N \alpha \subseteq N$$

The important thing about normal subgroups is that you can define the quotient group G/N.

Definition F.6.2 *Let N be a subgroup of G. Define an equivalence relation $\sim$ as follows.*

$$\alpha \sim \beta \text{ means } \alpha^{-1}\beta \in N$$

Why is this an equivalence relation? It is clear that $\alpha \sim \alpha$ because $\alpha^{-1}\alpha = \iota \in N$ since N is a subgroup. If $\alpha \sim \beta$, then $\alpha^{-1}\beta \in N$ and so, since N is a subgroup,

$$\left(\alpha^{-1}\beta\right)^{-1} = \beta^{-1}\alpha \in N$$

which shows that $\beta \sim \alpha$. Now suppose $\alpha\alpha \sim \beta$, then $\alpha^{-1}\beta \in N$ and so, since N is a subgroup,

$$\left(\alpha^{-1}\beta\right)^{-1} = \beta^{-1}\alpha \in N$$

which shows that $\beta \sim \alpha$. Now suppose $\alpha \sim \beta$ and $\beta \sim \gamma$. Then $\alpha^{-1}\beta \in N$ and $\beta^{-1}\gamma \in N$. Then since N is a subgroup

$$\alpha^{-1}\beta\beta^{-1}\gamma = \alpha^{-1}\gamma \in N$$

and so $\alpha \sim \gamma$ which shows that it is an equivalence relation as claimed. Denote by $[\alpha]$ the equivalence class determined by α.

Now in the case of N a **normal** subgroup, you can consider the quotient group.

Definition F.6.3 *Let N be a normal subgroup of a group G and define G/N as the set of all equivalence classes with respect to the above equivalence relation. Also define*

$$[\alpha]\,[\beta] \equiv [\alpha\beta]$$

Proposition F.6.4 *The above definition is well defined and it also makes G/N into a group.*

Proof: First consider the claim that the definition is well defined. Suppose then that $\alpha \sim \alpha'$ and $\beta \sim \beta'$. It is required to show that

$$[\alpha\beta] = \left[\alpha'\beta'\right]$$

But

$$\begin{aligned}(\alpha\beta)^{-1}\alpha'\beta' &= \beta^{-1}\alpha^{-1}\alpha'\beta' = \beta^{-1}\overbrace{\alpha^{-1}\alpha'}^{\in N}\beta' \\ &= \overbrace{\beta^{-1}\left(\alpha^{-1}\alpha'\right)\beta}^{\in N}\overbrace{\beta^{-1}\beta'}^{\in N} = n_1 n_2 \in N\end{aligned}$$

Thus the operation is well defined. Clearly the identity is $[\iota]$ where ι is the identity in G and the inverse is $\left[\alpha^{-1}\right]$ where α^{-1} is the inverse for α in G. The associative law is also obvious. ■

Note that it was important to have the subgroup be normal in order to have the operation defined on the quotient group.

F.7 Normal Extensions And Normal Subgroups

When $\mathbb{K}$ is a splitting field of a separable polynomial having coefficients in $\mathbb{F}$, the intermediate fields are each normal extensions from the above. If $\mathbb{L}$ is one of these, what about $G(\mathbb{L},\mathbb{F})$? is this a normal subgroup of $G(\mathbb{K},\mathbb{F})$? More generally, consider the following diagram which has now been established in the case that $\mathbb{K}$ is a splitting field of a separable polynomial in $\mathbb{F}[x]$.

$$\begin{array}{llllll} \mathbb{F}\equiv\mathbb{L}_0 & \subseteq\mathbb{L}_1 & \subseteq\mathbb{L}_2 & \cdots & \subseteq\mathbb{L}_{k-1} & \subseteq\mathbb{L}_k\equiv\mathbb{K} \\ G(\mathbb{F},\mathbb{F})=\{\iota\} & \subseteq G(\mathbb{L}_1,\mathbb{F}) & \subseteq G(\mathbb{L}_2,\mathbb{F}) & \cdots & \subseteq G(\mathbb{L}_{k-1},\mathbb{F}) & \subseteq G(\mathbb{K},\mathbb{F}) \end{array} \tag{6.20}$$

The intermediate fields $\mathbb{L}_i$ are each normal extensions of $\mathbb{F}$ each element of $\mathbb{L}_i$ being algebraic. As implied in the diagram, there is a one to one correspondence between the intermediate fields and the Galois groups displayed. Is $G(\mathbb{L}_{j-1},\mathbb{F})$ a normal subgroup of $G(\mathbb{L}_j,\mathbb{F})$?

Let $\sigma\in G(\mathbb{L}_j,\mathbb{F})$ and let $\eta\in G(\mathbb{L}_{j-1},\mathbb{F})$. Then is $\sigma^{-1}\eta\sigma\in G(\mathbb{L}_{j-1},\mathbb{F})$? Let $r=r_1$ be something in $\mathbb{L}_{j-1}$ and let $\{r_1,\cdots,r_m\}$ be the roots of the minimal polynomial of r denoted by $f(x)$, a polynomial having coefficients in $\mathbb{F}$. Then $0=\sigma f(r)=f(\sigma(r))$ and so $\sigma(r)=r_j$ for some j. Since $\mathbb{L}_{j-1}$ is normal, $\sigma(r)\in\mathbb{L}_{j-1}$. Therefore, it is fixed by η. It follows that

$$\sigma^{-1}\eta\sigma(r)=\sigma^{-1}\sigma(r)=r$$

and so $\sigma^{-1}\eta\sigma\in G(\mathbb{L}_{j-1},\mathbb{F})$. Thus $G(\mathbb{L}_{j-1},\mathbb{F})$ is a normal subgroup of $G(\mathbb{L}_j,\mathbb{F})$ as hoped.

This leads to the following fundamental theorem of Galois theory.

Theorem F.7.1 *Let $\mathbb{K}$ be a splitting field of a separable polynomial $p(x)$ having coefficients in a field $\mathbb{F}$. Let $\{\mathbb{L}_i\}_{i=0}^k$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$ as shown above in 6.20. Then each of these is a normal extension of $\mathbb{F}$ and the Galois group $G(\mathbb{L}_{j-1},\mathbb{F})$ is a normal subgroup of $G(\mathbb{L}_j,\mathbb{F})$. In addition to this,*

$$G(\mathbb{L}_j,\mathbb{F})\simeq G(\mathbb{K},\mathbb{F})/G(\mathbb{K},\mathbb{L}_j)$$

where the symbol $\simeq$ indicates the two spaces are isomorphic.

Proof: All that remains is to check that the above isomorphism is valid. Let

$$\theta : G(\mathbb{K},\mathbb{F})/G(\mathbb{K},\mathbb{L}_j) \to G(\mathbb{L}_j,\mathbb{F}),\ \theta[\sigma] \equiv \sigma|_{\mathbb{L}_j}$$

In other words, this is just the restriction of σ to $\mathbb{L}_j$. Is θ well defined? If $[\sigma_1] = [\sigma_2]$, then by definition, $\sigma_1\sigma_2^{-1} \in G(\mathbb{K},\mathbb{L}_j)$ and so $\sigma_1\sigma_2^{-1}$ fixes everything in $\mathbb{L}_j$. It follows that the restrictions of σ_1 and σ_2 to $\mathbb{L}_j$ are equal. Therefore, θ is well defined. It is obvious that θ is a homomorphism. Why is θ onto? This follows right away from Theorem F.4.5. Note that $\mathbb{K}$ is the splitting field of $p(x)$ over $\mathbb{L}_j$ since $L_j \supseteq \mathbb{F}$. Also if $\sigma \in G(\mathbb{L}_j,\mathbb{F})$ so it is an automorphism of $\mathbb{L}_j$, then, since it fixes $\mathbb{F}$, $p(x) = \bar{p}(x)$ in that theorem. Thus σ extends to ζ, an automorphism of $\mathbb{K}$. Thus $\theta\zeta = \sigma$. Why is θ one to one? If $\theta[\sigma] = \theta[\alpha]$, this means $\sigma = \alpha$ on $\mathbb{L}_j$. Thus $\sigma\alpha^{-1}$ is the identity on $\mathbb{L}_j$. Hence $\sigma\alpha^{-1} \in G(\mathbb{K},\mathbb{L}_j)$ which is what it means for $[\sigma] = [\alpha]$. ■

There is an immediate application to a description of the normal closure of an algebraic extension $\mathbb{F}[a_1, a_2, \cdots, a_m]$. To begin with, recall the following definition.

Definition F.7.2 *When you have* $\mathbb{F}[a_1, \cdots, a_m]$ *with each* a_i *algebraic so that* $\mathbb{F}[a_1, \cdots, a_m]$

is a field, you could consider

$$f(x) \equiv \prod_{i=1}^{m} f_i(x)$$

where $f_i(x)$ is the minimal polynomial of a_i. Then if $\mathbb{K}$ is a splitting field for $f(x)$, this $\mathbb{K}$ is called the normal closure. It is at least as large as $\mathbb{F}[a_1, \cdots, a_m]$ and it has the advantage of being a normal extension.

Let $G(\mathbb{K}, \mathbb{F}) = \{\eta_1, \eta_2, \cdots, \eta_m\}$. The conjugate fields are the fields

$$\eta_j(\mathbb{F}[a_1, \cdots, a_m])$$

Thus each of these fields is isomorphic to any other and they are all contained in $\mathbb{K}$. Let $\mathbb{K}'$ denote the smallest field contained in $\mathbb{K}$ which contains all of these conjugate fields. Note that if $k \in \mathbb{F}[a_1, \cdots, a_m]$ so that $\eta_i(k)$ is in one of these conjugate fields, then $\eta_j \eta_i(k)$ is also in a conjugate field because $\eta_j \eta_i$ is one of the automorphisms of $G(\mathbb{K}, \mathbb{F})$. Let

$$S = \left\{k \in \mathbb{K}' : \eta_j(k) \in \mathbb{K}' \text{ each } j\right\}.$$

Then from what was just shown, each conjugate field is in S. Suppose $k \in S$. What about k^{-1}?

$$\eta_j(k)\, \eta_j\left(k^{-1}\right) = \eta_j\left(kk^{-1}\right) = \eta_j(1) = 1$$

and so $\left(\eta_j(k)\right)^{-1} = \eta_j\left(k^{-1}\right)$. Now $\left(\eta_j(k)\right)^{-1} \in \mathbb{K}'$ because $\mathbb{K}'$ is a field. Therefore, $\eta_j\left(k^{-1}\right) \in \mathbb{K}'$. Thus S is closed with respect to taking inverses. It is also closed with respect to products. Thus it is clear that S is a field which contains each conjugate field. However, $\mathbb{K}'$ was defined as the smallest field which contains the conjugate fields. Therefore, $S = \mathbb{K}'$ and so this shows that each η_j maps $\mathbb{K}'$ to itself while fixing $\mathbb{F}$. Thus $G(\mathbb{K}, \mathbb{F}) \subseteq G(\mathbb{K}', \mathbb{F})$. However, since $\mathbb{K}' \subseteq \mathbb{K}$, it follows that also $G(\mathbb{K}', \mathbb{F}) \subseteq G(\mathbb{K}, \mathbb{F})$. Therefore, $G(\mathbb{K}', \mathbb{F}) = G(\mathbb{K}, \mathbb{F})$, and by the one to one correspondence between the intermediate fields and the Galois groups, it follows that $\mathbb{K}' = \mathbb{K}$. This proves the following lemma.

Lemma F.7.3 *Let $\mathbb{K}$ denote the normal extension of $\mathbb{F}[a_1, \cdots, a_m]$ with each a_i algebraic so that $\mathbb{F}[a_1, \cdots, a_m]$ is a field. Thus $\mathbb{K}$ is the splitting field of the product of the minimal polynomials of the a_i. Then $\mathbb{K}$ is also the smallest field containing the conjugate fields $\eta_j(\mathbb{F}[a_1, \cdots, a_m])$ for $\{\eta_1, \eta_2, \cdots, \eta_m\} = G(\mathbb{K}, \mathbb{F})$.*

F.8 Conditions For Separability

So when is it that a polynomial having coefficients in a field $\mathbb{F}$ is separable? It turns out that this is always the case for fields which are enough like the rational numbers. It involves considering the derivative of a polynomial. In doing this, there will be no analysis used, just the rule for differentiation which we all learned in calculus. Thus the derivative is defined as follows.

$$\begin{aligned} & \left(a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0\right)' \\ \equiv \quad & n a_n x^{n-1} + a_{n-1}(n-1)\, x^{n-2} + \cdots + a_1 \end{aligned}$$

This kind of formal manipulation is what most students do anyway, never thinking about where it comes from. Here na_n means to add a_n to itself n times. With this definition, it is clear that the usual rules such as the product rule hold. This discussion follows [17].

Definition F.8.1 *A field has characteristic 0 if $na \neq 0$ for all $n \in \mathbb{N}$ and $a \neq 0$. Otherwise a field $\mathbb{F}$ has characteristic p if $p \cdot 1 = 0$ for $p \cdot 1$ defined as 1 added to itself p times and p is the smallest positive integer for which this takes place.*

Note that with this definition, some of the terms of the derivative of a polynomial could vanish in the case that the field has characteristic p. I will go ahead and write them anyway. For example, if the field has characteristic p, then

$$(x^p - a)' = 0$$

because formally it equals $p \cdot 1x^{p-1} = 0x^{p-1}$, the 1 being the 1 in the field.

Note that the field $\mathbb{Z}_p$ does not have characteristic 0 because $p \cdot 1 = 0$. Thus not all fields have characteristic 0.

How can you tell if a polynomial has no repeated roots? This is the content of the next theorem.

Theorem F.8.2 *Let $p(x)$ be a monic polynomial having coefficients in a field $\mathbb{F}$, and let $\mathbb{K}$ be a field in which $p(x)$ factors*

$$p(x) = \prod_{i=1}^{n} (x - r_i), \quad r_i \in \mathbb{K}.$$

Then the r_i are distinct if and only if $p(x)$ and $p'(x)$ are relatively prime over $\mathbb{F}$.

Proof: Suppose first that $p'(x)$ and $p(x)$ are relatively prime over $\mathbb{F}$. Since they are not both zero, there exists polynomials $a(x), b(x)$ having coefficients in $\mathbb{F}$ such that

$$a(x)p(x) + b(x)p'(x) = 1$$

Now suppose $p(x)$ has a repeated root r. Then in $\mathbb{K}[x]$,

$$p(x) = (x - r)^2 g(x)$$

and so $p'(x) = 2(x - r)g(x) + (x - r)^2 g'(x)$. Then in $\mathbb{K}[x]$,

$$a(x)(x - r)^2 g(x) + b(x)\Big(2(x - r)g(x) + (x - r)^2 g'(x)\Big) = 1$$

Then letting $x = r$, it follows that $0 = 1$. Hence $p(x)$ has no repeated roots.

Next suppose there are no repeated roots of $p(x)$. Then

$$p'(x) = \sum_{i=1}^{n} \prod_{j \neq i} (x - r_j)$$

$p'(x)$ cannot be zero in this case because

$$p'(r_n) = \prod_{j=1}^{n-1} (r_n - r_j) \neq 0$$

because it is the product of nonzero elements of $\mathbb{K}$. Similarly no term in the sum for $p'(x)$ can equal zero because

$$\prod_{j \neq i} (r_i - r_j) \neq 0.$$

Then if $q(x)$ is a monic polynomial of degree larger than 1 which divides $p(x)$, then the roots of $q(x)$ in $\mathbb{K}$ are a subset of $\{r_1, \cdots, r_n\}$. Without loss of generality, suppose these roots of $q(x)$ are $\{r_1, \cdots, r_k\}$, $k \leq n-1$, since $q(x)$ divides $p'(x)$ which has degree at most $n-1$. Then $q(x) = \prod_{i=1}^{k}(x - r_i)$ but this fails to divide $p'(x)$ as polynomials in $\mathbb{K}[x]$ and so $q(x)$ fails to divide $p'(x)$ as polynomials in $\mathbb{F}[x]$ either. Therefore, $q(x) = 1$ and so the two are relatively prime. ■

The following lemma says that the usual calculus result holds in case you are looking at polynomials with coefficients in a field of characteristic 0.

Lemma F.8.3 *Suppose that $\mathbb{F}$ has characteristic 0. Then if $f'(x) = 0$, it follows that $f(x)$ is a constant.*

Proof: Suppose

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

Then take the derivative $n-1$ times to find that a_n multiplied by a positive integer ma_n equals 0. Therefore, $a_n = 0$ because, by assumption $ma_n \neq 0$ if $a_n \neq 0$. Now repeat the argument with

$$f_1(x) = a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

and continue this way to find that $f(x) = a_0 \in \mathbb{F}$. ■

Now here is a major result which applies to fields of characteristic 0.

Theorem F.8.4 *If $\mathbb{F}$ is a field of characteristic 0, then every polynomial $p(x)$, having coefficients in $\mathbb{F}$ is separable.*

Proof: It is required to show that the irreducible factors of $p(x)$ have distinct roots in $\mathbb{K}$ a splitting field for $p(x)$. So let $q(x)$ be an irreducible monic polynomial. If $l(x)$ is a monic polynomial of positive degree which divides both $q(x)$ and $q'(x)$, then since $q(x)$ is irreducible, it must be the case that $l(x) = q(x)$ which forces $q(x)$ to divide $q'(x)$. However, the degree of $q'(x)$ is less than the degree of $q(x)$ so this is impossible. Hence $l(x) = 1$ and so $q'(x)$ and $q(x)$ are relatively prime which implies that $q(x)$ has distinct roots. ■

It follows that the above theory all holds for any field of characteristic 0. For example, if the field is $\mathbb{Q}$ then everything holds.

Proposition F.8.5 *If a field $\mathbb{F}$ has characteristic p, then p is a prime.*

Proof: First note that if $n \cdot 1 = 0$, if and only if for all $a \neq 0, n \cdot a = 0$ also. This just follows from the distributive law and the definition of what is meant by $n \cdot 1$, meaning that you add 1 to itself n times. Suppose then that there are positive integers, each larger than 1 n, m such that $nm \cdot 1 = 0$. Then grouping the terms in the sum associated with $nm \cdot 1$, it follows that $n(m \cdot 1) = 0$. If the characteristic of the field is nm, this is a contradiction because then $m \cdot 1 \neq 0$ but n times it is, implying that $n < nm$ but $n \cdot a = 0$ for a nonzero a. Hence $n \cdot 1 = 0$ showing that mn is not the characteristic of the field after all. ■

Definition F.8.6 *A field $\mathbb{F}$ is called perfect if every polynomial $p(x)$ having coefficients in $\mathbb{F}$ is separable.*

The above shows that fields of characteristic 0 are perfect. The above theory about Galois groups and fixed fields all works for perfect fields. What about fields of characteristic p where p is a prime? The following interesting lemma has to do with a nonzero $a \in \mathbb{F}$ having a p^{th} root in $\mathbb{F}$.

Lemma F.8.7 *Let $\mathbb{F}$ be a field of characteristic p. Let $a \neq 0$ where $a \in \mathbb{F}$. Then either $x^p - a$ is irreducible or there exists $b \in \mathbb{F}$ such that $x^p - a = (x-b)^p$.*

Proof: Suppose that $x^p - a$ is not irreducible. Then $x^p - a = g(x) f(x)$ where the degree of $g(x)$, k is less than p and at least as large as 1. Then let b be a root of $g(x)$. Then $b^p - a = 0$. Therefore,

$$x^p - a = x^p - b^p = (x-b)^p .$$

That is right. $x^p - b^p = (x-b)^p$ just like many beginning calculus students believe. It happens because of the binomial theorem and the fact that the other terms have a factor of p. Hence

$$x^p - a = (x-b)^p = g(x) f(x)$$

and so $g(x)$ divides $(x-b)^p$ which requires that $g(x) = (x-b)^k$ since $g(x)$ has degree k. It follows, since $g(x)$ is given to have coefficients in $\mathbb{F}$, that $b^k \in \mathbb{F}$. Also $b^p \in \mathbb{F}$. Since k, p are relatively prime, due to the fact that $k < p$ with p prime, there are integers m, n such that

$$1 = mk + np$$

Then from what you mean by raising b to an integer power and the usual rules of exponents for integer powers,

$$b = \left(b^k\right)^m \left(b^p\right)^n \in \mathbb{F}.$$

■

So when is a field of characteristic p perfect? As observed above, for a field of characteristic p,

$$(a+b)^p = a^p + b^p.$$

Also,

$$(ab)^p = a^p b^p$$

It follows that $a \to a^p$ is a homomorphism. This is also one to one because, as mentioned above

$$(a-b)^p = a^p - b^p$$

Therefore, if $a^p = b^p$, it follows that $a = b$. Therefore, this homomorphism is also one to one.

Let $\mathbb{F}^p$ be the collection of a^p where $a \in \mathbb{F}$. Then clearly $\mathbb{F}^p$ is a subfield of $\mathbb{F}$ because it is the image of a one to one homomorphism. What follows is the condition for a field of characteristic p to be perfect.

Theorem F.8.8 *Let $\mathbb{F}$ be a field of characteristic p. Then $\mathbb{F}$ is perfect if and only if $\mathbb{F} = \mathbb{F}^p$.*

Proof: Suppose $\mathbb{F} = \mathbb{F}^p$ first. Let $f(x)$ be an irreducible polynomial over $\mathbb{F}$. By Theorem F.8.2, if $f'(x)$ and $f(x)$ are relatively prime over $\mathbb{F}$ then $f(x)$ has no repeated roots. Suppose then that the two polynomials are not relatively prime. If $d(x)$ divides both $f(x)$ and $f'(x)$ with degree of $d(x) \geq 1$. Then, since $f(x)$ is irreducible, it follows that $d(x)$ is a multiple of $f(x)$ and so $f(x)$ divides $f'(x)$ which is impossible unless $f'(x) = 0$. But if $f'(x) = 0$, then $f(x)$ must be of the form

$$a_0 + a_1 x^p + a_2 x^{2p} + \cdots + a_n x^{np}$$

since if it had some other nonzero term with exponent not a multiple of p then $f'(x)$ could not equal zero since you would have something surviving in the expression for the derivative after taking out multiples of p which is like

$$kax^{k-1}$$

where $a \neq 0$ and $k < p$. Thus $ka \neq 0$. Hence the form of $f(x)$ is as indicated above.

If $a_k = b_k^p$ for some $b_k \in \mathbb{F}$, then the expression for $f(x)$ is

$$\begin{aligned} & b_0^p + b_1^p x^p + b_2^p x^{2p} + \cdots + b_n^p x^{np} \\ = \quad & \left(b_0 + b_1 x + b_x x^2 + \cdots + b_n x^n\right)^p \end{aligned}$$

because of the fact noted earlier that $a \to a^p$ is a homomorphism. However, this says that $f(x)$ is not irreducible after all. It follows that there exists a_k such that $a_k \notin \mathbb{F}^p$ contrary to the assumption that $\mathbb{F} = \mathbb{F}^p$. Hence the greatest common divisor of $f'(x)$ and $f(x)$ must be 1.

Next consider the other direction. Suppose $\mathbb{F} \neq \mathbb{F}^p$. Then there exists $a \in \mathbb{F} \setminus \mathbb{F}^p$. Consider the polynomial $x^p - a$. As noted above, its derivative equals 0. Therefore, $x^p - a$ and its derivative cannot be relatively prime. In fact, $x^p - a$ would divide both. ■

Now suppose $\mathbb{F}$ is a finite field. If $n \cdot 1$ is never equal to 0 then, since the field is finite, $k \cdot 1 = m \cdot 1$, for some $k < m$. $m > k$, and $(m-k) \cdot 1 = 0$ which is a contradiction. Hence $\mathbb{F}$ is a field of characteristic p for some prime p, by Proposition F.8.5. The mapping $a \to a^p$ was shown to be a homomorphism which is also one to one. Therefore, $\mathbb{F}^p$ is a subfield of $\mathbb{F}$. It follows that it has characteristic q for some q a prime. However, this requires $q = p$ and so $\mathbb{F}^p = \mathbb{F}$. Then the following corollary is obtained from the above theorem.

Corollary F.8.9 *If $\mathbb{F}$ is a finite field, then $\mathbb{F}$ is perfect.*

With this information, here is a convenient version of the fundamental theorem of Galois theory.

Theorem F.8.10 *Let $\mathbb{K}$ be a splitting field of any polynomial $p(x) \in \mathbb{F}[x]$ where $\mathbb{F}$ is either of characteristic 0 or of characteristic p with $\mathbb{F}^p = \mathbb{F}$. Let $\{\mathbb{L}_i\}_{i=0}^k$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$. Then each of these is a normal extension of $\mathbb{F}$ and the Galois group $G(\mathbb{L}_{j-1}, \mathbb{F})$ is a normal subgroup of $G(\mathbb{L}_j, \mathbb{F})$. In addition to this,*

$$G(\mathbb{L}_j, \mathbb{F}) \simeq G(\mathbb{K}, \mathbb{F}) / G(\mathbb{K}, \mathbb{L}_j)$$

where the symbol $\simeq$ indicates the two spaces are isomorphic.

F.9 Permutations

Let $\{a_1, \cdots, a_n\}$ be a set of distinct elements. Then a permutation of these elements is usually thought of as a list in a particular order. Thus there are exactly $n!$ permutations of a set having n distinct elements. With this definition, here is a simple lemma.

Lemma F.9.1 *Every permutation can be obtained from every other permutation by a finite number of switches.*

Proof: This is obvious if $n = 1$ or 2. Suppose then that it is true for sets of $n-1$ elements. Take two permutations of $\{a_1, \cdots, a_n\}, P_1, P_2$. To get from P_1 to P_2 using switches, first make a switch to obtain the last element in the list coinciding with the last element of P_2. By induction, there are switches which will arrange the first $n-1$ to the right order. ■

It is customary to consider permutations in terms of the set $I_n \equiv \{1, \cdots, n\}$ to be more specific. Then one can think of a given permutation as a mapping σ from this set I_n to itself which is one to one and onto. In fact, $\sigma(i) \equiv j$ where j is in the i^{th} position. Often people write such a σ in the following form

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ i_1 & i_2 & \cdots & i_n \end{pmatrix} \tag{6.21}$$

An easy way to understand the above permutation is through the use of matrix multiplication by permutation matrices. The above vector $(i_1, \cdots, i_n)^T$ is obtained by

$$\begin{pmatrix} \mathbf{e}_{i_1} & \mathbf{e}_{i_2} & \cdots & \mathbf{e}_{i_n} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \tag{6.22}$$

This can be seen right away from looking at a simple example or by using the definition of matrix multiplication directly.

Definition F.9.2 *The sign of the permutation 6.21 is defined as the determinant of the above matrix in 6.22.*

In other words, the sign of the permutation

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ i_1 & i_2 & \cdots & i_n \end{pmatrix}$$

equals $\operatorname{sgn}(i_1, \cdots, i_n)$ defined earlier in Lemma 3.3.1.

Note that from the fact that the determinant is well defined and its properties, the sign of a permutation is 1 if and only if the permutation is produced by an even number of switches and that the number of switches used to produce a given permutation must be either even or odd. Of course a switch is a permutation itself and this is called a transposition. Note also that all these matrices are orthogonal matrices so to take the inverse, it suffices to take a transpose, the inverse also being a permutation matrix.

The resulting group consisting of the permutations of I_n is called S_n. An important idea is the notion of a cycle. Let σ be a permutation, a one to one and onto function defined on I_n. A cycle is of the form

$$\left(k, \sigma(k), \sigma^2(k), \sigma^3(k), \cdots, \sigma^{m-1}(k)\right), \ \sigma^m(k) = k.$$

The last condition must hold for some m because I_n is finite. Then a cycle can be considered as a permutation as follows. Let $(i_1, i_2, \cdots, i_m)$ be a cycle. Then define σ by $\sigma(i_1) = i_2, \sigma(i_2) = i_3, \cdots, \sigma(i_m) = i_1$, and if $k \notin \{i_1, i_2, \cdots, i_m\}$, then $\sigma(k) = k$.

Note that if you have two cycles, $(i_1, i_2, \cdots, i_m), (j_1, j_2, \cdots, j_m)$ which are disjoint in the sense that

$$\{i_1, i_2, \cdots, i_m\} \cap \{j_1, j_2, \cdots, j_m\} = \emptyset,$$

then they commute. It is then clear that every permutation can be represented in a unique way by disjoint cycles. Start with 1 and form the cycle determined by 1. Then start with the smallest $k \in I_n$ which was not included and begin a cycle starting with this. Continue this way. Use the convention that (k) is just the identity. This representation is unique up to order of the cycles which does not matter because they commute. Note that a transposition can be written as (a, b).

A cycle can be written as a product of non disjoint transpositions.

$$(i_1, i_2, \cdots, i_m) = (i_{m-1}, i_m) \cdots (i_2, i_m)(i_1, i_m)$$

Thus if m is odd, the permutation has sign 1 and if m is even, the permutation has sign -1. Also, it is clear that the inverse of the above permutation is $(i_1, i_2, \cdots, i_m)^{-1} = (i_m, \cdots, i_2, i_1)$.

Definition F.9.3 *A_n is the subgroup of S_n such that for $\sigma \in A_n$, σ is the product of an even number of transpositions. It is called the alternating group.*

The following important result is useful in describing A_n.

Proposition F.9.4 *Let $n \geq 3$. Then every permutation in A_n is the product of 3 cycles and the identity.*

Proof: In case $n = 3$, you can list all of the permutations in A_n

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

In terms of cycles, these are

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

You can easily check that they are inverses of each other. Now suppose $n \geq 4$. The permutations in A_n are defined as the product of an even number of transpositions. There are two cases. The first case is where you have two transpositions which share a number,

$$(a,c)(c,b) = (a,c,b)$$

Thus when they share a number, the product is just a 3 cycle. Next suppose you have the product of two transpositions which are disjoint. This can happen because $n \geq 4$. First note that

$$(a,b) = (c,b)(b,a,c) = (c,b,a)(c,a)$$

Therefore,

$$\begin{aligned} (a,b)(c,d) &= (c,b,a)(c,a)(a,d)(d,c,a) \\ &= (c,b,a)(c,a,d)(d,c,a) \end{aligned}$$

and so every product of disjoint transpositions is the product of 3 cycles. ■

Lemma F.9.5 *If $n \geq 5$, then if B is a normal subgroup of A_n, and B is not the identity, then B must contain a 3 cycle.*

Proof: Let α be the permutation in B which is "closest" to the identity without being the identity. That is, out of all permutations which are not the identity, this is one which has the most fixed points or equivalently moves the fewest numbers. Then α is the product of disjoint cycles. Suppose that the longest cycle is the first one and it has at least four numbers. Thus

$$\alpha = (i_1, i_2, i_3, i_4, \cdots, m) \gamma_1 \cdots \gamma_p$$

Since B is normal,

$$\alpha_1 \equiv (i_3, i_2, i_1)(i_1, i_2, i_3, i_4, \cdots, m)(i_1, i_2, i_3) \gamma_1 \cdots \gamma_p \in A_m$$

Then consider $\alpha_1 \alpha^{-1} =$

$$(i_3, i_2, i_1)(i_1, i_2, i_3, i_4, \cdots, m)(i_1, i_2, i_3)(m, \cdots i_4, i_3, i_2, i_1)$$

Then for this permutation, $i_1 \to i_3, i_2 \to i_2, i_3 \to i_4, i_4 \to i_1$. The other numbers not in $\{i_1, i_2, i_3, i_4\}$ are fixed, and in addition i_2 is fixed which did not happen with α. Therefore, this new permutation moves only 3 numbers. Since it is assumed that $m \geq 4$, this is a contradiction to α fixing the most points. It follows that

$$\alpha = (i_1, i_2, i_3) \gamma_1 \cdots \gamma_p \tag{6.23}$$

or else

$$\alpha = (i_1, i_2) \gamma_1 \cdots \gamma_p \tag{6.24}$$

In the first case, say $\gamma_1 = (i_4, i_5, \cdots)$. Multiply as follows $\alpha_1 =$

$$(i_4, i_2, i_1)(i_1, i_2, i_3)(i_4, i_5, \cdots) \gamma_2 \cdots \gamma_p (i_1, i_2, i_4) \in B$$

Then form $\alpha_1 \alpha^{-1} \in B$ given by

$$(i_4, i_2, i_1)(i_1, i_2, i_3)(i_4, i_5, \cdots) \gamma_2 \cdots \gamma_p (i_1, i_2, i_4) \gamma_p^{-1} \cdots \gamma_1^{-1} (i_3, i_2, i_1)$$

$$= (i_4, i_2, i_1)(i_1, i_2, i_3)(i_4, i_5, \cdots)(i_1, i_2, i_4)(\cdots, i_5, i_4)(i_3, i_2, i_1)$$

Then $i_1 \to i_4, i_2 \to i_3, i_3 \to i_5, i_4 \to i_2, i_5 \to i_1$ and other numbers are fixed. Thus $\alpha_1 \alpha^{-1}$ moves 5 points. However, α moves more than 5 if γ_i is not the identity for any $i \geq 2$. It follows that

$$\alpha = (i_1, i_2, i_3) \gamma_1$$

and γ_1 can only be a transposition. However, this cannot happen because then the above α would not even be in A_n. Therefore, $\gamma_1 = \iota$ and so

$$\alpha = (i_1, i_2, i_3)$$

Thus in this case, B contains a 3 cycle.

Now consider case 6.24. None of the γ_i can be a cycle of length more than 4 since the above argument would eliminate this possibility. If any has length 3 then the above argument implies that α equals this 3 cycle. It follows that each γ_i must be a 2 cycle. Say

$$\alpha = (i_1, i_2)(i_3, i_4) \gamma_2 \cdots \gamma_p$$

Thus it moves at least four numbers, greater than four if any of γ_i for $i \geq 2$ is not the identity. As before, $\alpha_1 \equiv$

$$\begin{aligned} & (i_4, i_2, i_1)(i_1, i_2)(i_3, i_4) \gamma_2 \cdots \gamma_p (i_1, i_2, i_4) \\ = \quad & (i_4, i_2, i_1)(i_1, i_2)(i_3, i_4)(i_1, i_2, i_4) \gamma_2 \cdots \gamma_p \in B \end{aligned}$$

Then $\alpha_1 \alpha^{-1} =$

$$\begin{aligned} & (i_4, i_2, i_1)(i_1, i_2)(i_3, i_4)(i_1, i_2, i_4) \gamma_2 \cdots \gamma_p \gamma_p^{-1} \cdots \gamma_2^{-1} \gamma_1^{-1} (i_3, i_4)(i_1, i_2) \\ = \quad & (i_4, i_2, i_1)(i_1, i_2)(i_3, i_4)(i_1, i_2, i_4)(i_3, i_4)(i_1, i_2) \in B \end{aligned}$$

Then $i_1 \to i_3, i_2 \to i_4, i_3 \to i_1, i_4 \to i_3$ so this moves exactly four numbers. Therefore, none of the γ_i is different than the identity for $i \geq 2$. It follows that

$$\alpha = (i_1, i_2)(i_3, i_4) \tag{6.25}$$

and α moves exactly four numbers. Then since B is normal, $\alpha_1 \equiv$

$$(i_5, i_4, i_3)(i_1, i_2)(i_3, i_4)(i_3, i_4, i_5) \in B$$

Then $\alpha_1\alpha^{-1} =$

$$(i_5, i_4, i_3)(i_1, i_2)(i_3, i_4)(i_3, i_4, i_5)(i_3, i_4)(i_1, i_2) \in B$$

Then $i_1 \to i_1, i_2 \to i_2, i_3 \to i_4, i_4 \to i_5, i_5 \to i_3$. Thus this permutation moves only three numbers and so α cannot be of the form given in 6.25. It follows that case 6.24 does not occur. ■

Definition F.9.6 *A group G is said to be simple if its only normal subgroups are itself and the identity.*

The following major result is due to Galois [17].

Proposition F.9.7 *Let $n \geq 5$. Then A_n is simple.*

Proof: From Lemma F.9.5, if B is a normal subgroup of A_n, $B \neq \{\iota\}$, then it contains a 3 cycle $\alpha = (i_1, i_2, i_3)$,

$$\begin{pmatrix} i_1 & i_2 & i_3 \\ i_2 & i_3 & i_1 \end{pmatrix}$$

Now let (j_1, j_2, j_3) be another 3 cycle.

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ j_2 & j_3 & j_1 \end{pmatrix}$$

Let σ be a permutation which satisfies

$$\sigma(i_k) = j_k$$

Then

$$\begin{aligned} \sigma\alpha\sigma^{-1}(j_1) &= \sigma\alpha(i_1) = \sigma(i_2) = j_2 \\ \sigma\alpha\sigma^{-1}(j_2) &= \sigma\alpha(i_2) = \sigma(i_3) = j_3 \\ \sigma\alpha\sigma^{-1}(j_3) &= \sigma\alpha(i_3) = \sigma(i_1) = j_1 \end{aligned}$$

while $\sigma\alpha\sigma^{-1}$ leaves all other numbers fixed. Thus $\sigma\alpha\sigma^{-1}$ is the given 3 cycle. It follows that B contains every 3 cycle. By Proposition F.9.4, this implies $B = A_n$. The only problem is that it is not know whether σ is in A_n. This is where $n \geq 5$ is used. You can modify σ on two numbers not equal to any of the $\{i_1, i_2, i_3\}$ by multiplying by a transposition so that the possibly modified σ is expressed as an even number of transpositions. ■

F.10 Solvable Groups

Recall the fundamental theorem of Galois theory which established a correspondence between the normal subgroups of $G(\mathbb{K},\mathbb{F})$ and normal field extensions. Also recall that if H is one of these normal subgroups, then there was an isomorphism between $G(\mathbb{K}_H,\mathbb{F})$ and the quotient group $G(\mathbb{K},\mathbb{F})/H$. The general idea of a solvable group is given next.

Definition F.10.1 *A group G is solvable if there exists a decreasing sequence of subgroups $\{H_i\}_{i=0}^m$ such that H^i is a normal subgroup of $H^{(i-1)}$,*

$$G = H_0 \supseteq H_1 \supseteq \cdots \supseteq H_m = \{\iota\},$$

and each quotient group H_{i-1}/H_i is Abelian. That is, for $[a],[b] \in H_{i-1}/H_i$,

$$[ab] = [a][b] = [b][a] = [ba]$$

Note that if G is an Abelian group, then it is automatically solvable. In fact you can just consider $H_0 = G, H_1 = \{\iota\}$. In this case H_0/H_1 is just the group G which is Abelian.

There is another idea which helps in understanding whether a group is solvable. It involves the commutator subgroup. This is a very good idea because this subgroup is defined in terms of the group G.

Definition F.10.2 *Let $a, b \in G$ a group. Then the commutator is*

$$aba^{-1}b^{-1}$$

The commutator subgroup, denoted by G', is the smallest subgroup which contains all the commutators.

The nice thing about the commutator subgroup is that it is a normal subgroup. There are also many other amazing properties.

Theorem F.10.3 *Let G be a group and let G' be the commutator subgroup. Then G' is a normal subgroup. Also the quotient group G/G' is Abelian. If H is any normal subgroup of G such that G/H is Abelian, then $H \supseteq G'$. If $G' = \{\iota\}$, then G must be Abelian.*

Proof: The elements of G' are just finite products of things like $aba^{-1}b^{-1}$. Note that the inverse of something like this is also one of these.

$$\left(aba^{-1}b^{-1}\right)^{-1} = bab^{-1}a^{-1}.$$

Thus the collection of finite products is indeed a subgroup. Now consider $h \in G$. Then

$$\begin{aligned} haba^{-1}b^{-1}h^{-1} &= hah^{-1}hbh^{-1}ha^{-1}h^{-1}hb^{-1}h^{-1} \\ &= hah^{-1}hbh^{-1}\left(hah^{-1}\right)^{-1}\left(hbh^{-1}\right)^{-1} \end{aligned}$$

which is another one of those commutators. Thus for c a commutator and $h \in G$,

$$hch^{-1} = c_1$$

another commutator. If you have a product of commutators $c_1 c_2 \cdots c_m$, then

$$hc_1c_2\cdots c_m h^{-1} = \prod_{i=1}^{m} hc_ih^{-1} = \prod_{i=1}^{m} d_i \in G'$$

where the d_i are each commutators. Hence G' is a normal subgroup.

Consider now the quotient group. Is $[g][h] = [h][g]$? In other words, is $[gh] = [hg]$? In other words, is $gh(hg)^{-1} = ghg^{-1}h^{-1} \in G'$? Of course. This is a commutator and $\acute{G}'$ consists of products of these things. Thus the quotient group is Abelian.

Now let H be a normal subgroup of G such that G/H is Abelian. Then if $g, h \in G$,

$$[gh] = [hg], \; gh(hg)^{-1} = ghg^{-1}h^{-1} \in H$$

Thus every commutator is in H and so $H \supseteq G$.

The last assertion is obvious because $G/\{\iota\}$ is isomorphic to G. Also, to say that $G' = \{\iota\}$ is to say that

$$aba^{-1}b^{-1} = \iota$$

which implies that $ab = ba$. ■

Let G be a group and let G' be its commutator subgroup. Then the commutator subgroup of G' is G'' and so forth. To save on notation, denote by $G^{(k)}$ the k^{th} commutator subgroup. Thus you have the sequence

$$G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq G^{(3)} \cdots$$

each $G^{(i)}$ being a normal subgroup of $G^{(i-1)}$ although it is possible that $G^{(i)}$ is not a normal subgroup of G. Then there is a useful criterion for a group to be solvable.

Theorem F.10.4 *Let G be a group. It is solvable if and only if $G^{(k)} = \{\iota\}$ for some k.*

Proof: If $G^{(k)} = \{\iota\}$ then G is clearly solvable because of Theorem F.10.3. The sequence of commutator subgroups provides the necessary sequence of subgroups.

Next suppose that you have

$$G = H_0 \supseteq H_1 \supseteq \cdots \supseteq H_m = \{\iota\}$$

where each is normal in the preceding and the quotient groups are Abelian. Then from Theorem F.10.3, $G^{(1)} \subseteq H_1$. Thus $H_1' \supseteq G^{(2)}$. But also, from Theorem F.10.3, since H_1/H_2 is Abelian,

$$H_2 \supseteq H_1' \supseteq G^{(2)}.$$

Continuing this way $G^{(k)} = \{\iota\}$ for some $k \leq m$. ■

Theorem F.10.5 *If G is a solvable group and if H is a homomorphic image of G, then H is also solvable.*

Proof: By the above theorem, it suffices to show that $H^{(k)} = \{\iota\}$ for some k. Let f be the homomorphism. Then $H' = f(G')$. To see this, consider a commutator of H, $f(a) f(b) f(a)^{-1} f(b)^{-1} = f\left(aba^{-1}b^{-1}\right)$. It follows that $H^{(1)} = f\left(G^{(1)}\right)$. Now continue this way, letting $G^{(1)}$ play the role of G and $H^{(1)}$ the role of H. Thus, since G is solvable, some $G^{(k)} = \{\iota\}$ and so $H^{(k)} = \{\iota\}$ also. ■

Now as an important example, of a group which is not solvable, here is a theorem.

Theorem F.10.6 *For $n \geq 5, S_n$ is not solvable.*

Proof: It is clear that A_n is a normal subgroup of S_n because if σ is a permutation, then it has the same sign as σ^{-1}. Thus $\sigma\alpha\sigma^{-1} \in A_n$ if $\alpha \in A_n$. If H is a normal subgroup of S_n, for which S_n/H is Abelian, then H contains the commutator G'. However, $\alpha\sigma\alpha^{-1}\sigma^{-1} \in A_n$ obviously so $A_n \supseteq S_n'$. By Proposition F.9.7, this forces $S_n' = A_n$. So what is S_n''? If it is S_n, then $S_n^{(k)} \neq \{\iota\}$ for any k and it follows that S_n is not solvable. If $S_n'' = \{\iota\}$, the only other possibility, then $A_n / \{\iota\}$ is Abelian and so A_n is Abelian, but this is obviously false because the cycles $(1,2,3), (2,1,4)$ are both in A_n. However, $(1,2,3)(2,1,4)$ is

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

while $(2,1,4)(1,2,3)$ is

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

■

Note that the above shows that A_n is not Abelian for $n = 4$ also.

F.11 Solvability By Radicals

First of all, there exists a field which has all the n^{th} roots of 1. You could simply define it to be the smallest sub field of $\mathbb{C}$ such that it contains these roots. You could also enlarge it by including some other numbers. For example, you could include $\mathbb{Q}$. Observe that if $\xi \equiv e^{i2\pi/n}$, then $\xi^n = 1$ but $\xi^k \neq 1$ if $k < n$ and that if $k < l < n$, $\xi^k \neq \xi^l$. Such a field has characteristic 0 because for m an integer, $m \cdot 1 \neq 0$. The following is from Herstein [13]. This is the kind of field considered here.

Lemma F.11.1 *Suppose a field $\mathbb{F}$ has all the n^{th} roots of 1 for a particular n and suppose there exists ξ such that the n^{th} roots of 1 are of the form ξ^k for $k = 1, \cdots, n$, the ξ^k being distinct. Let $a \in \mathbb{F}$ be nonzero. Let $\mathbb{K}$ denote the splitting field of $x^n - a$ over $\mathbb{F}$, thus $\mathbb{K}$ is a normal extension of $\mathbb{F}$. Then $\mathbb{K} = \mathbb{F}[u]$ where u is any root of $x^n - a$. The Galois group $G(\mathbb{K}, \mathbb{F})$ is Abelian.*

Proof: Let u be a root of $x^n - a$ and let $\mathbb{K}$ equal $\mathbb{F}[u]$. Then let ξ be the n^{th} root of unity mentioned. Then

$$\left(\xi^k u\right)^n = (\xi^n)^k u^n = a$$

and so each $\xi^k u$ is a root of $x^n - a$ and these are distinct. It follows that $\left\{u, \xi u, \cdots, \xi^{n-1} u\right\}$ are the roots of $x^n - a$ and all are in $\mathbb{F}[u]$. Thus $\mathbb{F}[u] = \mathbb{K}$. Let $\sigma \in G(\mathbb{K}, \mathbb{F})$ and observe that since σ fixes $\mathbb{F}$,

$$0 = \sigma\left(\left(\xi^k u\right)^n - a\right) = \left(\sigma\left(\xi^k u\right)\right)^n - a$$

It follows that σ maps roots of $x^n - a$ to roots of $x^n - a$. Therefore, if σ, α are two elements of $G(\mathbb{K}, \mathbb{F})$, there exist i, j each no larger than $n - 1$ such that

$$\sigma(u) = \xi^i u, \ \alpha(u) = \xi^j u$$

A typical thing in $\mathbb{F}[u]$ is $p(u)$ where $p(x) \in \mathbb{F}[x]$. Then

$$\begin{aligned} \sigma\alpha(p(u)) &= p\left(\xi^j \xi^i u\right) = p\left(\xi^{i+j} u\right) \\ \alpha\sigma(p(u)) &= p\left(\xi^i \xi^j u\right) = p\left(\xi^{i+j} u\right) \end{aligned}$$

Therefore, $G(\mathbb{K}, \mathbb{F})$ is Abelian. ■

Definition F.11.2 *For $\mathbb{F}$ a field, a polynomial $p(x) \in \mathbb{F}[x]$ is solvable by radicals over $\mathbb{F} \equiv \mathbb{F}_0$ if there is a sequence of fields $\mathbb{F}_1 = \mathbb{F}[a_1], \mathbb{F}_2 = \mathbb{F}_1[a_2], \cdots, \mathbb{F}_k = \mathbb{F}_{k-1}[a_k]$ such that for each $i \geq 1, a_i^{k_i} \in \mathbb{F}_{i-1}$ and $\mathbb{F}_k$ contains a splitting field $\mathbb{K}$ for $p(x)$ over $\mathbb{F}$.*

Lemma F.11.3 *In the above definition, you can assume that $\mathbb{F}_k$ is a normal extension of $\mathbb{F}$.*

Proof: First note that $\mathbb{F}_k = \mathbb{F}[a_1, a_2, \cdots, a_k]$. Let $\mathbb{G}$ be the normal extension of $\mathbb{F}_k$. By Lemma F.7.3, $\mathbb{G}$ is the smallest field which contains the conjugate fields

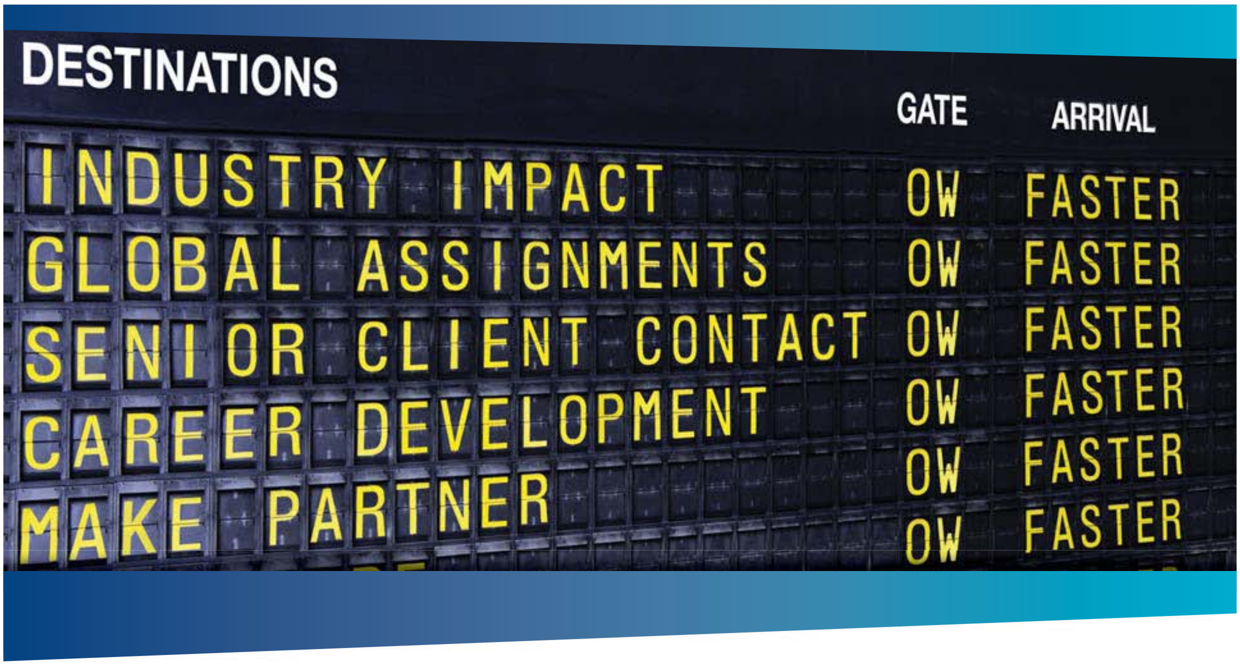
$$\eta_j(\mathbb{F}[a_1, a_2, \cdots, a_k]) = \mathbb{F}[\eta_j a_1, \eta_j a_2, \cdots, \eta_j a_k]$$

for $\{\eta_1, \eta_2, \cdots, \eta_m\} = G(\mathbb{F}_k, \mathbb{F})$. Also, $(\eta_j a_i)^{k_i} = \eta_j\left(a_i^{k_i}\right) \in \eta_j \mathbb{F}_{i-1}, \eta_j \mathbb{F} = \mathbb{F}$. Then

$$\mathbb{G} = \mathbb{F}[\eta_1(a_1), \eta_1(a_2), \cdots, \eta_1(a_k), \eta_2(a_1), \eta_2(a_2), \cdots, \eta_2(a_k) \cdots]$$

and this is a splitting field so is a normal extension. Thus $\mathbb{G}$ could be the new $\mathbb{F}_k$ with respect to a longer sequence but would now be a splitting field. ■

At this point, it is a good idea to recall the big fundamental theorem mentioned above which gives the correspondence between normal subgroups and normal field extensions since


DESTINATIONS
GATE
ARRIVAL
INDUSTRY IMPACT OW FASTER
GLOBAL ASSIGNMENTS OW FASTER
SENIOR CLIENT CONTACT OW FASTER
CAREER DEVELOPMENT OW FASTER
MAKE PARTNER OW FASTER

it is about to be used again.

$$\begin{array}{llllll}\mathbb{F}\equiv\mathbb{F}_0 & \subseteq\mathbb{F}_1 & \subseteq\mathbb{F}_2 & \cdots & \subseteq\mathbb{F}_{k-1} & \subseteq\mathbb{F}_k\equiv\mathbb{K} \\ G(\mathbb{F},\mathbb{F})=\{\iota\} & \subseteq G(\mathbb{F}_1,\mathbb{F}) & \subseteq G(\mathbb{F}_2,\mathbb{F}) & \cdots & \subseteq G(\mathbb{F}_{k-1},\mathbb{F}) & \subseteq G(\mathbb{F}_k,\mathbb{F})\end{array} \tag{6.26}$$

Theorem F.11.4 *Let $\mathbb{K}$ be a splitting field of any polynomial $p(x)\in\mathbb{F}[x]$ where $\mathbb{F}$ is either of characteristic 0 or of characteristic p with $\mathbb{F}^p=\mathbb{F}$. Let $\{\mathbb{F}_i\}_{i=0}^k$ be the increasing sequence of intermediate fields between $\mathbb{F}$ and $\mathbb{K}$. Then each of these is a normal extension of $\mathbb{F}$ and the Galois group $G(\mathbb{F}_{j-1},\mathbb{F})$ is a normal subgroup of $G(\mathbb{F}_j,\mathbb{F})$. In addition to this,*

$$G(\mathbb{F}_j,\mathbb{F})\simeq G(\mathbb{K},\mathbb{F})/G(\mathbb{K},\mathbb{F}_j)$$

where the symbol $\simeq$ indicates the two spaces are isomorphic.

Theorem F.11.5 *Let $f(x)$ be a polynomial in $\mathbb{F}[x]$ where $\mathbb{F}$ is a field of characteristic 0 which contains all n^{th} roots of unity for each $n\in\mathbb{N}$. Let $\mathbb{K}$ be a splitting field of $f(x)$. Then if $f(x)$ is solvable by radicals over $\mathbb{F}$, then the Galois group $G(\mathbb{K},\mathbb{F})$ is a solvable group.*

Proof: Using the definition given above for $f(x)$ to be solvable by radicals, there is a sequence of fields

$$\mathbb{F}_0=\mathbb{F}\subseteq\mathbb{F}_1\subseteq\cdots\subseteq\mathbb{F}_k,\ \mathbb{K}\subseteq\mathbb{F}_k,$$

where $\mathbb{F}_i=\mathbb{F}_{i-1}[a_i]$, $a_i^{k_i}\in\mathbb{F}_{i-1}$, and each field extension is a normal extension of the preceding one. You can assume that $\mathbb{F}_k$ is the splitting field of a polynomial having coefficients in $\mathbb{F}_{j-1}$. This follows from the Lemma F.11.3 above. Then starting the hypotheses of the theorem at $\mathbb{F}_{j-1}$ rather than at $\mathbb{F}$, it follows from Theorem F.11.4 that

$$G(\mathbb{F}_j,\mathbb{F}_{j-1})\simeq G(\mathbb{F}_k,\mathbb{F}_{j-1})/G(\mathbb{F}_k,\mathbb{F}_j)$$

By Lemma F.11.1, the Galois group $G(\mathbb{F}_j,\mathbb{F}_{j-1})$ is Abelian and so this requires that $G(\mathbb{F}_k,\mathbb{F})$ is a solvable group.

Of course $\mathbb{K}$ is a normal field extension of $\mathbb{F}$ because it is a splitting field. By Theorem F.10.5, $G(\mathbb{F}_k,\mathbb{K})$ is a normal subgroup of $G(\mathbb{F}_k,\mathbb{F})$. Also $G(\mathbb{K},\mathbb{F})$ is isomorphic to $G(\mathbb{F}_k,\mathbb{F})/G(\mathbb{F}_k,\mathbb{K})$ and so $G(\mathbb{K},\mathbb{F})$ is a homomorphic image of $G(\mathbb{F}_k,\mathbb{F})$ which is solvable. Here is why this last assertion is so. Define $\theta: G(\mathbb{F}_k,\mathbb{F})/G(\mathbb{F}_k,\mathbb{K})\to G(\mathbb{K},\mathbb{F})$ by $\theta[\sigma]\equiv\sigma|_{\mathbb{K}}$. Then this is clearly a homomorphism if it is well defined. If $[\sigma]=[\alpha]$ this means $\sigma\alpha^{-1}\in G(\mathbb{F}_k,\mathbb{K})$ and so $\sigma\alpha^{-1}$ fixes everything in $\mathbb{K}$ so that θ is indeed well defined. Therefore, by Theorem F.10.5, $G(\mathbb{K},\mathbb{F})$ must also be solvable. ■

Now this result implies that you can't solve the general polynomial equation of degree 5 or more by radicals. Let $\{a_1,a_2,\cdots,a_n\}\subseteq\mathbb{G}$ where $\mathbb{G}$ is some field which contains a field $\mathbb{F}_0$. Let

$$\mathbb{F}\equiv\mathbb{F}_0(a_1,a_2,\cdots,a_n)$$

the field of all rational functions in the numbers $a_1,a_2,\cdots,a_n$. I am using this notation because I don't want to assume the a_i are algebraic over $\mathbb{F}$. Now consider the equation

$$p(t)=t^n-a_1t^{n-1}+a_2t^{n-2}+\cdots\pm a_n.$$

and suppose that $p(t)$ has distinct roots, none of them in $\mathbb{F}$. Let $\mathbb{K}$ be a splitting field for $p(t)$ over $\mathbb{F}$ so that

$$p(t)=\prod_{k=1}^{n}(t-r_i)$$

Then it follows that

$$a_i=s_i(r_1,\cdots,r_n)$$

where the s_i are the elementary symmetric functions defined in Definition F.1.2. For $\sigma \in G(\mathbb{K}, \mathbb{F})$ you can define $\bar{\sigma} \in S_n$ by the rule

$$\bar{\sigma}(k) \equiv j \text{ where } \sigma(r_k) = r_j.$$

Recall that the automorphisms of $G(\mathbb{K}, \mathbb{F})$ take roots of $p(t)$ to roots of $p(t)$. This mapping $\sigma \to \bar{\sigma}$ is onto, a homomorphism, and one to one because the symmetric functions s_i are unchanged when the roots are permuted. Thus a rational function in $s_1, s_2, \cdots, s_n$ is unaffected when the roots r_k are permuted. It follows that $G(\mathbb{K}, \mathbb{F})$ cannot be solvable if $n \geq 5$ because S_n is not solvable.

For example, consider $3x^5 - 25x^3 + 45x + 1$ or equivalently $x^5 - \frac{25}{3}x^3 + 15x + \frac{1}{3}$. It clearly has no rational roots and a graph will show it has 5 real roots. Let $\mathbb{F}$ be the smallest field contained in $\mathbb{C}$ which contains the coefficients of the polynomial and all roots of unity. Then probably none of these roots are in $\mathbb{F}$ and they are all distinct. In fact, it appears that the real numbers which are in $\mathbb{F}$ are rational. Therefore, from the above, none of the roots are solvable by radicals involving numbers from $\mathbb{F}$. Thus none are solvable by radicals using numbers from the smallest field containing the coefficients either.

Bibliography

[1] **Apostol T.,** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Artin M.,** *Algebra,* Pearson *2011.*

[3] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[4] **Baker, A.** *Transcendental Number Theory*, Cambridge University Press 1975.

[5] **Chahal J.S.,** *Historical Perspective of Mathematics 2000 B.C. - 2000 A.D. Kendrick Press, Inc. (2007)*

[6] **Coddington and Levinson**, *Theory of Ordinary Differential Equations* McGraw Hill 1955.

[7] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[8] **Edwards C.H.,** *Advanced Calculus of several Variables,* Dover 1994.

[9] **Friedberg S. Insel A. and Spence L.**, *Linear Algebra*, Prentice Hall, 2003.

[10] **Golub, G. and Van Loan, C.**,*Matrix Computations*, Johns Hopkins University Press, 1996.

[11] **Gurtin M.,** *An introduction to continuum mechanics,* Academic press 1981.

[12] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[13] **Herstein I. N.**, *Topics In Algebra*, Xerox, 1964.

[14] **Hofman K. and Kunze R.,** *Linear Algebra,* Prentice Hall, 1971.

[15] **Householder A.** *The theory of matrices in numberical analysis* , Dover, 1975.

[16] **Horn R. and Johnson C.,** *matrix Analysis,* Cambridge University Press, 1985.

[17] **Jacobsen N.** *Basic Algebra* Freeman 1974.

[18] **Karlin S. and Taylor H.,** *A First Course in Stochastic Processes,* Academic Press, 1975.

[19] **Marcus M., and Minc H.,** *A Survey Of Matrix Theory and Matrix Inequalities,* Allyn and Bacon, INc. Boston, 1964

[20] **Nobel B. and Daniel J.,** *Applied Linear Algebra, Prentice Hall, 1977.*

[21] **E. J. Putzer,** American Mathematical Monthly, Vol. 73 (1966), pp. 2-7.

[22] **Rudin W.**, *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[23] **Rudin W.**, *Functional Analysis, McGraw Hill, 1991.*

[24] **Salas S. and Hille E.**, *Calculus One and Several Variables,* Wiley 1990.

[25] **Strang Gilbert**, *Linear Algebra and its Applications,* Harcourt Brace Jovanovich 1980.

[26] **Wilkinson, J.H.**, *The Algebraic Eigenvalue Problem, Clarendon Press Oxford 1965.*

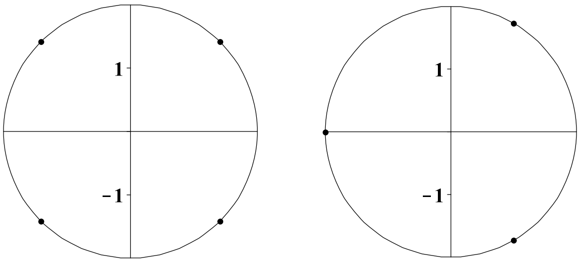
Selected Exercises

Exercises

1.6

1 $(5+i9)^{-1} = \frac{5}{106} - \frac{9}{106}i$

3 $-(1-i)\sqrt{2}, (1+i)\sqrt{2}.$



4

5 If $z \neq 0$, let $\omega = \dfrac{\overline{z}}{|z|}$

7 $\sin(5x) = 5\cos^4 x \sin x - 10\cos^2 x \sin^3 x + \sin^5 x$
$\cos(5x) = \cos^5 x - 10\cos^3 x \sin^2 x + 5\cos x \sin^4 x$

9 $(x+2)\left(x - \left(i\sqrt{3}+1\right)\right)\left(x - \left(1 - i\sqrt{3}\right)\right)$

11 $\left(x - \left((1-i)\sqrt{2}\right)\right)\left(x - \left(-(1+i)\sqrt{2}\right)\right) \cdot$
$\left(x - \left(-(1-i)\sqrt{2}\right)\right)\left(x - \left((1+i)\sqrt{2}\right)\right)$

15 There is no single $\sqrt{-1}$.

Exercises

1.11

1 $x = 2 - 4t, y = -8t, z = t.$

3 These are invalid row operations.

5 $x = 2, y = 0, z = 1.$

7 $x = 2 - 2t, y = -t, z = t.$

9 $x = t, y = s + 2, z = -s, w = s$

Exercises

1.14

4 This makes no sense at all. You can't add different size vectors.

Exercises

1.17

3 $\left|\sum_{k=1}^n \beta_k a_k b_k\right| \leq \left(\sum_{k=1}^n \beta_k |a_k|^2\right)^{1/2} \cdot$
$\left(\sum_{k=1}^n \beta_k |b_k|^2\right)^{1/2}$

4 The inequality still holds. See the proof of the inequality.

Exercises

2.2

2 $A = \frac{A+A^T}{2} + \frac{A-A^T}{2}$

3 You know that $A_{ij} = -A_{ji}$. Let $j = i$ to conclude that $A_{ii} = -A_{ii}$ and so $A_{ii} = 0$.

5 $0' = 0 + 0' = 0.$

6 $0A = (0+0)A = 0A + 0A$. Now add the additive inverse of $0A$ to both sides.

7 $0 = 0A = (1 + (-1))A = A + (-1)A$. Hence, $(-1)A$ is the unique additive inverse of A. Thus $-A = (-1)A$. The additive inverse is unique because if A_1 is an additive inverse, then $A_1 = A_1 + (A + (-A)) = (A_1 + A) + (-A) = -A$.

10 $(A\mathbf{x},\mathbf{y}) = \sum_i (A\mathbf{x})_i y_i = \sum_i \sum_k A_{ik}x_k y_i$
$\left(\mathbf{x},A^T\mathbf{y}\right) = \sum_k x_k \sum_i \left(A^T\right)_{ki} y_i = \sum_k \sum_i x_k A_{ik} y_i$, the same as above. Hence the two are equal.

11 $\left((AB)^T\mathbf{x},\mathbf{y}\right) \equiv$

$(\mathbf{x},(AB)\mathbf{y}) =$

$\left(A^T\mathbf{x},B\mathbf{y}\right) = \left(B^TA^T\mathbf{x},\mathbf{y}\right)$. Since this holds for every $\mathbf{x},\mathbf{y}$, you have for all $\mathbf{y}$, $\left((AB)^T\mathbf{x} - B^TA^T\mathbf{x},\mathbf{y}\right)$.

Let $\mathbf{y} = (AB)^T\mathbf{x} - B^TA^T\mathbf{x}$. Then since $\mathbf{x}$ is arbitrary, the result follows.

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

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

15 It appears that there are 8 ways to do this.

17 $ABB^{-1}A^{-1} = AIA^{-1} = I$

$B^{-1}A^{-1}AB = B^{-1}IB = I$

Then by the definition of the inverse and its uniqueness, it follows that $(AB)^{-1}$ exists and $(AB)^{-1} = B^{-1}A^{-1}$.

19 Multiply both sides on the left by A^{-1}.

21 $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

23 Almost anything works.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 11 & 6 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 7 & 10 \\ 2 & 4 \end{pmatrix}$$

25 $\begin{pmatrix} -z & -w \\ z & w \end{pmatrix}$, z, w arbitrary.

27 $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -2 & 4 & -5 \\ 0 & 1 & -2 \\ 1 & -2 & 3 \end{pmatrix}$

29 Row echelon form: $\begin{pmatrix} 1 & 0 & \frac{5}{3} \\ 0 & 1 & \frac{2}{3} \\ 0 & 0 & 0 \end{pmatrix}$. A has no inverse.

Exercises

2.7

1 Show the map $T : \mathbb{R}^n \to \mathbb{R}^m$ defined by $T(\mathbf{x}) = A\mathbf{x}$ where A is an $m \times n$ matrix and $\mathbf{x}$ is an $m \times 1$ column vector is a linear transformation.

This follows from matrix multiplication rules.

3 Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.

$$\begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}$$

5 Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$.

$$\begin{pmatrix} 2\cos(\pi/3) & -2\sin(\pi/3) \\ 2\sin(\pi/3) & 2\cos(\pi/3) \end{pmatrix} = \begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}$$

7 Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$ and then reflects across the x axis.

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \cos(2\pi/3) & -\sin(2\pi/3) \\ \sin(2\pi/3) & \cos(2\pi/3) \end{pmatrix}$$
$$= \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2}\sqrt{3} \\ -\frac{1}{2}\sqrt{3} & \frac{1}{2} \end{pmatrix}$$

9 Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$ and then reflects across the x axis.

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ -\frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{pmatrix}$$

11 Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/4$.

$$\begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{pmatrix}$$

13 Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/6$.

$$\begin{pmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2}\sqrt{3} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2}\sqrt{3} \end{pmatrix}$$

15 Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $5\pi/12$. **Hint:** Note that $5\pi/12 = 2\pi/3 - \pi/4$.

$$\begin{pmatrix} \cos(2\pi/3) & -\sin(2\pi/3) \\ \sin(2\pi/3) & \cos(2\pi/3) \end{pmatrix}$$
$$\cdot \begin{pmatrix} \cos(-\pi/4) & -\sin(-\pi/4) \\ \sin(-\pi/4) & \cos(-\pi/4) \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{4}\sqrt{2}\sqrt{3} - \frac{1}{4}\sqrt{2} & -\frac{1}{4}\sqrt{2}\sqrt{3} - \frac{1}{4}\sqrt{2} \\ \frac{1}{4}\sqrt{2}\sqrt{3} + \frac{1}{4}\sqrt{2} & \frac{1}{4}\sqrt{2}\sqrt{3} - \frac{1}{4}\sqrt{2} \end{pmatrix}$$

17 Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,5,3)^T$.

$$\frac{1}{35}\begin{pmatrix} 1 & 5 & 3 \\ 5 & 25 & 15 \\ 3 & 15 & 9 \end{pmatrix}$$

19 Give an example of a 2×2 matrix A which has all its entries nonzero and satisfies $A^2 = A$. Such a matrix is called idempotent.

You know it can't be invertible. So try this.

$$\begin{pmatrix} a & a \\ b & b \end{pmatrix}^2 = \begin{pmatrix} a^2+ba & a^2+ba \\ b^2+ab & b^2+ab \end{pmatrix}$$

Let $a^2+ab = a, b^2+ab = b$. A solution which yields a nonzero matrix is

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

21 $x_2 = -\frac{1}{2}t_1 - \frac{1}{2}t_2 - t_3, x_1 = -2t_1 - t_2 + t_3$ where the t_i are arbitrary.

23 $$\begin{pmatrix} -2t_1 - t_2 + t_3 \\ -\frac{1}{2}t_1 - \frac{1}{2}t_2 - t_3 \\ t_1 \\ t_2 \\ t_3 \end{pmatrix} + \begin{pmatrix} 4 \\ 7/2 \\ 0 \\ 0 \\ 0 \end{pmatrix}, t_i \in \mathbb{F}$$

That second vector is a particular solution.

25 Show that the function $T_{\mathbf{u}}$ defined by $T_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$ is also a linear transformation.

This is the sum of two linear transformations so it is obviously linear.

33 Let a basis for W be $\{\mathbf{w}_1, \cdots, \mathbf{w}_r\}$ Then if there exists $\mathbf{v} \in V \setminus W$, you could add in $\mathbf{v}$ to the basis and obtain a linearly independent set of vectors of V which implies that the dimension of V is at least $r+1$ contrary to assumption.

41 Obviously not. Because of the Coriolis force experienced by the fired bullet which is not experienced by the dropped bullet, it will not be as simple as in the physics books. For example, if the bullet is fired East, then $y' \sin\phi > 0$ and will contribute to a force acting on the bullet which has been fired which will cause it to hit the ground faster than the one dropped. Of course at the North pole or the South pole, things should be closer to what is expected in the physics books because there $\sin\phi = 0$. Also, if you fire it North or South, there seems to be no extra force because $y' = 0$.

Exercises

3.2

2 $1 = \det(AA^{-1}) = \det(A)\det(A^{-1})$.

3 $\det(A) = \det(A^T) = \det(-A) = \det(-I)\det(A) = (-1)^n \det(A) = -\det(A)$.

6 Each time you take out an a from a row, you multiply by a the determinant of the matrix which remains. Since there are n rows, you do this n times, hence you get a^n.

9 $\det A = \det(P^{-1}BP) = \det(P^{-1})\det(B)\det(P)$
$= \det(B)\det(P^{-1}P) = \det(B)$.

11 If that determinant equals 0 then the matrix $\lambda I - A$ has no inverse. It is not one to one and so there exists $\mathbf{x} \neq \mathbf{0}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$. Also recall the process for finding the inverse.

13 $$\begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-t}(\cos t + \sin t) & -(\sin t)e^{-t} \\ 0 & -e^{-t}(\cos t - \sin t) & (\cos t)e^{-t} \end{pmatrix}$$

15 You have to have $\det(Q)\det(Q^T) = \det(Q)^2 = 1$ and so $\det(Q) = \pm 1$.

Exercises

3.6

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

6 $\begin{pmatrix} \frac{1}{2}e^{-t} & 0 & \frac{1}{2}e^{-t} \\ \frac{1}{2}\cos t + \frac{1}{2}\sin t & -\sin t & \frac{1}{2}\sin t - \frac{1}{2}\cos t \\ \frac{1}{2}\sin t - \frac{1}{2}\cos t & \cos t & -\frac{1}{2}\cos t - \frac{1}{2}\sin t \end{pmatrix}$

8 $\det(\lambda I - A) = \det\left(\lambda I - S^{-1}BS\right)$

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

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

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

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

9 From the Cayley Hamilton theorem,$A^n + a_{n-1}A^{n-1} + \cdots + a_1A + a_0I = 0$. Also the characteristic polynomial is $\det(tI - A)$ and the constant term is

$(-1)^n \det(A)$. Thus $a_0 \neq 0$ if and only if $\det(A) \neq 0$ if and only if A^{-1} has an inverse. Thus if A^{-1} exists, it follows that

$a_0I = -\left(A^n + a_{n-1}A^{n-1} + \cdots + a_1A\right)$

$= A\left(-A^{n-1} - a_{n-1}A^{n-2} - \cdots - a_1I\right)$ and also

$a_0I = \left(-A^{n-1} - a_{n-1}A^{n-2} - \cdots - a_1I\right)A$ Therefore, the inverse is

$\frac{1}{a_0}\left(-A^{n-1} - a_{n-1}A^{n-2} - \cdots - a_1I\right)$

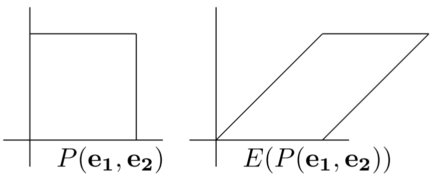
11 Say the characteristic polynomial is $q(t)$ which is of degree 3. Then if $n \geq 3, t^n = q(t)\,l(t) + r(t)$ where the degree of $r(t)$ is either less than 3 or it equals zero. Thus $A^n = q(A)\,l(A) + r(A) = r(A)$ and so all the terms A^n for $n \geq 3$ can be replaced with some $r(A)$ where the degree of $r(t)$ is no more than 2. Thus, assuming there are no convergence issues, the infinite sum must be of the form $\sum_{k=0}^{2} b_kA^k$.

Exercises

4.6

1 A typical thing in $\{A\mathbf{x} : \mathbf{x} \in P(\mathbf{u}_1, \cdots, \mathbf{u}_n)\}$ is $\sum_{k=1}^{n} t_kA\mathbf{u}_k : t_k \in [0,1]$ and so it is just $P(A\mathbf{u}_1, \cdots, A\mathbf{u}_n)$.

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


$P(\mathbf{e_1}, \mathbf{e_2})$
$E(P(\mathbf{e_1}, \mathbf{e_2}))$

5 Here they are.

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

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

So what is the dimension of the span of these? One way to systematically accomplish this is to unravel them and then use the row reduced echelon form. Unraveling these yields the column vectors

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

Then arranging these as the columns of a matrix yields the following along with its row reduced echelon form.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \text{ row echelon form:}$$

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

The dimension is 5.

10 It is because you cannot have more than $\min(m,n)$ nonzero rows in the row reduced echelon form. Recall that the number of pivot columns is the same as the number of nonzero rows from the description of this row reduced echelon form.

11 It follows from the fact that $\mathbf{e}_1, \cdots, \mathbf{e}_m$ occur as columns in row reduced echelon form that the dimension of the column space of A is n and so, since this column space is $A(\mathbb{R}^n)$, it follows that it equals $\mathbb{F}^m$.

12 Since $m > n$ the dimension of the column space of A is no more than n and so the columns of A cannot span $\mathbb{F}^m$.

15 If $\sum_i c_i \mathbf{z}_i = \mathbf{0}$, apply A to both sides to obtain $\sum_i c_i \mathbf{w}_i = \mathbf{0}$. By assumption, each $c_i = 0$.

19 There are more columns than rows and at most m can be pivot columns so it follows at least one column is a linear combination of the others hence A is not one too one.

21 $|\mathbf{b}-A\mathbf{y}|^2 = |\mathbf{b}-A\mathbf{x}+A\mathbf{x}-A\mathbf{y}|^2$

$= |\mathbf{b}-A\mathbf{x}|^2 + |A\mathbf{x}-A\mathbf{y}|^2 + 2(\mathbf{b}-A\mathbf{x}, A(\mathbf{x}-\mathbf{y}))$

$= |\mathbf{b}-A\mathbf{x}|^2 + |A\mathbf{x}-A\mathbf{y}|^2 + 2\left(A^T\mathbf{b}-A^TA\mathbf{x}, (\mathbf{x}-\mathbf{y})\right)$

$= |\mathbf{b}-A\mathbf{x}|^2 + |A\mathbf{x}-A\mathbf{y}|^2$ and so, $A\mathbf{x}$ is closest to $\mathbf{b}$ out of all vectors $A\mathbf{y}$.

27 No. $\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 7 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

29 Let A be an $m \times n$ matrix. Then $\ker(A)$ is a subspace of $\mathbb{F}^n$. Is it true that every subspace of $\mathbb{F}^n$ is the kernel or null space of some matrix? Prove or disprove.

Let M be a subspace of $\mathbb{F}^n$. If it equals $\{\mathbf{0}\}$, consider the matrix I. Otherwise, it has a basis $\{\mathbf{m}_1, \cdots, \mathbf{m}_k\}$. Consider the matrix

$$\begin{pmatrix} \mathbf{m}_1 & \cdots & \mathbf{m}_k & \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is either not there in case $k = n$ or has $n-k$ columns.

30 This is easy to see when you consider that P^{ij} is its own inverse and that P^{ij} multiplied on the right switches the i^{th} and j^{th} columns. Thus you switch the columns and then you switch the rows. This has the effect of switching A_{ii} and A_{jj}. For example,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a & b & c & d \\ e & f & z & h \\ j & k & l & m \\ n & t & h & g \end{pmatrix}.$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} a & d & c & b \\ n & g & h & t \\ j & m & l & k \\ e & h & z & f \end{pmatrix}$$

More formally, the ii^{th} entry of $P^{ij}AP^{ij}$ is

$$\sum_{s,p} P^{ij}_{is} A_{sp} P^{ij}_{pi} = P^{ij}_{ij} A_{jj} P^{ij}_{ji} = A_{ij}$$

31 If A has an inverse, then it is one to one. Hence the columns are independent. Therefore, they are each pivot columns. Therefore, the row reduced echelon form of A is I. This is what was needed for the procedure to work.

Exercises

5.8

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

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

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

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

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

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

9 $\begin{pmatrix} 1 & 2 & 1 & 0 \\ 3 & 0 & 1 & 1 \\ 1 & 0 & 2 & 1 \end{pmatrix}$

$$= \begin{pmatrix} \frac{1}{11}\sqrt{11} & \frac{1}{11}\sqrt{10}\sqrt{11} & 0 \\ \frac{3}{11}\sqrt{11} & -\frac{3}{110}\sqrt{10}\sqrt{11} & -\frac{1}{10}\sqrt{2}\sqrt{5} \\ \frac{1}{11}\sqrt{11} & -\frac{1}{110}\sqrt{10}\sqrt{11} & \frac{3}{10}\sqrt{2}\sqrt{5} \end{pmatrix} \cdot$$

$$\begin{pmatrix} \sqrt{11} & \frac{2}{11}\sqrt{11} & \frac{6}{11}\sqrt{11} & \frac{4}{11}\sqrt{11} \\ 0 & \frac{2}{11}\sqrt{10}\sqrt{11} & \frac{1}{22}\sqrt{10}\sqrt{11} & -\frac{2}{55}\sqrt{10}\sqrt{11} \\ 0 & 0 & \frac{1}{2}\sqrt{2}\sqrt{5} & \frac{1}{5}\sqrt{2}\sqrt{5} \end{pmatrix}$$

Exercises

6.6

1 The maximum is 7 and it occurs when $x_1 = 7, x_2 = 0, x_3 = 0, x_4 = 3, x_5 = 5, x_6 = 0$.

2 Maximize and minimize the following if possible. All variables are nonnegative.

(a) The minimum is -7 and it happens when $x_1 = 0, x_2 = 7/2, x_3 = 0$.

(b) The maximum is 7 and it occurs when $x_1 = 7, x_2 = 0, x_3 = 0$.

(c) The maximum is 14 and it happens when $x_1 = 7, x_2 = x_3 = 0$.

(d) The minimum is 0 when $x_1 = x_2 = 0, x_3 = 1$.

4 Find a solution to the following inequalities for $x, y \geq 0$ if it is possible to do so. If it is not possible, prove it is not possible.

(a) There is no solution to these inequalities with $x_1, x_2 \geq 0$.

(b) A solution is $x_1 = 8/5, x_2 = x_3 = 0$.

(c) There will be no solution to these inequalities for which all the variables are nonnegative.

(d) There is a solution when $x_2 = 2, x_3 = 0, x_1 = 0$.

(e) There is no solution to this system of inequalities because the minimum value of x_7 is not 0.

Exercises

7.3

1 Because the vectors which result are not parallel to the vector you begin with.

3 $\lambda \rightarrow \lambda^{-1}$ and $\lambda \rightarrow \lambda^m$.

5 Let $\mathbf{x}$ be the eigenvector. Then $A^m\mathbf{x} = \lambda^m\mathbf{x}, A^m\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x}$ and so

$$\lambda^m = \lambda$$

Hence if $\lambda \neq 0$, then

$$\lambda^{m-1} = 1$$

and so $|\lambda| = 1$.

7 $\begin{pmatrix} -1 & -1 & 7 \\ -1 & 0 & 4 \\ -1 & -1 & 5 \end{pmatrix}$, eigenvectors:

$\left\{\begin{matrix} 3 \\ 1 \\ 1 \end{matrix}\right\} \leftrightarrow 1, \left\{\begin{matrix} 2 \\ 1 \\ 1 \end{matrix}\right\} \leftrightarrow 2$. This is a defective matrix.

9 $\begin{pmatrix} -7 & -12 & 30 \\ -3 & -7 & 15 \\ -3 & -6 & 14 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix}\right\} \leftrightarrow -1, \left\{\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right\} \leftrightarrow 2$

This matrix is not defective because, even though $\lambda = 1$ is a repeated eigenvalue, it has a 2 dimensional eigenspace.

11 $\begin{pmatrix} 3 & -2 & -1 \\ 0 & 5 & 1 \\ 0 & 2 & 4 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -\frac{1}{2} \\ 1 \end{pmatrix}\right\} \leftrightarrow 3, \left\{\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}\right\} \leftrightarrow 6$

This matrix is not defective.

13 $\begin{pmatrix} 5 & 2 & -5 \\ 12 & 3 & -10 \\ 12 & 4 & -11 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} -\frac{1}{3} \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{5}{6} \\ 0 \\ 1 \end{pmatrix}\right\} \leftrightarrow -1$

This matrix is defective. In this case, there is only one eigenvalue, -1 of multiplicity 3 but the dimension of the eigenspace is only 2.

15 $\begin{pmatrix} 1 & 26 & -17 \\ 4 & -4 & 4 \\ -9 & -18 & 9 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} -\frac{1}{3} \\ \frac{2}{3} \\ 1 \end{pmatrix}\right\} \leftrightarrow 0, \left\{\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}\right\} \leftrightarrow -12,$

$\left\{\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right\} \leftrightarrow 18$

17 $\begin{pmatrix} -2 & 1 & 2 \\ -11 & -2 & 9 \\ -8 & 0 & 7 \end{pmatrix}$, eigenvectors: $\left\{\begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \\ 1 \end{pmatrix}\right\} \leftrightarrow 1$

This is defective.

19 $\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}$, eigenvectors:

$$\left\{\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}\right\} \leftrightarrow 4, \left\{\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2-2i,$$

$$\left\{\begin{pmatrix} i \\ i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2+2i$$

21 $\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}$, eigenvectors:

$$\left\{\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}\right\} \leftrightarrow 4, \left\{\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2-2i,$$

$$\left\{\begin{pmatrix} i \\ i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2+2i$$

23 $\begin{pmatrix} 1 & 1 & -6 \\ 7 & -5 & -6 \\ -1 & 7 & 2 \end{pmatrix}$, eigenvectors:

$$\left\{\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}\right\} \leftrightarrow -6, \left\{\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2-6i,$$

$$\left\{\begin{pmatrix} i \\ i \\ 1 \end{pmatrix}\right\} \leftrightarrow 2+6i$$

This is not defective.

25 First consider the eigenvalue $\lambda = 1$. Then you have $ax_2 = 0, bx_3 = 0$. If neither a nor $b = 0$ then $\lambda = 1$ would be a defective eigenvalue and the matrix would be defective. If $a = 0$, then the dimension of the eigenspace is clearly 2 and so the matrix would be nondefective. If $b = 0$ but $a \neq 0$, then you would have a defective matrix because the eigenspace would have dimension less than 2. If $c \neq 0$, then the matrix is defective. If $c = 0$ and $a = 0$, then it is non defective. Basically, if $a, c \neq 0$, then the matrix is defective.

27 $A(\mathbf{x} + i\mathbf{y}) = (a + ib)(\mathbf{x} + i\mathbf{y})$. Now just take complex conjugates of both sides.

29 Let A be skew symmetric. Then if $\mathbf{x}$ is an eigenvector for λ,

$$\lambda \mathbf{x}^T\bar{\mathbf{x}} = \mathbf{x}^T A^T \bar{\mathbf{x}} = -\mathbf{x}^T A \bar{\mathbf{x}} = -\mathbf{x}^T \bar{\mathbf{x}} \bar{\lambda}$$

and so $\lambda = -\bar{\lambda}$. Thus $a + ib = -(a - ib)$ and so $a = 0$.

31 This follows from the observation that if $A\mathbf{x} = \lambda\mathbf{x}$, then $A\bar{\mathbf{x}} = \overline{\lambda}\bar{\mathbf{x}}$

33 $\left(\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, 1\right), \left(\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix}, \frac{1}{2}\right), \left(\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{3}\right)$

35 $\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}(a\cos(t) + b\sin(t))$,

$\begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}\left(c\sin\left(\sqrt{2}t\right) + d\cos\left(\sqrt{2}t\right)\right)$,

$\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}(e\cos(2t) + f\sin(2t))$where a, b, c, d, e, f are scalars.

Exercises

7.10

1 To get it, you must be able to get the eigenvalues and this is typically not possible.

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

$A_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

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

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

$A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 2 & 0 \end{pmatrix}$. Now it is back to where you started. Thus the algorithm merely bounces between the two matrices $\begin{pmatrix} 0 & -1 \\ 2 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & -2 \\ 1 & 0 \end{pmatrix}$ and so it can't possibly converge.

15 $B(1 + 2i, 6), B(i, 3), B(7, 11)$

19 Gerschgorin's theorem shows that there are no zero eigenvalues and so the matrix is invertible.

21 $6x'^2 + 12y'^2 + 18z'^2$.

23 $(x')^2 + \frac{1}{3}\sqrt{3}x' - 2(y')^2 - \frac{1}{2}\sqrt{2}y' - 2(z')^2 - \frac{1}{6}\sqrt{6}z'$

25 $(0,-1,0)\,(4,-1,0)$ saddle point. $(2,-1,-12)$ local minimum.

27 $(1,1),(-1,1),(1,-1),(-1,-1)$ saddle points. $\left(-\frac{1}{6}\sqrt{5}\sqrt{6},0\right),\left(\frac{1}{6}\sqrt{5}\sqrt{6},0\right)$ Local minimums.

29 Critical points: $(0,1,0)$, Saddle point.

31 ± 1

Exercises

8.4

1 The first three vectors form a basis and the dimension is 3.

3 No. Not a subspace. Consider $(0,0,1,0)$ and multiply by -1.

5 NO. Multiply something by -1.

7 No. Take something nonzero in M where say $u_1 = 1$. Now multiply by 100.

9 Suppose $\{\mathbf{x}_1,\cdots,\mathbf{x}_k\}$ is a set of vectors from $\mathbb{F}^n$. Show that $\mathbf{0}$ is in span $(\mathbf{x}_1,\cdots,\mathbf{x}_k)$.

$\mathbf{0} = \sum_i 0\mathbf{x}_i$

11 It is a subspace. It is spanned by $\begin{pmatrix}3\\1\\1\\0\end{pmatrix},\begin{pmatrix}2\\1\\1\\0\end{pmatrix},\begin{pmatrix}1\\0\\0\\1\end{pmatrix}$. These are also independent so they constitute a basis.

13 Pick n points $\{x_1,\cdots,x_n\}$. Then let $e_i(x)=0$ unless $x=x_i$ when it equals 1. Then $\{e_i\}_{i=1}^n$ is linearly independent, this for any n.

15 $\{1,x,x^2,x^3,x^4\}$

17 $L\left(\sum_{i=1}^n c_i\mathbf{v}_i\right) \equiv \sum_{i=1}^n c_i\mathbf{w}_i$

19 No. There is a spanning set having 5 vectors and this would need to be as long as the linearly independent set.

23 No. It can't. It does not contain $\mathbf{0}$.

25 No. This would lead to $0=1$.The last one must not be a pivot column and the ones to the left must each be pivot columns.

43 Suppose $\sum_{i=1}^n a_i g_i = 0$. Then $0=\sum_i a_i\sum_j A_{ij}f_j = \sum_j f_j \sum_i A_{ij}a_i$. It follows that $\sum_i A_{ij}a_i = 0$ for each j. Therefore, since A^T is invertible, it follows that each $a_i=0$. Hence the functions g_i are linearly independent.

Exercises

9.5

1 This is because ABC is one to one.

7 In the following examples, a linear transformation, T is given by specifying its action on a basis β. Find its matrix with respect to this basis.

(a) $\begin{pmatrix}2&0\\1&1\end{pmatrix}$

(b) $\begin{pmatrix}2&1\\1&0\end{pmatrix}$

(c) $\begin{pmatrix}1&1\\2&-1\end{pmatrix}$

11 $A=\begin{pmatrix}0&1&0&0\\0&0&2&0\\0&0&0&3\\0&0&0&0\end{pmatrix}$

13 $\begin{pmatrix}1&0&2&0&0\\0&1&0&6&0\\0&0&1&0&12\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}$

15 You can see these are not similar by noticing that the second has an eigenspace of dimension equal to 1 so it is not similar to any diagonal matrix which is what the first one is.

19 This is because the general solution is $\mathbf{y}_p+\mathbf{y}$ where $A\mathbf{y}_p=\mathbf{b}$ and $A\mathbf{y}=\mathbf{0}$. Now $A\mathbf{0}=\mathbf{0}$ and so the solution is unique precisely when this is the only solution $\mathbf{y}$ to $A\mathbf{y}=\mathbf{0}$.

Exercises

10.6

2 Consider $\begin{pmatrix}1&1\\0&1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}$. These are both in Jordan form.

8 $\lambda^3 - \lambda^2 + \lambda - 1$

10 λ^2

11 $\lambda^3 - 3\lambda^2 + 14$

16 $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$

Exercises

10.9

4 $\lambda^3 - 3\lambda^2 + 14$

5 $\begin{pmatrix} 0 & 0 & -14 \\ 1 & 0 & 0 \\ 0 & 1 & 3 \end{pmatrix}$

6 $\begin{pmatrix} 0 & 0 & 0 & -3 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -11 \\ 0 & 0 & 1 & 8 \end{pmatrix}$

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

8 $\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbb{Q}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & i & 0 \\ 0 & 0 & -i \end{pmatrix}, \mathbb{Q} + i\mathbb{Q}$

Exercises

11.4

1 $\begin{pmatrix} .6 \\ .9 \\ 1 \end{pmatrix}$

6 The columns are
$\begin{pmatrix} \frac{1}{2^n} - (-1)^n + 1 \\ \frac{2}{2^n} - 3(-1)^n + 1 \\ \frac{1}{2^n} - 2(-1)^n + 1 \\ \frac{1}{2^n} - 2(-1)^n + 1 \end{pmatrix}, \begin{pmatrix} \frac{1}{2^n} - 1 \\ \frac{2}{2^n} - 1 \\ \frac{1}{2^n} - 1 \\ \frac{1}{2^n} - 1 \end{pmatrix},$
$\begin{pmatrix} 0 \\ 0 \\ \frac{1}{2^n} \\ 0 \end{pmatrix}, \begin{pmatrix} (-1)^n - \frac{2}{2^n} + 1 \\ 3(-1)^n - \frac{4}{2^n} + 1 \\ 2(-1)^n - \frac{3}{2^n} + 1 \\ 2(-1)^n - \frac{2}{2^n} + 1 \end{pmatrix}$

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

9 $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

12 Try $\begin{pmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{pmatrix}, \begin{pmatrix} -\frac{1}{2} & -1 \\ 1 & \frac{5}{3} \end{pmatrix}$

Exercises

12.7

1 $\begin{pmatrix} \frac{17}{15} \\ \frac{1}{45} \end{pmatrix}$

2 $\begin{pmatrix} \frac{1}{6}\sqrt{6} \\ \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{6} \end{pmatrix}, \begin{pmatrix} -\frac{1}{30}\sqrt{5}\sqrt{6} \\ \frac{1}{6}\sqrt{5}\sqrt{6} \\ -\frac{1}{15}\sqrt{5}\sqrt{6} \end{pmatrix}, \begin{pmatrix} \frac{2}{5}\sqrt{5} \\ 0 \\ -\frac{1}{5}\sqrt{5} \end{pmatrix}$

3 $|(A\mathbf{x}, \mathbf{y})| \leq (A\mathbf{x}, \mathbf{x})^{1/2} (A\mathbf{y}, \mathbf{y})$

9 $\left\{ \begin{array}{c} 1, \sqrt{3}\,(2x-1), 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \\ , 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \end{array} \right\}$

11 $2x^3 - \frac{9}{7}x^2 + \frac{2}{7}x - \frac{1}{70}$

14 $\begin{pmatrix} -\frac{9}{146}\sqrt{146} \\ \frac{2}{73}\sqrt{146} \\ \frac{7}{146}\sqrt{146} \\ 0 \end{pmatrix}$

16 $|x+y|^2 + |x-y|^2 = (x+y, x+y) + (x-y, x-y)$
$= |x|^2 + |y|^2 + 2(x,y) + |x|^2 + |y|^2 - 2(x,y)\,.$

21 Give an example of two vectors in $\mathbb{R}^4$ $\mathbf{x}, \mathbf{y}$ and a subspace V such that $\mathbf{x} \cdot \mathbf{y} = 0$ but $P\mathbf{x} {\cdot} P\mathbf{y} \neq 0$ where P denotes the projection map which sends $\mathbf{x}$ to its closest point on V.

Try this. V is the span of $\mathbf{e}_1$ and $\mathbf{e}_2$ and $\mathbf{x} = \mathbf{e}_3 + \mathbf{e}_1, \mathbf{y} = \mathbf{e}_4 + \mathbf{e}_1$.

$P\mathbf{x} = (\mathbf{e}_3 + \mathbf{e}_1, \mathbf{e}_1)\,\mathbf{e}_1 + (\mathbf{e}_3 + \mathbf{e}_1, \mathbf{e}_2)\,\mathbf{e}_2 = \mathbf{e}_1$

$P\mathbf{y} = (\mathbf{e}_4 + \mathbf{e}_1, \mathbf{e}_1)\,\mathbf{e}_1 + (\mathbf{e}_4 + \mathbf{e}_1, \mathbf{e}_2)\,\mathbf{e}_2 = \mathbf{e}_1$

$P\mathbf{x} {\cdot} P\mathbf{y} = 1$

22 $y = \frac{13}{5}x - \frac{2}{5}$

Exercises

12.9

1 volume is $\sqrt{218}$

3 0.

Exercises

13.12

13 This is easy because you show it preserves distances.

15 $(A\mathbf{x},\mathbf{x}) = (UDU^*\mathbf{x},\mathbf{x}) = (DU^*\mathbf{x},U^*\mathbf{x}) \geq \delta^2 \left|U^*\mathbf{x}\right|^2 = \delta^2 \left|\mathbf{x}\right|^2$

16 $0 > ((A + A^*)\,\mathbf{x},\mathbf{x}) = (A\mathbf{x},\mathbf{x}) + (A^*\mathbf{x},\mathbf{x})$
$= (A\mathbf{x},\mathbf{x}) + \overline{(A\mathbf{x},\mathbf{x})}$ Now let $A\mathbf{x} = \lambda\mathbf{x}$. Then you get $0 > \lambda\left|\mathbf{x}\right|^2 + \bar{\lambda}\left|\mathbf{x}\right|^2 = \operatorname{Re}(\lambda)\left|\mathbf{x}\right|^2$

19 If $A\mathbf{x} = \lambda\mathbf{x}$, then you can take the norm of both sides and conclude that $|\lambda| = 1$. It follows that the eigenvalues of A are $e^{i\theta}, e^{-i\theta}$ and another one which has magnitude 1 and is real. This can only be 1 or -1. Since the determinant is given to be 1, it follows that it is 1. Therefore, there exists an eigenvector for the eigenvalue 1.

Exercises

14.7

1 $\begin{pmatrix} 0.09 \\ 0.21 \\ 0.43 \end{pmatrix}$

3 $\begin{pmatrix} 4.2373 \times 10^{-2} \\ 7.6271 \times 10^{-2} \\ 0.71186 \end{pmatrix}$

28 You have $H = U^*DU$ where U is unitary and D is a real diagonal matrix. Then you have

$$e^{iH} = U^* \sum_{n=0}^{\infty} \frac{(iD)^n}{n!} U = U^* \begin{pmatrix} e^{i\lambda_1} & & \\ & \ddots & \\ & & e^{i\lambda_n} \end{pmatrix} U$$

and this is clearly unitary because each matrix in the product is.

Exercises

15.3

1 $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1.0 \\ 3 & 1 & 4 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} 0.53491 \\ 0.39022 \\ 0.7494 \end{pmatrix}\right\} \leftrightarrow 6.662,$

$\left\{\begin{pmatrix} 0.13016 \\ 0.83832 \\ -0.52942 \end{pmatrix}\right\} \leftrightarrow 1.6790,$

$\left\{\begin{pmatrix} 0.83483 \\ -0.38073 \\ -0.39763 \end{pmatrix}\right\} \leftrightarrow -1.341$

2 $\begin{pmatrix} 3 & 2 & 1.0 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} 0.57735 \\ 0.57735 \\ 0.57735 \end{pmatrix}\right\} \leftrightarrow 6.0,$

$\left\{\begin{pmatrix} 0.78868 \\ -0.21132 \\ -0.57735 \end{pmatrix}\right\} \leftrightarrow 1.7321,$

$\left\{\begin{pmatrix} 0.21132 \\ -0.78868 \\ 0.57735 \end{pmatrix}\right\} \leftrightarrow -1.7321$

3 $\begin{pmatrix} 3 & 2 & 1.0 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} 0.41601 \\ 0.77918 \\ 0.46885 \end{pmatrix}\right\} \leftrightarrow 7.8730,$

$\left\{\begin{pmatrix} 0.90453 \\ -0.30151 \\ -0.30151 \end{pmatrix}\right\} \leftrightarrow 2.0,$

$\left\{\begin{pmatrix} 9.3568 \times 10^{-2} \\ -0.54952 \\ 0.83022 \end{pmatrix}\right\} \leftrightarrow 0.12702$

4 $\begin{pmatrix} 0 & 2 & 1.0 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, eigenvectors:

$\left\{\begin{pmatrix} 0.28433 \\ 0.81959 \\ 0.49743 \end{pmatrix}\right\} \leftrightarrow 7.5146,$

$$\left\{\begin{pmatrix} 0.209\,84 \\ 0.453\,06 \\ -0.866\,43 \end{pmatrix}\right\} \leftrightarrow 0.189\,11,$$

$$\left\{\begin{pmatrix} 0.935\,48 \\ -0.350\,73 \\ 4.\,316\,8 \times 10^{-2} \end{pmatrix}\right\} \leftrightarrow -0.703\,70$$

5 $\begin{pmatrix} 0 & 2 & 1.0 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, eigenvectors:

$$\left\{\begin{pmatrix} 0.379\,2 \\ 0.584\,81 \\ 0.717\,08 \end{pmatrix}\right\} \leftrightarrow 4.\,975\,4,$$

$$\left\{\begin{pmatrix} 0.814\,41 \\ 0.156\,94 \\ -0.558\,66 \end{pmatrix}\right\} \leftrightarrow -0.300\,56,$$

$$\left\{\begin{pmatrix} 0.439\,25 \\ -0.795\,85 \\ 0.416\,76 \end{pmatrix}\right\} \leftrightarrow -2.\,674\,9$$

6 $|7.\,333\,3 - \lambda_q| \leq 0.471\,41$

7 $|7 - \lambda_q| = 2.\,449\,5$

8 $|\lambda_q - 8| \leq 3.\,266\,0$

9 $-10 \leq \lambda \leq 12$

10 $x^3 + 7x^2 + 3x + 7.0 = 0$, Solution is:

$$\left\{\begin{array}{c} [x = -0.145\,83 + 1.\,011i]\,, \\ [x = -0.145\,83 - 1.\,011i]\,, \\ [x = -6.\,708\,3] \end{array}\right\}$$

11 $-1.\,475\,5 + 1.\,182\,7i$,

$-1.\,475\,5 - 1.\,182\,7i, -0.024\,44 + 0.528\,23i$,

$-0.024\,44 - 0.528\,23i$

12 Let $Q^T AQ = H$ where H is upper Hessenberg. Then take the transpose of both sides. This will show that $H = H^T$ and so H is zero on the top as well.