# AMS792 Matrix Analysis Notes

by Chester Huynh (xhuynh1@jhu.edu)

## Contents

# 0 Chapter 0 – Review of Linear Algebra

## 0.1 Vector Spaces

We quickly define general algebraic structures of interest, paying particular attention to vector spaces and their properties.

**Definition 0.1** (Field)**.** *Fields* are algebraic structures endowed with addition and multiplication, which behave in the same way as we think of these operations for the real or rational numbers.

- Examples include $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Q}$.
- Fields contain the additive/multiplicative identities and additive/multiplicative inverses for each element in the field.
- The commutative, associative, and distributive properties hold.

**Definition 0.2** (Vector Space)**.** *Vector spaces* are defined over some field $\mathbb{F}$ and are endowed with addition and scalar multiplication. Elements of a vector space are called vectors.

- Examples include $\mathbb{R}^n$ over $\mathbb{R}$, $\mathbb{C}^n$ over $\mathbb{C}$, $\mathbb{F}^n$ over $\mathbb{F}$, $C^1[a,b]$ over $\mathbb{R}$, or real polynomials over $\mathbb{R}$.
- Vector spaces contain the contain the additive/multiplicative identities and additive/multiplicative inverses for each element in the vector space.
- The commutative, associative, and distributive properties hold.

**Definition 0.3** (Linear Combination)**.** For a vector space $V$ over a field $\mathbb{F}$, let $v_1, v_2, \dots, v_n \in V$, $a_1, a_2, \dots, a_n \in \mathbb{F}$. We say that

$$a_1v_1 + a_2v_2 + \cdots + a_nv_n$$

is a *linear combination* of $v_1, v_2, \dots, v_n$.

**Example 0.1.** An example of a linear combination:

$$2\begin{bmatrix}1\\2\\2\end{bmatrix} + 3\begin{bmatrix}-3\\2\\1\end{bmatrix} - 2\begin{bmatrix}6\\0\\1\end{bmatrix} = \begin{bmatrix}-19\\10\\5\end{bmatrix}$$

**Definition 0.4** (Span)**.** The *span* of a set of vectors $\{v_1, v_2, \dots, v_n\}$ is the set of all linear combinations of those vectors

$$\operatorname{span}\{v_1, v_2, \dots, v_n\} := \{\alpha_1v_1 + \alpha_2v_2 + \cdots + \alpha_nv_n | \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{F}\}.$$

*Remark.* The zero vector $\vec{0}$ is in the span of any set of vectors, that is, $\vec{0} \in \operatorname{span}\{v_1, v_2, \dots, v_n\}$ for any $\{v_1, v_2, \dots, v_n\}$ since

$$0v_1 + 0v_2 + \dots 0v_n = \vec{0} \qquad \text{(“trivial linear combination”)}$$

**Definition 0.5** (Linear Dependence/Independence)**.** We say that $v_1, v_2, \dots, v_n$ are *linearly dependent* if $\exists \alpha_1, \alpha_2, \dots, \alpha_n$ not all zero such that

$$\alpha_1v_1 + \alpha_2v_2 + \cdots + \alpha_nv_n = \vec{0}$$

*Remark.* Conversely, we say that $v_1, v_2, \dots, v_n$ are *linearly independent* if for any $\alpha_1, \alpha_2, \dots, \alpha_n$,

$$\alpha_1v_1 + \alpha_2v_2 + \cdots + \alpha_nv_n = \vec{0} \quad \Rightarrow \quad \alpha_1 = \alpha_2 = \cdots = \alpha_n$$

*Remark.* $v_1, v_2, \dots, v_n$ are linearly independent $\iff$ one of the $v_i$'s can be expressed as a linear combination of the others.

**Example 0.2.** An example of linear dependence:

$$\begin{aligned}
& 3v_1 + 6v_2 + 0v_3 + 2v_4 = \vec{0} \\
\implies \qquad & v_4 = -\frac{3}{2}v_1 - \frac{6}{2}v_2 - \frac{0}{2}v_3 \\
& \quad = -\frac{3}{2}v_1 - 3v_2 - 0v_3
\end{aligned}$$

In the context of span, $v_4$ is redundant.

**Definition 0.6** (Basis)**.** We say that $v_1, v_2, \ldots, v_n \in V$ form a *(Hamel) basis* for $V$ if

(i) $v_1, v_2, \ldots, v_n$ are linearly independent, and

(ii) $\operatorname{span}\{v_1, v_2, \ldots, v_n\} = V$.

*Remark.* If $v_1, v_2, \ldots, v_n$ are a basis for $V$, then $\forall v \in V$, $\exists! \alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{F}$ such that

$$v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n,$$

that is, every $v \in V$ is expressible as a unique linear combination of $v_1, v_2, \ldots, v_n$.

*Proof of Remark.* Let $v = \sum_{i=1}^n \alpha_i v_i = \sum_{i=1}^n \alpha_i' v_i$. This implies that

$$\vec{0} = v - v = \sum_{i=1}^n (\alpha_i - \alpha_i')v_i$$

By linear independence of a basis, this suggests

$$\begin{aligned}
& \alpha_i - \alpha_i' = 0 \qquad \forall i = 1, 2, \ldots, n \\
\Rightarrow \quad & \alpha_i = \alpha_i' \qquad \forall i = 1, 2, \ldots, n
\end{aligned}$$

□

**Definition 0.7** (Dimension)**.** The cardinality of a basis for $V$ is the *dimension* of $V$.

**Example 0.3.** Consider the following examples relating bases and dimension:

- Two possible bases in $\mathbb{R}^3$ (which has dimension 3):

$$\left\{ \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 7 \\ 9 \\ 13 \end{bmatrix} \right\},$$

$$\left\{ \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 7 \\ 9 \\ 13 \end{bmatrix} \right\}$$

- The vector space of all polynomials has bases which are countably infinite, so the dimension is also countably infinite:

$$\begin{array}{cc}
1 & 1 \\
t & -t \\
t^2 & 2t^2 - t \\
t^3 & 3t^3 - 6 \\
t^4 & -7t^4 + 5t + 3 \\
\vdots & \vdots
\end{array}$$

- $C^1[a, b]$ is an uncountably infinite-dimensional vector space.

We establish the dual structure of $M_{m,n}(\mathbb{F})$. We can think of $A \in M_{m,n}(\mathbb{F})$ as a function "$A$" : $\mathbb{F}^n \to \mathbb{F}^m$, whereby $\forall x \in \mathbb{F}^n$ we have that "$A$"$x \equiv Ax \in \mathbb{F}^m$. Such a function "$A$" is linear.

- Recall that $T : V \to W$ is *linear* if $\forall \alpha, \beta \in \mathbb{F}$, $x, y \in V$, we have that

$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y),$$

  that is all linear combinations are preserved by the function $T$.

- All such linear functions can be represented as matrices because they result in the same outputs given the same inputs

- To assure ourselves that there is a true 1-1 relationship between functions and matrices, we check that the functions give the same output given vectors from some basis set. We say that they are the "same given a basis". Because all other vectors are expressible as a linear combination of the basis vectors, we know the function and matrix operate the same, and therefore, are equiavlent.

**Definition 0.8** (Nullspace/Kernel)**.** For $A \in M_{m,n}(\mathbb{F})$, the *nullspace*, or *kernel*, of $A$ is

$$\text{null}(A) := \left\{ x \in \mathbb{F}^n | Ax = \vec{0} \right\}$$

*Remark.* The nullspace of $A \in M_{m,n}(\mathbb{F})$ is a subspace of $\mathbb{F}^n$.

*Remark.* The nullspace is closed under addition and scalar multiplication.

*Remark.* $\vec{0}$ is always in the nullspace.

**Definition 0.9** (Nullity)**.** The *nullity* of $A$ is dim(null($A$).

**Definition 0.10** (Range/Image)**.** For $A \in M_{m,n}(\mathbb{F})$, the *range*, or *image*, of $A$ is

$$\text{range}(A) := \{Ax | x \in \mathbb{F}^n\}$$

*Remark.* The range of $A \in M_{m,n}(\mathbb{F})$ is a subspace of $\mathbb{F}^m$.

*Remark.* The range is closed under addition and scalar multiplication.

*Remark.* $\vec{0}$ is always in the range.

**Definition 0.11** (Rank)**.** The *rank* of $A$ is dim(range($A$).

**Proposition 0.1.** *Suppose $A \in M_{m,n}(\mathbb{F})$. Then, the columns of $A$ are linearly independent if and only if null(A) $= \{\vec{0}\}$, if and only if $A$ is 1-1 (injective) as a function.*

*Proof of Proposition 0.1.* We first show that null(A) $= \{\vec{0}\}$ if and only if $A$ is 1-1 (injective) as a function. ($\Longrightarrow$) Suppose null$(A) = \{\vec{0}\}$. Let $x, y \in F^n$ such that $Ax = Ay$. We want to show that this implies $x = y$, to show that $A$ is 1-1.

$$\begin{aligned} & Ax - Ay = \vec{0} \\ \Longrightarrow \quad & A(x - y) = \vec{0} \quad \text{(by linearity of } A) \\ \Longrightarrow \quad & x - y = \vec{0} \\ \Longrightarrow \quad & x = y \end{aligned}$$

($\Longleftarrow$) This is clear. If $A$ is 1-1, then we know only the zero vector $\vec{0} \in \mathbb{F}^n$ is mapped to $\vec{0} \in \text{null}(A) \subseteq \mathbb{F}^m$.

Now we prove null$(A) = \{\vec{0}\}$ if and only if the columns of $A$ are linearly independent.

($\Longrightarrow$) Suppose $\text{null}(A) = \{\vec{0}\}$. This means that

$$\begin{aligned}
&\forall x \in \mathbb{F}^n, \quad Ax = \vec{0} \Rightarrow x = \vec{0} \\
\Longrightarrow &\forall x \in \mathbb{F}^n, \quad \begin{bmatrix} A_1 & A_2 & \dots & A_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \vec{0} \qquad (A_i \in \mathbb{F}^m \text{ is the } i\text{th column of A}) \\
\Longrightarrow &\forall x \in \mathbb{F}^n, \quad x_1 = x_2 = \cdots = x_n = 0.
\end{aligned}$$

This shows that the only linear combination of the columns of $A$ that equate to $\vec{0}$ is the trivial combination $x_1 = x_2 = \cdots = x_n = 0$. By definition, this demonstrates the columns of $A$ are linearly independent.

($\Longleftarrow$) This is clear. If the columns of $A$ are linearly independent, then only the trivial linear combination will yield $\vec{0}$. This means that the only vector $x \in \mathbb{F}^n$ that is in $\text{null}(A)$ is $\vec{0} \in \mathbb{F}^n$. □

## 0.2 Linear Systems

The system of linear equations

$$\begin{aligned}
a_{11}x_1 + a_{11}x_2 + \cdots + a_{1n}x_n &= b_1 \qquad\qquad \forall i, j \quad a_{ij}, b_i \in \mathbb{F} \\
a_{21}x_1 + a_{21}x_2 + \cdots + a_{2n}x_n &= b_1 \\
\vdots \qquad\qquad &= \vdots \\
a_{m1}x_1 + a_{m1}x_2 + \cdots + a_{mn}x_n &= b_1
\end{aligned}$$

can be compactly expressed as "$Ax = b$", where $A \in M_{m,n}(\mathbb{F})$ and $b \in \mathbb{F}^m$. We often store this as an *augmented* matrix $\begin{bmatrix} A \mid b \end{bmatrix}$.

There are three types of *row operations* that can be performed on $\begin{bmatrix} A \mid b \end{bmatrix}$ without affecting the solution set of the linear system:

(i) Swapping two rows

(ii) Multiplying a single row by a nonzero scalar $c \in \mathbb{F}$

(iii) Adding one row to another

These row operations are useful in converting matrices or linear systems into Reduced Row Echelon Form (RREF).

**Definition 0.12** (Reduced Row Echelon Form (RREF))**.** The *reduced row echelon form* (RREF) of a matrix $A$ is the unique row equivalent matrix such that

(i) The leading entry of every row is a 1 ("pivot") unless the row consists only of 0's

(ii) Each pivot has 0's above and below it

(iii) The pivot moves strictly to the right, and the rows of 0's are at the bottom

**Example 0.4.** Consider the following example of a matrix in RREF:

$$A = \begin{bmatrix} \color{red}{1} & 5 & 0 & 4 \\ 0 & 0 & \color{red}{1} & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Note the red 1's are the pivots. Above and below each of them are 0's and the 0 row is at the bottom.

The strategy for solving a linear system of equations is to row reduce $\begin{bmatrix} A \mid b \end{bmatrix}$ to RREF and then deduce the solution. In particular, suppose that $A \in M_{m,n}(\mathbb{F})$, $b \in \mathbb{F}^n$, and if $A \overset{\text{RR}}{\sim} I$ then the solutions for $Ax = b$ are the solutions to $Ix = d$, which are unique, that is

$$\begin{bmatrix} A \mid b \end{bmatrix} \overset{\text{RR}}{\sim} \begin{bmatrix} I \mid d \end{bmatrix}$$

If we have multiple linear systems $Ax = b^{(1)}, Ax = b^{(2)}, \ldots, Ax = b^{(n)}$ simultaneously under consideration, we can solve them in parallel by row reducing the augmented matrix

$$\left[\begin{array}{c|c|c|c|c} A & b^{(1)} & b^{(2)} & \ldots & b^{(n)} \end{array}\right] \overset{\text{RR}}{\sim} \left[\begin{array}{c|c|c|c|c} I & d^{(1)} & d^{(2)} & \ldots & d^{(n)} \end{array}\right],$$

where each $Ix = d^{(i)}$ has the same solution set as $Ax = b^{(i)}$.

**Definition 0.13** (Determinant (Laplace expansion))**.** For a matrix $A \in M_n(\mathbb{F})$, the *determinant* can be defined inductively in the following way:

$$\begin{aligned} \det A &= \sum_{k=1}^{n} (-1)^{i+k} a_{ik} M_{ik} \\ &= \sum_{k=1}^{n} (-1)^{k+j} a_{kj} M_{kj}, \end{aligned}$$

where $M_{ij}$ is the minor determinant, *i.e.* the determinant of $A$ when row $i$ and column $j$ are deleted.

**Definition 0.14** (Determinant (Alternating sums and permutations))**.** A permutation of $\{1, \ldots, n\}$ is a bijective function $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. There are $n!$ distinct permutations of $\{1, \ldots, n\}$. The alternative presentation of the determinant is

$$\det A = \sum_{\sigma} \left( \operatorname{sign}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \right)$$

## 0.3 Properties of Determinants

The three types of row operations that can be performed on a matrix $A \in M_n(\mathbb{F})$ scale the determinant.

| | **Row operation** | **Effect on Determinant** |
|---|---|---|
| (i) | Swapping two rows | Multiplies determinant by negative 1 |
| (ii) | Multiplying a single row by a nonzero scalar $c \in \mathbb{F}$ | Multiplies determinant by $c$ |
| (iii) | Adding one row to another | No change to determinant |

**Proposition 0.2.** *If $A, B \in M_n(\mathbb{F})$, the following determinant properties hold:*

*(i) If $A$ is upper triangular, then its determinant is the product of its diagonal entries.*

$$\det \begin{bmatrix} a_{11} & & & * \\ & a_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & a_{nn} \end{bmatrix} = \prod_{i=1}^{n} a_{ii}$$

*(ii) If $A$ has a row or column of zeros, then* $\det A = 0$.

*(iii)* $\det AB = (\det A)(\det B)$

*Note.* The strategy for computing $\det A$ is to compute $\det [\text{RREF}(A)]$ and track the row operations to scale back to $\det A$.

**Definition 0.15** (Inverse)**.** If $A, B \in M_n(\mathbb{F})$ such that $AB = BA = I$, then we call $B$ the *inverse* of $A$.

*Remark.* If such a $B$ truly exists, then it is unique.

*Remark.* We call "$B$" : $\mathbb{F}^n \to \mathbb{F}^n$ the *inverse function* of "$A$" since $\forall x \in \mathbb{F}^n$, $A(Bx) = (AB)x = Ix = x$ and $B(Ax) = (BA)x = Ix = x$.

**Lemma 0.1.** *Let $A \in M_n(\mathbb{F})$. Denote the row reduced form of $A$ as $RREF(A)$. Then either $RREF(A) = I$ or $RREF(A) \neq I$. In the first case, then $\det[RREF(A)] = 1 \implies \det A \neq 0$. In the second case, then $\det[RREF(A)] = 0 \implies \det A = 0$.*

*Proof sketch of 0.1.* In the first case, then RREF($A$) looks like

$$\begin{bmatrix} 1 & & & & & & \mathbf{0} \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ \mathbf{0} & & & & & & 1 \end{bmatrix}$$

which has determinant 1 by part (i) of Proposition 0.2. In the second case, without loss of generality, then RREF($A$) looks like

$$\begin{bmatrix} 1 & & & & & & * \\ & 1 & & & & & \\ & & 0 & 1 & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

which has determinant 0 by part (ii) of Proposition 0.2. $\square$

**Theorem 0.1.** *Let $A \in M_n(\mathbb{F})$. Then the following statements are equivalent:*

*(i)* $\exists B \in M_n(\mathbb{F})$ *such that* $AB = BA = I$

*(ii)* $\det A \neq 0$

*(iii)* $\ker(A) = \{\vec{0}\} \iff$ *"$A$" is 1-1* $\iff$ *the columns of $A$ are linearly independent*

*(iv)* $\operatorname{rank}(A) = n \iff$ *"$A$" is onto* $\iff$ *the columns of $A$ span $\mathbb{F}^n$*

*Proof of Theorem 0.1.* We showed the equivalencies of (iii) in Proposition 0.2.

We first argue that the equivalencies of (iv) are true. Saying that $\operatorname{rank}(A) = n$ is equivalent to saying that the dimension of any basis set for range($A$) has $n$ linearly independent vectors in $\mathbb{F}^n$. This means that a given basis set is also a basis for $\mathbb{F}^n$, which is the codomain of the function "$A$". So, $y$ in the codomain of "$A$" has a corresponding $x$ in the domain of "$A$" such that $A(x) = y$ since "$A$" is linear and there is a non-trivial linear combination of the basis vectors in $\mathbb{F}^n$ that give $y$. This means "$A$" is onto. Equivalently, the columns of $A$ span all of $\mathbb{F}^n$.

*(iii)* $\iff$ *(iv).* We show that the columns of $A$ are linearly independent if and only if the columns of $A$ span $\mathbb{F}^n$. If they are linearly independent, then the columns form a basis for $\mathbb{F}^n$ and thus span $\mathbb{F}^n$. If the columns of $A$ span $\mathbb{F}^n$, then they form a basis for $\mathbb{F}^n$, which implies each column is linearly independent. For the sake of contradiction, if either linear independence or full spanning of $\mathbb{F}^n$ did not hold, then we would be able to shrink or grow the basis set to be smaller or larger than $n$, respectively, which would break the dimensionality theorem.

*(i)* $\implies$ *(ii).* Suppose that $A, B \in M_n(\mathbb{F})$ and $AB = BA = I$. Then

$$1 = \det I = \det AB = (\det A)(\det B) \implies \det A \neq 0$$

*(ii)* $\Longrightarrow$ *(i).* Suppose that $\det A \neq 0$. We wish to show $AB = BA = I$ for some $B \in M_n(\mathbb{F})$. Since $\det A \neq 0$, by Lemma 0.1, then we know there is some nonzero $c \in \mathbb{F}$ such that

$$0 \neq \det A = c \cdot \det[\mathrm{RREF}(A)] \Longrightarrow \mathrm{RREF}(A) = I.$$

To solve for $X$ in $AX = I$, we row reduce $\begin{bmatrix} A \mid I \end{bmatrix}$, which gives $\begin{bmatrix} I \mid B \end{bmatrix}$, where $B \in M_n(\mathbb{F})$ is some matrix resulting from the row operations to get $A \overset{\mathrm{RR}}{\sim} I$. This means that $X = B$ and that $AB = I$. Now we show that $BA =$ I.

$$\begin{aligned} & \begin{bmatrix} A \mid I \end{bmatrix} \overset{\mathrm{RR}}{\sim} \begin{bmatrix} I \mid B \end{bmatrix} \\ \Longrightarrow\ & \begin{bmatrix} I \mid B \end{bmatrix} \overset{\mathrm{RR}}{\sim} \begin{bmatrix} A \mid I \end{bmatrix} \\ \Longrightarrow\ & \begin{bmatrix} B \mid I \end{bmatrix} \overset{\mathrm{RR}}{\sim} \begin{bmatrix} I \mid A \end{bmatrix} \\ \Longrightarrow\ & BA \qquad\qquad = I \end{aligned}$$

*(i)* $\Longrightarrow$ *(iii)* Suppose $\exists B \in M_n(\mathbb{F})$ where $AB = BA = I$. If we have $x \in \mathbb{F}^n$ such that $Ax = \vec{0}$, then we show that it must be the case that $x = \vec{0}$:

$$\begin{aligned} & Ax = \vec{0} \\ \Longrightarrow\ & BAx = B\vec{0} = \vec{0} \\ \Longrightarrow\ & Ix \qquad = \vec{0} \\ \Longrightarrow\ & x \qquad = \vec{0} \end{aligned}$$

So the only vector that is in the nullspace of $A$ is $\vec{0}$, i.e. $\mathrm{null}(A) = \{\vec{0}\}$.

*(iii), (iv)* $\Longrightarrow$ *(i)* Suppose "$A$" is 1-1 and onto. Then "$A$" has an inverse function "$B$". Note that $B$ is linear. This means it can be expressed as $B \in M_n(\mathbb{F})$. Then $\forall x \in F^n$, we have that

$$\begin{aligned} A(B(x)) &= B(A(x)) \\ &= I(x) \\ \Longrightarrow \qquad AB &= BA = I. \end{aligned}$$

□

*Remark.* If $A, B \in M_n(\mathbb{F})$ such that $AB = I$, then $B = A^{-1}$, i.e. a matrix $B$ that turns $A$ into $I$ is called the inverse of $A$, further, the inverse is unique.

*Proof of Remark.* Since $AB = I$, we have that

$$1 = \det I = \det AB = (\det A)(\det B) \to \det A \neq 0.$$

By Theorem 0.1, $\exists$ an inverse matrix $A^{-1}$, so

$$\begin{aligned} & AB = I \\ \Longrightarrow\ & A^{-1}(AB) = A^{-1}I \\ \Longrightarrow\ & (A^{-1}A)B = A^{-1} \\ \Longrightarrow\ & \qquad B = A^{-1} \end{aligned}$$

□

# 1 Chapter 1 – Eigenvalues and Similarity of Matrices

## 1.1 Similarity

**Definition 1.1** (Similar)**.** Let $A, B \in M_n(\mathbb{F})$. We say that "$A$ is *similar* to $B$", denoted $A \sim B$ if there exists an $S \in M_n(\mathbb{F})$ such that $A = SBS^{-1}$.

*Remark.* The $\sim$ operation is an equivalence relation on $M_n(\mathbb{F})$.

- Reflexive: $A = IAI^{-1}$
- Symmetric: $A = SBS^{-1} \Longrightarrow S^{-1}AS = B$
- Transitive: $A = SBS^{-1}$ and $B = TCT^{-1} \Longrightarrow A = (ST)C(T^{-1}S^{-1})$ since $(ST)^{-1} = T^{-1}S^{-1}$

This gives us equivalence classes. As functions we can think of "$A$" as applying a transformation onto some input $x$. Now consider $B$ which is similar to $A$, so that $A = SBS^{-1}$. We can think of $S$ as a change of basis or change of coordinates. $A$ and $B$ apply the same transformation or function, just from a different perspectives or basis sets. This tells us that we can think of similarity as the "sameness" of transformations up to a change of basis.

Note that just because $A$ is a rotation matrix does not mean that $B$ is also a rotation matrix because $A$ and $B$ do not necessarily respect the same metric or isometry.

*Note.* Suppose $A, B \in M_n(\mathbb{F})$ are similar. Say that $A = SBS^{-1}$ for some $S \in M_n(\mathbb{F})$ which is invertible. Then,

$$\det A = \det SBS^{-1} = \det S \det B \det S^{-1} = \det S \det S^{-1} \det B = \det SS^{-1} \det B = \det B$$

That is, the determinant is preserved between similar matrices.

## 1.2 Eigenvectors and Eigenvalues

**Definition 1.2** (Eigenvector, Eigenvalue)**.** Let $A \in M_n(\mathbb{F})$. If $x \in \mathbb{F}^n$ nonzero and we have $\lambda \in \mathbb{F}$ such that $Ax = \lambda x$ (which serves as a scaling operation along the direction of $x$), then we call $x$ and *eigenvector* associated with *eignevalue* $\lambda$.

*Note.* For $A \in M_n(\mathbb{F})$, when is $\lambda \in \mathbb{F}$ an eigenvalue? If and only if there exists $x \in \mathbb{F}^n$ nonzero such that

$$\begin{aligned}
& Ax = \lambda x \\
\iff\quad & Ax = \lambda I x \\
\iff\quad & (\lambda I - A)x = \vec{0} \\
\iff\quad & (\lambda I - A) \text{ has a nontrivial nullspace} \\
\iff\quad & \det(\lambda I - A) = \vec{0}
\end{aligned}$$

**Definition 1.3** (Characteristic Polynomial)**.** We denote $\det(\lambda I - A)$ as $p_A(\lambda)$ and call it the *characteristic polynomial*, which is a monic (leading term is 1) polynomial of degree $n$ in variable $\lambda$.

**Example 1.1.** Suppose $A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$. Then

$$p_A(\lambda) = \det(\lambda I - A) = \det \begin{bmatrix} \lambda - 2 & 2 \\ -1 & \lambda - 3 \end{bmatrix} = (\lambda - 2)(\lambda - 3) - 2 = \lambda^2 - 5\lambda + 4 = (\lambda - 4)(\lambda - 1)$$

$p_A(\lambda) = 0$ when $\lambda = 1$ or $\lambda = 4$, so our eigenvalues are 1 and 4.

**Definition 1.4** (Eigenspace)**.** For a matrix $A \in M_n(\mathbb{F})$, given an eigenvalue $\lambda$, the associated eigenvectors form the associated *eigenspace* $\{x | (\lambda I - A)x = \vec{0}\}$, which is equivalently the nullspace of $\lambda I - A$. Note, here we consider $\vec{0}$ as an "honorary eigenvector" to make the eigenspace a valid nullspace.

**Example 1.2.** Consider an upper triangular matrix $T$:

$$T = \begin{bmatrix} t_{11} & & & * \\ & t_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & t_{nn} \end{bmatrix}$$

Then its eigenvalues can be immediately read from its characteristic polynomial:

$$\begin{aligned} p_T(\lambda) &= \det(\lambda I - T) \\ &= \det \begin{bmatrix} \lambda - t_{11} & & & * \\ & \lambda - t_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda - t_{nn} \end{bmatrix} \\ &= (\lambda - t_{11})(\lambda - t_{22}) \cdots (\lambda - t_{nn}) \end{aligned}$$

So the eigenvalues of an upper triangular matrix is its diagonal entries.

**Definition 1.5** (Spectrum)**.** If $A \in M_n(\mathbb{C})$, then by the Fundamental Theorem of Algebra, there exists exactly $n$ roots of $p_A(\lambda)$, the multiset of which is called the *spectrum* of $A$, denoted $\sigma(A)$.

**Definition 1.6** (Algebraic Multiplicity, Geometric Multiplicity)**.** If $\lambda \in \sigma(A)$, then

- The *algebraic multiplicity of* $\lambda$ is the number of times $\lambda$ appears as a root of $p_A(\lambda)$, i.e. the number of times $\lambda$ appears in $\sigma(A)$.
- The *geometric multiplicity of* $\lambda$ is the dimension of the eigenspace for $\lambda$.

**Example 1.3.** Consider the following matrix $A$:

$$A = \begin{bmatrix} 7 & & & & & & \mathbf{0} \\ & 7 & & & & & \\ & & 7 & & & & \\ & & & 8 & 1 & & \\ & & & & 8 & & \\ & & & & & 8 & 1 \\ \mathbf{0} & & & & & & 8 \end{bmatrix}$$

Then $\sigma(A) = \{7, 7, 7, 8, 8, 8, 8\}$. For $\lambda = 7$, the algebraic multiplicity is 3 and the geometric multiplicity is 3 since $e_1, e_2, e_3$ span its eigenspace. For $\lambda = 8$, the algebraic multiplicity is 4 and the geometric multiplicity is 2 since only $e_4, e_6$ span its eigenspace.

*Remark.* For all $A \in M_n(\mathbb{F})$ and $\lambda \in \sigma(A)$,

$$1 \leq \text{geometric multiplicity of } \lambda \leq \text{algebraic multiplicity of } \lambda \leq n$$

**Proposition 1.1.** *Suppose that $A, B \in M_n(\mathbb{F})$ are similar. Then $p_A(\lambda) = p_B(\lambda)$, hence $\sigma(A) = \sigma(B)$.*

*Proof of Proposition 1.1.* Say $A = SBS^{-1}$ for some invertible $S \in M_n(\mathbb{F})$. Then

$$\begin{aligned} p_A(\lambda) &= \det(\lambda I - A) \\ &= \det(\lambda I - SBS^{-1}) \\ &= \det(\lambda SIS^{-1} - SBS^{-1}) \\ &= \det S(\lambda I - B)S^{-1} \\ &= \det SS^{-1} \det(\lambda I - B) \\ &= \det(\lambda I - B) \\ &= p_B(\lambda) \end{aligned}$$

□

**Definition 1.7** (Diagonalizable)**.** $A \in M_n(\mathbb{F})$ is *diagonalizable* if $A$ is similar to a diagonal matrix.

*Note.* If $A \in M_n(\mathbb{F})$ is diagonalizable, say for some invertible $S \in M_n(\mathbb{F})$

$$A = S \begin{bmatrix} d_{11} & & & \mathbf{0} \\ & d_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & d_{nn} \end{bmatrix} S^{-1}$$

then $\sigma(A) = \{d_{11}, d_{22}, \ldots, d_{nn}\}$. This allows us to decompose $A$ into its eigenvectors and eigenvalues. In fact, $S$ is the matrix with columns being the eigenvectors corresponding to each entry in the diagonal matrix.

**Example 1.4.** $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is not diagonalizable. For the sake of contradiction (FSOC), suppose it was the case that $A = SDS^{-1}$ for some invertible $S \in M_n(\mathbb{F})$ and diagonal $D \in M_n(\mathbb{F})$. Since $A$ is upper triangular, then $\sigma(A) = \{0, 0\}$. This implies that the diagonal entries of $D$ are 0, i.e. $D = \mathbf{0}$. This suggests $\mathbf{0} \neq A = SDS^{-1} = \mathbf{0}$. Therefore, $A$ is not diagonalizable.

**Lemma 1.1.** *Suppose $A \in M_n(\mathbb{C})$. Say $\mathcal{F} \subseteq \mathbb{C}^n$ is a collection of eigenvectors associated with distinct eigenvalues. Then $\mathcal{F}$ is linearly independent.*

*Proof of Lemma 1.1.* FSOC, suppose $\mathcal{F}$ is not linearly independent. For $k \geq 2$, let $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}$ be the smallest linearly dependent subset of $\mathcal{F}$ and $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{F}$ be the associated eigenvalues, which are all distinct. Then $\exists \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{F}$ not all zero such that

$$\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \cdots + \alpha_k x^{(k)} = \vec{0} \tag{1}$$

In fact, since $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ is the smallest linearly dependent subset, then all $\alpha_1, \alpha_2, \ldots, \alpha_k$ must be nonzero, else there would be at least one $\alpha_i = 0$ and we could further reduce the subset.
Applying $A$ to equation (1), we have

$$\begin{aligned} A(\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \cdots + \alpha_k x^{(k)}) = A\vec{0} = \vec{0} \\ \lambda_1 \alpha_1 x^{(1)} + \lambda_2 \alpha_2 x^{(2)} + \cdots + \lambda_k \alpha_k x^{(k)} = \vec{0} \end{aligned} \tag{2}$$

Subtracting equation (2) from $\lambda_1 \cdot$ equation (1) to get

$$\underbrace{(\lambda_1 - \lambda_1)}_{=0} \alpha_1 x^{(1)} + \underbrace{(\lambda_1 - \lambda_2)}_{\neq 0} \alpha_2 x^{(2)} + \cdots + \underbrace{(\lambda_1 - \lambda_k)}_{\neq 0} \alpha_k x^{(k)} = \vec{0}$$

This gives a strictly smaller linearly dependent subset of $\mathcal{F}$. $\Rightarrow\Leftarrow$ So $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}$ must be linearly independent. $\square$

**Theorem 1.1.** *$A \in M_n(\mathbb{F})$ is diagonalizable if and only if $A$ has $n$ linearly independent eigenvectors.*

*Proof of Theorem 1.1.* $A$ being diagonalizable means the following are all equivalent

- $\exists S \in M_n(\mathbb{F})$ invertible and $D \in M_n(\mathbb{F})$ diagonal such that $A = SDS^{-1}$.
- $\exists S \in M_n(\mathbb{F})$ invertible and $D \in M_n(\mathbb{F})$ diagonal such that $AS = SD$.
- For $S^{(1)}, S^{(2)}, \ldots, S^{(n)}$ linearly independent (which is equivalent to $S$ being invertible) and $d_{11}, d_{22}, \ldots, d_{nn} \in \mathbb{F}$, then $A\ [S^{(1)} \mid S^{(2)} \mid \ldots \mid S^{(n)}] = [S^{(1)} \mid S^{(2)} \mid \ldots \mid S^{(n)}] \begin{bmatrix} d_{11} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & d_{nn} \end{bmatrix}$
- For $S^{(1)}, S^{(2)}, \ldots, S^{(n)}$ linearly independent, $d_{11}, d_{22}, \ldots, d_{nn} \in \mathbb{F}$, then $AS^{(1)} = d_{11}S^{(1)}$, $AS^{(2)} = d_{22}S^{(2)}$, $\ldots$, $AS^{(n)} = d_{nn}S^{(n)}$, so $S^{(1)}, S^{(2)}, \ldots S^{(n)}$ are $n$ eigenvectors of $A$, which form a linearly independent set.

$\square$

Note that Theorem 1.1 suggests the following ideas:

- The eigenvectors form a basis over $\mathbb{F}^n$ since they are a set of $n$ linearly independent vectors
- The matrix of eigenvectors serves as a change of basis matrix, as introduced in the concept of similar matrices. The operator $A$ just acts like a diagonal matrix that scales each coordinate independently in the new space.

**Corollary 1.1.** *If $A$ has $n$ distinct eigenvalues (which implies a linear independent set of $n$ eigenvectors), then $A$ is diagonalizable.*

**Definition 1.8** (Principal Submatrix)**.** For $A \in M_n(\mathbb{F})$, an $r \times r$ *principal submatrix* is a matrix constructed from the rows and columns corresponding to an arbitrary index set $\{i_1, i_2, \ldots, i_r\}$ of size $r$.

**Definition 1.9.** Suppose $A \in M_n(\mathbb{C})$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. For $i = 1, 2, \ldots, n$,

$$S_i := \sum_{\substack{\text{all } \binom{n}{i} \text{ sets } U \\ \text{of } i \text{ eigenvalues}}} \prod_{\lambda \in U} \lambda$$

$$E_i := \sum_{\substack{\text{all } \binom{n}{i} \\ \text{principal } i \times i \\ \text{submatrices } M \\ \text{of } A}} \det M$$

- The $S_i$'s look like the following

$$\begin{aligned}
S_1 &= \lambda_1 + \lambda_2 + \cdots + \lambda_n \\
S_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_1\lambda_n + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \cdots + \lambda_{n-1}\lambda_n \\
S_3 &= \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \cdots + \lambda_1\lambda_{n-1}\lambda_n + \cdots + \lambda_2\lambda_3\lambda_4 + \cdots + \lambda_{n-2}\lambda_{n-1}\lambda_n \\
&\vdots \\
S_n &= \lambda_1\lambda_2\lambda_3 \ldots \lambda_n
\end{aligned}$$

- The $E_i$'s look like the following

$$\begin{aligned}
E_1 &= a_{11} + a_{22} + \cdots + a_{nn} = \operatorname{tr}(A) \\
&\vdots \\
E_n &= \det A
\end{aligned}$$

**Proposition 1.2.** *For all $A \in M_n(\mathbb{C})$, then*

$$\begin{aligned}
p_A(\lambda) &= \lambda^n - S_1\lambda^{n-1} + S_2\lambda^{n-2} - S_3\lambda^{n-3} + \cdots \pm S_n\lambda^0 \\
&= \lambda^n - E_1\lambda^{n-1} + E_2\lambda^{n-2} - E_3\lambda^{n-3} + \cdots \pm E_n\lambda^0
\end{aligned}$$

For the $S_i$, we can expand $\prod_{i=1}^n (\lambda - \lambda_i)$. For the $E_i$, we can use induction with the Laplace expansion for $\det(\lambda I - A)$.

*Note.* For all $i = 1, 2, \ldots, n$, $S_i = E_i$. Thus, $\operatorname{tr}(A) = \sum_{i=1}^n \lambda_i$ and $\det A = \prod_{i=1}^n \lambda_i$, which implies that the $\lambda_i$'s are nonzero when $\det A \neq 0$.

**Lemma 1.2.** *Multiplying partition matrices. Consider $A$ and $B$ to be matrices with the following structure:*

$$A = \begin{bmatrix} A_{11} & \ldots & A_{1s} \\ \vdots & \ddots & \vdots \\ A_{r1} & \ldots & A_{rs} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & \ldots & B_{1t} \\ \vdots & \ddots & \vdots \\ B_{s1} & \ldots & B_{st} \end{bmatrix}$$

*where each $A_{ij}$ is itself a $m_i \times n_j$ matrix and $B_{ij}$ is itself a $n_i \times p_j$ matrix. That is $A$ and $B$ are block matrices. Then for $AB = C$, it holds that $C_{ij} = \sum_{k=1}^s A_{ik}B_{kj}$, where each $A_{ik}$ is a $m_i \times n_k$ matrix, $B_{kj}$ is a $n_k \times p_j$ matrix, and $C_{ij}$ is a $m_i \times p_j$ matrix.*

## 1.3 Properties of Diagonalizabile Matrices

**Definition 1.10** (Permutation Matrices)**.** A *permutation matrix* is a square matrix such that every row has a single 1, every column has a single 1, and the rest are zeros.

**Example 1.5.** Consider $P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$. When applied to some matrix $A$, then

$$PA = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \text{row } 1 \\ \text{row } 2 \\ \text{row } 3 \end{bmatrix} = \begin{bmatrix} \text{row } 2 \\ \text{row } 3 \\ \text{row } 1 \end{bmatrix}$$

$$AP^T = \begin{bmatrix} \text{col } 1 & \text{col } 2 & \text{col } 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \text{col } 2 & \text{col } 3 & \text{col } 1 \end{bmatrix}$$

*Remark.* $PP^T = I$ and so $P^T = P^{-1}$.

*Proof of Remark.* If the $i$th row of $P$ has a 1 at the $j$th entry, then the $i$th column of $P^T$ has a 1 at the $j$th entry. All other columns of $P^T$ will yield a zero in the matrix multiplication when involving the $i$th row of $P$. Thus, only the terms $(PP^T)_{ii}$ will be nonzero, specifically they will be 1. Thus, $PP^T = I$. By definition, $P^T = P^{-1}$. □

*Note.* If $D$ is diagonal, then $PDP^T$ is diagonal and the diagonals of $D$ have been simply rearranged. Thus, permutation matrices $P$ give a class of similarity transformations for diagonal matrices.

Suppose that $C_{ii} \in M_{n_i}(\mathbb{F})$. We can construct the block matrix $C$ by putting the $C_{ii}$ on the diagonals. Further, we generalize so that $C$ also has nonzero upper triangular entries. $C$ has the following properties:

- $\sigma(C) = \sigma\left(\begin{bmatrix} C_{11} & & & * \\ & C_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & C_{kk} \end{bmatrix}\right) = \bigcup_{i=1}^{k} \sigma(C_{ii})$
- $\det(\lambda I - C) = \det\left(\begin{bmatrix} \lambda I - C_{11} & & & * \\ & \lambda I - C_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda I - C_{kk} \end{bmatrix}\right) = \prod_{i=1}^{k} \det(\lambda I - C_{ii})$

**Theorem 1.2.** *Let $A \in M_{m,n}(\mathbb{F}), B \in M_{n,m}(\mathbb{F})$. Without loss of generality (WLOG), say that $m \leq n$. Then,*

$$p_{BA}(\lambda) = \lambda^{n-m} p_{AB}(\lambda).$$

*In particular, the nonzero eigenvalues of $AB$ and $BA$ and their multiplicities are the same. This also tells us that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ (see Proposition 1.1).*

*Proof of Theorem 1.2.* Observe that

$$\underbrace{\begin{bmatrix} AB & \mathbf{0} \\ B & \mathbf{0} \end{bmatrix}}_{\mathbb{A}} \underbrace{\begin{bmatrix} I & A \\ \mathbf{0} & I \end{bmatrix}}_{\mathbb{S}} = \begin{bmatrix} AB & ABA \\ B & BA \end{bmatrix} = \underbrace{\begin{bmatrix} I & A \\ \mathbf{0} & I \end{bmatrix}}_{\mathbb{S}} \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B & BA \end{bmatrix}}_{\mathbb{B}}$$

This tell us that

$$\begin{aligned} & \mathbb{A} \sim \mathbb{B} \\ \implies \quad & \begin{bmatrix} AB & \mathbf{0} \\ B & \mathbf{0} \end{bmatrix} \sim \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B & BA \end{bmatrix} \\ \implies \quad & \sigma\left(\begin{bmatrix} AB & \mathbf{0} \\ B & \mathbf{0} \end{bmatrix}\right) \sim \sigma\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B & BA \end{bmatrix}\right) \end{aligned}$$

The final similarity relation tells us that $\sigma(AB) \cup \{n \text{ zeros}\} = \sigma(BA) \cup \{m \text{ zeros}\}$. This means that the nonzero components of the spectra for $AB$ and $BA$ are equivalent. □

**Lemma 1.3.** *If a block matrix $B$ is diagonalizable, then the blocks $B_{11}$, $B_{22}$, ..., $B_{rr}$ are each diagonal. Say that for all $i = 1, 2, \dots, r$, each of the $B_{ii}$ are diagonalizations, i.e.*

$$B_{ii} = S_i D_i S_i^{-1}.$$

*Then,*

$$\begin{aligned} B &= \begin{bmatrix} B_{11} & & & \mathbf{0} \\ & B_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & B_{rr} \end{bmatrix} \\ &= \begin{bmatrix} S_1 & & & \mathbf{0} \\ & S_2 & & \\ & & \ddots & \\ \mathbf{0} & & & S_r \end{bmatrix} \begin{bmatrix} D_1 & & & \mathbf{0} \\ & D_2 & & \\ & & \ddots & \\ \mathbf{0} & & & D_r \end{bmatrix} \begin{bmatrix} S_1^{-1} & & & \mathbf{0} \\ & S_2^{-1} & & \\ & & \ddots & \\ \mathbf{0} & & & S_r^{-1} \end{bmatrix} \\ &= \begin{bmatrix} S_1 & & & \mathbf{0} \\ & S_2 & & \\ & & \ddots & \\ \mathbf{0} & & & S_r \end{bmatrix} \begin{bmatrix} D_1 & & & \mathbf{0} \\ & D_2 & & \\ & & \ddots & \\ \mathbf{0} & & & D_r \end{bmatrix} \begin{bmatrix} S_1 & & & \mathbf{0} \\ & S_2 & & \\ & & \ddots & \\ \mathbf{0} & & & S_r \end{bmatrix}^{-1} \end{aligned}$$

**Theorem 1.3.** *Let $A, B \in M_n(\mathbb{C})$ be diagonalizable. $AB = BA$ if and only if they are simultaneously diagonalizable, i.e. $\exists S \in M_n(\mathbb{C})$ invertible such that $A = SD_1S^{-1}$, $B = SD_2S^{-1}$. Note $S$ is a matrix of eigenvectors, so this tell us that $A, B$ have common "eigen-structure", that is they are both scaling operations under the same basis set of eigenvectors.*

*Proof of Theorem 1.3.* ($\Longleftarrow$) Suppose $A, B$ are simultaneously diagonalizable, i.e. $A = SD_1S^{-1}$ and $B = SD_1S^{-1}$. Then we show $AB = BA$ directly:

$$\begin{aligned} AB &= (SD_1S^{-1})(SD_2S^{-1}) \\ &= SD_1D_2S^{-1} \\ &= SD_2D_1S^{-1} \\ &= SD_2S^{-1}SD_1S^{-1} \\ &= BA \end{aligned}$$

($\Longrightarrow$) Suppose $AB = BA$ and $A, B$ diagonalizable. We want to show that $A, B$ are simultaneously diagonalizable.

- Special Case ($*$): Suppose that we have a particularly structured

$$A = \begin{bmatrix} \lambda_1 I & & & \mathbf{0} \\ & \lambda_2 I & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_r I \end{bmatrix}$$

where the $\lambda_i$ are distinct and they are ordered and grouped together within $A$, where each $\lambda_i$ occurs $n_i$ times along the diagonal. We first show that $B$ must be a block matrix, i.e.

$$B = \begin{bmatrix} B_{11} & & & \mathbf{0} \\ & B_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & B_{rr} \end{bmatrix}$$

where each $B_{ii} \in M_{n_i}$. Let $i$ be in the $n_s$ rows and $j$ be in the $n_t$ columns, where $s \neq t$. Since $AB = BA$, then

$$\begin{aligned} & (AB)_{ij} = (BA)_{ij} \\ \implies & \lambda_s b_{ij} = \lambda_t b_{ij} \\ \implies & (\lambda_s - \lambda_t) b_{ij} = 0 \\ \implies & b_{ij} = 0 \end{aligned}$$

By Lemma 1.3, we have that

$$\begin{aligned} B &= \begin{bmatrix} S_1 & & & 0 \\ & S_2 & & \\ & & \ddots & \\ 0 & & & S_r \end{bmatrix} \begin{bmatrix} D_1 & & & 0 \\ & D_2 & & \\ & & \ddots & \\ 0 & & & D_r \end{bmatrix} \begin{bmatrix} S_1^{-1} & & & 0 \\ & S_2^{-1} & & \\ & & \ddots & \\ 0 & & & S_r^{-1} \end{bmatrix} \\ A &= \begin{bmatrix} \lambda_1 I & & & 0 \\ & \lambda_2 I & & \\ & & \ddots & \\ 0 & & & \lambda_r I \end{bmatrix} \\ &= \begin{bmatrix} S_1 & & & 0 \\ & S_2 & & \\ & & \ddots & \\ 0 & & & S_r \end{bmatrix} \begin{bmatrix} \lambda_1 I & & & 0 \\ & \lambda_2 I & & \\ & & \ddots & \\ 0 & & & \lambda_r I \end{bmatrix} \begin{bmatrix} S_1^{-1} & & & 0 \\ & S_2^{-1} & & \\ & & \ddots & \\ 0 & & & S_r^{-1} \end{bmatrix} \end{aligned}$$

So $A$ and $B$ are simultaneously diagonalizable.

- For general $A$: Let $S \in M_n(\mathbb{C})$ be invertible such that $S^{-1}AS$ is diagonal. Let $P$ be a permutation matrix such that $PS^{-1}ASP^{-1}$ is of the form in the special case $(*)$. Note that $PS^{-1}ASP^{-1}$ and $PS^{-1}BSP^{-1}$ are both diagonalizable and these are similar to $A$ and $B$, respectively. They both also commute since $AB = BA$. By $(*)$, then they are simultaneously diagonalizable. Say $V^{-1}PS^{-1}ASP^{-1}V$ and $V^{-1}PS^{-1}BSP^{-1}V$. Then both are diagonal!

□

**Proposition 1.3.** *For a collection diagonal matrices, they commute pairwise with each other if and only if they are simultaneously diagonalizable.*

# 2 Chapter 2 – Unitary Similarity and Unitary Equivalence

## 2.1 Unitary Matrices

Recall the properties of complex conjugates:

$$\begin{aligned}
\alpha, \beta \in \mathbb{R} \qquad & \overline{\alpha + \beta i} = \alpha - \beta i \\
\rho, \theta \in \mathbb{R}_{\geq 0} \qquad & \overline{\rho e^{\theta i}} = \rho e^{-\theta i} \\
y, z \in \mathbb{C} \qquad & \overline{y + z} = \bar{y} + \bar{z} \\
& \overline{yz} = \overline{yz} \\
y_1, y_2, \dots, z_1, z_2, \dots \in \mathbb{C} \qquad & \overline{y_1 z_1 + y_2 z_z + \dots} = \bar{y_1}\bar{z_1} + \bar{y_2}\bar{z_2} + \dots
\end{aligned}$$

**Definition 2.1** (Conjugate Transpose)**.** For $A \in M_{m,n}(\mathbb{C})$, the *conjugate transpose* of $A$ is

$$\begin{aligned}
A^* &:= \overline{A^T} = \bar{A}^T \\
A^*_{ij} &= \overline{A_{ji}} \qquad \forall i, j
\end{aligned}$$

*Note.* For $A \in M_{n,p}$ and $B \in M_{p,m}$,

- $(AB)^T = B^T A^T$
- $(AB)^* = \overline{(AB)^T} = \overline{B^T A^T} = \bar{B^T}\bar{A^T} = B^* A^*$
- $(A^*)^* = A$

**Definition 2.2** (Orthogonal, Orthogonal Set, Orthonormal)**.** Suppose $x, y \in \mathbb{C}^n$. We say that $x$ is *orthogonal* to $y$, denoted $x \perp y$, when $y^* x = 0$. For a set $\{x^{(1)}, x^{(2)}, \dots, x^{(k)}\} \subseteq \mathbb{C}^n$, we call it an *orthogonal set* if the vectors are pairwise orthogonal. We say that a set is *orthonormal* if the set is an orthogonal set and all vectors are normal ($\left\|x^{(i)}\right\|_2 = \sqrt{x^{(i)*} x^{(i)}} = 1$ for $i = 1, 2, \dots, k$).

Our notion of conjugation is nice because we can extend $(\alpha + \beta i) + \overline{(\alpha + \beta i)} = \alpha^2 + \beta^2$ to matrices.

**Definition 2.3** (Unitary)**.** For $U \in M_n$, we say that $U$ is *unitary* if $U^* U = I$, that is $U^{-1} = U^*$ and $UU^* = I$. This is a generalization of orthonormality to $\mathbb{C}$.

**Definition 2.4** (Real Orthogonal)**.** For $Q \in M_n(\mathbb{R})$, we say that it is *real orthogonal* if $Q^T Q = I$. In this notion of orthogonal is an overloaded definition. For matrices, this just means that the matrix is real and unitary.

**Proposition 2.1.** *$U \in M_n$ is unitary if and only if the columns of $U$ are orthonormal.*

*Proof of Proposition 2.1.*

$$\begin{aligned}
& U^* U = I \\
\iff & \begin{bmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{bmatrix} \begin{bmatrix} u_1 \mid u_2 \mid \dots \mid u_n \end{bmatrix} = I \\
\iff & u_i^* u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad \forall i, j \\
\iff & \{u_1, u_2, \dots, u_n\} \text{ are orthonormal}
\end{aligned}$$

By symmetry, the columns of $U^*$ are also orthonormal. □

*Note.* If $U, V \in M_n$ are unitary, then $UV$ is also unitary. Check that $(UV)^*(UV) = V^* U^* UV = I$.

*Note.* If $U \in M_n$ is unitary, then it is an isometry, that is, it preserves lengths and distances, i.e. $\forall x \in \mathbb{C}^n$, $\|Ux\|_2 = \|x\|_2$. This is true since

$$\|Ux\|_2 = \sqrt{(Ux)^*Ux} = \sqrt{x^*U^*Ux} = \sqrt{x^*x} = \|x\|_2 .$$

This also tells us that $\forall x, y \in \mathbb{C}^n$, $\|Ux - Uy\|_2 = \|U(x-y)\|_2 = \|x-y\|_2$.

In fact, it turns out that matrices that are an isometry are also unitary. We can think of unitary matrices as a class of operators that rotate or reflect a vector space.

**Definition 2.5** (Unitarily Similar/Unitarily Equivalent)**.** $A, B \in M_n$ are *unitarily similar* if $\exists U \in M_n$ unitary such that $A = UBU^*$.

*Note.* Unitary similarity is an equivalence relation.

- Reflexive: $A = IAI^*$
- Symmetric: $A = UBU^* \implies U^*AU = B$
- Transitive: $A = UBU^*$ and $B = WCW^* \implies A = UWCW^*U*$.

**Proposition 2.2.** *Suppose $A, B \in M_n$ are unitarily similar. Then $\|A\|_F = \|B\|_F$, where*

$$\|A\|_F := \sqrt{\sum_i \sum_j |a_{ij}|^2}$$

*Proof of Proposition 2.2.* Say that $A = UBU^*$ for a unitary $U \in M_n$. We claim that $\|A\|_F^2 = \operatorname{tr}(A^*A)$. This is direct since

$$(A^*A)_{ii} = \sum_{k=1}^n (A^*)_{ik}A_{ki} = \sum_{k=1}^n \overline{A_{ki}}A_{ki} = \sum_{k=1}^n |A_{ki}|^2.$$

The last equality is from the fact that for $z = \alpha + \beta i \in \mathbb{C}$, then $\bar{z}z = (\overline{\alpha + \beta i})(\alpha + \beta i) = \alpha^2 + \beta^2 = |z|^2$. This means that

$$\operatorname{tr}(A^*A) = \sum_{i=1}^n (A^*A)_{ii} = \sum_{i=1}^n \sum_{k=1}^n |A_{ki}|^2 = \|A\|_F^2 .$$

From here, we can show our proposition directly

$$\begin{aligned}
\|A\|_F^2 = \operatorname{tr}(A^*A) &= \operatorname{tr}\left((UBU^*)^* UBU^*\right) \\
&= \operatorname{tr}\left(UB^*U^*UBU^*\right) \\
&= \operatorname{tr}\left(UB^*BU^*\right) \\
&= \operatorname{tr}\left(B^*B\right) \\
&= \|B\|_F^2
\end{aligned}$$

□

**Definition 2.6** (Hermitian)**.** We say that a matrix $A \in M_n$ is *Hermitian* if $A = A^*$.

**Definition 2.7** (Householder transformation)**.** Let $w \in \mathbb{C}^n$ such that $\|w\|_2 = 1$. We define the *Householder transformation* is

$$H_w := I - 2ww^*.$$

*Remark.* Observe that $H_w$ is unitary and Hermitian.

- <u>Hermitian</u>: $H_w^* = (I - 2ww^*)^* = I^* - 2(ww^*)^* = I - 2ww^* = H_w$
- <u>Unitary</u>: $H_w^*H_w = (I - 2ww^*)^*(I - 2ww^*) = I - 2ww^* - 2ww^* + 4ww^*ww^* = I - 4ww^* + 4ww^* = I$

*Note.* For any $x, y \in \mathbb{R}^n$ such that $x \neq y$ and $\|x\|_2 = \|y\|_2$, if we set

$$w := \frac{1}{\|x-y\|_2}(x-y),$$

then $H_w x = y$ and $H_w y = x$.

**Lemma 2.1.** *Given any $x \in \mathbb{C}^n$ such that $\|x\|_2 = 1$, $\exists U \in M_n$ unitary such that $x$ is the first column of $U$.*

*Proof of Lemma 2.1.* The simplest approach is the apply the Gram-Schmidt algorithm to extend $x$ to an orthonormal basis of $\mathbb{C}^n$. These are the columns of $U$. Computationally, there is an $O(n^2)$ algorithm. If $x \in \mathbb{R}^n$, then

- If $x = \mathrm{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, then take $U$ to be $I$.
- Otherwise, take $w := \frac{1}{\|x-\mathrm{e}_1\|}(x - \mathrm{e}_1)$ and $U = H_w$. Then by Lemma 2.1,

  $$H_w \mathrm{e}_1 = \mathrm{x}.$$

  Since $\mathrm{e}_1$ selects the first column of any matrix with which it is left multiplied, then the above equality tells us that the first column of $H_w$ is $x$.

□

## 2.2 Schur's Theorem

**Theorem 2.1** (Schur)**.** *Let $A \in M_n$ have eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (in any order). Then $\exists U \in M_n$ unitary and $T \in M_n$ upper triangular such that $A = UTU^*$, where*

$$T = \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n \end{bmatrix}.$$

*That is, every matrix in $M_n$ is unitarily similar to an upper triangular matrix. If $A$ is real and its eigenvalues are real, then $U, T$ may be chosen to be real.*

*Proof of Theorem 2.1.* Suppose $A \in M_n$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. We argue the claim inductively.

Let $x \in \mathbb{C}^n$ be a normalized ($\|x\|_2 = 1$) eigenvector of $A$ associated with eigenvalue $\lambda_1$ and let $U \in M_n$ be a unitary matrix such that the first column of $U$ is $x$, which is guaranteed to exist by Lemma 2.1.

$$\begin{aligned} U^*AU = \begin{bmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{bmatrix} A \, [u_1 \mid u_2 \mid \ldots \mid u_n] &= \begin{bmatrix} x_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{bmatrix} A \, [x_1 \mid u_2 \mid \ldots \mid u_n] \\ &= \begin{bmatrix} x_1^* \\ u_2^* \\ \vdots \\ u_n^* \end{bmatrix} [\lambda_1 x_1 \mid Au_2 \mid \ldots \mid Au_n] \\ &= \begin{bmatrix} \lambda_1 & * \\ 0 & \\ \vdots & B \\ 0 & \end{bmatrix} \end{aligned}$$

Observe that $\sigma(B) = \{\lambda_2, \lambda_3, \ldots, \lambda_n\}$ since $A \sim T$, i.e. $\sigma(B)$ must be the remainder of $\sigma(A) = \sigma(T)$.

Let $y \in \mathbb{C}^{n-1}$ be a normalized ($\|y\|_2 = 1$) eigenvector of $B$ associated with eigenvalue $\lambda_2$ and let $V \in M_{n-1}$ be a unitary matrix such that the first column of $V$ is $y$, which is guaranteed to exist by Lemma 2.1.

$$\begin{aligned}
\left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\right)^* A \left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\right) &= \begin{bmatrix}1 & 0\\ 0 & V^*\end{bmatrix} U^*AU \begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\\
&= \begin{bmatrix}1 & 0\\ 0 & V^*\end{bmatrix} \begin{bmatrix}\lambda_1 & * \\ 0 & \\ \vdots & B \\ 0 & \end{bmatrix} \begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\\
&= \begin{bmatrix}\lambda_1 & * \\ 0 & \\ \vdots & V^*BV \\ 0 & \end{bmatrix}\\
&= \begin{bmatrix}\lambda_1 & * & \\ 0 & \lambda_2 & * \\ \vdots & \vdots & C \\ 0 & 0 & \end{bmatrix}
\end{aligned}$$

Observe that $\sigma(C) = \{\lambda_3, \lambda_4, \ldots, \lambda_n\}$ and $\left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\right)$ is unitary.

Let $z \in \mathbb{C}^{n-2}$ be a normalized ($\|z\|_2 = 1$) eigenvector of $C$ associated with eigenvalue $\lambda_3$ and let $W \in M_{n-2}$ be a unitary matrix such that the first column of $W$ is $z$, which is guaranteed to exist by Lemma 2.1.

$$\left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & W\end{bmatrix}\right)^* A \left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & W\end{bmatrix}\right) = \begin{bmatrix}\lambda_1 & * & * & & * \\ 0 & \lambda_2 & * & & \\ \vdots & 0 & \lambda_3 & & \\ 0 & \vdots & 0 & \ddots & \\ 0 & 0 & 0 & & \end{bmatrix}$$

Observe again that $\left(U\begin{bmatrix}1 & 0\\ 0 & V\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & W\end{bmatrix}\right)$ is unitary. By induction, we will have unitary $\Lambda$ such that

$$\Lambda * A\Lambda = \begin{bmatrix}\lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n\end{bmatrix}$$

This means $\Lambda^*A\Lambda = T \implies A = \Lambda T\Lambda^*$. If $A$ is real, then the eigenvalues are real and all the above steps may be done as if all values are real. $\square$

**Theorem 2.2.** *If $\mathcal{F} \subseteq M_n$ is a set of commuting matrices, then they are simultaneously, unitarily upper triangularizable, meaning $\exists U \in M_n$ unitary such that*

$$\forall A \in \mathcal{F}, \quad U^*AU \text{ is upper triangular.}$$

**Corollary 2.1.** *If $A, B \in M_n$ commute, and say $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ in any order and $\sigma(B) = \{\tau_1, \tau_2, \ldots, \tau_n\}$. Then there exists a bijection $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ such that*

$$\sigma(A + B) = \{\lambda_i + \tau_{\pi(i)} : i = 1, 2, \ldots, n\}$$
$$\sigma(AB) = \{\lambda_i\tau_{\pi(i)} : i = 1, 2, \ldots, n\}$$

*Proof of Corollary 2.1.* Say that $A = UT_AU^*$, $B = UT_BU^*$ for $U \in M_n$ unitary, $T_A, T_B \in M_n$ upper triangular, then

$$\begin{aligned} A + B &= UT_AU^* + UT_BU^* \\ &= U(T_A + T_B)U^* \\ A + B &\sim T_A + T_B \text{ unitarily} \end{aligned}$$

Note that each diagonal entry of $T_A + T_B$ is $\lambda_i + \tau_{\pi(i)}$.

$$AB = (UT_AU^*)(UT_BU^*) = U(T_AT_B)U^*$$

□

**Fact 2.1.** For two upper triangular matrices $T, R \in M_n$, the diagonal entries of $TR$ are the product of the corresponding diagonal entries of $T$ and $R$:

$$TR = \begin{bmatrix} t_{11} & & & * \\ & t_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & t_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & & & * \\ & r_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & r_{nn} \end{bmatrix} = \begin{bmatrix} t_{11}r_{11} & & & * \\ & t_{22}r_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & t_{nn}r_{nn} \end{bmatrix}$$

**Definition 2.8** (Matrix Polynomial)**.** If $q(t) = a_mt^m + a_{m-1}t^{m-1} + \dots a_1t + a_0$ for $a_i \in \mathbb{C}$. Suppose that $A \in M_n$. We define the *matrix polynomial* $q(A)$ as

$$q(A) := a_mA^m + a_{m-1}A^{m-1} + a_1A + a_0I.$$

*Remark.* Suppose that $q(t) = a_m(t - \tau_1)(t - \tau_2)\cdots(t - \tau_m)$, then the $q(A)$ also factors similarly

$$q(A) = a_m(A - \tau_1I)(A - \tau_2I)\cdots(A - \tau_mI).$$

*Note.* Suppose $A = UTU^*$ is a Schur decomposition, where $U \in M_n$ unitary, $T \in M_n$ upper triangular. If $k \geq 0$, then

$$A^k = (UTU^*)(UTU^*)\cdots(UTU^*) = UT^kU^*.$$

Observe that $T^k$ is also upper triangular and the diagonals are $t_{11}^k, t_{22}^k, \dots, t_{nn}^k$.

*Remark.* Suppose $A = UTU^*$ is a Schur decomposition, where $U \in M_n$ unitary, $T \in M_n$ upper triangular. Let $q(A)$ be the matrix polynomial of $A$. Then

$$\begin{aligned} q(A) &= Uq(T)U^* \\ &= U \begin{bmatrix} q(t_{11}) & & & * \\ & q(t_{22}) & & \\ & & \ddots & \\ \mathbf{0} & & & q(t_{nn}) \end{bmatrix} U^* \end{aligned}$$

**Example 2.1.** Suppose $q(t) = t^5 - 11t^4 + t + 2$. Then,

$$\begin{aligned} A^5 - 11A^4 + A + 2I &= U[T^5 - 11T^4 + T + 2I]U^* \\ &= U \begin{bmatrix} t_{11}^5 - 11t_{11}^4 + t_{11} + 2 & & & * \\ & t_{22}^5 - 11t_{22}^4 + t_{22} + 2 & & \\ & & \ddots & \\ \mathbf{0} & & & t_{nn}^5 - 11t_{nn}^4 + t_{nn} + 2 \end{bmatrix} U^* \end{aligned}$$

*Remark.* The spectrum of a matrix polynomial is the polynomial transformation applied to each of the eigenvalues of the matrix, including all the multiplities.

$$\sigma(q(A)) := \{q(\lambda) : \lambda \in \sigma(A)\}$$

With a slight misuse of notation, we denote this $q(\sigma(A))$.

## 2.3 Cayley-Hamilton Theorem

**Lemma 2.2.**

$$
\begin{bmatrix} \mathbf{0}_{k\times k} & & * \\ & & \\ \mathbf{0}_{k\times(n-k)} & & T \end{bmatrix}
\begin{bmatrix} * & & * & & \\ & * & & \ast & \\ & & \ddots & & \\ \mathbf{0} & & * & & \\ & & \mathbf{0} & & * \\ & & & * & \\ & \mathbf{0}_{k\times(n-k)} & & & \ddots \\ & & & & * \end{bmatrix}
=
\begin{bmatrix} & 0 & \\ \mathbf{0}_{k\times(n-k)} & 0 & * \\ & 0 & \\ & 0 & \\ \mathbf{0}_{k\times(n-k)} & 0 & \mathcal{T} \\ & 0 & \end{bmatrix}
$$

**Theorem 2.3** (Cayley-Hamilton)**.** *For $A \in M_n$, $p_A(A) = \mathbf{0}$.*

*Proof of Theorem 2.3.* Let $A = UTU^*$ be a Schur decomposition and

$$
T = \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n \end{bmatrix}
$$

and say that $p_A(t) = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_n)$. Then,

$$
\begin{aligned}
p_A(A) &= Up_A(T)U^* \\
&= U[(T-\lambda_1 I)(T-\lambda_2 I)\cdot(T-\lambda_3 I)]U^*
\end{aligned}
$$

Observe that each sub-component $(T-\lambda_i I)$ of $p_A(T)$ has a 0 as its $i^{\text{th}}$ diagonal entry. By Lemma 2.2, we build up columns of zero, so $p_A(T) = 0$, meaning

$$
p_A(A) = U\mathbf{0}U^* = \mathbf{0}.
$$

□

**Corollary 2.2.** *Suppose $A \in M_n$ invertible with*

$$
p_A(t) = t^n + a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_0.
$$

*Then, the inverse of $A$ is*

$$
A^{-1} = \frac{(-1)^{n+1}}{\det A}[A^{n-1} + a_{n-1}A^{n-2} + a_{n-2}A^{n-3} + \cdots + a_1 I] = q(A)
$$

*for some polynomial function $q$. This shows that the inverse of $A$ is a polynomial of $A$ with degree less than $n$.*

*Proof of Corollary 2.2.* Recall that $\det A = (-1)^n a_0$ from Definition 1.9. By Theorem 2.3 (Cayley-Hamilton), then

$$
\begin{aligned}
& p_A(t) = A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots + a_1 A + a_0 I = 0 \\
\Longrightarrow \quad & A[A^{n-1} + a_{n-1}A^{n-2} + \cdots + a_1 I] + a_0 I = 0 \\
\Longrightarrow \quad & A[A^{n-1} + a_{n-1}A^{n-2} + \cdots + a_1 I] = -a_0 I \\
\Longrightarrow \quad & A[\underbrace{-\frac{1}{a_0}\left(A^{n-1} + a_{n-1}A^{n-2} + \cdots + a_1 I\right)}_{A^{-1}}] = I
\end{aligned}
$$

□

**Theorem 2.4** ("Every matrix is almost diagonalizable")**.** *Let $A \in M_n$. $\forall \epsilon > 0$, $\exists S \in M_n$ invertible, $D \in M_n$ diagonal, and $E \in M_n$ such that $\|E\|_F < \epsilon$ such that*

$$A = S(D+E)S^{-1}.$$

*Proof of Theorem 2.4.* Let $A = UTU^*$ be a Schur decomposition. For all $\delta > 0$, we consider the following matrix product,

$$\begin{bmatrix} \delta^{-1} & & & \mathbf{0} \\ & \delta^{-2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{-n} \end{bmatrix} T \begin{bmatrix} \delta^{1} & & & \mathbf{0} \\ & \delta^{2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{n} \end{bmatrix} = \begin{bmatrix} t_{11}\delta^0 & t_{12}\delta^1 & \dots & t_{1n}\delta^{n-1} \\ t_{21}\delta^{-1} & t_{22}\delta^0 & \dots & t_{2n}\delta^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1}\delta^{-(n-1)} & t_{n2}\delta^{-(n-2)} & \dots & t_{nn}\delta^0 \end{bmatrix}$$

Observe that the $(i,j)$-entry of this matrix product is $t_{ij}\delta^{j-i}$. In the limit as $\delta \to 0$, all non-diagonal entries converge component-wise to 0. Thus,

$$A = U \underbrace{\begin{bmatrix} \delta^{1} & & & \mathbf{0} \\ & \delta^{2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{n} \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} \delta^{-1} & & & \mathbf{0} \\ & \delta^{-2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{-n} \end{bmatrix} T \begin{bmatrix} \delta^{-1} & & & \mathbf{0} \\ & \delta^{-2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{-n} \end{bmatrix}}_{\begin{bmatrix} t_{11} & & & \mathbf{0} \\ & t_{22} & & \\ & & \ddots & \\ 0 & & & t_{nn} \end{bmatrix} + O(\delta)} \underbrace{\begin{bmatrix} \delta^{1} & & & \mathbf{0} \\ & \delta^{2} & & \\ & & \ddots & \\ \mathbf{0} & & & \delta^{n} \end{bmatrix}}_{S^{-1}} U^*$$

□

**Theorem 2.5** ("Every matrix is almost diagonalizable")**.** . *Let $A \in M_n$. $\forall \epsilon > 0$, $\exists E \in M_n$ such that $\|E\|_F < \epsilon$ such that $A+E$ is diagonalizable.*

*Proof of 2.5.* Let $A = UTU^*$ be a Schur decomposition. $\exists D \in M_n$ diagonal such that $T + D$ has distinct diagonal entries and $\|D\|_F < \epsilon$.

Define $E := UDU^*$. Note that $\|E\|_F = \|UDU^*\|_F = \|D\|_F < \epsilon$. Also, note that

$$A + E = UTU^* + UDU^* = U(T+D)U^*$$

Since $(A+E)$ has $n$ distinct eigenvalues, it is also diagonalizable. □

## 2.4 Normal Matrices

**Definition 2.9** (Normal Matrix)**.** $A \in M_n$ is *normal* if $AA^* = A^*A$. Some notable examples are

- Diagonal matrices
- Hermitian matrices ($A = A^*$, see Definition 2.6)
- Unitary matrices

A non-example is $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

**Lemma 2.3.** *Let $T \in M_n$ be upper triangular. $T$ is normal if and only if $T$ is diagonal.*

*Proof of Lemma 2.3.* ($\Longleftarrow$) This is trivial.

($\Longrightarrow$) Suppose $T$ is normal, then $TT^* = T^*T$. Observe the following equalities:

$$(T^*T)_{ij} = \sum_{k=1}^{n} (T^*)_{ik} T_{kj} = \|T_{:,j}\|_2^2 = \|T_{i,:}\|_2^2 = \sum_{k=1}^{n} T_{ik}(T^*_{kj}) = (TT^*)_{ij}.$$

Thus, $(T^*T)_{ij}$ is the squared-length of the $j^{\text{th}}$ column of $T$ and $(TT^*)_{ij}$ is the squared-length of the $i^{\text{th}}$ row of $T$. We show that $T$ is diagonal in an inductive manner by considering the diagonal of the product matrices

Consider $(T^*T)_{11} = (TT^*)_{11}$. The squared length of the $1^{\text{st}}$ column of $T$ is just $(T^*T)_{11} = |t_{11}|^2$ since $T$ is upper triangular. But, the squared length of the $1^{\text{st}}$ row of $T$ is $(TT^*)_{11} = |t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2$. Thus, $t_{1j} = 0$ for all $j > 1$.

Consider $(T^*T)_{22} = (TT^*)_{22}$. The squared length of the $2^{\text{nd}}$ column of $T$ is just $(T^*T)_{22} = |t_{22}|^2$ since $T$ is upper triangular and we argued above that $t_{12} = 0$. But, the squared length of the $2^{\text{nd}}$ row of $T$ is $(TT^*)_{22} = |t_{22}|^2 + |t_{23}|^2 + \cdots + |t_{2n}|^2$. Thus, $t_{2j} = 0$ for all $j > 2$.

Consider $(T^*T)_{33} = (TT^*)_{33}$. The squared length of the $3^{\text{rd}}$ column of $T$ is just $(T^*T)_{33} = |t_{33}|^2$ since $T$ is upper triangular and we argued above that $t_{13} = t_{23} = 0$. But, the squared length of the $3^{\text{rd}}$ row of $T$ is $(TT^*)_{33} = |t_{33}|^2 + |t_{34}|^2 + \cdots + |t_{3n}|^2$. Thus, $t_{3j} = 0$ for all $j > 3$.

Continuing this argument, we inductively determine that $T$ is diagonal. □

**Lemma 2.4.** *Let $A, B \in M_n$ be unitarily similar. Then $A$ is normal if and only if $B$ is normal.*

*Proof of Lemma 2.4.* Say $A = UBU^*$ for some unitary $U \in M_n$. Since $A$ is normal then

$$\begin{aligned} & AA^* = A^*A \\ \Longrightarrow \quad & (UBU^*)(UBU^*)^* = (UBU^*)^*(UBU^*) \\ \Longrightarrow \quad & UBB^*U^* = UB^*BU^* \\ \Longrightarrow \quad & BB^* = B^*B \end{aligned}$$

Thus, $B$ is normal. □

**Definition 2.10** (Unitarily Diagonalizable)**.** We say $A \in M_n$ is *unitarily diagonalizable* if $A$ is unitarily similar to a diagonal matrix, that is

$$A = UDU^*.$$

Like how being similar to a diagonalizable matrix gives an invertible matrix $S \in M_n$ of the eigenvectors of $A$, the columns of unitary matrix $U \in M_n$ form an orthonormal set of eigenvectors for $A$.

**Theorem 2.6** (Spectral Theorem for Normal Matrices)**.** *Let $A \in M_n$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The following are equivalent:*

*(1) $A$ is normal*

*(2) $A$ is unitarily diagonalizable*

*(3) $\|A\|_F^2 = \sum_{i=1}^n |\lambda_i|^2$*

*Proof of Theorem 2.6.* [(2) $\Longrightarrow$ (1) and (3)] Suppose $A$ is unitarily diagonalizable. Say that $A = VDV*$, where $V \in M_n$ unitary, $D \in M_n$ diagonal. By Lemma 2.3, then $D$ is also normal. By Lemma 2.4, then $A$ is also normal, giving condition (1). Since $A$ and $D$ are unitarily similar, then by Proposition 2.2, $\|A\|_F^2 = \left\|VDV^{-1}\right\|_F^2 = \|D\|_F^2 = \sum_{i=1}^n |\lambda_i|^2$.

[(1) $\Longrightarrow$ (2)] Suppose $A$ is normal. Let $A = UTU^*$ by a Schur decomposition. By Lemma 2.4 since $A$ is normal, then $T$ is also normal. By Lemma 2.3, since $T$ is normal and upper triangular, then $T$ is also diagonal, giving condition (2).

[(3) $\implies$ (2)] Suppose $\|A\|_F^2 = \sum_{i=1}^n |\lambda_i|^2$. Let $A = UTU^*$ be a Schur decomposition. Since $T$ is upper triangular, then its eigenvalues are on its diagonal. Then,

$$\sum_{i=1}^n |t_{ii}|^2 = \sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2 = \|UTU^*\|_F^2 = \|T\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n |t_{ij}|^2.$$

So for $i \neq j$, then $t_{ij} = 0$, meaning $T$ is diagonal, giving us condition (2).

□

**Theorem 2.7** (Spectral Theorem of Hermitian Matrices)**.** *$A \in M_n$ is Hermitian if and only if*

*(1) $A$ is unitarily diagonalizable, and*

*(2) the eigenvalues of $A$ are real*

*Proof of Theorem 2.7.* ($\implies$) Suppose $A \in M_n$ is Hermitian. Then $A$ is also normal (see Definition 2.9), and therefore also unitarily diagonalizable by Theorem 2.6, giving us condition (1). Suppose $A$ is unitarily diagonalizable into $UDU^*$. Now

$$A = A^* \implies UDU^* = (UDU^*)^* \implies D = D^* \implies D \in M_n(\mathbb{R}).$$

Since $D$ is real, then the eigenvalues of $A$ are all real, giving condition (1).

($\impliedby$) Suppose that $A$ is unitarily diagonalizable, say $A = UDU^*$ for $U \in M_n$ unitary, $D \in M_n$ diagonal, and the eigenvalues of $A$ are all real. Then

$$A^* = (UDU^*)^* = UD^*U^* = UDU^* = A,$$

so $A$ is Hermitian. □

**Theorem 2.8.** *Suppose $A \in M_n(\mathbb{R})$. Then $A$ is symmetric if and only if $A$ is (real) orthogonally diagonalizable.*

*Proof of Theorem 2.8.* ($\impliedby$) Suppose $A$ is orthogonally diagonalizable (see Definition 2.4). Specifically, suppose that $A = QDQ^T$ for $Q \in M_n(\mathbb{R})$ orthogonal and $D \in M_n(\mathbb{R})$ diagonal. Then,

$$A^T = (QDQ^T)^T = QD^TQ^T = QDQ^T = A,$$

so $A$ is symmetric.

($\implies$) $A$ is symmetric tells us that $A$ is Hermitian, and therefore $A$ has real eigenvalues by Theorem 2.7. Since $A$ is real and has real eigenvalues, then $A$ has a real Schur decomposition, say $A = QTQ^T$. Now, since $A$ is Hermitian, then $A$ is also normal (see Definition 2.9). By Lemma 2.4, $T$ is normal by unitary similarity. By Lemma 2.3 $T$ is diagonal, so $A = QTQ^T$ is real orthogonally diagonalizable. □

# 3 Chapter 3 – Canonical Forms

## 3.1 Jordan Matrices and Jordan Canonical Form

**Definition 3.1** (Jordan block)**.** A "$k \times k$ *Jordan block* with eigenvalue $\lambda$", denoted "$J_k(\lambda)$" is the matrix

$$\begin{bmatrix} \lambda & 1 & & & \mathbf{0} \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ \mathbf{0} & & & & \lambda \end{bmatrix} \in M_n,$$

which has $k$ $\lambda$'s on the main diagonal and $k-1$ 1's on the superdiagonal. The eigenvectors are directly derived from $J_k(\lambda)$:

$$\begin{aligned} \vec{x} : (J_k(\lambda) - \lambda I)\vec{x} = \vec{0} \quad & \implies \quad \begin{bmatrix} \lambda & 1 & & & \mathbf{0} \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ \mathbf{0} & & & & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{bmatrix} = \vec{0} \\ & \implies \quad x_1 = \text{anything}, x_2 = x_3 = \cdots = x_k = 0 \\ & \implies \quad \vec{x} \in \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\} \end{aligned}$$

*Remark.* Observe that powers of $J_k(0)$ have decreasing rank and that it shifts the entries of a column vector up by one entry.

- $J_4(0)J_4(0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ has rank 2
- $J_4(0)J_4(0)J_4(0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ has rank 1
- $J_4(0)J_4(0)J_4(0)J_4(0) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
  has rank 0
- $J_4(0) \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} b \\ c \\ d \\ 0 \end{bmatrix}$
- $[J_k^l(0)]_{ij} = \begin{cases} 1 & j - i = l \ (l^{\text{th}} \text{ superdiagonal}) \\ 0 & \text{else} \end{cases}$
- $\operatorname{rank} J_k^l(0) = (k-l)_{\geq 0}$ where the subscript $\geq 0$ denotes the ReLU function.

**Definition 3.2** (Jordan Matrix)**.** A *Jordan matrix* is the direct sum of Jordan blocks:

$$J = \bigoplus_{i=1}^{s} J_{n_i}(\lambda_i) = J_{n_1}(\lambda_1) \oplus J_{n_2}(\lambda_2) \oplus \cdots \oplus J_{n_s}(\lambda_s) = \begin{bmatrix} J_{n_1}(\lambda_1) & & & \mathbf{0} \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ \mathbf{0} & & & J_{n_s}(\lambda_s) \end{bmatrix}$$

As an example,

$$J_5(i) \oplus J_3(2) \oplus J_1(3) = \begin{bmatrix} i & 1 & 0 & 0 & 0 & & & & \mathbf{0} \\ 0 & i & 1 & 0 & 0 & & & & \\ 0 & 0 & i & 1 & 0 & & & & \\ 0 & 0 & 0 & i & 1 & & & & \\ 0 & 0 & 0 & 0 & i & & & & \\ & & & & & 2 & 1 & 0 & \\ & & & & & 0 & 2 & 1 & \\ & & & & & 0 & 0 & 2 & \\ & & & & & & & & 3 \end{bmatrix} \sim J_3(2) \oplus J_1(3) \oplus J_5(i).$$

As a note, Jordan matrices whose blocks are rearrangements to another Jordan matrix are similar by a permutation matrix $P$.

**Theorem 3.1** (Jordan Canonical Form (JCF))**.** *For any $A \in M_n$, there exists a Jordan matrix $J$ such that $A \sim J$ and $J$ is unique up to a rearrangement of the Jordan blocks. This creates similarity classes, each identifiable by a Jordan matrix, and all that we need to know about $A$ is characterized by $J$.*

*Note.* A diagonal matrix is a Jordan matrix with all blocks $1 \times 1$. So a matrix is diagonalizable if and only if its Jordan Form $\bigoplus_{i=1}^{s} J_{n_i}(\lambda_i)$ has $n_i = 1$ for all $i = 1, 2, \ldots, s$.

**Example 3.1.** Suppose $A \in M_n$ has eigenvalues 0, 0, …, 0. Then $A \sim \bigoplus_{i=1}^{s} J_{n_i}(0)$ and $\sum_{i=1}^{s} n_i = n$. In particular suppose that $A \sim J$, where $J = J_5(0) \oplus J_3(0) \oplus J_3(0) \oplus J_2(0)$ is

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

Observe the following pattern about the rank of $A$, and consequently, the rank of $J$ since $A^k = SJ^kS^{-1}$:

- $\operatorname{rank} A^0 = \operatorname{rank} J^0 = 5 + 3 + 3 + 2$
- $\operatorname{rank} A^1 = \operatorname{rank} J^1 = 4 + 2 + 2 + 1$
- $\operatorname{rank} A^2 = \operatorname{rank} J^2 = 3 + 1 + 1 + 0$
- $\operatorname{rank} A^3 = \operatorname{rank} J^3 = 2 + 0 + 0 + 0$
- $\operatorname{rank} A^4 = \operatorname{rank} J^4 = 1 + 0 + 0 + 0$

- $\operatorname{rank} A^5 = \operatorname{rank} J^5 = 0+0+0+0$

**Definition 3.3** (Ferrers Diagram, Dual Sequence)**.** Suppose we have a sequence of nonincreasing positive integers, e.g. $(5,3,3,2)$. The *dual sequence* of a *Ferrers diagaram*, involves constructing stacks of blocks with heights $(5,3,3,2)$. Then the widths of the stacks going horizontally gives the dual sequence. Formally, $(5,3,3,2)^* = (4,4,3,1,1)$. Note, that the $l^{\text{th}}$ entry of the dual sequence gives the difference: $\operatorname{rank} A^{l-1} - \operatorname{rank} A^l$.

**Proposition 3.1.** *If $A \in M_n$ has eigenvalues 0, 0, …, 0, then for $l = 1,2,3,\dots$, let $s_k := \operatorname{rank} A^{l-1} - \operatorname{rank} A^l$. Then $(t_1,t_2,t_3,\dots) = (s_1,s_2,s_3,\dots)^*$ is the block sequence of the Jordan matrix, i.e. $A \sim \bigoplus_i J_{t_i}(0)$.*

## 3.2 Jordan Canonical Forms and Multiplicities of Eigenvalues

Suppose that $A \in M_n$ with JCF $A = S\left(\bigoplus_{i=1}^k J_{n_i}(\lambda_i)\right)S^{-1}$. Then, for all $\gamma \in \mathbb{C}$,

$$A - \gamma I = S\left(\bigoplus_{i=1}^k J_{n_i}(\lambda_i) - \gamma I\right)S^{-1} = S\left(\bigoplus_{i=1}^k J_{n_i}(\lambda_i - \gamma)\right)S^{-1}$$

In particular, if we take $\gamma = \lambda_1$, and say $\lambda_1 = \lambda_2 = \cdots = \lambda_q \neq \lambda_i$ for $i > q$, then $\operatorname{rank}(A-\gamma I)^l = \operatorname{rank} S(J-\gamma I)^l S^{-1} = \operatorname{rank}(J-\gamma I)^l$ since invertible matrices do not affect the rank of a matrix. So, we have the following

$$\operatorname{rank}(A-\lambda_1 I)^l = \operatorname{rank}\left(\bigoplus_i J_{n_i}(\lambda_i - \lambda_1)^l\right) = \operatorname{rank}\begin{bmatrix} J_{n_1}(0) & & & & & & \mathbf{0} \\ & J_{n_2}(0) & & & & & \\ & & \ddots & & & & \\ & & & J_{n_q}(0) & & & \\ & & & & J_{n_{q+1}}(\neq 0) & & \\ & & & & & \ddots & \\ \mathbf{0} & & & & & & J_{n_k}(\neq 0) \end{bmatrix}$$

From here, we can deduce the Jordan structure of the matrix. Let $z = \sum_{i=q+1}^k n_i$. Since the Jordan blocks corresponding to $n_{q+1}$ to $n_k$ have non-zero eigenvalues, then they are invertible, and therefore have the same rank as before subtracting $\lambda_1$. Thus, $s_l = \operatorname{rank} A^l - \operatorname{rank} A^{l-1}$ helps us deduce the size of the Jordan blocks of the Jordan matrix of $A$.
We can use this procedure for all eigenvalues $\lambda \in \sigma(A)$ of $A$ to get block structure and the Jordan matrix $\bigoplus_{i=1}^k J_{n_i}(\lambda_i)$ by taking

$$\operatorname{rank}(A - \lambda I)^l$$

for $l = 0,1,2,\dots$ until the rank of the matrix stops dropping, which gives us a sequence of $s_l$ and we can get the sizes of the Jordan blocks by taking the dual sequence $(t_1,t_2,t_3,\dots) = (s_1,s_2,s_3,\dots)^*$. This also shows us that each matrix has a unique Jordan form.

For $A \in M_n$ similar to $\bigoplus_{i=1}^k J_{n_i}(\lambda_i)$, say $\lambda_1 = \lambda - 2 = \cdots = \lambda_q$, $\lambda_i \neq \lambda_q$ for $i > q$, then the algebraic multiplicity of $\lambda_1$ is $n_1 + n_2 + \cdots + n_q$ and the geometric multiplicity of $\lambda_1$ is $1+1+\cdots+1 = q$. This is because each block only contributes 1 to the geometric multiplicity of $\lambda_1$.

**Example 3.2.** Suppose that $A \in M_n$ is similar to the following Jordan matrix:

$$J = J_3(\pi) \oplus J_2(\pi) \oplus J_1(17) = \begin{bmatrix} \pi & 1 & 0 & & & \\ 0 & \pi & 1 & & & \\ 0 & 0 & \pi & & & \\ & & & \pi & 1 & \\ & & & 0 & \pi & \\ & & & & & 17 \end{bmatrix}$$

Then the algebraic multiplicity of $\pi$ is 5 while the geometric multiplicity of $\pi$ is 2 since each block for $\pi$ only contributes 1 to its geometric multiplicity.

If we look at $A - \lambda I$, the eigenspace is $\{x : (A - \lambda I)x = \vec{0}\}$, which is the nullspace of $A - \lambda I$. If we compute $J - \pi I$, then we have

$$\begin{bmatrix} 0 & 1 & 0 & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & & & \\ & & & 0 & 1 & \\ & & & 0 & 0 & \\ & & & & & 17-\pi \end{bmatrix}$$

We see that each Jordan block has a rank defect of 1 (it has rank 1 less than full rank, or order), so by the rank-nullity theorem the nullity of $J - \pi I$ is 2. So each Jordan block with eigenvalue $\pi$ corresponds to a single linearly independent eigenvector. More explicitly, observe that if $A = SJS^{-1}$, then $AS = SJ$, so

$$A\left[\begin{array}{c|c|c|c} S_1 & S_2 & \dots & S_6 \end{array}\right] = \left[\begin{array}{c|c|c|c} S_1 & S_2 & \dots & S_6 \end{array}\right] \begin{bmatrix} \pi & 1 & 0 & & & \\ 0 & \pi & 1 & & & \\ 0 & 0 & \pi & & & \\ & & & \pi & 1 & \\ & & & 0 & \pi & \\ & & & & & 17 \end{bmatrix}$$

Then we have

$$\begin{aligned} AS_1 &= \pi S_1 \\ AS_4 &= \pi S_4 \\ AS_6 &= 17 S_6 \end{aligned}$$

So the columns of $S$ corresponding to the first columns of each of the Jordan blocks is are linearly independent eigenvectors.

**Definition 3.4** (Spectral radius)**.** For a matrix $A \in M_n$, the *spectral radius* $\rho(A)$ is

$$\rho(A) := \max_{\lambda \in \sigma(A)} |\lambda|$$

**Definition 3.5** (Nilpotent)**.** We say that a matrix $A \in M_n$ is *nilpotent* if there exists a positive integer $k$ such that $A^k = \mathbf{0}$.

**Proposition 3.2.** *A matrix $A \in M_n$ is nilpotent if and only if $A$ has eigenvalues all 0's, that is $\rho(A) = 0$.*

*Proof of Proposition 3.2.* Let $A = SJS^{-1}$ be a Jordan canonical form. Then if $A^l = \mathbf{0}$, then $J^l = \mathbf{0}$. $J^l = \mathbf{0}$ if and only if $J = \mathbf{0}$ since any block in $J$ corresponding to a non-zero eigenvalue will never zero out for any power $k > 0$ (only the Jordan block with eigenvalue 0 has the property of shifting columns up by one entry for each power).

So nilpotency implies $A^l = \mathbf{0}$, which implies $J^l = \mathbf{0}$ and so $J$ only consists of Jordan blocks with eigenvalues 0, so $A$ has eigenvalues 0. If $A$ has all zero eigenvalues, then $J$ clearly must only consist of Jordan blocks with eigenvalues 0, which will result in the zero matrix for some $k > 0$. □

**Proposition 3.3** ("Every matrix is almost diagonalizable")**.** *For all $A \in M_n$, $A$ can be expressed as a diagonal matrix plus a nilpotent matrix.*

*Proof of Proposition 3.3.* Say $A = SJS^{-1}$ is a Jordan canonical form. $J$ only has nonzeros on its diagonal and superdiagonal, which we can decompose into a diagonal matrix $D = \operatorname{diag}(J)$ and nilpotent matrix $N = \operatorname{superdiag}(J)$ (it is a superdiagonal matrix, so its spectrum is only 0's). We consider $N$ small because it has eigenvalues 0 and will eventually go to the zero matrix for some power, so

$$A = S(D+N)S^{-1} = SDS^{-1} + SNS^{-1}$$

□

**Corollary 3.1.** *For any $A \in M_n$, $A \sim A^T$. This is because they are similar to the same Jordan matrix.*

*Proof of Corollary 3.1.* Observe that any Jordan block $J_k(\lambda)$ is similar to its transpose $J_k^T(\lambda)$ by a permutation matrix.

$$\begin{bmatrix} \mathbf{0} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \mathbf{0} \end{bmatrix} \begin{bmatrix} \lambda & 1 & & \mathbf{0} \\ & \lambda & 1 & \\ & & \ddots & 1 \\ \mathbf{0} & & & \lambda \end{bmatrix} \begin{bmatrix} \mathbf{0} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \lambda & & & \mathbf{0} \\ 1 & \lambda & & \\ & 1 & \ddots & \\ \mathbf{0} & & 1 & \lambda \end{bmatrix}$$

By extension, for any Jordan matrix $J$, $J \sim J^T$ (by taking the direct sum of permutation matrices in the same way the Jordan blocks are directly summed). So say $A = SJS^{-1}$ is JCF, then $A^T = S^{-T}J^TS^T$, so $A \sim J \sim J^T \sim A^T$.

Another way to look at it, for every $\lambda \in \sigma(A) = \sigma(A^T)$, so $p_A(t) = p_{A^T}(t)$. Also, notice for every $l$, $\operatorname{rank}(A - \lambda I)^l = \operatorname{rank}[(A - \lambda I)^l]^T = \operatorname{rank}([A^T - \lambda I]^l)$ because row rank is equal to column rank. This tells us that they have the same JCF since they have the same eigenvalues and same ranks for each $A - \lambda I$. $\square$

## 3.3 The Minimal Polynomial

**Definition 3.6** (Annihilating polynomial)**.** Suppose $A \in M_n$. A complex polynomial $p(t) = a_s t^s + a_{s-1}t^{s-1} + \cdots + a_1 t + a_0$ is called an *annihilating polynomial* of $A$ if $p(A) = \mathbf{0}$. An example of an annihilating polynomial of $A$ is $p_A(t)$ by Cayley-Hamilton.

(Preliminaries on polynomials.) Let $m$ be the least non-negative degree such that there exists an annihilating polynomial of that degree.

- The zero polynomial is an annihilating polynomial, but by convention, its degree is $-\infty$.
- Observe that any polynomial of degree 0 is not annihilating as $\alpha t^0 \implies \alpha I \neq \mathbf{0}$ since $\alpha \neq 0$. So $m \geq 1$. Further, by Cayley-Hamilton, $m \leq n$.

**Proposition 3.4.** *There exists a unique monic polynomial of degree $m$ that annihilates $A$. We call this the minimal polynomial and denote it $q_A(t)$.*

*Proof of Proposition 3.4 (Uniqueness).* Suppose $q, q'$ are monic annihilating polynomials of degree $m$. Define $g(t) := q(t) - q'(t)$ is an annihilating polynomial of degree less than $m$. By minimality of $m$, then $g$ can only be the zero polynomial, i.e. $g(t) \equiv 0$. So $q = q'$. $\square$

**Proposition 3.5.** *Let $A \in M_n$. Then for any annihilating polynomial $p(t)$, then $q_A(t)|p_A(t)$ ($q_A(t)$ divides $p_A(t)$), that is, there exists a polynomial $d(t)$ such that $p(t) = d(t)q_A(t)$.*

*Proof of Proposition 3.5.* By division, there exists $d(t), r(t)$ such that $p(t) = d(t)q_A(t) + r(t)$, where $r(t)$ has degree smaller than the degree of $q_A(t)$. Then $p(A) = d(A)q_A(A) + r(A) \implies r(A) \equiv 0$, so $r(t)$ also annihilates $A$. Since $r(t)$ has degree smaller than degree of $q_A(t)$, then for $r(t)$ to annihilate A, it must be that $r(t) \equiv 0$. So, $p(t) = d(t)q_A(t)$. $\square$

*Note.* In particular, $\forall A \in M_n$, $q_A(t)|p_A(t)$ by Cayley-Hamilton. So, for any root of $q_A(t)$ with multiplicity $k$ is a root of $p_A(t)$ of multiciplicity $\geq k$.

**Theorem 3.2.** *$\forall A \in M_n$, $\lambda \in \sigma(A)$, $\lambda$ is a root of $q_A(t)$.*

*Proof of Theorem 3.2.* Let $\lambda \in \sigma(A)$ with associated eigenvector $x \neq \vec{0}$. Say $q_A(t) = \sum_{i=0}^m c_i t^i$. Then,

$$\vec{0} = 0x = q_A(A)x = \sum_{i=0}^m c_i A^i x = \sum_{i=0}^m c_i \lambda^i x = q_A(\lambda)x.$$

So it must be that $q_A(\lambda) = 0$, i.e. $\lambda$ is a root of $q_A(t)$. $\square$

**Corollary 3.2.** *$\forall A \in M_n$, if*

$$p_A(t) = \prod_{i=1}^{r} (t - \lambda_i)^{n_i}$$

*for $\lambda_i$ distinct. Then,*

$$q_A(t) = \prod_{i=1}^{r} (t - \lambda_i)^{m_i}$$

*where $\forall i$, we have $1 \leq m_i \leq n_i$.*

*Proof of Corollary 3.2.* With the same preceding notation, $m_i \geq 1$ by Theorem 3.2. By the preceding note note, $m_i \leq n_i$. $\square$

**Theorem 3.3.** *With the same preceding notation, $\forall i$, $m_i$ is the order of the largest Jordan block associated with $\lambda_i$.*

*Proof of Theorem 3.3.* Let $A = SJS^{-1}$ be JCF. Say

$$J = \begin{bmatrix} \bigoplus \text{J blocks with eigenvalue } \lambda_1 & & & \mathbf{0} \\ & \bigoplus \text{J blocks with eigenvalue } \lambda_2 & & \\ & & \ddots & \\ & & & \bigoplus \text{J blocks with eigenvalue } \lambda_n \end{bmatrix}$$

We know that the form a minimal polynomial $q_A(t)$ is $\prod_{i=1}^{r}(t-\lambda_1)^{s_i}$ We need to find the minimum $s_i$ for which $q_A(t)$ is annihilating. The above form tells us that $q_A(t)$ is an annihilating polynomial of $A$ if and only if on input $A$, $q_A(A)$ evaluates to $\mathbf{0}$. So, this tell us,

$$(A-\lambda_1 I)^{s_1}(A-\lambda_1 I)^{s_2}\cdots(A-\lambda_r I)^{s_r} = S[\underbrace{(J-\lambda_1 I)^{s_1}(J-\lambda_2 I)^{s_2}\cdots(J-\lambda_r I)^{s_r}}_{(*)}]S^{-1} = \mathbf{0} \iff (*) = \mathbf{0}.$$

Writing out $(*)$ explicitly, we have

$$(*) = \begin{bmatrix} \bigoplus \text{J blocks with eigenvalue } 0^{s_1} & & & \\ & \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_1} & & \\ & & \ddots & \\ & & & \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_1} \end{bmatrix}$$
$$\begin{bmatrix} \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_2} & & & \\ & \bigoplus \text{J blocks with eigenvalue } 0^{s_2} & & \\ & & \ddots & \\ & & & \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_2} \end{bmatrix}$$
$$\vdots$$
$$\begin{bmatrix} \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_r} & & & \\ & \bigoplus \text{J blocks with eigenvalue } \neq 0^{s_r} & & \\ & & \ddots & \\ & & & \bigoplus \text{J blocks with eigenvalue } 0^{s_r} \end{bmatrix}$$

For each Jordan block with eigenvalue 0 in the product, we get to $\mathbf{0}$ if and only if $s_i$ is greater than or equal to the order of the largest Jordan block with the corresponding to $\lambda_i$ because the each power of a Jordan block with eigenvalue 0 shift the entries of each column up by 1. So, we choose the minimum $s_i$ which are exactly the order of the largest Jordan block with eigenvalue $\lambda_i$. $\square$

**Corollary 3.3.** *Any matrix $A \in M_n$ is diagonalizable if and only if $q_A(t)$ is a product of distinct linear factors, or simple roots, i.e. $m_i = 1$ for all $i$.*

## 3.4 Non-derogatory Matrices and Companion Matrices

**Definition 3.7** (Non-derogatory)**.** We say that a matrix $A \in M_n$ is *non-derogatory* if any of the following equivalent statements are true:

- $\forall \lambda \in \sigma(A)$, there is only one Jordan block associated with $\lambda$
- $\forall \lambda \in \sigma(A)$, the geometric multiplicity of $\lambda$ is 1 (each block contributes 1 geometric multiplicity)
- $p_A(t) = q_A(t)$
- $m_i = n_i \ \forall i$

**Proposition 3.6.** *A matrix $A \in M_n$ is diagonalizable and non-derogatory if and only if all $n$ eigenvalues are distinct.*

**Definition 3.8** (Companion matrix)**.** Let $g(t) = t^n + a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_1 t + a_0$ be a complex polynomial. The associated *companion matrix* is

$$C_g := \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix} \in M_n,$$

is the matrix with 1's on the superdiagonal and the negative coefficients of $g(t)$.

*Remark.* $\lambda$ is an eigenvalue of $C_g$ if and only if $\lambda$ is a root of $g(t)$.

*Proof of Remark.* Suppose that $\lambda$ is an eigenvalue of $C_g$ with associated eigenvector $x$. Then,

$$C_g x = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \\ \lambda x_3 \\ \lambda x_4 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

The top $n-1$ equations tell us

$$\begin{aligned} x_2 &= \lambda x_1 \\ x_3 &= \lambda x_2 \\ \vdots &= \ \vdots \\ x_n &= \lambda x_{n-1} \end{aligned}$$

Choosing $x_1 = \alpha \neq 0$, then

$$\begin{aligned} x_2 &= \lambda \alpha \\ x_3 &= \lambda^2 \alpha \\ \vdots &= \ \vdots \\ x_n &= \lambda^{n-1} \alpha, \end{aligned}$$

which tells us that $x = \alpha \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{n-1} \end{bmatrix} \in \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{n-1} \end{bmatrix} \right\}$. The bottommost equation in $C_g x = \lambda x$ tells us

$$\begin{aligned}
-\sum_{i=0}^{n-1} a_i x_{i+1} &= -\sum_{i=0}^{n-1} a_i \lambda^i \alpha && \text{(from top } n-1 \text{ equations)} \\
&= \lambda x_n && \text{(from last equation)} \\
&= \lambda(\lambda^{n-1}\alpha) && \text{(from last equation of top } n-1 \text{ equations)} \\
&= \lambda^n \alpha
\end{aligned}$$

$$\begin{aligned}
\iff \quad & \lambda^n \alpha + \sum_{i=0}^{n-1} a_i \lambda^i \alpha = 0 \\
\iff \quad & \alpha \left( \lambda^n + \sum_{i=0}^{n-1} a_i \lambda^i \right) = 0 \\
\iff \quad & \lambda^n + \sum_{i=0}^{n-1} a_i \lambda^i = 0 \\
\iff \quad & g(\lambda) = 0.
\end{aligned}$$

Thus, $\lambda$ is an eigenvalue of $C_g$ if and only if $\lambda$ is a root of $g$. Note that the geometric multiplicity of such an eigenvalue is 1, so $C_g$ is non-derogatory. □

**Proposition 3.7.** *For all monic polynomials $g$, $g(t) = p_{C_g}(t) = q_{C_g}(t)$.*

**Theorem 3.4.** *A matrix $A \in M_n$ is non-derogatory if and only if $\{B \in M_n : AB = BA\} = \{p(A) :$ $p$ is a complex polynomial$\}$. Note that the sets $\{B \in M_n : AB = BA\} = \{p(A) : p$ is a complex polynomial$\}$ are equal. We have $\supseteq$ because every polynomial of $A$ trivially commutes with $A$. We have $\subseteq$ as a consequence of $A$ being non-derogatory.*

*Proof of Theorem 3.4 (A is diagonalizable case).* ($\Longrightarrow$) If $A$ is diagonalizable and non-derogatory, say $A = SDS^{-1}$ is a diagonalization with $D$ having distinct diagonal entries (by non-derogatory). Suppose $B \in M_n$ such that $AB = BA$, then $A$ and $B$ are simultaneously diagonalizable. Say $B = S\hat{D}S^{-1}$

Let $p(t)$ be an interpolating polynomial such that for all $i$, we have $p(\lambda_i) = \hat{D}_{ii}$. Such an interpolating polynomial exists since the $\lambda_i$ are distinct. Then,

$$p(A) = p(SDS^{-1}) = Sp(D)S^{-1} = S\hat{D}S^{-1} = B.$$

($\Longleftarrow$) Suppose $A$ is diagonalizable and suppose $A$ is not non-derogatory. Then we set a matrix $B$ that commutes with $A$ and show that no polynomial of $A$ can equal $B$. Say $ADS^{-1}$ is a diagonalization with $D$ diagonals not all distinct. Set $B = S\hat{D}S^{-1}$ for any particular diagonal matrix $\hat{D}$ with distinct diagonals. Note, $A$ commutes with $B$ since they are simultaneously diagonalizable. No polynomial of $A$ can equal $B$ since no polynomial of $D$ can equal $\hat{D}$ since each of $\hat{D}_{ii}$ are distinct, while some of the $D_{ii}$ are repeated. There is no polynomial (which is a function) that sends the same input to different outputs. □

# 4 Chapter 4 – Hermitian Matrices, Symmetric Matrices, and Congruence

## 4.1 Field of Values and Characterization of Hermitian Matrices

**Definition 4.1** (Field of values)**.** Suppose $A \in M_n$. The *field of values* $F(A)$ of $A$ is defined as

$$\begin{aligned} F(A) &:= \left\{ \frac{x^*Ax}{x^*x} : x \in \mathbb{C}^n, x \neq \vec{0} \right\} \\ &= \left\{ x^*Ax : x \in \mathbb{C}^n, \|x\|_2 = 1 \right\}. \end{aligned}$$

The second equality comes from the fact that $\|x\|_2^2 = x^*x$ and

$$\left( \frac{x^*}{\|x\|_2} \right) A \left( \frac{x}{\|x\|_2} \right) = \frac{x^*Ax}{\|x\|_2^2} = \frac{x^*Ax}{x^*x}$$

**Theorem 4.1.** *Let $A \in M_n$ be a normal matrix. Its field of values $F(A)$ is the same as the convex hull of its eigenvalues $\mathcal{H}(\sigma(A))$, that is,*

$$F(A) = \mathcal{H}(\sigma(A)) = \left\{ \sum_{i=1}^n \alpha_i \lambda_i : \forall i, \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1 \right\}.$$

*Note that $\mathcal{H}$ is the smallest convex set containing the eigenvalues of $A$.*

*Proof of 4.1.* Since $A$ is normal, then it is unitarily diagonalizable as $A = UDU^*$ for some unitary $U \in M_n$ and diagonal matrix $D \in M_n$ with $\lambda_1, \lambda_2, \ldots, \lambda_n$ along its diagonal. Then we have

$$\begin{aligned} F(A) &= \{x^*Ax : x \in \mathbb{C}^n, \|x\|_2 = 1\} \\ &= \{x^*UD\underbrace{U^*x}_{y} : x \in \mathbb{C}^n, \|x\|_2 = 1\} \\ &= \{y^*Dy : y \in \mathbb{C}^n, \|y\|_2 = 1\} \qquad \text{(since } U \text{ is unitary, and is an isometry)} \\ &= \left\{ \begin{bmatrix} \bar{y_1} & \bar{y_2} & \ldots & \bar{y_n} \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \ddots \\ y_n \end{bmatrix} : \|y\|_2 = 1 \right\} \\ &= \left\{ \sum_{i=1}^n \underbrace{|y_i|^2}_{\text{“}\alpha_i\text{”}} \lambda_i : \|y\|_2 = 1 = \|y\|_2^2 = \sum_{i=1}^n |y_i|^2 = 1 \right\} \\ &= \mathcal{H}(\sigma(A)). \end{aligned}$$

**Example 4.1.** This theorem does not hold for non-normal $A \in M_n$. Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then, $\sigma(A) = \{0, 0\}$, so $\mathcal{H}(\sigma(A)) = \{0\}$. But consider a vector $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Then $F(A)$ contains the value $\frac{1}{2}$, but

$$\frac{x^*Ax}{x^*x} = \frac{\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}} = \frac{1}{2} \notin \mathcal{H}(\sigma(A)).$$

□

From here on, we keep in mind that Hermitian matrices are normal matrices with real eigenvalues. Further, we consider the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ in nondecreasing order, and call $\frac{x^*Ax}{x^*x}$ the “*Rayleigh-Ritz ratios*”.

## 4.2 Variational Characterizations

**Theorem 4.2** (Rayleigh-Ritz)**.** *Suppose $A \in M_n$ is Hermitian. Then,*

$$\min_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{x^* A x}{x^* x} = \lambda_1(A), \qquad \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{x^* A x}{x^* x} = \lambda_n(A).$$

*Proof of Theorem 4.2.* Since $A$ is Hermitian, then $A$ is also normal and $\lambda_i(A) \in \mathbb{R}$ for all $i = 1, 2, \ldots, n$. Since $\lambda_i(A)$ are all real, then the convex hull is just the closed interval from the smallest eigenvalue to the largest eigenvalue. By Theorem 4.1, then this closed interval is also $F(A)$. The left side of the claim minimizes $F(A)$, which gives you $\lambda_1$, and the right side maximizes $F(A)$, which gives you $\lambda_n$. □

**Definition 4.2** (Positive definite, Positive semidefinite, Negative definite, Negative semidefinite, Indefinite)**.** Suppose that $A \in M_n$ is Hermitian. Then we have several similar definitions:

- $A$ is *positive definite* if $\forall x \in \mathbb{C}^n$ nonzero, then $x^* A x > 0$.
- $A$ is *positive semidefinite* if $\forall x \in \mathbb{C}^n$, then $x^* A x \geq 0$.
- $A$ is *negative definite* if $\forall x \in \mathbb{C}^n$ nonzero, then $x^* A x < 0$.
- $A$ is *negative semidefinite* if $\forall x \in \mathbb{C}^n$, then $x^* A x \leq 0$.
- $A$ is *indefinite* otherwise.

**Theorem 4.3.** *Suppose $A \in M_n$ is Hermitian. Then we have several similar equivalencies:*

- *$A$ is positive definite if and only if $\sigma(A) \subseteq \mathbb{R}_{>0}$.*
- *$A$ is positive semidefinite if and only if $\sigma(A) \subseteq \mathbb{R}_{\geq 0}$.*
- *$A$ is negative definite if and only if $\sigma(A) \subseteq \mathbb{R}_{<0}$.*
- *$A$ is negative semidefinite if and only if $\sigma(A) \subseteq \mathbb{R}_{\leq 0}$.*

*Proof of Theorem 4.3.* Use Theorem 4.2 (Rayleigh-Ritz). The signs of Rayleigh-Ritz ratios do not change regardless of whether you divide by $x^* x$. □

**Theorem 4.4.** *Suppose $A \in M_n$ is Hermitian with orthonormal eigenvectors $\mu_1, \mu_2, \ldots, \mu_n$ with associated eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then for all $k = 1, 2, \ldots, n$,*

$$\min_{\substack{x \in \mathbb{C}^n \setminus \{\vec{0}\} \\ x \perp \mu_1, \mu_2, \ldots, \mu_{k-1}}} \frac{x^* A x}{x^* x} = \lambda_k(A), \qquad \max_{\substack{x \in \mathbb{C}^n \setminus \{\vec{0}\} \\ x \perp \mu_{k+1}, \mu_{k+2}, \ldots, \mu_n}} \frac{x^* A x}{x^* x} = \lambda_k(A).$$

On the min side, if $k = 1$, the orthogonality conditions are vacuously satisfied by any $x \in \mathbb{C}^n \setminus \{\vec{0}\}$, so by Theorem 4.2 (Rayleigh-Ritz), this is just $\lambda_1$. If $k = n$, then $x \perp \mu_1, \mu_2, \ldots, \mu_{n-1}$, so $x \in \operatorname{Span}\{\mu_n\}$.

Similarly, on the max side, if $k = 1$, the orthogonality conditions are vacuously satisfied by any $x \in \mathbb{C}^n \setminus \{\vec{0}\}$, so by Theorem 4.2 (Rayleigh-Ritz), this is just $\lambda_n$. If $k = n$, then $x \perp \mu_2, \mu_3, \ldots, \mu_n$, so $x \in \operatorname{Span}\{\mu_1\}$.

*Proof of Theorem 4.4 (min side).* Say $A = UDU^*$ for $U = \left[\ \mu_1 \mid \mu_2 \mid \cdots \mid \mu_n\ \right]$ is a unitary matrix and $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix. Then,

$$
\begin{aligned}
\min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\}\\ x\perp\mu_1,\mu_2,\ldots,\mu_{k-1}}} \frac{x^*Ax}{x^*x} &= \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\}\\ x\in\operatorname{Span}\{\mu_k,\mu_{k+1},\ldots,\mu_n\}}} \frac{x^*UDU^*x}{x^*x} \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \frac{\left(\sum_{i=k}^n \delta_i\mu_i\right)^* UDU^* \left(\sum_{i=k}^n \delta_i\mu_i\right)}{\left(\sum_{i=k}^n \delta_i\mu_i\right)^* \left(\sum_{i=k}^n \delta_i\mu_i\right)} \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \frac{\begin{bmatrix}0 & \cdots & 0 & \bar{\delta}_k & \cdots & \bar{\delta}_n\end{bmatrix} \begin{bmatrix}\lambda_1 & & & 0\\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n\end{bmatrix} \begin{bmatrix}0\\ \cdots\\ 0\\ \bar{\delta}_k\\ \vdots\\ \bar{\delta}_n\end{bmatrix}}{\sum_{i=k}^n \bar{\delta}_i\delta_i} \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \frac{\sum_{j=k}^n |\delta_j|^2\lambda_j}{\sum_{i=k}^n |\delta_i|^2} \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \sum_{j=k}^n \underbrace{\left(\frac{|\delta_j|^2}{\sum_{i=k}^n |\delta_i|^2}\right)}_{\text{“}\alpha_j\geq 0\text{”}} \lambda_j \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \sum_{j=k}^n \alpha_j\lambda_j \\
&= \min_{\substack{\delta_k,\delta_{k+1},\ldots,\delta_n\in\mathbb{C}^n\\ \text{not all } 0}} \mathcal{H}(\lambda_k,\lambda_{k+1},\ldots,\lambda_n) \\
&= \lambda_k.
\end{aligned}
$$

□

## 4.3 Courant-Fischer Theorem

**Theorem 4.5.** *Let $A \in M_n$ be Hermitian. Then for all $k = 1, 2, \ldots, n$,*

$$
\max_{y_1,y_2,\ldots,y_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\}\\ x\perp y_1,y_2,\ldots,y_{k-1}}} \frac{x^*Ax}{x^*x} = \lambda_k(A), \qquad \min_{y_{k+1},y_{k+2},\ldots,y_n\in\mathbb{C}^n} \max_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\}\\ x\perp y_{k+1},y_{k+2},\ldots,y_n}} \frac{x^*Ax}{x^*x} = \lambda_k(A).
$$

Some things to note are that we can consider the maximin problem as a maximization of a function $\phi(y_1, y_2, \ldots, y_{k-1})$, where $\phi(\cdot)$ evaluates the minimum value over all choices of $x$ which are orthogonal for specified input $y_1$, $y_2$, $\ldots$, $y_{k-1} \in \mathbb{C}^n$.

Similarly, for the minimax problem, we can consider it as the minimization of a function $\theta(y_{k+1}, y_{k+2}, \ldots, y_n)$, where $\theta(\cdot)$ evaluates the maximum value over all choices of $x$ which are orthogonal for specified input $y_{k+1}$, $y_{k+2}$, $\ldots$, $y_n \in \mathbb{C}^n$.

*Proof of Theorem 4.5 (min side).* Let $A = UDU^*$ where $U = [\ \mu_1 \mid \mu_2 \mid \cdots \mid \mu_k\ ] \in M_n$ is a unitary matrix and $D = \operatorname{diag}(\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A))$ is a diagonal matrix.

For any $y_1, y_2, \ldots, y_{k-1} \in \mathbb{C}^n$,

$$
\underbrace{\overset{\text{“}V_1\text{”}}{\operatorname{Span}\{\mu_1,\mu_2,\ldots,\mu_k\}}}_{\dim V_1 = k} \cap \underbrace{\overset{\text{“}V_2\text{”}}{\operatorname{Span}\{y_1,y_2,\ldots,y_{k-1}\}^\perp}}_{\dim V_2 \geq n-(k-1)} \neq \{\vec{0}\}.
$$

From inclusion-exclusion, then

$$\begin{aligned}\dim V_1 \cap V_2 &= \dim V_1 + \dim V_2 - \dim V_1 \cup V_2 \\ &\geq k + n - (k-1) - n \\ &= 1,\end{aligned}$$

where the inequality is because $\dim V_1 \cup V_2 \leq n$. Let $w \in \mathbb{C}^n \setminus \{\vec{0}\}$ be some vector in this intersection. Then $\frac{w^*Aw}{w^*w} \in F(A) = \mathcal{H}(\sigma(A))$, meaning it is also a convex combination of $\lambda_1(A), \ldots, \lambda_k(A)$ since $w \in \mathrm{Span}\{\mu_1, \mu_2, \ldots, \mu_k\}$. This means that $\frac{w^*Aw}{w^*w} \leq \lambda_k(A)$. This tell us that

$$\min_{\substack{x \in \mathbb{C}^n \setminus \{\vec{0}\} \\ x \perp y_1, y_2, \ldots y_{k-1}}} \frac{x^*Ax}{x^*x} \leq \lambda_k(A)$$

with equality only when $y_1 = \mu_1, y_2 = \mu_2, \ldots, y_{k-1} = \mu_{k-1}$ by Theorem 4.2, i.e. the above minimization problem for a given choice of $y_1, y_2, \ldots y_{k-1}$ achieves the largest value of $\lambda_k(A)$ when we have $y_1 = \mu_1, y_2 = \mu_2, \ldots, y_{k-1} = \mu_{k-1}$. So,

$$\max_{y_1, y_2, \ldots, y_{k-1} \in \mathbb{C}^n} \min_{\substack{x \in \mathbb{C}^n \setminus \{\vec{0}\} \\ x \perp y_1, y_2, \ldots, y_{k-1}}} \frac{x^*Ax}{x^*x} = \lambda_k(A).$$

□

*Note.* Consider a set of indices $\mathcal{I}$ and sequences $a = \{a_i\}_{i \in \mathcal{I}} \subseteq \mathbb{R}$ and $b = \{b_i\}_{i \in \mathcal{I}} \subseteq \mathbb{R}$. Suppose that $\forall i \in \mathcal{I}$, $a_i \leq b_i$, that is, the sequence $b$ dominates $a$. Then, the following inequalities hold:

$$\min_{i \in \mathcal{I}} a_i \leq \min_{i \in \mathcal{I}} b_i, \qquad \max_{i \in \mathcal{I}} a_i \leq \max_{i \in \mathcal{I}} b_i.$$

Let $\check{f}$, $\hat{f}$ be functions where $\hat{f}$ dominates $\check{f}$, i.e. $\check{f} \leq \hat{f}$. Also, let $f$ be some arbitrary function. Then, the following inequalities are true:

$$\begin{aligned}&\max\min \check{f} \leq \max\min \hat{f} \\ &\min\max \check{f} \leq \min\max \hat{f} \\ &\max\min f \leq \max \min_{\text{more restrictions}} f \\ &\min\max f \geq \min \max_{\text{more restrictions}} f\end{aligned}$$

## 4.4 Eigenvalue Inequalities for Hermitian Matrices

**Theorem 4.6** (Weyl)**.** *Let $A, B \in M_n$ be Hermitian matrices. For all $k = 1, 2, \ldots, n$,*

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_n(B).$$

*Proof of Theorem 4.6.* We show this claim directly using Theorem 4.5 (Courant-Fischer) and Theorem 4.2 (Rayleigh-Ritz).

$$\begin{aligned}\lambda_k(A+B) &\underset{\text{C-F}}{=} \min_{\text{C-F conditions}} \max \frac{x^*(A+B)x}{x^*x} \\ &= \min_{\text{C-F conditions}} \max \frac{x^*Ax}{x^*x} + \frac{x^*Bx}{x^*x} \\ &\underset{\text{R-R}}{\leq} \min_{\text{C-F conditions}} \max \left(\frac{x^*Ax}{x^*x} + \lambda_n(B)\right) \\ &= \min_{\text{C-F conditions}} \max \left(\frac{x^*Ax}{x^*x}\right) + \lambda_n(B) \\ &\underset{\text{C-F}}{=} \lambda_k(A) + \lambda_n(B).\end{aligned}$$

□

**Corollary 4.1.** *Let $A, B \in M_n$ be Hermitian matrices and $B$ be positive semidefinite. Then for all $k = 1, 2, \ldots, n$,*

$$\lambda_n(A) \leq \lambda_n(A+B).$$

*Proof of Corollary 4.1.* Since $B$ is positive semidefinite, then $\lambda_1(B) \geq 0$, so $\lambda_k(A) \leq \lambda_k(A) + \lambda_1(B)$. $\square$

**Theorem 4.7** (Interlacing I)**.** *Suppose $A \in M_n$ Hermitian, $y \in \mathbb{C}^n$, $a \in \mathbb{R}$. Then for all $k = 1, 2, \ldots, n-1$,*

$$\lambda_k(A + ayy^*) \leq \lambda_{k+1}(A).$$

Observe that $yy^*$ is a rank 1 matrix, which can be thought of as a perturbation matrix to $A$.

*Proof of Theorem 4.7.* We show this directly:

$$\begin{aligned}
\lambda_k(A+ayy^*) &= \max_{z_1,z_2,\ldots,z_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp z_1,z_2,\ldots,z_{k-1}}} \frac{x^*(A+ayy^*)x}{x^*x} \\
&\leq \max_{z_1,z_2,\ldots,z_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp z_1,z_2,\ldots,z_{k-1} \\ x\perp y}} \frac{x^*(A+ayy^*)x}{x^*x} \\
&= \max_{z_1,z_2,\ldots,z_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp z_1,z_2,\ldots,z_{k-1},y}} \frac{x^*Ax}{x^*x} \\
&\leq \max_{z_1,z_2,\ldots,z_{k-1},z_k\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp z_1,z_2,\ldots,z_{k-1},z_k}} \frac{x^*Ax}{x^*x} \\
&\underset{\text{C-F}}{=} \lambda_{k+1}(A).
\end{aligned}$$

$\square$

To establish notational convention, for $A \in M_n$ Hermitian, then

$$\lambda_k(A) = \begin{cases} \infty & \text{if } i > n \\ -\infty & \text{if } i < 1 \end{cases}$$

and recall that $\lambda_k(A + ayy^*) \leq \lambda_{k+1}(A)$ for Theorem 4.7.

**Corollary 4.2.** *Let $A, B \in M_n$ be Hermitian with* $\operatorname{rank} B = r$*. Then for all $k = 1, 2, \ldots, n$,*

$$\lambda_k(A+B) \leq \lambda_{k+r}(A).$$

*Proof of Corollary 4.2.* Let $B = UDU^*$ for unitary $U = \left[\begin{array}{c|c|c|c} \mu_1 & \mu_2 & \cdots & \mu_n \end{array}\right] \in M_n$, $D \in M_n$ diagonal with $d_{11}, d_{22}, \ldots, d_{rr} \neq 0$ and the other diagonal entries being 0. Then, $B = \sum_{i=1}^r d_{ii}\mu_i\mu_i^* = \sum_{i=1}^{r-1} d_{ii}\mu_i\mu_i^* + d_{rr}UU^*$ and by

$$\begin{aligned}
\lambda_k(A+B) = \lambda_{k+r-r}\left(A + \sum_{i=1}^r d_{ii}\mu_i\mu_i^*\right) &\leq \lambda_{k+r-(r-1)}\left(A + \sum_{i=1}^{r-1} d_{ii}\mu_i\mu_i^*\right) \\
&\leq \lambda_{k+r-(r-2)}\left(A + \sum_{i=1}^{r-2} d_{ii}\mu_i\mu_i^*\right) \\
&\;\vdots \\
&\leq \lambda_{k+r-1}\left(A + d_{11}\mu_i\mu_i^*\right) \\
&\leq \lambda_{k+r}(A).
\end{aligned}$$

$\square$

This gives us the following bounds:

- $\lambda_{k-r}(A) \leq \lambda_k(A+B) \leq \lambda_{k+r}(A)$
- $\lambda_{k-r}(A+B) \leq \lambda_k(A) \leq \lambda_{k+r}(A+B)$.

If we also define $\mathbb{A} := A+B$ and $\mathbb{B} := -B$, then

$$\lambda_k(\mathbb{A}+\mathbb{B}) \leq \lambda_{k+r}(\mathbb{A}).$$

**Theorem 4.8** (Interlacing II, Inclusion Principle)**.** *Let $A \in M_n$ be Hermitian and $B \in M_r$ be a principal submatrix of $A$. Then for all $k = 1, 2, \ldots, r$, then*

$$\lambda_k(A) \leq \lambda_k(B) \leq \lambda_{k+n-r}(A)$$

*Proof of Theorem 4.8.* Say the row and column indices deleted from $A$ to construct $B$ are $i_1$, $i_2$, $\ldots$, $i_{n-r}$. By Theorem 4.5 (Courant-Fischer), then

$$\begin{aligned}
\lambda_k(A) &= \max_{y_1,y_2,\ldots,y_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp y_1,y_2,\ldots,y_{k-1}}} \frac{x^*Ax}{x^*x} \\
&\leq \max_{y_1,y_2,\ldots,y_{k-1}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n\setminus\{\vec{0}\} \\ x\perp y_1,y_2,\ldots,y_{k-1},e_{i_1},e_{i_2},\ldots,e_{i_{n-r}}}} \frac{x^*Ax}{x^*x} \\
&= \max_{w_1,w_2,\ldots,w_{k-1}\in\mathbb{C}^n} \min_{\substack{z\in\mathbb{C}^n\setminus\{\vec{0}\} \\ z\perp w_1,w_2,\ldots,w_{k-1}}} \frac{z^*Bz}{z^*z} \\
&\underset{\text{C-F}}{=} \lambda_k(B),
\end{aligned}$$

where $z \in \mathbb{C}^r$ is the vector constructed from $x$ where the zero entries are removed and the $w_i \in \mathbb{C}^r$ are vectors constructed from the $y_i$ where the corresponding entries are removed. ☐

**Corollary 4.3.** *For $A \in M_n$ Hermitian and $B \in M_{n-1}$ being a principal submatrix of $A$, then*

$$\lambda_1(A) \leq \hat{\lambda}_1(B) \leq \lambda_2(A) \leq \hat{\lambda}_2(B) \leq \cdots \leq \lambda_{n-1}(A) \leq \hat{\lambda}_{n-1}(B) \leq \lambda_n(A).$$

**Corollary 4.4.** *If $A \in M_n$ Hermitian, then for every diagonal $a_{ii}$,*

$$\lambda_1(A) \leq a_{ii} \leq \lambda_n(A),$$

*where we can consider $a_{ii}$ to be a $1 \times 1$ principal submatrix of $A$.*

**Definition 4.3** (Majorize)**.** We say that a vector $x \in \mathbb{R}^n$ *majorizes* $y \in \mathbb{R}^n$ if when the components each vector are ordered as $x_{l_1} \leq x_{l_2} \leq \cdots \leq x_{l_n}$, $y_{m_1} \leq y_{m_2} \leq \cdots \leq y_{m_n}$, we have for all $k = 1, 2, \ldots, n$

$$\sum_{i=1}^{k} x_{l_i} \geq \sum_{i=1}^{k} y_{m_i}$$

and exact equality when $k = n$. This means the partial sums of the ordered components of $x$ dominate the partial sums of the ordered components of $y$ and we have equality for full sum ($k = n$).

**Example 4.2.** If $x = \begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix}$ and $y = \begin{bmatrix} 9 \\ 1 \\ 2 \end{bmatrix}$, then $x$ majorizes $y$. The sorted $x$ is $\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$ and the sorted $y$ is $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$. The sequence of partial sums for $x$ is 3, 7, and 12, and the partial sums for $y$ is 1, 3, 12.

**Example 4.3.** If $x = \begin{bmatrix} 6 \\ 2 \\ -3 \end{bmatrix}$ and $y = \begin{bmatrix} 3 \\ -5 \\ 7 \end{bmatrix}$, then $x$ majorizes $y$. The sorted $x$ is $\begin{bmatrix} -3 \\ 2 \\ 6 \end{bmatrix}$ and the sorted $y$ is $\begin{bmatrix} -5 \\ 3 \\ 7 \end{bmatrix}$. The sequence of partial sums for $x$ is -3, -1, and 5, and the partial sums for $y$ is -5, -2, 5.

**Theorem 4.9.** *If $A \in M_n$ Hermitian, then* $\operatorname{diag}(A) = \begin{bmatrix} a_{11} \\ a_{22} \\ \vdots \\ a_{nn} \end{bmatrix}$ *majorizes* $\lambda(A) = \begin{bmatrix} \lambda_1(A) \\ \lambda_2(A) \\ \vdots \\ \lambda_n(A) \end{bmatrix}$.

*Proof of Theorem 4.9.* We prove by induction on $n$. This is trivially true when $n = 1$. Suppose the claim holds true for $n - 1$. Consider any arbitrary $A \in M_n$ Hermitian. Let $B \in M_{n-1}$ be a principal submatrix of $A$ obtained by deleting row and column $l_n$ where the diagonals of $A$ are ordered

$$a_{l_1,l_1} \leq a_{l_2,l_2} \leq a_{l_3,l_3} \leq \cdots \leq a_{l_n,l_n}.$$

Since $A$ is Hermitian, then $B$ is also Hermitian. Then for all $k = 1, 2, \ldots, n-1$, then

$$\sum_{i=1}^{k} \lambda_i(A) \leq \sum_{i=1}^{k} \lambda_i(B) \leq \sum_{i=1}^{k} a_{l_i,l_i},$$

where the first inequality is from Theorem 4.8 (Interlacing II) and the second inequality is by our induction hypothesis. Then when $k = n$, we have

$$\sum_{i=1}^{n} \lambda_i(A) = \operatorname{tr}(A) = \sum_{i=1}^{n} a_{l_i,l_i}.$$

□

**Theorem 4.10.** *Let $A \in M_n$ Hermitian, and $r$ such that $1 \leq r \leq n$, then*

$$\begin{aligned} \sum_{k=1}^{r} \lambda_k(A) &= \min_{\substack{U \in M_{n,r} \\ \text{orthonormal columns}}} \operatorname{tr}(U^*AU) \\ \sum_{k=0}^{r-1} \lambda_{n-k}(A) &= \max_{\substack{U \in M_{n,r} \\ \text{orthonormal columns}}} \operatorname{tr}(U^*AU) \end{aligned}$$

*Proof of Theorem 4.10.* Consider any matrix $U \in M_{n,r}$ with orthonormal columns. Extend this to a matrix $V = [U|*] \in M_n$ unitary and apply Gram-Schmidt to maintain orthonormality of the columns of $V$. Then

$$V^*AV = \begin{bmatrix} U^* \\ * \end{bmatrix} A \left[U|*\right] = \begin{bmatrix} U^*AU & * \\ * & * \end{bmatrix},$$

so for all since $V^*AV \sim A$ $k = 1, 2, \ldots, r$, $\lambda_k(A) = \lambda_k(V^*AV) \leq \lambda_k(U^*AU)$ by Theorem 4.8.

Summing over $k$ from 1 to $r$, we have

$$\sum_{k=1}^{r} \lambda_k(A) \leq \operatorname{tr}(U^*AU)$$

with equality when $U = [\ u_1 \mid u_2 \mid \cdots \mid u_r\ ]$, where $u_1, u_2, \dots u_r$ are orthonormal eigenvectors associated with $\lambda_1(A), \lambda_2(A), \dots, \lambda_r(A)$. Then,

$$\begin{aligned}
\operatorname{tr}(U^*AU) &= \operatorname{tr}\left(\begin{bmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_r^* \end{bmatrix} A[\ u_1 \mid u_2 \mid \cdots \mid u_r\ ]\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_r^* \end{bmatrix} [\ \lambda_1 u_1 \mid \lambda_2 u_2 \mid \cdots \mid \lambda_r u_r\ ]\right) \\
&= \operatorname{tr}\begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_r \end{bmatrix} \\
&= \sum_{i=1}^{r} \lambda_i(A)
\end{aligned}$$

□

**Corollary 4.5.** *Let $A, B \in M_n$ be Hermitian. Then*

$$\lambda(A+B) = \begin{bmatrix} \lambda_1(A+B) \\ \lambda_2(A+B) \\ \vdots \\ \lambda_n(A+B) \end{bmatrix} \text{ majorizes } \lambda(A) + \lambda(B) = \begin{bmatrix} \lambda_1(A) + \lambda_1(B) \\ \lambda_2(A) + \lambda_2(B) \\ \vdots \\ \lambda_n(A) + \lambda_n(B) \end{bmatrix}$$

*Proof of Corollary 4.5.* For all $k = 1, 2, \dots, n$,

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_i(A+B) &= \min_{\substack{U \in M_{n,k} \\ \text{orthonormal columns}}} \operatorname{tr}(U^*(A+B)U) \\
&\geq \min_{\substack{U \in M_{n,k} \\ \text{orthonormal columns}}} \operatorname{tr}(U^*AU) + \min_{\substack{U \in M_{n,k} \\ \text{orthonormal columns}}} \operatorname{tr}(U^*BU) \\
&= \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B),
\end{aligned}$$

where the last equality is by Theorem 4.10. We also have

$$\sum_{i=1}^{n} \lambda_i(A+B) = \operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B = \sum_{i=1}^{n} \lambda_i(A) + \sum_{i=1}^{n} \lambda_i(B).$$

□

**Theorem 4.11** (Hadamard's Inequality)**.** *Let $A \in M_n$ be positive semidefinite. Then* $\det A \leq \prod_{i=1}^{n} a_{ii}$.

*Proof of Theorem 4.11.* For all $i = 1, 2, \dots, n$, $0 \leq \lambda_i(A) \leq a_{ii}$. Thus, $\prod_{i=1}^{n} a_{ii} \geq 0$. If $A$ is singular, the result is trivial. Otherwise $A$ is positive definite and $0 < \lambda_1(A) \leq a_{ii}$. Thus, define $D := \operatorname{diag}(\frac{1}{\sqrt{a_{11}}}, \frac{1}{\sqrt{a_{22}}}, \dots, \frac{1}{\sqrt{a_{nn}}})$. Then, $DAD$ is Hermitian and positive definite since for all $x \in \mathbb{C}^n \setminus \{\vec{0}\}$, we have $x^*DADx = y^*Ay > 0$, where $y = Dx$. Then,

$$\frac{\det A}{\prod_{i=1}^{n} a_{ii}} = \det D \det A \det D = \det DAD = \prod_{i=1}^{n} \lambda_i(DAD) \leq \left(\frac{1}{n}\sum_{i=1}^{n} \lambda_i(DAD)\right)^n = \left(\frac{1}{n}\operatorname{tr}(DAD)\right)^n \leq 1.$$

□

# 5 Chapter 7 – Positive Definite and Semidefinite Matrices

## 5.1 Introduction of Singular Value Decomposition

**Theorem 5.1.** *Let $A \in M_{m,n}$ such that $m \leq n$. Then there exist $U \in M_m$ unitary, $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m) \in M_m$ diagonal with $\sigma_i \geq 0$ and $\sigma_i$'s in nonincreasing order ($\sigma_i \geq \sigma_{i+1}$ for all $i$), $W \in M_{m,n}$ with orthonormal rows such that $A = U\Sigma W$.*

*Proof of Theorem 5.1.* Observe that $AA^*$ is Hermitian and positive semidefinite since for all $x \in \mathbb{C}^n$, $x^*AA^*x = \|Ax\|_2^2 \geq 0$, so there exist $U = [\ \mu_1 \mid \mu_2 \mid \cdots \mid \mu_m\ ] \in M_m$ unitary and $D = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ with $\sigma_i \geq 0$ in nonincreasing order, such that $AA^* = UDU^*$. Define $\Sigma := \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m) \in M_m$.

Let $\operatorname{rank} A = k$ and say that $\sigma_k \neq 0$ and $\sigma_{k+1} = 0$. For $i = 1, 2, \ldots, k$, define the $i^{\text{th}}$ row of $W$ to be $\frac{1}{\sigma_i}\mu_i^* A$. It is easy to see that the columns are all orthonormal. For $i = k+1, k+2, \ldots, m$, define the $i^{\text{th}}$ row of $W$ in any way so long as they are orthonormal to all rows of $W$ (which can be done by Gram-Schmidt).

<u>Claim:</u> $U^*A = \Sigma W \implies A = U\Sigma W$. For rows $i = 1, 2, \ldots, k$ equality holds by definition since $\mu_i^* A = \sigma_i$, the $i^{\text{th}}$ row of $W$. We also have

$$AA^*\mu_i = \vec{0} \implies \mu_i^* AA^* \mu_i = \|A^*\mu_i\|_2^2 \implies A^*\mu_i = \vec{0}^T,$$

so the $i^{\text{th}}$ of the LHS is also the $i^{\text{th}}$ row of the RHS, which is $\vec{0}^T$. □

*Remark.* Note the following:

- If $A$ is real, then $U, \Sigma, W$ may be chosen real,
- $\sigma_i$ is uniquely determined.

**Theorem 5.2** (Singular Value Decomposition). *For all $A \in M_{m,n}$, there exist $U \in M_m$ unitary, $V \in M_n$ unitary, and $\Sigma \in M_{m,n}$ "diagonal" such that $A = U\Sigma V^*$.*

*Proof of Theorem 5.2.* If $m \leq n$, then by Theorem 5.1, then $A = U\Sigma W = U[\ \Sigma \mid 0\ ]\underbrace{\begin{bmatrix} W \\ * \end{bmatrix}}_{V^*}$, where $*$ is chosen to maintain orthonormality. If $m > n$, then say $A^* = U\Sigma V^*$ is an SVD of $A^*$. Then $V\Sigma^* U^*$ is an SVD of $A$. □

**Corollary 5.1.** *For all $A \in M_n$, there exist $P \in M_n$ Hermitian and $W \in M_n$ unitary such that $A = PW$. This is called a polar decomposition of $A$.*

*Proof of Corollary 5.1.* Say $A = U\Sigma V^*$ is an SVD. Then $A = \underbrace{U\Sigma U^*}_{P}\underbrace{UV^*}_{W}$. Since $U^*U = I$. □

*Note.* Let $A = U\Sigma V^*$ be an SVD with $U = [\ \mu_1 \mid \mu_2 \mid \cdots \mid \mu_m\ ]$, $V = [\ v_1 \mid v_2 \mid \cdots \mid v_m\ ]$ and $\sigma_k \neq 0$ and $\sigma_{k+1} = 0$. Then the following are true:

- $A = U\Sigma V^* = \sum_{i=1}^k \sigma_i \mu_i v_i^*$.
- $\operatorname{rank} A = \operatorname{rank} \Sigma = k$.
- $\operatorname{range} A = \operatorname{Span}\{\mu_1, \mu_2, \ldots, \mu_k\}$, so $Ax = \sum_{i=1}^k \sigma_i \mu_i (v_i^* x) = \sum_{i=1}^k (\sigma_i v_i^* x)\mu_i$.
- $\ker A = \operatorname{Span}\{v_{k+1}, v_{k+2}, \ldots, v_n\}$, so $Ax = A\sum_{j=1}^n c_j v_j = \sum_{i=1}^k \sigma_i \mu_i v_i^* \sum_{j=1}^n c_j v_j = \sum_{i=1}^k \sigma_i c_i \mu_i = \vec{0}$ if and only if $c_i = 0$ for all $i = 1, 2, \ldots, n$.
- $\operatorname{range} A^* = \operatorname{Span}\{v_1, v_2, \ldots, v_k\}$.
- $\ker A^* = \operatorname{Span}\{\mu_{k+1}, \mu_{k+2}, \ldots, \mu_m\}$, and $A^* = V\Sigma^* U^*$.

## 5.2 Consequences of Singular Value Decomposition and Generalized Inverses

**Definition 5.1** (Matrix 2-norm)**.** For all $A \in M_{m,n}$ define the *matrix 2-norm* as

$$\|A\|_2 := \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{\|Ax\|_2}{\|x\|_2}.$$

**Proposition 5.1.** *For all $A \in M_{m,n}$, $\|A\|_2 = \sigma_1(A)$, which is the largest singular value of $A$.*

*Proof of Proposition 5.1.* We show this directly:

$$\begin{aligned}\|A\|_2^2 &= \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2 \\ &= \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{x^* A^* A x}{x^* x} \\ &\underset{\text{R-R}}{=} \lambda_n(A^*A) \\ &= \sigma_1^2(A),\end{aligned}$$

where we use the fact that $A^*A = U\Sigma V^* V \Sigma^* U^* = U\Sigma\Sigma^* U^*$ is Hermitian and has eigenvalues $\sigma_i^2$. □

**Proposition 5.2.** *For all $A \in M_{m,n}$, $\|A\|_F = \sqrt{\sum_i \sigma_i^2(A)}$.*

*Proof of Proposition 5.2.*

$$\|A\|_F^2 = \|U\Sigma V^*\|_F^2 = \|\Sigma\|_F^2 = \sum_i \sigma_i^2.$$

□

**Definition 5.2** (C-generalized inverses)**.** Let $A \in M_{m,n}$ and $C \subseteq \{1,2,3,4\}$. The matrix $B \in M_{n,m}$ is called a *$C$-generalized inverse* of $A$ if

(i) $1 \in C \implies ABA = A$

(ii) $2 \in C \implies AB$ Hermitian

(iii) $3 \in C \implies BAB = B$

(iv) $4 \in C \implies BA$ Hermitian.

*Remark.* Here are some examples and about $C$-generalized inverses:

- {1,2}-generalized inverse means $A, B$ satisfy conditions (i) and (ii). It is possible the others are satisfied, but not guaranteed to hold always.
- An example of a {2,3,4}-generalized inverse is the 0 matrix.
- Note, if $A \in M_n$ invertible, then $A^{-1}$ is uniquely the $\{1,2,3,4\}$-generalized inverse of $B$

**Proposition 5.3.** *Let $A \in M_{m,n}$, $b \in \mathbb{C}^m$ be given. Suppose $B \in M_{n,m}$ is a 1-generalized inverse for $A$. If $Ax = b$ is consistent (there is at least 1 solution to the system), then $x = Bb$ is a solution.*

*Proof of Proposition 5.3.* Say $Az = b$ for $z \in \mathbb{C}^n$ since the system is consistent. Then,

$$A(Bb) = ABAz = Az = b.$$

□

**Theorem 5.3.** *Let $A \in M_{m,n}$ and $b \in C^m$ be given. Suppose $B \in M_{m,n}$ is a {1,2}-generalized inverse of $A$. Then $x = Bb$ solves $Ax = b$ in a least squares sense, that is,*

$$\min_{x \in \mathbb{C}^n} \|Ax - b\|_2$$

*has optimal solution $\hat{x} = Bb$. Note, there is no assumption on consistency.*

*Proof of Theorem 5.3.* Any vector in $\mathbb{C}^n$ can be expressed as $Bb + y$, where $y \in \mathbb{C}^n$. We will show that $\|A(Bb+y) - b\|_2^2$ is minimized when $y = \vec{0}$. Then,

$$\begin{aligned}\|A(Bb+y)-b\|_2^2 &= [(AB-I)b + Ay]^*[(AB-I)b+Ay] \\ &= \|(AB-I)b\|_2^2 + \|Ay\|_2^2 + y^*A^*(AB-I)b + b^*(AB-I)^*Ay \\ &= \|(AB-I)b\|_2^2 + \|Ay\|_2^2,\end{aligned}$$

which is optimal when $y = \vec{0}$. The third equality comes from the fact that $A^*(AB-I) = 0$ since $[A^*(AB-I)]^* = (AB-I)A = ABA - A = 0$.

Note, the set of solutions to $\min_{x \in \mathbb{C}^n} \|Ax - b\|_2$ are $Bb \oplus \ker A$, which is an affine space. □

## 5.3 The Moore-Penrose Inverse

**Theorem 5.4.** *For every $A \in M_{m,n}$, there exists a unique $\{1,2,3,4\}$-generalized inverse of $A$. It is called the Moore-Penrose generalized inverse.*

*Proof of Theorem 5.4.* We first show existence, consider the first special case where $\Sigma \in M_{m,n}$ is diagonal (i.e. $\forall i \neq j, \Sigma_{ij} = 0$). Define $\Sigma^\dagger \in M_{n,m}$ diagonal such that for all $i$,

$$\Sigma^\dagger_{ii} = \begin{cases} 1/\Sigma_{ii} & \text{if } \Sigma_{ii} \neq 0 \\ 0 & \text{if } \Sigma_{ii} = 0. \end{cases}$$

By inspection, then $\Sigma^\dagger$ is the {1,2,3,4}-generalized inverse of $\Sigma$:

(i) $\Sigma\Sigma^\dagger\Sigma = \Sigma$

(ii) $\Sigma\Sigma^\dagger = \begin{bmatrix} 0 & & & & 0 \\ & 0 & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 0 \end{bmatrix}$

(iii) $\Sigma^\dagger\Sigma\Sigma^\dagger = \Sigma^\dagger$

(iv) $\Sigma^\dagger\Sigma = \begin{bmatrix} 1 & & & & 0 \\ & 0 & & & \\ & & 0 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}$

Next, consider any $E \in M_{m,n}$ with a {1,2,3,4}-generalized inverse $F \in M_{n,m}$. Let $U \in M_m$ unitary, $V \in M_n$ unitary. Then $UEV^*$ has a {1,2,3,4}-generalized inverse $VFU^*$ since $(UEV^*)(VFU^*)(UEV^*) = UEV^*$, so $UEFU^*$ is Hermitian since $EF$ is Hermitian. Thus, for $A$, say $A = U\Sigma V^*$ is an SVD. By the above, the {1,2,3,4}-generalized inverse is $V\Sigma^\dagger U^* = A^\dagger$. Note, that if $A$ is real, then $A^\dagger$ is real. Also, note that $\left(A^\dagger\right)^\dagger = A$.

We now show uniqueness. Suppose $B, C \in M_{m,n}$ such that $B, C$ are $\{1,2,3,4\}$-generalized inverses of $A$. We show that $AB = AC$ and $BA = CA$:

$$AB = ACAB = C^*A^*B^*A^* = C^*(ABA)^* = C^*A^* = AC$$

$$BA = BACA = A^*B^*A^*C^* = (ABA)^*C^* = A^*C^* = CA$$

Thus, $B = BAB = CAB = CAC = C$. □

**Theorem 5.5.** *Let $A \in M_{m,n}$ and $b \in \mathbb{C}^n$ be given. Among all least square solutions to $\min_{x \in \mathbb{C}^n} \|Ax - b\|_2$, $A^\dagger b$ is the unique solution of minimum 2-norm.*

*Proof of Theorem 5.5.* The solutions to the least squares problem are $\{A^\dagger b + y : y \in \ker A\}$. Let $A = U\Sigma V^*$ be an SVD with $k = \operatorname{rank} A$. Then the following are true:

- $\ker A = \operatorname{Span}\{v_{k+1}, v_{k+2}, \dots, v_n\}$
- $\operatorname{range} A^\dagger = \operatorname{Span}\{v_1, v_2, \dots, v_k\}$ since $A^\dagger = V\Sigma^\dagger U^*$ is (almost) an SVD (since $\sigma_i$'s are out of order).

Thus, $\forall y \in \ker A$, $y \perp A^\dagger b$ by orthonormality of $V$. Consequently, for any $y \in \ker A$,

$$\left\|A^\dagger b + y\right\|_2^2 = (A^\dagger b + y)^*(A^\dagger b + y) = \left\|A^\dagger b\right\|_2^2 + \|y\|_2^2 + 0 + 0$$

The minimum is clearly obtained when $y = \vec{0}$ uniquely. Thus, $A^\dagger b$ is the unique solution of minimum 2-norm. ☐

*Note.* If $A \in M_{m,n}$ has full column rank, then $A^\dagger = (A^*A)^{-1}A^*$. Similarly, if $A \in M_{m,n}$ has full row rank, then $A^\dagger = A^*(AA^*)^{-1}$. In either of these cases, $AA^*$ is invertible since, if $A = U\Sigma V^*$ is an SVD, then

$$AA^* = V\Sigma^* U U^* \Sigma V^* = V^*\Sigma^*\Sigma V^* = V \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} V^*,$$

where none of the $\sigma_i^2$ terms are 0 by full rank. Then,

$$\begin{aligned} AA^\dagger &= AA^*(AA^*)^{-1} = I \\ A^\dagger A &= (A^*A)^{-1}A^*A = I \end{aligned}$$

*Note.* $A^\dagger A$ and $AA^\dagger$ are orthogonal projections onto the range of $A$ and $A^*$, respectively. This is because the two matrices are also Hermitian by the properties of a {1,2,3,4}-generalized inverse and they are idempotent since $(A^\dagger A)^2 = A^\dagger A A^\dagger A = A^\dagger A$.

# 6 Chapter 5 – Norms for Vectors and Matrices

## 6.1 Inner Product and Normed Linear Spaces

In this chapter, we denote $V$ as a vector space over $\mathcal{K}$, where $\mathcal{K}$ is either $\mathbb{R}$ or $\mathbb{C}$.

**Definition 6.1** (Inner product)**.** Suppose $V$ is a vector space over $\mathcal{K}$. We define an *inner product* to be a function $\langle \cdot, \cdot \rangle : V \times V \to \mathcal{K}$ such that $\forall x, y, z \in V$, $c \in \mathcal{K}$, the following axioms hold:

(i) $\langle x, x \rangle$ is real and non-negative, with $\langle x, x \rangle = 0$ if and only if $x = \vec{0}$

(ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(iii) $\langle cx, z \rangle = c\langle x, z \rangle$

(iv) $\langle x, z \rangle = \overline{\langle z, x \rangle}$

**Definition 6.2** (Inner product space)**.** We call the pair $(V, \langle \cdot, \cdot \rangle)$ an *inner product space* (IPS).

**Example 6.1.** In $\mathbb{C}^n$ over $\mathbb{C}$, an example of an inner product is $\forall x, y$ we have $\langle x, y \rangle := y^* x$.

**Example 6.2.** In $\mathbb{C}^n$ over $\mathbb{C}$, and $A \in M_n$ positive semidefinite, an example of an inner product is $\forall x, y$ we have $\langle x, y \rangle_A := y^* A x$.

*Note.* For all $x, y, z, w \in V$ and $a, b, c, d \in \mathcal{K}$,

(1) $\langle x, cz \rangle = \overline{\langle cz, x \rangle} = \overline{c\langle z, x \rangle} = \bar{c}\overline{\langle z, x \rangle} = \bar{c}\langle x, z \rangle$

(2) $\langle x, y + z \rangle = \overline{\langle y + z, x \rangle} = \overline{\langle y, x \rangle + \langle z, x \rangle} = \langle x, y \rangle + \langle x, z \rangle$

(3) $\langle ax + by, cz + dw \rangle = a\bar{c}\langle x, z \rangle + a\bar{d}\langle x, w \rangle + b\bar{c}\langle y, z \rangle + b\bar{d}\langle y, w \rangle$

**Definition 6.3** (Vector norm)**.** Suppose $V$ is a vector space over $\mathcal{K}$. We define a *(vector) norm* to be a function $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ such that $\forall x, y, \in V$, $c \in \mathcal{K}$, the following axioms hold:

(i) $\|x\| = 0$ if and only if $x = \vec{0}$ (positivity)

(ii) $\|cx\| = |c| \, \|x\|$ (homogeneity)

(iii) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

**Definition 6.4** (Normed linear space)**.** We call the pair $(V, \|\cdot\|)$ a *normed linear space* (NLS).

**Example 6.3.** In $\mathcal{K}^n$ over $\mathcal{K}$, given a positive integer $p$, $\forall x \in \mathcal{K}^n$, the $L_p$-norm is $\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. When $p = 1$, we call the $L_1$-norm the *Manhattan norm*. We can also define $\|x\|_\infty := \max_i |x_i|$, which is indeed a norm (the triangle inequality requires utilizing Holder's inequality).

**Example 6.4.** Suppose we have the NLS $(\mathcal{K}^n, \|\cdot\|)$ over $\mathcal{K}$ and $A \in M_n(\mathcal{K})$ invertible. Define $\forall x \in \mathcal{K}^n$, $\|x\|_A := \|Ax\|$ is a norm. Observe that the triangle inequality holds since

$$\|x + y\|_A = \|A(x + y)\| = \|Ax + Ay\| \leq \|Ax\| + \|Ay\| = \|x\|_A + \|y\|_A .$$

**Theorem 6.1** (Cauchy-Schwarz Inequality)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be an IPS. Then, $\forall x, y \in V$,*

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

*Proof of Theorem 6.1.* For $\mathcal{K} = \mathbb{C}$, then $\forall x, y \in V$ and $\forall t, \theta \in \mathbb{R}$,

$$\begin{aligned}
0 &\leq \langle te^{i\theta}x + y, te^{i\theta}x + y \rangle \\
&= t^2 e^{i\theta} e^{-i\theta} \langle x, x \rangle + te^{i\theta} \langle x, y \rangle + te^{-i\theta} \langle y, x \rangle + \langle y, y \rangle \\
&= \langle x, x \rangle t^2 + 2\Re(e^{i\theta} \langle x, y \rangle)t + \langle y, y \rangle
\end{aligned}$$

Observe that the final line is a quadratic polynomial in $t$. Then, choosing $\theta$ such that $\Re(e^{i\theta}\langle x,y\rangle) = |\langle x,y\rangle|$ $\langle x,x\rangle t^2 + 2|\langle x,y\rangle|t + \langle y,y\rangle \geq 0$. This only occurs when the quadratic polynomial has no real roots in $t$, i.e. when the discriminant is negative:

$$\begin{aligned} &(2|\langle x,y\rangle|)^2 - 4\langle x,x\rangle\langle y,y\rangle \leq 0 \\ \iff \quad & |\langle x,y\rangle|^2 \leq \langle x,x\rangle\langle y,y\rangle. \end{aligned}$$

□

**Theorem 6.2.** *If we have an IPS $(V, \langle\cdot,\cdot\rangle)$ over $\mathcal{K}$, it induces a norm $\|\cdot\|$. In particular, $\forall x \in V$, $\|x\| := \sqrt{\langle x,x\rangle}$.*

Observe that we have positivity of $\|\cdot\|$ by positivity of $\langle\cdot,\cdot\rangle$. We also have homogeneity. It suffices to show triangle inequality.

*Proof of Theorem 6.2.* We show the triangle inequality of the induced norm $\|\cdot\|$ is satisfied. Indeed, $\forall x,y \in V$,

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y\rangle \\ &= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle \\ &= \|x\|^2 + \|y\|^2 + 2\Re(\langle x,y\rangle) \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x,y\rangle| \\ &\underset{\text{C-S}}{\leq} \|x\|^2 + \|y\|^2 + 2\,\|x\|\,\|y\| \\ &= (\|x\| + \|y\|)^2 \end{aligned}$$

□

## 6.2 Hilbert Spaces and Banach Spaces

**Definition 6.5** (Metric space)**.** Recall from real analysis that we define a *metric space* to be a pair $(S,d)$, where $S$ is a set and $d : S \times S \to \mathbb{R}_{\geq 0}$ is a function (i.e. a distance metric) such that $\forall x,y,z \in S$, the following axioms hold:

(i) $d(x,y) = 0$ if and only if $x = y$

(ii) $d(x,y) = d(y,x)$

(iii) $d(x,z) \leq d(x,y) + d(y,z)$

*Note.* If $(V, \|\cdot\|)$ is a NLS, then define the distance metric

$$d(x,y) := \|x-y\| \quad \forall x,y \in V.$$

Indeed, this induces a metric space $(V,d)$ since

(i) $\|x-y\| = 0$ if and only if $x - y = \vec{0}$ if and only if $x = y$

(ii) $\|y-x\| = \|(-1)(x-y)\| = |-1|\,\|x-y\| = \|x-y\|$

(iii) $\|x-z\| = \|x-y+y-z\| \leq \|x-y\| + \|y-z\|$,

where the final inequality in (iii) is due to the triangle inequality of the norm $\|\cdot\|$.

**Definition 6.6** (Banach Space)**.** We saw that an NLS $(V, \|\cdot\|)$ induces a metric $d(x,y) := \|x-y\|$. If the metric space $(V,d)$ is also complete, that is, if all Cauchy sequences converge in $V$, then we call $V$ a *Banach space*. More succinctly, a Banach space is a complete normed vector space.

**Definition 6.7** (Hilbert Space)**.** We saw that an IPS $(V, \langle\cdot,\cdot\rangle$ induces a norm $\|x\| = \sqrt{\langle x, x\rangle}$, which induces a distance metric $d(x, y) := \|x - y\|$. If the metric space $(V, d)$ is also complete, that is, if all Cauchy sequences converge in $V$, then we call $V$ a *Hilbert space.* More succinctly, a Hilbert space is a complete inner product space.

**Theorem 6.3.** *If we have a NLS $(V, \|\cdot\|)$ with $V$ being finite-dimensional, then $V$ is also a metric space with an associated distance metric $d(x, y) := \|x - y\|$, and in particular, $V$ is complete. If we have an IPS $(V, \langle\cdot,\cdot\rangle)$ with $V$ being finite-dimensional, then $V$ is also a NLS with an associated norm $\|x\| = \sqrt{\langle x, x\rangle}$ and a metric space with an associated distance metric $d(x, y) := \|x - y\|$, and in particular, $V$ is complete.*

To give some intuition to the above theorem, consider an NLS $(V, \|\cdot\|)$ and observe that

$$|\,\|x\| - \|y\|\,| \leq \|x - y\| \quad \forall x, y \in V.$$

This tell us that $\|\cdot\|$ is a (Lipschitz) continuous function since $\|y + x - y\| \leq \|y\| + \|x - y\|$.

**Theorem 6.4.** *Let $(V, \|\cdot\|)$ be a NLS. The following are equivalent:*

*(1) $V$ is finite dimensional*

*(2) The unit sphere $\{x \in V : \|x\| = 1\}$ is compact, that is, every sequence has a converging subsequence*

*(3) The unit ball $\{x \in V : \|x\| \leq 1\}$ is compact, that is, every sequence has a converging subsequence*

*(4) $S \subseteq V$ is compact if and only if $S$ is closed and bounded.*

**Theorem 6.5.** *On a finite-dimensional vector space, all norms are equivalent. That is, for $(V, \|\cdot\|)$ and $(V, \|\cdot\|')$, where $V$ is finite-dimensional, then there exist positive real numbers $m, M$ (which can depend on the chosen pair of norms) such that*

$$m\,\|x\|' \leq \|x\| \leq M\,\|x\|' \quad \forall x \in V.$$

**Corollary 6.1.** *Suppose we have a finite-dimensional vector space $V$ and normed linear spaces $(V, \|\cdot\|)$ and $(V, \|\cdot\|')$. Let $\{x^{(k)}\}_{k=1}^{\infty} \subseteq V$ be a sequence in $V$ and $x \in V$. Then $x^{(k)} \underset{\|\cdot\|}{\longrightarrow} x$ if and only if $x^{(k)} \underset{\|\cdot\|'}{\longrightarrow} x$. This holds for all notions of convergence.*

*Proof Sketch of Corollary 6.1.* By the squeeze theorem, if $\left\|x^{(k)} - x\right\|' \longrightarrow 0$, then $\left\|x^{(k)} - x\right\| \longrightarrow 0$ since $\left\|x^{(k)} - x\right\| \leq M\left\|x^{(k)} - x\right\|'$. Similarly, by the squeeze theorem, if $\left\|x^{(k)} - x\right\| \longrightarrow 0$, then $\left\|x^{(k)} - x\right\|' \longrightarrow 0$ since $\left\|x^{(k)} - x\right\|' \leq \frac{1}{m}\left\|x^{(k)} - x\right\|$. □

**Corollary 6.2.** *Consider the NLS $(\mathcal{K}, \|\cdot\|)$ over $\mathcal{K}$. For a sequence $\{x^{(k)}\} \subseteq \mathcal{K}^n$ and $x \in \mathcal{K}$, $x^{(k)} \underset{\|\cdot\|}{\longrightarrow} x$ if and only if $[x^{(k)}]_j \longrightarrow x_j$ for all $j = 1, 2, \ldots, n$. That is, we have component-wise convergence.*

*Proof of Corollary 6.2.* $\left\|x^{(k)} - x\right\| \longrightarrow 0$ if and only if $\left\|x^{(k)} - x\right\|_\infty \longrightarrow 0$ if and only if $|x_j^{(k)} - x_j| = \max_j |x_j^{(k)} - x_j| \longrightarrow 0$ for all $j = 1, 2, \ldots, n$. □

**Definition 6.8** (Linear transformation)**.** Recall from linear algebra, for two vector spaces $V, W$ over $\mathcal{K}$, a *linear transformation* $T : V \to W$ is a function such that for all $x, y \in V$ and $\alpha, \beta \in \mathcal{K}$,

$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y).$$

**Example 6.5.** Any matrix $A \in M_{m,n}(\mathcal{K}) : \mathcal{K}^n \to \mathcal{K}^m$ is a linear transformation.

**Example 6.6.** The derivative operator $\frac{d}{dt} : c^1[a, b] \to c^0[a, b]$ is a linear transformation since for $f, g \in c^1[a, b]$ (the set of continuous and once-differentiable functions over $[a, b]$), $(\alpha f + \beta g)'(x) = \alpha f'(x) + \beta g'(x)$ for all $x \in [a, b]$.

Recall that if $V$ is a finite-dimensional vector space over $\mathcal{K}$, say with $\dim V = n$, then $V$ is isomorphic to $\mathcal{K}^n$ over $\mathcal{K}$, that is, there is a 1-1 correspondence between elements of the vector space $V$ and the $n$-vectors in $\mathcal{K}^n$. Suppose that $\mathcal{B} = \{b^{(1)}, b^{(2)}, \ldots, b^{(n)}\}$ is a basis for $V$. Then for all $x \in V$, there exist unique $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathcal{K}$ such that $x = \sum_{i=1}^n \alpha_i b^{(i)}$. This gives the following correspondence:

$$x \longleftrightarrow \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = x_{\mathcal{B}},$$

where $x_{\mathcal{B}}$ is the representation of the object $x \in V$ under the basis $\mathcal{B}$.

Recall that if we have finite-dimensional vector spaces $V, W$ which have associated bases $\mathcal{B} = \{b^{(1)}, b^{(2)}, \ldots, b^{(n)}\}$ for $V$ and $\mathcal{B}' = \{b'^{(1)}, b'^{(2)}, \ldots, b'^{(n)}\}$ for $W$, then $T : V \to W$ is a linear transformation if and only if there exists $A \in M_{m,n}(\mathcal{K})$ such that for every $x \in V, y \in W$, $T(x) = y$ if and only if $Ax_{\mathcal{B}} = y'_{\mathcal{B}}$.

## 6.3 Dual Spaces and Operator Norms

**Definition 6.9** (Linear functional)**.** Let $V$ be a vector space over $\mathcal{K}$. A linear transformation $T : V \to \mathcal{K}$ is a *linear functional* because it maps a vector to its underlying field of scalars.

*Note.* The linear functionals on $\mathcal{K}^n$ over $\mathcal{K}$ are precisely of the form $T : \mathcal{K}^n \to K$. A linear functional applied to a vector $x \in \mathcal{K}^n$ can be thought of as left multiplying $x$ by a $1 \times n$ matrix.

**Definition 6.10** (Linear functional in $\mathcal{K}^n$)**.** Let $w \in \mathcal{K}^n$. Then $\hat{w}$ is a linear functional $\hat{w} : \mathcal{K}^n \to \mathcal{K}$ whereby for all $x \in \mathcal{K}^n$, $\hat{w}(x) := w^* x$.

**Theorem 6.6.** *Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two normed linear spaces, and let $T : V \to W$ be a linear operator. $T$ is continuous if and only if*

$$\sup_{x \in V \setminus \{\vec{0}\}} \frac{\|Tx\|_W}{\|x\|_V} < \infty.$$

*The following two observations follow from linearity and a re-expression of the operator norm expression:*

*(i) $T(\vec{0}) = \vec{0}$ since $T(\vec{0}) = T(0 \cdot \vec{0}) = 0 \cdot T(\vec{0}) = \vec{0}$, and*

*(ii) $\sup_{x \in V \setminus \{\vec{0}\}} \frac{\|Tx\|_W}{\|x\|_V} = \sup_{x \in V \setminus \{\vec{0}\}} \left\| T\left(\frac{1}{\|x\|_V} x\right) \right\|_W = \sup_{\substack{z \in V \setminus \{\vec{0}\} \\ \|z\|_V = 1}} \|Tz\|_W$.*

*Proof of Theorem 6.6.* ($\Longleftarrow$) Suppose that for all $x \in V \setminus \{\vec{0}\}$, $\frac{\|Tx\|}{\|x\|} \leq M$ for some $M < \infty$. In particular, for all $y, z \in V$,

$$\frac{\|T(y-z)\|}{\|y-z\|} \leq M \quad \Longrightarrow \quad \|Ty - Tz\| \leq M \|y - z\|.$$

The last inequality tells us that $T$ is Lipschitz continuous.

($\Longrightarrow$) Suppose $T$ is continuous. In particular, there exists $\delta > 0$ such that for all $y$ such that $\left\|y - \vec{0}\right\| \leq \delta$ implies $\left\|Ty - T\vec{0}\right\| \leq \epsilon = 1$. Thus, for all $x \in V \setminus \{\vec{0}\}$,

$$\frac{\|Tx\|}{\|x\|} = \frac{1}{\delta} \left\| T\left(\delta \frac{1}{\|x\|} x\right) \right\| \leq \frac{1}{\delta} \cdot 1 < \infty.$$

□

**Theorem 6.7.** *Suppose $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ are two normed linear spaces with $V$ being finite-dimensional, and $T : V \to W$ is linear. Then $T$ is continuous.*

**Example 6.7.** An example of a discontinuous linear operator in infinite-dimensional space. Consider the derivative operator $\frac{d}{dt} : c^1[0,1] \to c^0[0,1]$. Clearly $\frac{d}{dt}$ is linear since for any $f, g \in c^1[0,1]$ and $\alpha, \beta \in \mathcal{K}$, $(\alpha f + \beta g)' = \alpha f' + \beta g'$. Define $\|f\| := \max_{t\in[0,1]} |f(t)|$ which is a norm in both the domain and codomain of $\frac{d}{dt}$. For $t^k$ on $[0,1]$, $\|t^k\| = 1$ and $\left\|\frac{d}{dt}t^k\right\| = k$. Since $\frac{\|\frac{d}{dt}t^k\|}{\|t^k\|} = k$ is unbounded above (take $k$ as large as you like), then the derivative operator is not continuous relative to $\|\cdot\|$.

**Definition 6.11** (Operator norm)**.** Let $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ are normed linear spaces, $T : V \to W$ be linear. If $T$ is continuous, then the *operator norm* of $T$ is

$$\|T\|_{(V,W)} := \sup_{x\in V\setminus\{\vec{0}\}} \frac{\|Tx\|_W}{\|x\|_V}.$$

*Note.* If $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ are normed linear spaces, $T, S : V \to W$ are linear, and $\alpha, \beta \in \mathcal{K}$, then define $\alpha T + \beta S : V \to W$ as

$$[\alpha T + \beta S](x) = \alpha T(x) + \beta S(x) \qquad \forall x \in V,$$

which shows that $\alpha T + \beta S$ is linear.

If $T, S$ are also continuous, then

(i) $\|T\| = 0$ if and only if $T \equiv 0$, i.e. $T$ is the zero function

(ii) $\|\alpha T\| = |\alpha| \, \|T\|$

(iii) $\|T + S\| \le \|T\| + \|S\|$

Observe that the above three consequences from the continuity of $T, S$ are true and tell us that the operator norm is indeed a norm since

(i) If $T \not\equiv 0$, then there exists $x \neq \vec{0}$ such that $T(x) \neq \vec{0}$, and so $\frac{\|Tx\|}{\|x\|} > 0$, so the supremum is positive as well, and therefore, $\|T\| > 0$.

(ii) $\sup \frac{\|\alpha T(x)\|}{\|x\|} = \sup |\alpha| \frac{\|Tx\|}{\|x\|} = |\alpha| \sup \frac{\|Tx\|}{\|x\|} = |\alpha| \, \|T\|$

(iii) $\sup \frac{\|[T+S](x)\|}{\|x\|} = \sup \frac{\|Tx+Sx\|}{\|x\|} \le \sup \frac{\|Tx\|+\|Sx\|}{\|x\|} \le \sup \frac{\|Tx\|}{\|x\|} + \sup \frac{\|Sx\|}{\|x\|}$

**Definition 6.12** (Continuous linear functionals)**.** Given normed linear spaces $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$, the set $\mathcal{B}(V, W)$ is the *set of continuous linear functions* $V \to W$. This set is a NLS with operator norm $\|\cdot\|$.

**Definition 6.13** (Dual space)**.** Given $(V, \|\cdot\|)$ NLS, $\mathcal{B}(V, \mathcal{K})$, which is the set of linear functionals on $V$, sometimes denoted $V^*$, is the *dual* of $V$.

*Note.* If we have $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ normed linear spaces with $T : V \to W$ continuous, that is $T \in \mathcal{B}(V, W)$, then for all $x \in V$,

$$\|Tx\| \le \|T\| \, \|x\|.$$

This is clear since $\sup \frac{\|Tx\|}{\|x\|} = \|T\|$, so $\frac{\|Tx\|}{\|x\|} \le \|T\|$.

*Note.* Consider normed linear spaces $(V, \|\cdot\|)$, $(W, \|\cdot\|)$, $(U, \|\cdot\|)$, and $T \in \mathcal{B}(V, W)$, $S \in \mathcal{S}(W, V)$ then $S \circ T \in \mathcal{B}(V, U)$. Also,

$$\|S \circ T\| \le \|S\| \, \|T\|$$

since for all $x \in V \setminus \{\vec{0}\}$

$$\|[S \circ T](x)\| = \|S(T(x))\| \le \|S\| \, \|Tx\| \le \|S\| \, \|T\| \, \|x\|$$

by the preceding note.

*Note.* If $A \in M_{m,n}(\mathcal{K}) : \mathcal{K}^n \to \mathcal{K}^m$, $B \in M_{m,n}(\mathcal{K}) : \mathcal{K}^n \to \mathcal{K}^m$, $C \in M_{p,m}(\mathcal{K}) : \mathcal{K}^m \to \mathcal{K}^p$, then $CA \in M_{p,n} : \mathcal{K}^n \to \mathcal{K}^p$ consisting of $C \circ A$. For any $\alpha, \beta \in \mathcal{K}$,

$$\alpha A + \beta B \in M_{m,n} : \mathcal{K}^n \to \mathcal{K}^m$$

is represented by a matrix $\alpha A + \beta B$.

## 6.4 Dual Norms and Algebraic Properties of Norms

*Note.* For all $A \in M_{m,n}$, $\|A\|_{2,2}$ is the max singular value $\sigma_1(A)$ since

$$\|A\|_{2,2} = \max_x \frac{\|Ax\|}{\|x\|}.$$

**Proposition 6.1.** *For all $A \in M_{m,n}$,*

$$\|A\|_{1,1} = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$$

$$\|A\|_{\infty,\infty} = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|,$$

*that is, $\|A\|_{1,1}$ is the largest column sum of $A$ and $\|A\|_{\infty,\infty}$ is the largest row sum of $A$.*

*Proof of Propositon 6.1.* Let $x \in \mathbb{C}^n$ be nonzero. Then

$$\|Ax\|_1 = \sum_{i=1}^{m} |\sum_{j=1}^{n} a_{ij}x_j| \le \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{m} |a_{ij}| \le \sum_{j=1}^{n} |x_j| \left( \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}| \right) = \|x\|_1 \left( \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}| \right),$$

and so, $\frac{\|Ax\|_1}{\|x\|_1} \le \max_j \sum_{i=1}^{m} |a_{ij}|$. Since this holds for any arbitrary $x$, then it holds for all $x$. Let $\hat{j} = \arg\max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$. Observe that $\left\|e_{\hat{j}}\right\| = 1$, then

$$\left\|Ae_{\hat{j}}\right\|_1 = \left\|A_{\hat{j}}\right\| = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|,$$

so $e_{\hat{j}}$ achieves equality. Then the operator norm of $A$ maximizes the quantity, which can be achieved. □

*Note.* Let $(V, \|\cdot\|)$ be a finite-dimensional NLS. So all linear functionals for $V$ are continuous. Then, $V^* \underset{\text{isom.}}{\sim} V$ as a vector space. If we fix a basis $B$ for $V$, then there is a one-to-one correspondence – for each $y \in V$, there is a corresponding linear functional that maps $x \mapsto y_B^T x_B$ for all $x \in V$.

**Definition 6.14** (Dual norm)**.** Let $\|\cdot\|$ be a norm on $\mathcal{K}^n$. The *dual norm* $\|\cdot\|^D$ on $\mathcal{K}^n$ is defined as

$$\|y\|^D := \max_{x \in \mathcal{K}^n \setminus \{\vec{0}\}} \frac{|y^*x|}{\|x\|} = \max_{\substack{x \in \mathcal{K}^n \\ \|x\|=1}} |y^*x|.$$

*Note.* Indeed $\|\cdot\|^D$ is a norm on $\mathcal{K}^n$ since it is an operator norm (in the dual space) of a linear functional $\hat{y}$, where $\hat{(y)}(x) = y^*x$. To see this, simply substitute $T$ into the definition of the dual norm.

**Lemma 6.1.** *Let $(V, \|\cdot\|)$ be a finite-dimensional NLS. For all $x, y \in \mathcal{K}^n$, $|y^*x| \le \|x\| \, \|y\|^D$.*

*Proof of Lemma 6.1.* This is direct:

$$\max_x \frac{|y^*x|}{\|x\|} = \|y\|^D \quad \Longrightarrow \quad \frac{|y^*x|}{\|x\|} \le \|y\|^D \quad \Longrightarrow \quad |y^*x| \le \|x\| \, \|y\|^D.$$

□

**Fact 6.1** (Holder's inequality)**.** For all $x, y \in \mathcal{K}^n$, $|y^*x| \le \|x\|_1 \|y\|_\infty$ since

$$|y^*x| = |\sum_i \bar{y}_i x_i| \le \sum_i |y_i||x_i| \le \sum_i (\max_j |y_j|)|x_i| = \max_j |y_j| \sum_i |x_i| = \|y\|_\infty \|x\|_1.$$

**Proposition 6.2.** *On $\mathcal{K}^n$, $\|\cdot\|_1^D = \|\cdot\|_\infty$, $\|\cdot\|_\infty^D = \|\cdot\|_1$, and $\|\cdot\|_2^D = \|\cdot\|_2$.*

*Proof of Proposition 6.2.* We leverage Holder's inequality.

- Given any $y \in \mathcal{K}^n$, if we restrict to $x$ such that $\|x\|_1 = 1$, then we have that $|y^*x| \leq \|y\|_\infty$ by Holder's inequality with equality for unit length $x$ with 1 in the component of $\arg\max_i |y_i|$ and 0 elsewhere. Then,
$$\|y\|_1^D = \max_{x:\|x\|=1} |y^*x| \leq \|y\|_\infty \qquad \forall y \in \mathcal{K}^n$$
with equality when $x = e_{\hat{j}}$, where $\hat{j} = \arg\max_i |y_i|$.

- Given any $x \in \mathcal{K}^n$, if we restrict to $y$ such that $\|y\|_\infty = 1$, then we have that $|y^*x| \leq \|x\|_1$ by Holder's inequality with equality for unit length $y$ with all components as unit complex numbers/rotations $e^{i\theta}$. Then,
$$\|x\|_\infty^D = \max_{y:\|y\|_\infty=1} |x^*y| \leq \|x\|_1 \qquad \forall x \in \mathcal{K}^n$$
with equality when $y$ has all unit complex components.

- Given any $y \in \mathcal{K}^n$, if we restrict to $x$ such that $\|x\|_2 = 1$, then we have that $|y^*x| \leq \|y\|_2$ by Cauchy-Schwarz with equality for unit length $x = \frac{1}{\|y\|_2} y$. Then,
$$\|y\|_2^D = \max_{x:\|x\|_2=1} |y^*x| \leq \|y\|_2 \qquad \forall y \in \mathcal{K}^n$$
with equality when $x = \frac{1}{\|y\|_2} y$.

□

**Fact 6.2** (Hahn-Banach Theorem)**.** Let $(V, \|\cdot\|)$ be a NLS over $\mathcal{K}$. If $x \in V$ is nonzero, then there exists $f \in V^*$ such that $\|f\|_{V^*} = 1$ and $f(x) = \|x\|_V$.

**Corollary 6.3** (Corollary of Hahn-Banach Theorem)**.** *Let $\|\cdot\|$ be a norm of $\mathcal{K}^n$. Then, $\left(\|\cdot\|^D\right)^D = \|\cdot\|$.*

*Proof of Corollary 6.3.* Let $x \in \mathcal{K}^n \setminus \{\vec{0}\}$. Then
$$\left(\|x\|^D\right)^D = \max_{y:\|y\|^D=1} |y^*x| \leq \max_{y:\|y\|^D=1} \|y\|^D \|x\| = \|x\|.$$
By the Hahn-Banach Theorem, there exists $y \in \mathcal{K}^n$ such that $\|y\|^D = 1$ and $y^*x = \|x\|$. Thus, we have equality. □

## 6.5 Induced Matrix Norms

**Theorem 6.8.** *Let $\|\cdot\|$ be a norm on $\mathcal{K}^n$. For all $A \in M_n(\mathcal{K})$,*
$$\|A\|_{\|\cdot\|,\|\cdot\|} = \|A^*\|_{\|\cdot\|^D,\|\cdot\|^D}$$

*Proof of Theorem 6.8.* For all $A \in M_n(\mathcal{K})$,
$$\begin{aligned}
\|A^*\|_{\|\cdot\|^D,\|\cdot\|^D} &= \max_{x:\|x\|^D=1} \|A^*x\|^D \\
&= \max_{x:\|x\|^D=1} \max_{y:\|y\|=1} |(A^*x)^*y| \\
&= \max_{y:\|y\|=1} \max_{x:\|x\|^D=1} |(Ay)^*x| \\
&= \max_{y:\|y\|=1} \left(\|Ay\|^D\right)^D \\
&= \max_{y:\|y\|=1} \|Ay\| \\
&= \|A\|_{\|\cdot\|,\|\cdot\|}.
\end{aligned}$$

□

**Definition 6.15** (Matrix norm)**.** A norm $\|\cdot\|$ on $M_n(\mathcal{K})$ over $\mathcal{K}$ is called a *matrix norm* if $\forall A, B \in M_n(\mathcal{K})$, $\|AB\| \leq \|A\| \|B\|$. This guarantees the following properties:

(i) $A \neq \mathbf{0} \Longrightarrow \|A\| > 0$

(ii) $\|\alpha A\| = |\alpha| \|A\|$

(iii) $\|A + B\| \leq \|A\| + \|B\|$

(iv) $\|AB\| \leq \|A\| \|B\|$

**Example 6.8.** In the following examples, consider the norm of any arbitrary matrix $A \in M_n$:

- The $l_1$ norm on matrices defined as $||A||_1 := \sum_{i,j} |a_{ij}|$ is a matrix norm.
- The $l_2$ norm on matrices defined as $||A||_2 = ||A||_F := \sqrt{\sum_{i,j} |a_{ij}|^2}$ is a matrix norm.
- The $l_\infty$ norm on matrices defined as $||A||_\infty := \max_{i,j} |a_{i,j}|$ is *not* a matrix norm. In particular, the submultiplicative property does not hold:

$$\left\| \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_\infty = \left\| \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \right\|_\infty \not\leq \left\| \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_\infty \left\| \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_\infty$$

Note that if we have a NLS $\mathcal{K}^n, \|\cdot\|'$ and consider matrices $M_n(\mathcal{K}) : \mathcal{K}^n, \|\cdot\|' \to \mathcal{K}^n, \|\cdot\|'$, then the operator norm $\|\cdot\|_{\|\cdot\|', \|\cdot\|'}$ is also a matrix norm.

**Definition 6.16** (Induced matrix norm)**.** Let $\|\cdot\|'$ be a norm on $\mathcal{K}^n$. Then $\|\cdot\|'$ induces a matrix norm $\|\cdot\|$ defined as

$$\|A\| := \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{\|Ax\|'}{\|x\|'}.$$

We call $\|\cdot\|$ an *induced matrix norm.*

- A necessary condition of an induced matrix norm $\|\cdot\|$ is that $\|I\| = 1$:

$$\|I\| = \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{\|Ix\|'}{\|x\|'} = \max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{\|x\|'}{\|x\|'} = 1.$$

- The following are induced matrix norms: $\|\cdot\|_{1,1}$, $\|\cdot\|_{2,2}$, and $\|\cdot\|_{\infty,\infty}$.

**Theorem 6.9.** *Let $\|\cdot\|$ be a matrix norm on $M_n$. Then for all $A \in M_n$, $\rho(A) \leq \|A\|$.*

*Proof of Theorem 6.9.* Let $A \in M_n$ and $x$ be an eigenvector associated with eigenvalue $\lambda$ of maximum modulus. Then define $B := [\ x \mid x \mid \cdots \mid x\ ] \in M_n$. Then,

$$|\lambda| \|B\| = \|\lambda B\| = \left\|[\ Ax \mid Ax \mid Ax \mid Ax\ ]\right\| = \|AB\| \leq \|A\| \|B\|.$$

Since $x \neq \vec{0}$, then $B \neq \mathbf{0}$, so $\|B\| > 0$. Then $\rho(A) = |\lambda| \leq \|A\|$. □

**Lemma 6.2.** *Let $\|\cdot\|$ be a matrix norm on $M_n(\mathcal{K})$ and $S \in M_n$ be an invertible matrix. Then $\|\cdot\|_S$ defined by $\forall A \in M_n(\mathcal{K})$, $\|A\|_S := \left\|S^{-1}AS\right\|$ is a matrix norm.*

*Proof of Lemma 6.2.* We check that $\|\cdot\|_S$ has all the properties of a matrix norm. $\forall A, B \in M_n$ and $\alpha \in \mathbb{C}$,

(i) $A \neq \mathbf{0} \Longrightarrow S^{-1}AS \neq \mathbf{0} \Longrightarrow \|A\|_S = \left\|S^{-1}AS\right\| > 0$

(ii) $\|\alpha A\|_S = \left\|S^{-1}\alpha AS\right\| = |\alpha| \left\|S^{-1}AS\right\| = |\alpha| \|A\|_S$

(iii) $\|A + B\|_S = \left\|S^{-1}(A + B)S\right\| = \left\|S^{-1}AS + S^{-1}BS\right\| \leq \left\|S^{-1}AS\right\| + \left\|S^{-1}BS\right\| = \|A\|_S + \|B\|_S$

(iv) $\|AB\|_S = \|S^{-1}ABS\| = \|S^{-1}ASS^{-1}BS\| \leq \|S^{-1}AS\| \|S^{-1}BS\| = \|A\|_S \|B\|_S$.

Indeed $\|\cdot\|_S$ is a matrix norm on $M_n(\mathcal{K})$. $\square$

**Theorem 6.10.** *Let $A \in M_n$ be fixed and $\epsilon > 0$. Then there exists a matrix norm $\|\cdot\|$ on $M_n$ such that $\|A\| \leq \rho(A) + \epsilon$.*

*Proof of 6.10.* Let $A \in M_n$ be given and $\epsilon > 0$ be given. Let $A = SJS^{-1}$ be a JCF. WLOG suppose $J$ has the form

$$J = \begin{bmatrix} \lambda_1 & 1 & & 0 \\ & \lambda_2 & 1 & \\ & & \ddots & 1 \\ & & & \lambda_n \end{bmatrix}$$

Define

$$D := \begin{bmatrix} \epsilon & & & 0 \\ & \epsilon^2 & & \\ & & \ddots & \\ & & & \epsilon^n \end{bmatrix}$$

and observe that this turns any 1's on the superdiagonal of $J$ into $\epsilon$'s:

$$D^{-1}JD = \begin{bmatrix} \lambda_1 & \epsilon & & 0 \\ & \lambda_2 & \epsilon & \\ & & \ddots & \epsilon \\ & & & \lambda_n \end{bmatrix}$$

Note that $\|D^{-1}JD\|_{1,1} \leq \rho(A) + \epsilon$, where $\|\cdot\|_{1,1}$ is the maximum column sum of a matrix. Then $\|A\|_{1,1_{SD}} = \|D^{-1}S^{-1}ASD\|_{1,1} = \|D^{-1}JD\| \leq \rho(A) + \epsilon$. $\square$

**Example 6.9.** For the matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, it has a positive matrix norm by the positivity property of any matrix norm, but $\rho(A) = 0$. In this case, equality of $\rho(A)$ and any matrix norm $\|A\|$ cannot be achieved.

**Corollary 6.4.** *For all $A \in M_n$, $\inf_{\|\cdot\| \text{ matrix norm}} \|A\| = \rho(A)$.*

## 6.6 Analytical Properties of Matrix Norms

**Theorem 6.11.** *For $A \in M_n$, $\lim_{k\to\infty} A^k = \mathbf{0}$ if and only if $\rho(A) < 1$. I.e. $\|A^k - \mathbf{0}\| \to 0$.*

- *This describes nilpotency in the limit*
- *Recall, a nilpotent matrix has $\rho(A) = 0$.*

*Proof of Theorem 6.11.* ($\Longleftarrow$) If $\rho(A) < 1$, then there exists a matrix norm $\|\cdot\|$ such that $\|A\| < 1$. Now observe that $\|A^k\| \leq \|A\|^k$ by submultiplicativity. Since $\|A\| < 1$, then as $k \to \infty$, $\|A\|^k \to 0$, so $\|A^k\| \to 0$ as well.

($\Longrightarrow$) Suppose $A^k \to \mathbf{0}$ as $k \to \infty$. Let $x$ be an eigenvector associated with eigenvalue $\lambda$ of maximum modulus. Then $A^k x = \lambda^k x \to \vec{0}$ as $k \to \infty$. Then $|\lambda| < 1$ because $\lambda^k$ must exponentially decay to 0. $\square$

**Theorem 6.12.** *For any matrix norm $\|\cdot\|$ on $M_n$ and any $A \in M_n$, $\lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \rho(A)$.*

*Proof of Theorem 6.12.* First note that $[\rho(A)]^k = \rho(A^k)$ since $A^k$ has the same eigenvalues as $A$ but raised to the $k^{\text{th}}$ power. So, $\rho(A)^k = \rho(A^k) \leq \|A^k\| \Longrightarrow \rho(A) \leq \|A^k\|^{\frac{1}{k}}$. Let $\epsilon > 0$ be given. Then $\rho\left(\frac{1}{\rho(A)+\epsilon}A\right) < 1$. Then by Theorem 6.11, $\left(\frac{1}{\rho(A)+\epsilon}A\right)^k \to \mathbf{0}$. This means that $\exists M$ such that $\forall k \geq M$, $\left\|\left(\frac{1}{\rho(A)+\epsilon}A\right)^k - \mathbf{0}\right\| < 1$. This implies that $\left\|\left(\frac{1}{\rho(A)+\epsilon}A\right)^k\right\| = \frac{1}{(\rho(A)+\epsilon)^k} \|A^k\| < 1$, i.e. $\|A^k\|^{\frac{1}{k}} < \rho(A) + \epsilon$. Since $\epsilon > 0$ was chosen arbitrarily, the result follows. $\square$

**Theorem 6.13.** *A NLS $(V, \|\cdot\|)$ over $\mathcal{K}$ is complete if and only if all absolutely convergent series converge. That is, for all $\{x^{(i)}\}_{i=0}^{\infty} \subseteq V$, $\sum_{i=0}^{\infty} \|x^{(i)}\| < \infty \iff \sum_{i=0}^{\infty} x^{(i)} \in V$. Note, absolute convergent series means refers to taking the absolute value of terms in the series. This means that a complete vector space is equivalent to the condition that having all absolute series, which is a real series, converge imply all series of vectors $x^{(i)} \in V$ are in $V$.*

**Definition 6.17** (Matrix exponential)**.** $\forall A \in M_n$, $e^A := \sum_{i=0}^{\infty} \frac{1}{i!} A^i$ is well-defined.

- Let $\|\cdot\|$ be any matrix norm.

$$\sum_{i=0}^{\infty} \left\| \frac{1}{i!} A^i \right\| \leq \sum_{i=0}^{\infty} \frac{1}{i!} \|A\|^i = e^{\|A\|} \implies \sum_{i=0}^{\infty} \frac{1}{i!} A^i \text{ converges since all NLS are complete.}$$

- The specific case of $i = 0$ where we use the submultiplicative property in the inequality does not always hold, that is, it is not always true that $\|A^0\| = \|I\| \leq \|A\|^0 = 1$ holds for every matrix norm (e.g. $\|\cdot\|_F$). But we just need to argue convergence in a single matrix norm to argue convergence in all matrix norms. So we can argue convergence for any induced matrix norm.

**Theorem 6.14.** *Let $B \in M_n$ and $\|\cdot\|$ be a matrix norm on $M_n$ such that $\|B\| < 1$. Then $I - B$ is invertible and $(I - B)^{-1} = \sum_{i=0}^{\infty} B^i$.*

*Proof of Theorem 6.14.* Since $\|B\| < 1$, then we have the following absolute converging series:

$$\sum_{i=0}^{\infty} \|B^i\| \leq \sum_{i=0}^{\infty} \|B\|^i = \frac{1}{1 - \|B\|} < \infty.$$

This implies that $\sum_{i=0}^{\infty} B^i$ converges, and so

$$(I - B) \sum_{i=0}^{N} B^i = \left(I + B + B^2 + B^3 + \cdots + B^N\right) - \left(B - B^2 - B^3 \cdots - B^N - B^{N+1}\right) = I - B^{N+1}.$$

$\|B\| < 1 \implies \|B^{N+1}\| \leq \|B\|^{N+1} \to 0$ as $N \to \infty$. So $(I - B) \sum_{i=0}^{\infty} B^i = I$, i.e. $\sum_{i=0}^{\infty} B^i = (I - B)^{-1}$. ☐

*Note.* By Theorem 6.14, if $A \in M_n$ and $\|\cdot\|$ is a matrix norm on $M_n$ such that $\|I - A\| < 1$, then $A$ is invertible and $A^{-1} = \sum_{i=0}^{\infty} (I - A)^i$.

*Note.* If $B \in M_n$, $\rho(B) < 1$, then $I - B$ invertible and $(I - B)^{-1} = \sum_{i=0}^{\infty} B^i$ since there exists a matrix norm $\|\cdot\|$ such that $\|B\| < 1$, and so the result directly follows from Theorem 6.14.

## 6.7 Applications of Matrix Norms

**Definition 6.18** (Compatibility)**.** The matrix norm $\|\cdot\|$ on $M_n$ is *compatible* with vector norm $\|\cdot\|$ on $\mathbb{C}^n$ if $\forall A \in M_n, x \in \mathbb{C}^n$, $\|Ax\| \leq \|A\| \|x\|$, i.e. the subordinate property of the matrix norm holds.

- If $\|\cdot\|$ is an induced matrix norm by vector norm $\|\cdot\|$, then they are compatible.

**Definition 6.19** (Condition number)**.** Let $\|\cdot\|$ be a matrix norm on $M_n$. $\forall A \in M_n$ invertible, the *condition number* of $A$ is $\kappa_{\|\cdot\|}(A) := \|A\| \|A^{-1}\|$.

*Note.* $\|I \cdot I\| \leq \|I\| \|I\| \implies \|I\| \geq 1$. If $\|\cdot\|$ is an induced matrix norm, then $\|I\| = 1$. Then, $\kappa(A) \geq 1$ since $\kappa(A) = \|A\| \|A^{-1}\| \geq \|AA^{-1}\| = \|I\| \geq 1$. If $\|\cdot\|$ is induced, then $\kappa(I) = \|I\| \|I^{-1}\| = 1$.

**Theorem 6.15.** *Suppose $A \in M_n$ invertible, $b, x, \Delta b, \Delta x \in \mathbb{C}^n$ with $b, x$ nonzero, and $\|\cdot\|$ is a matrix norm on $M_n$ that is compatible with the vector norm $\|\cdot\|$ on $\mathbb{C}^n$. Further, suppose $Ax = b$ and $A(x + \Delta x) = b + \Delta b$. Then,*

$$\frac{1}{\kappa(A)} \frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|}.$$

*Proof of Theorem 6.15.* Subtracting, we have $A\Delta x = \Delta b$. Thus, by the subordinate property

(1) $\|b\| \le \|A\| \, \|x\| \implies \frac{1}{\|x\|} \le \|A\| \frac{1}{\|b\|}$

(2) $\|x\| \le \left\|A^{-1}\right\| \|b\| \implies \frac{1}{\|A^{-1}\|} \frac{1}{\|b\|} \le \frac{1}{\|x\|}$

(3) $\|\Delta x\| \le \left\|A^{-1}\right\| \|b\|$

(4) $\|\Delta b\| \le \|A\| \, \|\Delta x\| \implies \frac{\|\Delta b\|}{\|A\|} \le \|\Delta x\|$

Multiplying (1) and (3) together and (2) and (4) together, we get $\frac{\|\Delta x\|}{\|x\|} \le \|A\| \left\|A^{-1}\right\| \frac{\|\Delta b\|}{\|b\|} = \kappa(A) \frac{\|\Delta b\|}{\|b\|}$ and $\frac{1}{\|A\|\|A^{-1}\|} \frac{\|\Delta b\|}{\|b\|} = \frac{1}{\kappa(A)} \frac{\|\Delta b\|}{\|b\|} \le \frac{\|\Delta x\|}{\|x\|}$. ☐

**Theorem 6.16.** *Suppose $\|\cdot\|$ is a matrix norm on $M_n$. Let $A, \Delta A \in M_n$ with $A$ being invertible and $\left\|A^{-1}\right\| \|\Delta A\| < 1$. Then $A + \Delta A$ is invertible (preserves invertibility). Define $\Delta(A^{-1}) := A^{-1} - (A + \Delta A)^{-1}$ to be the error in the inverse $A^{-1}$. Then, we can form a bound about stability:*

$$\frac{\left\|\Delta(A^{-1})\right\|}{\|A^{-1}\|} \le \frac{\kappa(A) \frac{\|\Delta A\|}{\|A\|}}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}}$$

- *Note, $\frac{\|\Delta A\|}{\|A\|}$ is the relative error for $A$.*
- *Typically, $\kappa(A) \frac{\|\Delta A\|}{\|A\|}$ is very small, so the denominator is not very significant.*

*Proof of Theorem 6.16.* Observe that

$$\left\|-A^{-1}\Delta A\right\| \le |-1| \left\|A^{-1}\right\| \|A\| < 1.$$

Then by Theorem 6.14, $I - (-A^{-1}\Delta A)$ is invertible and $[I - (-A^{-1}\Delta A)]^{-1} = \sum_{i=0}^{\infty} \left(-A^{-1}\Delta A\right)^i$. So, $A + \Delta A = A(I + A^{-1}\Delta A)$ is invertible and $(A + \Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1} A^{-1} = \sum_{i=0}^{\infty} (-A^{-1}\Delta A)^i A^{-1}$. Then,

$$\Delta(A^{-1}) = A^{-1} - (A + \Delta A)^{-1} = -\sum_{i=1}^{\infty} (-A^{-1}\Delta A)^i A^{-1}$$

$$\left\|\Delta(A^{-1})\right\| = \left\| -\sum_{i=1}^{\infty} (-A^{-1}\Delta A)^i A^{-1} \right\| \le \sum_{i=1}^{\infty} \left\|A^{-1}\Delta A\right\|^i \left\|A^{-1}\right\|$$

$$\frac{\left\|\Delta(A^{-1}\right\|}{\|A^{-1}\|} \le \sum_{i=1}^{\infty} \left\|A^{-1}\Delta A\right\|^i = \frac{\left\|A^{-1}\Delta A\right\|}{1 - \|A^{-1}\Delta A\|} \le \frac{\left\|A^{-1}\right\| \|\Delta A\|}{1 - \|A^{-1}\| \|\Delta A\|} = \frac{\|A\| \left\|A^{-1}\right\| \frac{\|\Delta A\|}{\|A\|}}{1 - \|A\| \left\|A^{-1}\right\| \frac{\|\Delta A\|}{\|A\|}}.$$

☐

## 6.8 Consequences of Absolute & Monotone Norms

**Definition 6.20** (Absolute vector norm)**.** A vector norm $\|\cdot\|$ on $\mathbb{C}^n$ is called *absolute* if $\forall x \in \mathbb{C}^n$, $\||x|\| = \|x\|$, where $|\cdot|$ is the component-wise absolute value, for instance,

$$\left| \begin{bmatrix} 3 \\ -4i \\ 3 + 4i \end{bmatrix} \right| = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}.$$

**Definition 6.21** (Monotone vector norm)**.** A vector norm $\|\cdot\|$ on $\mathbb{C}^n$ is called *monotone* if $\forall x, y \in \mathbb{C}^n$, $|x| \le |y|$ implies $\|x\| \le \|y\|$.

**Theorem 6.17.** *Let $\|\cdot\|$ be a vector norm on $\mathbb{C}^n$. The following are equivalent:*

*(i)* $\|\cdot\|$ *is monotone.*

*(ii)* $\|\cdot\|$ *is absolute.*

*(iii) The matrix norm* $\|\cdot\|'$ *on* $M_n$ *induced by* $\|\cdot\|$ *satisfies the "diagonal property"* $\forall D \in M_n, \|D\|' = \max_i |d_{ii}|$.

*Proof of Theorem 6.17.* $[(i) \implies (ii)]$ Suppose $\|\cdot\|$ is monotone. Then $|x| \leq \Big||x|\Big|$ implies $\|x\| \leq \||x|\|$ and $\Big||x|\Big| \leq |x|$ implies $\||x|\| \leq \|x\|$, so $\|x\| = \||x|\|$.
$[(ii) \implies (i)]$ See text.
$[(i) \implies (iii)]$ Suppose $\|\cdot\|$ is monotone. Let $D \in M_n$ be diagonal and $x \in \mathbb{C}^n$ be nonzero. Then

$$Dx = \begin{bmatrix} d_{11}x_1 \\ d_{22}x_2 \\ \vdots \\ d_{nn}x_n \end{bmatrix},$$

so $|Dx| \leq |(\max_i |d_{ii}|)\, x|$. By monotonicity, $\|Dx\| \leq \|(\max_i |d_{ii}|)\, x\| = \max_i |d_{ii}| \, \|x\|$, i.e. $\frac{\|Dx\|}{\|x\|} \leq \max_i |d_{ii}|$, where equality is achieved for $x = \mathrm{e}_k$ for $k = \arg\max_i |d_{ii}|$. Then $\max_{x \in \mathbb{C}^n \setminus \{\vec{0}\}} \frac{\|Dx\|}{\|x\|} = \max_i |d_{ii}|$.
$[(iii) \implies (i)]$ Suppose $\|\cdot\|$ induces $\|\cdot\|'$ with the diagonal property. Let $x, y \in \mathbb{C}^n$ such that $|x| \leq |y|$. For $i = 1, \dots, n$, define $d_{ii} := \begin{cases} x_i/y_i & \text{if } y_i \neq 0 \\ 0 & \text{if } y_i = 0 \end{cases}$, and let $D := \operatorname{diag}(d_{11}, d_{22}, \dots, d_{nn})$. Note that $Dy = x$ and $\|D\|' = \max_i |d_{ii}| \leq 1$ by domination of $y$ to $x$. Thus, $\|x\| = \|Dy\| \leq \|D\|' \, \|y\| \leq 1 \cdot \|y\|$, where we have the first inequality due to compatibility for any induced matrix norm. Thus, $\|x\| \leq \|y\|$. □

**Example 6.10.** Note that any $l_p$ norm is absolute since we immediately take the absolute value anyway in the operation. They are also clearly monotone since if $x$ is dominated by $y$ component-wise, then clearly $\|x\|_p \leq \|y\|_p$. Further for any diagonal matrix $D \in M_n$, any induced $l_p$ matrix norm will yield the maximum modulus diagonal entry, e.g. $\|\cdot\|_{1,1}$ gives the maximum column sum, $\|\cdot\|_{2,2}$ gives the maximum singular value, and $\|\cdot\|_{\infty,\infty}$ gives the maximum row sum, which all give the maximum modulus diagonal entry of $D$.

**Theorem 6.18** (Bauer-Fike)**.** *Let* $\|\cdot\|$ *be a matrix norm on* $M_n$ *induced by a monotone norm. Let* $A, \Delta A \in M_n$ *be such that* $A$ *is diagonalizable, say* $A = SDS^{-1}$*. Then* $\forall \lambda \in \sigma(A + \Delta A)$*, there exists* $\tau \in \sigma(A)$ *such that* $|\lambda - \tau| \leq \kappa(S) \|\Delta A\|$*. Further, if* $A$ *is normal, then there exists* $\tau \in \sigma(A)$ *such that* $|\lambda - \tau| \leq \|\Delta A\|_{2,2}$*.*

*Proof of Theorem 6.18.* If $A$ is normal, then it is unitarily diagonalizable, so $S$ can be chosen to be unitary. Then $\|S\|_{2,2} = \sqrt{\rho(S^*S)} = 1$ and $\kappa(S) = \|S\|_{2,2} \|S^*\|_{2,2} = 1 \cdot 1 = 1$. Note, $\|\cdot\|_{2,2}$ is induced by $\|\cdot\|_2$, which in fact, a monotone vector norm.

For the general case, say $\lambda \in \sigma(A + \Delta A)$. If $\lambda \in \sigma(A)$, then the result is trivial, so suppose $\lambda \in \sigma(A)$. By definition of the characteristic polynomial, $\lambda I - (A + \Delta A)$ is singular. Pre-multiplying by $S^{-1}$ and post-multiplying by $S$, we have that $S^{-1}(\lambda I - (A + \Delta A))S = \lambda I - D - S^{-1} \Delta A S$ is singular. Now, pre-multiplying by $(\lambda I - D)^{-1}$, we have that $I - (\lambda I - D)^{-1} S^{-1} \Delta A S$ is singular. By the contrapositive of Theorem 6.14, $\left\|(\lambda I - D)^{-1} S^{-1} \Delta A S\right\| \geq 1$, so $\left\|(\lambda I - D)^{-1}\right\| \left\|S^{-1}\right\| \|\Delta A\| \|S\| \geq \left\|(\lambda I - D)^{-1} S^{-1} \Delta A S\right\| \geq 1$. Since $\|\cdot\|$ is induced by a monotone vector norm, then $\|\cdot\|$ has the diagonal property, so $\left\|(\lambda I - D)^{-1}\right\| = \max_i \left|\frac{1}{\lambda - d_{11}}\right| = \frac{1}{|\lambda - \tau|}$ for some $\tau \in \sigma(A)$. Then directly, $|\lambda - \tau| \leq \left\|S^{-1}\right\| \|\Delta A\| \|S\| = \kappa(S) \|\Delta A\|$. □

*Note.* If $A, \Delta A \in M_n$ are Hermitian with ordered eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, then by Weyl's Theorem (Theorem 4.6), for all $k$ $\lambda_1(\Delta A) + \lambda_k(A) \leq \lambda(A + \Delta A) \leq \lambda_n(\Delta A) + \lambda_k(A)$, or equivalently $\lambda_1(\Delta A) \leq \lambda_k(A + \Delta A) - \lambda_k(A) \leq \lambda_n(\Delta A)$. This gives

$$|\lambda_k(A + \Delta A) - \lambda_k(A)| \leq \rho(\Delta A) \leq \|A\|_{2,2},$$

where we have equality of $\rho(\Delta A) = \|A\|_{2,2}$ since $\Delta A$ is Hermitian. This is exactly gives the Bauer-Fike relationship: $|\lambda - \tau| \leq \|A\|_{2,2}$, but now we know $k^{\text{th}}$ eigenvalue of $A + \Delta A$ is near the $k^{\text{th}}$ eigenvalue of $A$.

# 7 Chapter 6 – Gerschgorin Theorem and Eigenvalue Perturbations

## 7.1 Gerschgorin Discs

**Definition 7.1** (Gerschgorin Disc and Gerschgorin Region)**.** Consider $A \in M_n$. For $i = 1, \ldots, n$, the $i^{\text{th}}$ *Gerschgorin disc* is

$$G_i(A) := \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\},$$

that is, the $i^{\text{th}}$ Gerschgorin disc (G-disc) is the disc in $\mathbb{C}$ centered at $a_{ii}$ with radius equal to the row sum of the remaining entries. If the radius of a G-disc is 0, then the disc is called a *degenerate disc*. The *Gerschgorin region* for $A$ is the union of its G-discs

$$G(A) := \bigcup_{i=1}^{n} G_i(A).$$

**Example 7.1.** Consider $A = \begin{bmatrix} 2+i & i & -1 \\ 0.01 & 4 & 0 \\ 0.5 & 0 & 3 \end{bmatrix}$. The first G-disc of $A$ is centered at $2+i$ with radius 2. The second G-disc is centered at $4+0i$ with radius 0.01. The third G-disc is centered at $3+0i$ with radius 0.5.

**Theorem 7.1** (Gerschgorin)**.** *For all $A \in M_n$, $\sigma(A) \subseteq G(A)$. Furthermore, if a connected component of $G(A)$ consists of, say, $k$ G-discs, then it contains exactly $k$ eigenvalues.*

*Proof of Theorem 7.1.* Let $\lambda \in \sigma(A)$ with eigenvector $x$. Let $k \in \arg\max_i |x_i|$ i.e. $|x_k| = \|x\|_\infty > 0$. Since $Ax = \lambda x$, then looking at the $k^{\text{th}}$ component of both sides gives $\lambda x_k = \sum_j a_{kj} x_j$, then $(\lambda - a_{kk})x_k = \sum_{j \neq k} a_{kj} x_j$. By the triangle inequality,

$$|(\lambda - a_{kk})x_k| = |\lambda - a_{kk}||x_k| = \left|\sum_{j \neq k} a_{kj} x_j\right| \leq \sum_{j \neq k} |a_{kj}||x_j| \leq \left(\sum_{j \neq k} |a_{kj}|\right) |x_k|.$$

Since $x \neq \vec{0}$, then $|x_k| \neq 0$, and so $|\lambda - a_{kk}| \leq \sum_{j \neq k} |a_{kj}|$. So $\lambda \in G_k(A)$.

Let $A \equiv D + N$ where $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$. Define for $\epsilon > 0$, $A_\epsilon := D + \epsilon N$. Note, $\sigma(A_\epsilon)$ is continuous in $\epsilon$, that is, the roots of the characteristic polynomial are continuous in its coefficients, which are continuous in the entries of the matrix. As $\epsilon$ goes from 0 to 1, the discs proportionally expand and the eigenvalues cannot jump to disconnected regions by continuity. □

*Note.* For all $A \in M_n$, $\sigma(A) = \sigma(A^T) \subseteq G(A^T)$, i.e. we have column G-discs. Further for any invertible $S \in M_n$, $\sigma(A) = \sigma(S^{-1}AS) \subseteq G(S^{-1}AS)$.

**Example 7.2.** Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Then $\sigma(A)$ lies in the intersection of the disc centered at 0 with radius 1 (from from $G(A^T)$) and the disc centered at 1 with radius 1 (from $G(A^T)$).

**Example 7.3.** Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Now consider pre-multiplying $A$ by $\begin{bmatrix} \epsilon^1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}^{-1}$ and post-multiplying $A$ by $\begin{bmatrix} \epsilon^1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}$, which gives $\begin{bmatrix} 0 & \epsilon \\ 0 & 1 \end{bmatrix}$. The first G-disc for this matrix is a disc centered at 0 with radius $\epsilon$ and the other G-disc is a disc centered at 1 with radius 0, which gives a tight region for eigenvalues.

**Example 7.4.** Note that $\sigma(A) = \bigcap_{S \in M_n \text{ invertible}} G(S^{-1}AS)$. Clearly, $\sigma(A) \subseteq \bigcap_{S \in M_n \text{ invertible}} G(S^{-1}AS)$. Consider the JCF of $A$ and then pre-multiplying by the inverse of $D = \operatorname{diag}(\epsilon^1, \ldots, \epsilon^n)$ and post-multiplying by $D = \operatorname{diag}(\epsilon^1, \ldots, \epsilon^n)$. Then we can control the radius of the G-discs for any $\epsilon$.

**Definition 7.2** (Diagonally dominant)**.** We say that $A \in M_n$ is *diagonally dominant* if for all $i$, $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$. $A$ is *strictly diagonally dominant* if the inequality is strict.

**Theorem 7.2.** *If $A \in M_n$ is strictly diagonally dominant, then $A$ is invertible.*

*Proof of Theorem 7.2.* We prove the contrapositive. If $A$ is singular, then $0 \in \sigma(A)$, so there exists $i$ such that $|0 - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|$, so $A$ is not strictly diagonally dominant. $\square$

## 7.2 Gerschgorin Theorem – A Closer Look

**Theorem 7.3.** *Suppose $A \in M_n$ is diagonally dominant, and strictly so in all rows except, say, the $k^{th}$ row, and suppose $a_{kk} \neq 0$. Then $A$ is invertible.*

*Proof of Theorem 7.3.* FSOC, suppose the above condition holds, but $A$ is singular. Let $D = \operatorname{diag}(1, \dots, 1, 1+\epsilon, 1, \dots, 1)$, where the $1+\epsilon$ entry is the $k^{\text{th}}$ diagonal entry of $D$ and $\epsilon$ is sufficiently small. Specifically, choose $\epsilon$ such that $0 < \epsilon < \min_i \frac{|a_{ii}| - \sum_{j \neq i} |a_{ij}|}{|a_{ik}|}$. Note that $D^{-1}AD$ is the matrix $A$ but with the $k^{\text{th}}$ row multiplied by $(1+\epsilon)^{-1}$ and the $k^{\text{th}}$ column multiplied by $(1+\epsilon)$. For every $i \neq k$, $G_i(D^{-1}AD) = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \left( \sum_{j \neq i} |a_{ij}| + \epsilon |a_{ik}| \right) \right\}$. If $0 \in G_i(D^{-1}AD)$, then this contradicts $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ since $\epsilon$ was chosen sufficiently small. Also, $G_k(D^{-1}AD) = \{ z \in \mathbb{C} : |z - a_{kk}| \leq \frac{1}{1+\epsilon} \sum_{j \neq k} |a_{kj}| \}$. If $0 \in G_i(D^{-1}AD)$, then $|a_{kk}| < (1+\epsilon)|a_{kk}| \leq \sum_{j \neq k} |a_{kj}|$, contradicting the diagonal dominance in row $k$. Thus, $A$ is invertible. $\square$

**Definition 7.3** (Entry digraph)**.** Let $A \in M_n$. We associate with $A$ the *entry digraph* of $A$, denoted as $\Gamma(A)$. Note digraph is an abbreviation for directed graph. The vertex set is $\{1, 2, \dots, n\}$ and for all $i, j \in \{1, 2, \dots, n\}$, the edge $(i, j) \in \Gamma(A)$ if and only if $a_{ij} \neq 0$.

**Definition 7.4** (Strongly connected)**.** $\Gamma(A)$ is *strongly connected* if for all $i, j \in \{1, 2, \dots, n\}$, there exists an $(i, j)$-directed walk. A directed walk exists if there is a sequence of directed edges that starts from $i$ and ends at $j$. As an exception, we only have a single node 1, then we only consider it to be strongly connected if $(1, 1) \in \Gamma(A)$, else it is not strongly connected.

**Definition 7.5** (Reducible)**.** A matrix $A \in M_n$ is *reducible* if $n = 1$ and $A = \mathbf{0}$, or $n \geq 2$ and there exists a permutation matrix $P \in M_n$ and $r$ such that $1 \leq r \leq n-1$ and

$$PAP^T = \left[ \begin{array}{c|c} * & * \\ \hline \mathbf{0}_{(n-r) \times r} & * \end{array} \right],$$

i.e. a rectangular block of zeros exists in the lower left partition of $PAP^T$. Else, we say $A$ is *irreducible.*

**Theorem 7.4.** *Let $A \in M_n$. The following are equivalent:*

*(i) $A$ is irreducible.*

*(ii) $\Gamma(A)$ is strongly connected.*

*(iii) $(I + |A|)^{-1} > \mathbf{0}$ component-wise.*

*Proof of Theorem 7.4.* Note that if $A$ induces $\Gamma(A)$, then $\Gamma(PAP^T)$ is simply a relabeling of the vertices. If $A$ is reducible, then $A$ is re-indexed as $PAP^T = \left[ \begin{array}{c|c} * & * \\ \hline \mathbf{0}_{(n-r) \times r} & * \end{array} \right]$. Note, from any vertex in the "$r+1$-to-$n$-relabeled", there is no directed walk to any of the "1-to-$r$-relabeled" vertices, so $\Gamma(A)$ is not strongly connected.

Conversely, suppose $\Gamma$ is not strongly connected, i.e. $\exists (i, j)$ such that there is no $(i, j)$-directed walk. Define $\Lambda := \{\text{vertices reachable from } i \text{ by a directed walk}\}$. Note $i \in \Lambda$ implies $j \notin \Lambda$. Then, re-index $A$ via permutation $P$ so that the first $r$ rows and columns are the edges going from vertices in $\Lambda^c$ to vertices in $\Lambda^c$ and the last $n-r$ rows and first $r$ columns correspond to edges going from vertices in $\Lambda$ to vertices in $\Lambda^c$. This lower left partition of $PAP^T$ is all zeros, else there would be a walk from a vertex reachable from $i$ to a vertex not reachable from $i$. which would be a contradiction of membership in $\Lambda^c$. Thus, $A$ reducible.

Define $Z \in M_n$ such that for all $i, j$, $Z_{ij} = \begin{cases} 1 & \text{if } a_{ij} \neq 0 \\ 0 & \text{if } a_{ij} = 0 \end{cases}$. We claim that for any $i, j, k$, $Z^k_{ij}$ is the number of direct walks from $i$ to $j$ of length $k$ in $\Gamma(A)$. This is seen by induction, where $Z^0 = I$, the walks from anywhere to itself has only 1 walk using 0 edges. Otherwise, there are 0 possible walks using 0 edges. Suppose the claim holds for $Z^k$. Consider $Z_{ij}^{k+1} = \sum_\ell Z_{i\ell}^k Z_{\ell j}^1$. Wherever $Z_{\ell j}^1$ is 1, then there is a walk that allows you to take any other path one edge longer, so indeed $Z_{ij}^k$ counts the number of walks from $i$ to $j$ in $\Gamma(A)$. Now consider $|A|^k$. Note that $(|A|^k)_{ij} \neq 0$ if and only if $Z_{ij}^k \neq 0$. Then we have strong connectivity if and only if for each $i$ and $j$, there exists a $k \in \{0, \dots, n-1\}$ such that $Z_{ij}^k \neq 0$. Then performing a binomial expansion of $(I + |A|)^{n-1}$ we have that $(I + |A|)^{n-1} = \sum_{k=0}^{n-1} (> 0)|A|^k > \mathbf{0}$ if and only if for every pair $i, j$ there exists a $k \in \{0, \dots, n-1\}$ such that $Z_{ij}^k \neq 0$. $\square$

## 7.3 Guarantees of Irreducibility

**Definition 7.6** (Interior of G-discs)**.** For $A \in M_n$, we say that $\lambda \in \sigma(A)$ is *interior* of some G-disc if $\exists i$ such that $|\lambda - a_{ii}| < \sum_{j \neq i} |a_{ij}|$. We say that $\lambda \in \sigma(A)$ is *not interior* of any G-disc if $\forall i$, $|\lambda - a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$.

**Theorem 7.5.** *Let $A \in M_n$ and $\lambda \in \sigma(A)$ be not interior of any G-disc. Let $x$ be an associated eigenvector. Set $\mathcal{I} := \arg\max_i |x_i|$. Then, the following are true:*

*(i) $\forall i \in \mathcal{I}$, $\lambda \in \partial G_i(A)$, where $\partial G_i(A)$ is the boundary of $G_i(A)$. That is, $|\lambda - a_{ii}| = \sum_{j \neq i} |a_{ij}|$*

*(ii) $\forall i \in \mathcal{I}$, $(i, j) \in \Gamma(A)$ implies $j \in \mathcal{I}$.*

*Proof of Theorem 7.5.* For all $i \in \mathcal{I}$, i.e. $|x_i| = \|x\|_\infty$. Then,

$$\begin{aligned}
&& Ax &= \lambda x \\
\implies && \lambda x_i &= \sum_j a_{ij} x_j \\
\implies && (\lambda - a_{ii}) x_i &= \sum_{j \neq i} a_{ij} x_j \\
\implies && |\lambda - a_{ii}||x_i| &\leq \sum_{j \neq i} |a_{ij}||x_j| \\
&& &\leq \left( \sum_{j \neq i} |a_{ij}| \right) |x_i| \\
\implies && |\lambda - a_{ii}| &\leq \sum_{j \neq i} |a_{ij}|.
\end{aligned}$$

But since $\lambda$ is not interior of any G-disc, then for all $i$, we have $|\lambda - a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$. Combining these two inequalities we have $|\lambda - a_{ii}| = \sum_{j \neq i} |a_{ij}|$, i.e. $\lambda \in \partial G_i(A)$ and we have exact equality in the above derivation. Thus, $\sum_{j \neq i} |a_{ij}|(\|x\|_\infty - |x_j|) = 0$. Since $|a_{ij}| \geq 0$ and $(\|x\|_\infty - |x_j|) \geq 0$, then for any $j$ such that $a_{ij} \neq 0$, then $(\|x\|_\infty - |x_j|) = 0$, so $j \in \mathcal{I}$. $\square$

**Corollary 7.1.** *Let $A \in M_n$ be irreducible. If $\lambda \in \sigma(A)$ is not interior of any G-disc, then $\lambda$ is on the boundary of all G-discs.*

*Proof of Corollary 7.1.* Recall by Theorem 7.4, $A$ being irreducible implies that $\Gamma(A)$ is not strongly connected. Let $x$ be an eigenvector associated with $\lambda$. Note that $\mathcal{I} := \arg\max_i |x_i|$ is not empty. Say $i' \in \mathcal{I}$. For all $i = 1, \dots, n$, there exists a directional path from $i'$ to $i$ by strong connectivity. Hence, by Theorem 7.5 and $\lambda$ not being interior of any G-disc, we have $i \in \mathcal{I}$ (the immediate neighbors of $i'$ are clearly in $\mathcal{I}$, but this recursively implies the neighbors of those neighbors are also in $\mathcal{I}$, thus any $i$ reachable from $i'$ is also in $\mathcal{I}$). Thus, for all $i = 1, \dots, n$, $\lambda \in \partial G_i(A)$. This means all components of $x_i$ are of maximum modulus. $\square$

**Corollary 7.2** (Tausky)**.** *Let $A \in M_n$ be irreducible, diagonally dominant, and strict diagonal dominance in one row, that is, $\exists k$ such that $|a_{kk}| > \sum_{j \neq k} |a_{kj}|$. Then $A$ is invertible.*

*Proof of Corollary 7.2.* Suppose $A \in M_n$ is irreducible, diagonally dominant, and $\exists k$ such that $|a_{kk}| > \sum_{j \neq k} |a_{kj}|$. FSOC, suppose $A$ is singular i.e. $0 \in \sigma(A)$. By diagonal dominance, then

$$\forall i = 1, 2, \dots, n, \qquad |0 - a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$$

implies 0 is not interior of any G-disc. Thus, since $A$ is irreducible then 0 is on the boundary of all G-discs by Corollary 7.1. So, in particular, $|0 - a_{kk}| = \sum_{j \neq k} |a_{kj}|$, which is a contradiction to the fact that we have strict diagonal dominance in row $k$ of $A$. Thus, $A$ is invertible. □

## 7.4 Norms and Gerschgorin Discs

**Corollary 7.3.** *Let $A \in M_n$ be irreducible and suppose $\exists k$ such that $\sum_j |a_{kj}| < \|A\|_{\infty,\infty}$. Then $\rho(A) < \|A\|_{\infty,\infty}$. That is, if there exists a row such that its row sum is smaller than the maximum row sum of $A$, then $\rho(A)$ is smaller than the maximum row sum of $A$.*

*Proof of Corollary 7.3.* FSOC, suppose $\exists \lambda \in \sigma(A)$ such that $|\lambda| = \|A\|_{\infty,\infty}$. For all $i = 1, \dots, n$,

$$|\lambda - a_{ii}| \geq |\lambda| - |a_{ii}| = \|A\|_{\infty,\infty} - |a_{ii}| \geq \sum_j |a_{ij}| - |a_{ii}| = \sum_{j \neq i} |a_{ij}|.$$

This implies that $\lambda$ is not interior of any G-disc. By Corollary 7.1, since $A$ is irreducible, we have $\lambda$ on the boundary of all G-discs. Thus, we have equality in the above inequalities, and so

$$\|A\|_{\infty,\infty} = \sum_j |a_{kj}|.$$

This contradicts the assumption that there is a $k$ such that $\sum_j |a_{kj}| < \|A\|_{\infty,\infty}$. □

**Theorem 7.6.** *Let $A, \Delta A \in M_n$ be such that $A = SDS^{-1}$ is a diagonalization. Then $\forall \lambda \in \sigma(A + \Delta A)$, there exists $\tau \in \sigma(A)$ such that $|\lambda - \tau| \leq \kappa_{\|\cdot\|_{\infty,\infty}}(S) \|\Delta A\|_{\infty,\infty}$.*

*Proof of Theorem 7.6.* This proof will be done using the Gerschgorin framework, but has been proved earlier with matrix norm properties.

Let $\lambda \in \sigma(A + \Delta A) = \sigma(S^{-1}(A + \Delta A)S) = \sigma(D + S^{-1}\Delta A S)$. By Gerschgorin's Theorem (Theorem 7.1), there exists an $i$ such that

$$|\lambda - d_{ii}| - \left|[S^{-1}\Delta A S]_{ii}\right| \leq \left|\lambda - (d_{ii} + [S^{-1}\Delta A S]_{ii})\right| \leq \sum_{j \neq i} \left|[S^{-1}\Delta A S]_{ij}\right|.$$

The first inequality is by the reverse triangle inequality and the second inequality is by the Gerschgorin Theorem applied to $S^{-1}\Delta A S$. Then,

$$|\lambda - d_{ii}| \leq \sum_{j=1}^{n} [S^{-1}\Delta A S]_{ij} \leq \left\|S^{-1}\Delta A S\right\|_{\infty,\infty} \leq \left\|S^{-1}\right\|_{\infty,\infty} \|\Delta A\|_{\infty,\infty} \|S\|_{\infty,\infty} = \kappa_{\|\cdot\|_{\infty,\infty}}(S) \|\Delta A\|_{\infty,\infty}.$$

□

# 8 Chapter 8 – Positive and Nonnegative Matrices

## 8.1 Nonnegative Matrices

We begin with notation. Let $A \in M_n$. Then,

- $A \geq \mathbf{0}$ means that $\forall i, j,\ A_{ij} \geq 0$ entrywise
- $A > \mathbf{0}$ means that $\forall i, j,\ A_{ij} > 0$ entrywise
- $A \leq \mathbf{0}$ means that $\forall i, j,\ A_{ij} \leq 0$ entrywise
- $A < \mathbf{0}$ means that $\forall i, j,\ A_{ij} < 0$ entrywise

Observe that

- By the triangle inequality, $|AB| \leq |A||B|$, which is an entrywise inequality
- Inductively, $|A^k| \leq |A|^k$
- $\mathbf{0} \leq A \leq B$ implies that $\mathbf{0} \leq A^k \leq B^k$.
- For $x \in \mathbb{C}^n$, $|x|$ is the coordinate-wise absolute value of $x$
- $|Ax| \leq |A||x|$ By the triangle inequality applied componentwise.

**Proposition 8.1.** *Let $A \in M_n$ such that $A \geq \mathbf{0}$ and $\forall i$, $\sum_j a_{ij} = \alpha$. Then, $\rho(A) = \|A\|_{\infty,\infty} = \alpha$.*

*Proof of Proposition 8.1.* $A\mathrm{e} = \alpha \mathrm{e}$, so $\alpha \in \sigma(A)$. Thus, we have $\alpha \leq \rho(A) \leq \|A\|_{\infty,\infty} = \alpha$. $\square$

**Theorem 8.1.** *Let $A, B \in M_n$ such that $|A| \leq B$. Then, $\rho(A) \leq \rho(|A|) \leq \rho(B)$.*

*Proof of Theorem 8.1.* For all $k$, $|A^k| \leq |A|^k \leq B^k$. Thus,

$$\begin{aligned} \left\| |A^k| \right\|_F &\leq \left\| |A|^k \right\|_F \leq \left\| B^k \right\|_F \\ \implies \left\| A^k \right\|_F^{\frac{1}{k}} &\leq \left\| |A|^k \right\|_F^{\frac{1}{k}} \leq \left\| B^k \right\|_F^{\frac{1}{k}} \end{aligned}$$

The second line holds because of the Frobenius norm being a monotone and absolute norm. Then, as $k \to \infty$, we have convergence to the spectral radii:

$$\rho(A) \leq \rho(|A|) \leq \rho(B).$$

$\square$

**Corollary 8.1.** *Let $A \in M_n$ such that $A \geq \mathbf{0}$. Then $\min_i \sum_j a_{ij} \leq \rho(A)$. Contrast this with $\rho(A) \leq \|A\|_{\infty,\infty} = \max_i \sum_j a_{ij}$.*

*Proof of Corollary 8.1.* If $\min_i \sum_j a_{ij} = 0$ then this is trivial. Else, obtain $\tilde{A}$ from $A$ in the following way: for each row $k = 1, 2, \ldots, n$, multiply row $k$ of $A$ by $\frac{\min_i \sum_j a_{ij}}{\sum_j a_{kj}} \leq 1$ to obtain row $k$ of $\tilde{A}$. So, $0 \leq \tilde{A} \leq A$ and each row sum of $\tilde{A}$ is $\min_i \sum_j a_{ij}$. So, $\min_i \sum_j a_{ij} = \rho(\tilde{A}) \leq \rho(A)$ by Proposition 8.1 and Theorem 8.1. $\square$

*Note.* By Corollary 8.1, $A > \mathbf{0}$ implies that $\rho(A) > 0$. $A \geq \mathbf{0}$ being irreducible implies $\rho(A) > \mathbf{0}$.

**Theorem 8.2.** *Let $A \in M_n$, $x \in \mathbb{C}^n$, $A \geq \mathbf{0}$, and $x > \vec{0}$. Then the following hold:*

- *If $\exists \alpha \geq 0$ such that $Ax > \alpha x$, then $\rho(A) > \alpha$.*
- *If $\exists \alpha \geq 0$ such that $Ax \geq \alpha x$, then $\rho(A) \geq \alpha$.*
- *If $\exists \alpha \geq 0$ such that $Ax < \alpha x$, then $\rho(A) < \alpha$.*
- *If $\exists \alpha \geq 0$ such that $Ax \leq \alpha x$, then $\rho(A) \leq \alpha$.*

*Proof of Theorem 8.2.* Let $X \in M_n$ be defined as $X = \operatorname{diag}(x_1, \ldots, x_n)$, note that $X\mathrm{e} = x$. Thus,

$$\begin{aligned} & Ax \geq \alpha x \\ \implies \quad & AX\mathrm{e} \geq \alpha X\mathrm{e} \\ \implies \quad & X^{-1}AX\mathrm{e} \geq \alpha X^{-1}X\mathrm{e} \\ & = \alpha\mathrm{e}, \end{aligned}$$

i.e. the row sums of $X^{-1}AX$ are all at least $\alpha$. By Corollary 8.1, $\rho(A) = \rho(X^{-1}AX) \geq \min_i \sum_j a_{ij} = \alpha$.

Similarly, if $Ax \leq \alpha x$, then

$$\begin{aligned} & Ax \leq \alpha x \\ \implies \quad & AX\mathrm{e} \leq \alpha X\mathrm{e} \\ \implies \quad & X^{-1}AX\mathrm{e} \leq \alpha X^{-1}X\mathrm{e} \\ & = \alpha\mathrm{e} \end{aligned}$$

i.e. all row sums of $X^{-1}AX$ are at most $\alpha$. Then, $\rho(A) = \rho(X^{-1}AX) \leq \left\|X^{-1}AX\right\|_{\infty,\infty} \leq \alpha$.

Similarly, if $Ax > \alpha x$, then $\exists\epsilon > 0$ such that $Ax \geq (\alpha + \epsilon)x$, which implies $\rho(A) \geq \alpha + \epsilon > \alpha$. If $Ax < \alpha x$, then $\exists\epsilon > 0$ such that $Ax \leq (\alpha - \epsilon)x$, which implies $\rho(A) \leq \alpha - \epsilon < \alpha$. ☐

**Corollary 8.2.** *Suppose $A \in M_n$ such that $A \geq \mathbf{0}$. If $A$ has a positive eigenvector, then its associated eigenvalue is $\rho(A)$.*

*Proof of Corollary 8.2.* If $x > \vec{0}$ such that $Ax = \lambda x$, where $\lambda \in \sigma(A)$, then $\lambda \in \mathbb{R}_{\geq 0}$, because the LHS has all non-negative real values. So,

- $Ax \geq \lambda x$ implies $\rho(A) \geq \lambda$
- $Ax \leq \lambda x$ implies $\rho(A) \leq \lambda$.

Together, these imply that $\rho(A) = \lambda$. ☐

## 8.2 Positive Matrices

**Lemma 8.1.** *Let $A \in M_n$, $A > \mathbf{0}$. If $x$ is an eigenvector with associated eigenvalue $\lambda$ such that $|\lambda| = \rho(A)$, then $A|x| = |\lambda||x|$ and $|x| > \vec{0}$. That is, the eigenvector associated with the eigenvalue of maximum modulus has an eigenvector which is strictly nonzero.*

*Proof of Lemma 8.1.* Suppose $A > \mathbf{0}$. Then $\rho(A) > 0$. Next, $A > \mathbf{0}$ and $|x| \neq \vec{0}$ implies $A|x| > \vec{0}$. Note, $|x| \neq \vec{0}$ means $|x|$ is nonnegative, but also nonzero. Now,

$$A|x| = |A||x| \geq |Ax| = |\lambda x| = |\lambda||x|,$$

and so, $A|x| - \lambda|x| \geq \vec{0}$.

If $A|x| - |\lambda||x| \neq \vec{0}$, then $A\left(A|x| - \lambda|x|\right) > \vec{0}$. Thus, $AA|x| > |\lambda|(A|x|)$ and by Theorem 8.2, $\rho(A) > |\lambda|$, which is a contradiction. Thus, $A|x| = |\lambda||x|$, and so $\frac{1}{|\lambda|}A|x| = |x|$. ☐

**Theorem 8.3** (Perron)**.** *If $A \in M_n$ and $A > \mathbf{0}$, then the following are equivalent:*

*(i)* $\rho(A) > 0$.

*(ii)* $\rho(A)$ *is an eigenvalue of* $A$.

*(iii)* $\exists x \in \mathbb{C}^n$, $x > \vec{0}$ *such that* $Ax = \rho(A)x$.

*(iv) The algebraic multiplicity of $\rho(A)$ is 1 (and so the geometric multiplicity of $\rho(A)$ is also 1).*

*(v)* $\forall\lambda \in \sigma(A)$ *such that* $\lambda \neq \rho(A)$, *then* $|\lambda| < \rho(A)$.

**Definition 8.1.** If $A > \mathbf{0}$, then $\exists! x > \vec{0}$ such that $Ax = \rho(A)x$ where $\|x\|_1 = 1$. We call $x$ a *Perron vector.*

*Note.* In a Markov model, the Perron vector for a transition matrix $M$ is the limiting distribution of the states. Further, if $M$ is doubly stochastic, then the Perron vector is $\vec{p} = \frac{1}{n}\mathrm{e}$.

## 8.3 Consequences of Perron's Theorem

**Lemma 8.2.** *Suppose $A \in M_n$, $A > \mathbf{0}$ and $\lambda$ is an eigenvalue of maximum modulus with associated eigenvector $x$. Then there exists an angle $\theta$ such that $e^{i\theta}x = |x| > \vec{0}$, i.e. rotate each component of the vector $x$ so that it lies on the real axes.*

*Proof of Lemma 8.2.* By Lemma 8.1, $|A||x| = |\lambda||x|$ and $|x| > \vec{0}$. We also know that $|\lambda||x| = |\lambda x| = |Ax|$. So for every $i$, $(|A||x|)_i = (|Ax|)_i$, i.e. $\sum_j |a_{ij}||x_j| = |\sum_j a_{ij}x_j|$, which is an equality in the triangle inequality. Thus, the angle for each of the $x_j$ are all equal. □

*Proof of Theorem 8.3 (v).* If $Ax = \lambda x$, $|\lambda| = \rho(A)$, $x \neq \vec{0}$, then by Lemma 8.2, then $\exists\theta$ such that $e^{i\theta}x = |x| > \vec{0}$. By Lemma 8.1, $A|x| = |\lambda||x|$, so

$$\lambda e^{i\theta}x = A(e^{i\theta}x) = A|x| = |\lambda||x| = |\lambda|e^{i\theta}x,$$

and so $\lambda = |\lambda| = \rho(A)$. □

**Theorem 8.4** (Perron-Frobenius)**.** *Let $A \in M_n$, $A \geq \mathbf{0}$, and $A$ irreducible. Then the following are equivalent:*

*(i)* $\rho(A) > 0$.

*(ii)* $\rho(A) \in \sigma(A)$.

*(iii)* $\exists$ *positive eigenvector associated with eigenvalue* $\rho(A)$.

*(iv)* *The algebraic multiplicity of* $\rho(A)$ *is 1.*