Math 504: Advanced Linear Algebra

Hugo Woerdeman and Darij Grinberg*

January 4, 2022 (unfinished!)

Contents

*Drexel University, Korman Center, 15 S 33rd Street, Philadelphia PA, 19104, USA

Preface

These are lecture notes originally written by Hugo Woerdeman and edited by myself for the Math 504 (Advanced Linear Algebra) class at Drexel University in Fall 2021. The website of this class can be found at

`http://www.cip.ifi.lmu.de/~grinberg/t/21fala` .

This document is a work in progress.
Please report any errors you find to `darijgrinberg@gmail.com` **.**

What is this?

This is a second course on linear algebra, meant for (mostly graduate) students that are already familiar with matrices, determinants and vector spaces. Much of the prerequisites (but also some of our material, and even some content that goes beyond our course) is covered by textbooks like [Heffer20], [LaNaSc16], [Taylor20], [Treil15], [Strick20], [GalQua20, Part I], [Loehr14], [Woerde16][1]. The text we will follow the closest is [HorJoh13].

We will freely use the basic theory of complex numbers, including the Fundamental Theorem of Algebra. See [LaNaSc16, Chapters 2–3] or [Korner20, Chapters 9–10] for an introduction to these matters.

Notations

- We let $\mathbb{N} := \{0, 1, 2, \ldots\}$.

[1] This list is nowhere near complete. (It is biased towards freely available sources, but even in that category it is probably far from comprehensive.)

- For any $n \in \mathbb{N}$, we let $[n]$ denote the n-element set $\{1, 2, \ldots, n\}$.
- If $\mathbb{F}$ is a field, and $n, m \in \mathbb{N}$, then $\mathbb{F}^{n \times m}$ denotes the set (actually, an $\mathbb{F}$-vector space) of all $n \times m$-matrices over $\mathbb{F}$.
- If $\mathbb{F}$ is a field, and $n \in \mathbb{N}$, then the space $\mathbb{F}^{n \times 1}$ of all $n \times 1$-matrices over $\mathbb{F}$ (that is, column vectors of size n) is also denoted by $\mathbb{F}^n$.
- The $n \times n$ identity matrix is denoted by I_n or by I if the n is clear from the context.
- The transpose of a matrix A is denoted by A^T.
- Zero vectors and zero matrices will be denoted by 0, no matter what their sizes or ambient spaces are.
- If A is an $n \times m$-matrix, and if $i \in [n]$ and $j \in [m]$, then:
  - we let $A_{i,j}$ denote the (i,j)-th entry of A (that is, the entry of A in the i-th row and the j-th column);
  - we let $A_{i,\bullet}$ denote the i-th row of A;
  - we let $A_{\bullet,j}$ denote the j-th column of A.
- The letter i usually denotes the complex number $\sqrt{-1}$. Sometimes (e.g. in the bullet point just above) it also stands for something else (usually an **i**ndex that is an **i**nteger). I'll do my best to avoid the latter meaning when there is any realistic chance that it be confused for the former.
- We use the notation $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ for the diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$.

0.1. Remark on exercises

Each exercise gives a number of "experience points", which roughly corresponds to its difficulty (with some adjustment for its relevance). This is the number in the square (like $\boxed{3}$ or $\boxed{5}$). The harder or more important the exercise, the larger is the number in the square. A $\boxed{1}$ is a warm-up question whose solution you will probably see right after reading; a $\boxed{3}$ typically requires some thinking or work; a $\boxed{5}$ requires both; higher values tend to involve some creativity or research.

0.2. Scribes

Parts of these notes were scribed by Math 504 students. I thank the following students for their help:

scribe	sections
Hunter Wages	proof of Theorem 2.8.2

1. Unitary matrices ([HorJoh13, §2.1])

In this chapter, n will usually denote a nonnegative integer.

Lecture 1 starts here.

1.1. Inner products

We recall a basic definition regarding complex numbers:

Definition 1.1.1. Let $z \in \mathbb{C}$ be a complex number. Then, the *complex conjugate* of z means the complex number $a - bi$, where z is written in the form $z = a + bi$ for some $a, b \in \mathbb{R}$. In other words, the complex conjugate of z is obtained from z by keeping the real part unchanged but flipping the sign of the imaginary part.

The complex conjugate of z is denoted by $\overline{z}$.

Complex conjugation is known to preserve all arithmetic operations: i.e., for any complex numbers z and w, we have

$$\overline{z+w} = \overline{z} + \overline{w} \qquad \text{and} \qquad \overline{z-w} = \overline{z} - \overline{w} \qquad \text{and}$$
$$\overline{z \cdot w} = \overline{z} \cdot \overline{w} \qquad \text{and} \qquad \overline{z/w} = \overline{z}/\overline{w}.$$

Also, a complex number z satisfies $\overline{z} = z$ if and only if $z \in \mathbb{R}$. Finally, if z is any complex number, then $z\overline{z} = |z|^2$ is a nonnegative real.

Definition 1.1.2. For any two vectors $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{C}^n$ and $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{C}^n$,

we define the scalar

$$\langle x, y \rangle := x_1\overline{y_1} + x_2\overline{y_2} + \cdots + x_n\overline{y_n} \in \mathbb{C} \tag{1}$$

(where $\overline{z}$ denotes the complex conjugate of a $z \in \mathbb{C}$). This scalar $\langle x, y \rangle$ is called the *inner product* (or *dot product*) of x and y.

Example 1.1.3. If $x = \begin{pmatrix} 1+i \\ 2+3i \end{pmatrix} \in \mathbb{C}^2$ and $y = \begin{pmatrix} -i \\ 4+i \end{pmatrix} \in \mathbb{C}^2$, then

$$\begin{aligned} \langle x, y \rangle &= (1+i)\left(\overline{-i}\right) + (2+3i)\left(\overline{4+i}\right) \\ &= (1+i)\, i + (2+3i)\,(4-i) \\ &= i - 1 + 8 - 2i + 12i + 3 = 10 + 11i. \end{aligned}$$

Some warnings about the literature are in order:

- Some authors (e.g., Treil in [Treil15]) write (x, y) instead of $\langle x, y \rangle$ for the inner product of x and y. This can be rather confusing, since (x, y) also means the pair consisting of x and y.
- The notation $\langle x, y \rangle$, too, can mean something different in certain texts (namely, the span of x and y); however, it won't have this second meaning in our course.
- If I am not mistaken, Definition 1.1.2 is also not the only game in town. Some authors follow a competing standard, which causes their $\langle x, y \rangle$ to be what we would denote $\langle y, x \rangle$.
- Finally, the word "dot product" often means the analogue of $\langle x, y \rangle$ that does not use complex conjugation (i.e., that replaces (1) by $\langle x, y \rangle := x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$). This convention is used mostly in abstract algebra, where complex conjugation is not considered intrinsic to the number system. We will not use this convention. For vectors with real entries, the distinction disappears, since $\overline{\lambda} = \lambda$ for any $\lambda \in \mathbb{R}$.

Definition 1.1.4. For any column vector $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{C}^n$, we define the row vector

$$y^* := \begin{pmatrix} \overline{y_1} & \overline{y_2} & \cdots & \overline{y_n} \end{pmatrix} \in \mathbb{C}^{1 \times n}.$$

Proposition 1.1.5. Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$. Then:

(a) We have $\langle x, y \rangle = y^* x$.

(b) We have $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

(c) We have $\langle x + x', y \rangle = \langle x, y \rangle + \langle x', y \rangle$ for any $x' \in \mathbb{C}^n$.

(d) We have $\langle x, y + y' \rangle = \langle x, y \rangle + \langle x, y' \rangle$ for any $y' \in \mathbb{C}^n$.

(e) We have $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$ for any $\lambda \in \mathbb{C}$.

(f) We have $\langle x, \lambda y \rangle = \overline{\lambda} \langle x, y \rangle$ for any $\lambda \in \mathbb{C}$.

(g) We have $\langle x - x', y \rangle = \langle x, y \rangle - \langle x', y \rangle$ for any $x' \in \mathbb{C}^n$.

(h) We have $\langle x, y - y' \rangle = \langle x, y \rangle - \langle x, y' \rangle$ for any $y' \in \mathbb{C}^n$.

(i) We have $\left\langle \sum_{i=1}^{k} \lambda_i x_i, y \right\rangle = \sum_{i=1}^{k} \lambda_i \langle x_i, y \rangle$ for any $k \in \mathbb{N}$, any $x_1, x_2, \ldots, x_k \in \mathbb{C}^n$ and any $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{C}$.

(j) We have $\left\langle x, \sum_{i=1}^{k} \lambda_i y_i \right\rangle = \sum_{i=1}^{k} \overline{\lambda_i} \langle x, y_i \rangle$ for any $k \in \mathbb{N}$, any $y_1, y_2, \ldots, y_k \in \mathbb{C}^n$ and any $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{C}$.

Proof. Parts **(a)** till **(h)** are straightforward computations using Definition 1.1.2, since

- the multiplication in $\mathbb{C}$ is commutative;
- we have $\overline{\overline{z}} = z$ for any $z \in \mathbb{C}$.

Parts **(i)** and **(j)** follow from parts **(c)**, **(d)**, **(e)** and **(f)** by induction on k. □

Proposition 1.1.6. Let $x \in \mathbb{C}^n$. Then:

(a) The number $\langle x, x \rangle$ is a nonnegative real.

(b) We have $\langle x, x \rangle > 0$ whenever $x \neq 0$.

Proof. Write x as $x = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$. Then, the definition of $\langle x, x \rangle$ yields

$$\begin{aligned} \langle x, x \rangle &= x_1 \overline{x_1} + x_2 \overline{x_2} + \cdots + x_n \overline{x_n} \\ &= |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2, \end{aligned} \tag{2}$$

since any complex number z satisfies $z\overline{z} = |z|^2$. Since all the absolute values $|x_1|, |x_2|, \ldots, |x_n|$ are real, this yields immediately that $\langle x, x \rangle$ is a nonnegative real. Thus, Proposition 1.1.6 **(a)** is proved.

(b) Assume that $x \neq 0$. Thus, at least one $i \in [n]$ satisfies $x_i \neq 0$ and therefore $|x_i|^2 > 0$. This entails $\langle x, x \rangle = |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 > 0$ (because a sum of nonnegative reals that has at least one positive addend is always > 0). In view of (2), this rewrites as $\langle x, x \rangle > 0$. This proves Proposition 1.1.6 **(b)**. □

Definition 1.1.7. Let $x \in \mathbb{C}^n$. We define the *length* of x to be the nonnegative real number

$$||x|| := \sqrt{\langle x, x \rangle}.$$

This is well-defined, since Proposition 1.1.6 **(a)** says that $\langle x, x \rangle$ is a nonnegative real.

Example 1.1.8. If $x = \begin{pmatrix} 1+i \\ 3-2i \end{pmatrix} \in \mathbb{C}^2$, then

$$\begin{aligned} \langle x, x \rangle &= (1+i)\left(\overline{1+i}\right) + (3-2i)\left(\overline{3+2i}\right) = (1+i)(1-i) + (3-2i)(3+2i) \\ &= 1+1+9+4 = 15 \end{aligned}$$

and thus $||x|| = \sqrt{\langle x, x \rangle} = \sqrt{15}$.

The length $||x||$ of a vector $x \in \mathbb{C}^n$ is sometimes also called the *norm* of x (but beware that other things are called "norms" as well).

A vector $x \in \mathbb{C}^n$ has zero length if and only if it is 0:

Proposition 1.1.9. Let $x \in \mathbb{C}^n$. Then, $||x|| = 0$ if and only if $x = 0$.

Proof. The "if" part follows from $||0|| = 0$, which is obvious. To prove the "only if" part, we assume that $||x|| = 0$. Thus, $\langle x, x \rangle = 0$ (since $||x|| = \sqrt{\langle x, x \rangle}$). However, if we had $x \neq 0$, then Proposition 1.1.6 **(b)** would yield $\langle x, x \rangle > 0$, which would contradict $\langle x, x \rangle = 0$. Thus, we cannot have $x \neq 0$. Hence, $x = 0$. Thus, the "only if" part is proven. □

Proposition 1.1.10. For any $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n$, we have $||\lambda x|| = |\lambda| \cdot ||x||$.

Proof. Straightforward. □

Exercise 1.1.1. [3] Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$. Prove that

$$||x + y||^2 - ||x||^2 - ||y||^2 = \langle x, y \rangle + \langle y, x \rangle = 2 \cdot \operatorname{Re} \langle x, y \rangle .$$

Here, $\operatorname{Re} z$ denotes the real part of any complex number z.

One of the most famous properties of the inner product is the *Cauchy–Schwarz inequality* (see [Steele04] for various applications):

Theorem 1.1.11 (Cauchy–Schwarz inequality)**.** Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$ be two vectors. Then:

(a) The inequality

$$||x|| \cdot ||y|| \geq |\langle x, y \rangle|$$

holds.

(b) This inequality becomes an equality if and only if the pair (x, y) of vectors is linearly dependent.

Proof of Theorem 1.1.11. If $x = 0$, then Theorem 1.1.11 is obvious (because the inequality in part **(a)** simplifies to $0 \geq 0$, and since the pair $(0, y)$ of vectors is always linearly dependent). Hence, for the rest of this proof, we WLOG assume that $x \neq 0$.

Thus, Proposition 1.1.6 **(a)** yields that $\langle x, x \rangle$ is a nonnegative real, and Proposition 1.1.6 **(b)** yields $\langle x, x \rangle > 0$. Let $a := \langle x, x \rangle$. Then, $a = \langle x, x \rangle > 0$. Furthermore, let $b := \langle y, x \rangle \in \mathbb{C}$. Thus, $\overline{b} = \overline{\langle y, x \rangle} = \langle x, y \rangle$ (by Proposition 1.1.5 **(b)**).

Now, Proposition 1.1.6 **(a)** (applied to $bx - ay$ instead of x) yields that

$$\langle bx - ay, bx - ay \rangle \geq 0. \tag{3}$$

Since

$$\begin{aligned}
&\langle bx-ay, bx-ay\rangle \\
&= \underbrace{\langle bx, bx-ay\rangle}_{=\langle bx,bx\rangle-\langle bx,ay\rangle \text{ (by Proposition 1.1.5 \textbf{(h)})}} - \underbrace{\langle ay, bx-ay\rangle}_{=\langle ay,bx\rangle-\langle ay,ay\rangle \text{ (by Proposition 1.1.5 \textbf{(h)})}} \qquad \text{(by Proposition 1.1.5 \textbf{(g)})} \\
&= \langle bx, bx\rangle - \langle bx, ay\rangle - (\langle ay, bx\rangle - \langle ay, ay\rangle) \\
&= \underbrace{\langle bx, bx\rangle}_{=b\langle x,bx\rangle \text{ (by Proposition 1.1.5 \textbf{(e)})}} + \underbrace{\langle ay, ay\rangle}_{=a\langle y,ay\rangle \text{ (by Proposition 1.1.5 \textbf{(e)})}} \\
&\quad - \underbrace{\langle bx, ay\rangle}_{=b\langle x,ay\rangle \text{ (by Proposition 1.1.5 \textbf{(e)})}} - \underbrace{\langle ay, bx\rangle}_{=a\langle y,bx\rangle \text{ (by Proposition 1.1.5 \textbf{(e)})}} \\
&= b\underbrace{\langle x, bx\rangle}_{=\overline{b}\langle x,x\rangle \text{ (by Proposition 1.1.5 \textbf{(f)})}} + a\underbrace{\langle y, ay\rangle}_{=\overline{a}\langle y,y\rangle \text{ (by Proposition 1.1.5 \textbf{(f)})}} \\
&\quad - b\underbrace{\langle x, ay\rangle}_{=\overline{a}\langle x,y\rangle \text{ (by Proposition 1.1.5 \textbf{(f)})}} - a\underbrace{\langle y, bx\rangle}_{=\overline{b}\langle y,x\rangle \text{ (by Proposition 1.1.5 \textbf{(f)})}} \\
&= b\overline{b}\underbrace{\langle x, x\rangle}_{=a} + a\underbrace{\overline{a}}_{=a \text{ (since } a\in\mathbb{R})}\langle y, y\rangle - b\underbrace{\overline{a}}_{=a \text{ (since } a\in\mathbb{R})}\underbrace{\langle x, y\rangle}_{=\overline{b}} - a\overline{b}\underbrace{\langle y, x\rangle}_{=b} \\
&= b\overline{b}a + \underbrace{aa}_{=a^2}\langle y, y\rangle - \underbrace{ba}_{=ab}\overline{b} - \underbrace{a\overline{b}b}_{=b\overline{b}a} \\
&= b\overline{b}a + a^2\langle y, y\rangle - ab\overline{b} - b\overline{b}a = a^2\langle y, y\rangle - ab\overline{b} = a\left(a\langle y, y\rangle - b\overline{b}\right),
\end{aligned}$$

we can rewrite this as

$$a\left(a\langle y, y\rangle - b\overline{b}\right) \geq 0.$$

We can divide both sides of this inequality by a (since $a > 0$). Thus, we obtain

$$a\langle y, y\rangle - b\overline{b} \geq 0.$$

In other words,

$$a\langle y, y\rangle \geq b\overline{b}.$$

In view of

$$a = \langle x, x\rangle = ||x||^2 \qquad \left(\text{since } ||x|| = \sqrt{\langle x, x\rangle} \text{ (by the definition of } ||x||\text{)}\right)$$

and

$$\langle y, y\rangle = ||y||^2 \qquad \left(\text{since } ||y|| = \sqrt{\langle y, y\rangle} \text{ (by the definition of } ||y||\text{)}\right)$$

and

$$b\overline{b} = \overline{b}\underbrace{b}_{=\overline{\overline{b}}} = \overline{b}\overline{b} = \left|\overline{b}\right|^2 \qquad \left(\text{because } z\overline{z} = |z|^2 \text{ for any } z \in \mathbb{C}\right),$$

we can rewrite this as $||x||^2 \, ||y||^2 \geq \left|\overline{b}\right|^2$. Since $||x||$ and $||y||$ and $\left|\overline{b}\right|$ are nonnegative reals, we can take square roots on both sides of this inequality, and obtain $||x|| \cdot ||y|| \geq \left|\overline{b}\right|$. In other words, $||x|| \cdot ||y|| \geq |\langle x, y\rangle|$ (since $\overline{b} = \langle x, y\rangle$). This proves Theorem 1.1.11 **(a)**.

(b) Our above proof of the inequality $||x|| \cdot ||y|| \geq |\langle x, y\rangle|$ shows that this inequality can only become an equality if $\langle bx - ay, bx - ay\rangle = 0$ (since it was obtained by a chain of reversible transformations from the inequality (3)). But this happens if and only if $bx - ay = 0$ (since Proposition 1.1.6 **(b)** shows that $\langle bx - ay, bx - ay\rangle > 0$ in any other case). In turn, $bx - ay = 0$ entails that the pair (x, y) is linearly dependent (since $a > 0$). Thus, the inequality $||x|| \cdot ||y|| \geq |\langle x, y\rangle|$ can only become an equality if the pair (x, y) is linearly dependent. Conversely, it is easy to see that if the pair (x, y) is linearly dependent, then the inequality $||x|| \cdot ||y|| \geq |\langle x, y\rangle|$ indeed becomes an equality (because in light of $x \neq 0$, the linear dependence of the pair (x, y) yields that $y = \lambda x$ for some $\lambda \in \mathbb{C}$). Thus, Theorem 1.1.11 **(b)** is proven. □

Using Theorem 1.1.11 and Exercise 1.1.1, we can easily obtain the following:

Theorem 1.1.12 (triangle inequality)**.** Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$. Then:

(a) The inequality $||x|| + ||y|| \geq ||x + y||$ holds.

(b) This inequality becomes an equality if and only if we have $y = 0$ or $x = \lambda y$ for some nonnegative real λ.

Exercise 1.1.2. [3] Prove Theorem 1.1.12.

Theorem 1.1.12 **(a)** is the reason why the map $\mathbb{C}^n \to \mathbb{R}$, $x \mapsto ||x||$ is called a "norm".

1.2. Orthogonality and orthonormality

We shall now define orthogonality first for two vectors, then for any tuple of vectors.

Definition 1.2.1. Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$ be two vectors. We say that x is *orthogonal* to y if and only if $\langle x, y\rangle = 0$. The shorthand notation for this is "$x \perp y$".

The relation $\perp$ is symmetric:

Proposition 1.2.2. Let $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^n$ be two vectors. Then, $x \perp y$ holds if and only if $y \perp x$.

Proof. Follows from Proposition 1.1.5 **(b)**. □

Definition 1.2.3. Let $(u_1, u_2, \ldots, u_k)$ be a tuple of vectors in $\mathbb{C}^n$. Then:

(a) We say that the tuple $(u_1, u_2, \ldots, u_k)$ is *orthogonal* if we have

$$u_p \perp u_q \qquad \text{whenever } p \neq q.$$

(b) We say that the tuple $(u_1, u_2, \ldots, u_k)$ is *orthonormal* if it is orthogonal **and** satisfies

$$||u_1|| = ||u_2|| = \cdots = ||u_k|| = 1.$$

(c) We note that the orthogonality and the orthonormality of a tuple are preserved when the entries of the tuple are permuted. Thus, we can extend both notions ("orthogonal" and "orthonormal") to finite sets of vectors in $\mathbb{C}^n$: A set $\{u_1, u_2, \ldots, u_k\}$ of vectors in $\mathbb{C}^n$ (with $u_1, u_2, \ldots, u_k$ being distinct) is said to be *orthogonal* (or *orthonormal*, respectively) if and only if the tuple $(u_1, u_2, \ldots, u_k)$ is orthogonal (resp., orthonormal).

(d) Sometimes, we (sloppily) say "the vectors $u_1, u_2, \ldots, u_k$ are orthogonal" when we mean "the tuple $(u_1, u_2, \ldots, u_k)$ is orthogonal". The same applies to "orthonormal".

Example 1.2.4. (a) The tuple $\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right)$ of vectors in $\mathbb{C}^3$ is orthonormal. It is also a basis of $\mathbb{C}^3$, and known as the *standard basis*.

(b) More generally: Let $n \in \mathbb{N}$. Let $e_1, e_2, \ldots, e_n \in \mathbb{C}^n$ be the vectors defined by

$$e_i = \underbrace{\begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix}^T}_{\substack{\text{the 1 is in the } i\text{-th position;} \\ \text{all other entries are 0}}}.$$

Then, $(e_1, e_2, \ldots, e_n)$ is an orthonormal basis of $\mathbb{C}^n$, and is known as the *standard basis* of $\mathbb{C}^n$.

(c) The pair $\left(\begin{pmatrix} 1 \\ -i \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2i \\ 1 \end{pmatrix} \right)$ of vectors in $\mathbb{C}^3$ is orthogonal (but not orthonormal). Indeed,

$$\left\langle \begin{pmatrix} 1 \\ -i \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2i \\ 1 \end{pmatrix} \right\rangle = 1 \cdot \overline{0} + (-i) \cdot \overline{2i} + 2 \cdot \overline{1} = 0 - 2 + 2 = 0.$$

(d) The pair $\left(\frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -i \\ 2 \end{pmatrix}, \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 2i \\ 1 \end{pmatrix} \right)$ of vectors in $\mathbb{C}^3$ is orthonormal. (This is just the previous pair, with each vector scaled so that its length becomes 1.)

Proposition 1.2.5. Let $(u_1, u_2, \ldots, u_k)$ be an orthogonal tuple of nonzero vectors in $\mathbb{C}^n$. Then, the tuple

$$\left(\frac{1}{||u_1||} u_1, \ \frac{1}{||u_2||} u_2, \ \ldots, \ \frac{1}{||u_k||} u_k \right)$$

is orthonormal.

Proof. Straightforward. (Observe that any two orthogonal vectors remain orthogonal when they are scaled by scalars.) □

Proposition 1.2.6. Any orthogonal tuple of nonzero vectors in $\mathbb{C}^n$ is linearly independent.

Proof. Let $(u_1, u_2, \ldots, u_k)$ be an orthogonal tuple of nonzero vectors in $\mathbb{C}^n$. We must prove that it is linearly independent.

Indeed, for any $i \in [k]$ and any $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{C}$, we have

$$\begin{aligned} &\langle \lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_k u_k, u_i \rangle \\ &= \lambda_1 \langle u_1, u_i \rangle + \lambda_2 \langle u_2, u_i \rangle + \cdots + \lambda_k \langle u_k, u_i \rangle \\ &\qquad \text{(by parts \textbf{(c)} and \textbf{(e)} of Proposition 1.1.5)} \\ &= \lambda_i \langle u_i, u_i \rangle + \sum_{\substack{j \in [k]; \\ j \neq i}} \lambda_j \underbrace{\langle u_j, u_i \rangle}_{\substack{=0 \\ \text{(since } u_j \perp u_i \\ \text{(because } (u_1, u_2, \ldots, u_k) \text{ is} \\ \text{an orthogonal tuple))}}} \\ &= \lambda_i \langle u_i, u_i \rangle . \end{aligned} \tag{4}$$

For any $i \in [k]$, we have $u_i \neq 0$ (since $(u_1, u_2, \ldots, u_k)$ is a tuple of nonzero vectors) and thus

$$\langle u_i, u_i \rangle > 0 \tag{5}$$

(by Proposition 1.1.6 **(b)**, applied to $x = u_i$).

Now, let $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{C}$ be such that $\lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_k u_k = 0$. Then, for each $i \in [k]$, we have

$$\begin{aligned} \lambda_i \langle u_i, u_i \rangle &= \left\langle \underbrace{\lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_k u_k}_{=0}, u_i \right\rangle \qquad \text{(by (4))} \\ &= \langle 0, u_i \rangle = 0 \end{aligned}$$

and therefore $\lambda_i = 0$ (indeed, we can divide by $\langle u_i, u_i \rangle$, because of (5)).

Forget that we fixed $\lambda_1, \lambda_2, \ldots, \lambda_k$. We thus have shown that if $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{C}$ are such that $\lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_k u_k = 0$, then we have $\lambda_i = 0$ for each $i \in [k]$. In other words, $(u_1, u_2, \ldots, u_k)$ is linearly independent. This proves Proposition 1.2.6. □

The following simple lemma will be used further below:

Lemma 1.2.7. Let $k < n$. Let $a_1, a_2, \ldots, a_k$ be k vectors in $\mathbb{C}^n$. Then, there exists a nonzero vector $b \in \mathbb{C}^n$ that is orthogonal to each of $a_1, a_2, \ldots, a_k$.

Proof. Write each vector a_i as $a_i = \begin{pmatrix} a_{i,1} & a_{i,2} & \cdots & a_{i,n} \end{pmatrix}^T$. Now, consider an arbitrary vector $b = \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix}^T \in \mathbb{C}^n$, whose entries $b_1, b_2, \ldots, b_n$ are so far undetermined. This new vector b is orthogonal to each of $a_1, a_2, \ldots, a_k$ if and only if it satisfies

$$\langle b, a_i \rangle = 0 \qquad \text{for all } i \in [k].$$

In other words, this new vector b is orthogonal to each of $a_1, a_2, \ldots, a_k$ if and only if it satisfies

$$b_1 \overline{a_{i,1}} + b_2 \overline{a_{i,2}} + \cdots + b_n \overline{a_{i,n}} = 0 \qquad \text{for all } i \in [k]$$

(since $\langle b, a_i \rangle = b_1 \overline{a_{i,1}} + b_2 \overline{a_{i,2}} + \cdots + b_n \overline{a_{i,n}}$ for each $i \in [k]$). In other words, this new vector b is orthogonal to each of $a_1, a_2, \ldots, a_k$ if and only if it satisfies the system of equations

$$\begin{cases} b_1 \overline{a_{1,1}} + b_2 \overline{a_{1,2}} + \cdots + b_n \overline{a_{1,n}} = 0; \\ b_1 \overline{a_{2,1}} + b_2 \overline{a_{2,2}} + \cdots + b_n \overline{a_{2,n}} = 0; \\ \qquad \cdots; \\ b_1 \overline{a_{k,1}} + b_2 \overline{a_{k,2}} + \cdots + b_n \overline{a_{k,n}} = 0. \end{cases}$$

But this is a system of k homogeneous linear equations in the n unknowns $b_1, b_2, \ldots, b_n$, and thus (by a classical fact in linear algebra[2]) has at least one nonzero solution (since $k < n$). In other words, there exists at least one nonzero vector $b = \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix}^T \in \mathbb{C}^n$ that is orthogonal to each of $a_1, a_2, \ldots, a_k$. This proves Lemma 1.2.7.

Here is a neater way to state the same argument: We define a map $f : \mathbb{C}^n \to \mathbb{C}^k$ by setting

$$f(w) = \begin{pmatrix} \langle w, a_1 \rangle \\ \langle w, a_2 \rangle \\ \vdots \\ \langle w, a_k \rangle \end{pmatrix} \qquad \text{for each } w \in \mathbb{C}^n.$$

[2]The fact we are using here is the following: If p and q are two integers such that $0 \leq p < q$, then any system of p homogeneous linear equations in q unknowns has at least one nonzero solution. Rewritten in terms of matrices, this is saying that if p and q are two integers such that $0 \leq p < q$, then any $p \times q$-matrix has a nonzero vector in its kernel (= nullspace). For a proof, see, e.g., [Strick20, Remark 8.9] or (rewritten in the language of linear maps) [LaNaSc16, Corollary 6.5.3 item 1].

It is easy to see that this map f is $\mathbb{C}$-linear. (Indeed, Proposition 1.1.5 **(c)** shows that every two vectors $x, x' \in \mathbb{C}^n$ and every $i \in [k]$ satisfy $\langle x + x', a_i \rangle = \langle x, a_i \rangle + \langle x', a_i \rangle$; therefore, every two vectors $x, x' \in \mathbb{C}^n$ satisfy $f(x + x') = f(x) + f(x')$. Similarly, Proposition 1.1.5 **(e)** can be used to show that $f(\lambda x) = \lambda f(x)$ for each $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n$. Hence, f is $\mathbb{C}$-linear.)

Now, we know that f is a $\mathbb{C}$-linear map from $\mathbb{C}^n$ to $\mathbb{C}^k$. Hence, the rank-nullity theorem (see, e.g., [Treil15, Chapter 2, Theorem 7.2] or [Knapp16, Chapter II, Corollary 2.15] or [Goodma15, Proposition 3.3.35]) yields that

$$n = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f),$$

where $\operatorname{Ker} f$ denotes the kernel of f (that is, the subspace of $\mathbb{C}^n$ that consists of all vectors $v \in \mathbb{C}^n$ satisfying $f(v) = 0$), and where $\operatorname{Im} f$ denotes the image[3] of f (that is, the subspace of $\mathbb{C}^k$ consisting of all vectors of the form $f(v)$ with $v \in \mathbb{C}^n$). Therefore,

$$\dim(\operatorname{Ker} f) = n - \dim(\operatorname{Im} f).$$

However, $\operatorname{Im} f$ is a vector subspace of $\mathbb{C}^k$, and thus has dimension $\leq k$. Thus, $\dim(\operatorname{Im} f) \leq k < n$, so that

$$\dim(\operatorname{Ker} f) = n - \underbrace{\dim(\operatorname{Im} f)}_{<n} > n - n = 0.$$

This shows that the vector space $\operatorname{Ker} f$ contains at least one nonzero vector b. Consider this b. Thus, $b \in \operatorname{Ker} f \subseteq \mathbb{C}^n$.

However, $b \in \operatorname{Ker} f$ shows that $f(b) = 0$. But the definition of f yields $f(b) = \begin{pmatrix} \langle b, a_1 \rangle \\ \langle b, a_2 \rangle \\ \vdots \\ \langle b, a_k \rangle \end{pmatrix}$. Thus, $\begin{pmatrix} \langle b, a_1 \rangle \\ \langle b, a_2 \rangle \\ \vdots \\ \langle b, a_k \rangle \end{pmatrix} = f(b) = 0$. In other words, each $i \in [k]$ satisfies $\langle b, a_i \rangle = 0$. In other words, each $i \in [k]$ satisfies $b \perp a_i$. In other words, b is orthogonal to each of $a_1, a_2, \ldots, a_k$. Thus, we have found a nonzero vector $b \in \mathbb{C}^n$ that is orthogonal to each of $a_1, a_2, \ldots, a_k$. This proves Lemma 1.2.7. □

Corollary 1.2.8. Let $(u_1, u_2, \ldots, u_k)$ be an orthogonal k-tuple of nonzero vectors in $\mathbb{C}^n$. Then, we have $k \leq n$, and we can find $n - k$ further nonzero vectors $u_{k+1}, u_{k+2}, \ldots, u_n$ such that $(u_1, u_2, \ldots, u_n)$ is an orthogonal basis of $\mathbb{C}^n$.

Exercise 1.2.1. [2] Prove Corollary 1.2.8.

Corollary 1.2.9. Let $(u_1, u_2, \ldots, u_k)$ be an orthonormal k-tuple of vectors in $\mathbb{C}^n$. Then, we have $k \leq n$, and we can find $n - k$ further nonzero vectors $u_{k+1}, u_{k+2}, \ldots, u_n$ such that $(u_1, u_2, \ldots, u_n)$ is an orthonormal basis of $\mathbb{C}^n$.

[3] also known as "range"

Proof. The k-tuple $(u_1, u_2, \ldots, u_k)$ is an orthogonal tuple of nonzero vectors (since it is orthonormal). Hence, Corollary 1.2.8 yields that we can find $n-k$ further nonzero vectors $u_{k+1}, u_{k+2}, \ldots, u_n$ such that $(u_1, u_2, \ldots, u_n)$ is an orthogonal basis of $\mathbb{C}^n$. Consider these $n-k$ vectors $u_{k+1}, u_{k+2}, \ldots, u_n$, and replace them by

$$\frac{1}{||u_{k+1}||}u_{k+1}, \quad \frac{1}{||u_{k+2}||}u_{k+2}, \quad \ldots, \quad \frac{1}{||u_n||}u_n,$$

respectively. Then, the orthogonal basis $(u_1, u_2, \ldots, u_n)$ becomes an orthonormal basis (since an orthogonal basis remains orthogonal when we scale its entries, and since the first k vectors $u_1, u_2, \ldots, u_k$ already have length 1 by assumption). Thus, Corollary 1.2.9 is proved. □

1.3. Conjugate transposes

The following definition generalizes Definition 1.1.4:

Definition 1.3.1. Let $A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{pmatrix} \in \mathbb{C}^{n \times m}$ be any $n \times m$-matrix. Then, we define the $m \times n$-matrix

$$A^* := \begin{pmatrix} \overline{a_{1,1}} & \cdots & \overline{a_{n,1}} \\ \vdots & \ddots & \vdots \\ \overline{a_{1,m}} & \cdots & \overline{a_{n,m}} \end{pmatrix} \in \mathbb{C}^{m \times n}.$$

This matrix A^* is called the *conjugate transpose* of A.

This conjugate transpose A^* can thus be obtained from the usual transpose A^T by conjugating all entries.

Example 1.3.2. $\begin{pmatrix} 1+i & 2-3i & 5i \\ 6 & 2+4i & 10-i \end{pmatrix}^* = \begin{pmatrix} 1-i & 6 \\ 2+3i & 2-4i \\ -5i & 10+i \end{pmatrix}$.

In the olden days, the conjugate transpose of a matrix was also known as the "adjoint" of A. Unsurprisingly, this word has at least one other meaning, which opens the door to a lot of unwanted confusion; thus we will speak of the "conjugate transpose" instead.

Some authors use the alternative notation $A^\dagger$ (read "A dagger") for A^*. (The Wikipedia suggests calling it the "bedaggered matrix A", although I am not aware of anyone using this terminology outside of the Wikipedia.)

The following rules for conjugate transposes are straightforward to check:

Proposition 1.3.3. (a) If $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{n\times m}$ are two matrices, then $(A+B)^* = A^* + B^*$.

(b) If $A \in \mathbb{C}^{n\times m}$ and $\lambda \in \mathbb{C}$, then $(\lambda A)^* = \overline{\lambda} A^*$.

(c) If $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{m\times k}$ are two matrices, then $(AB)^* = B^* A^*$.

(d) If $A \in \mathbb{C}^{n\times m}$, then $(A^*)^* = A$.

1.4. Isometries

Definition 1.4.1. An $n \times k$-matrix A is said to be an *isometry* if $A^*A = I_k$.

Proposition 1.4.2. An $n \times k$-matrix A is an isometry if and only if its columns form an orthonormal tuple of vectors.

Proof. Let A be an $n \times k$-matrix with columns $a_1, a_2, \ldots, a_k$ from left to right. Therefore,

$$A = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_k \\ | & & | \end{pmatrix} \qquad \text{and thus} \qquad A^* = \begin{pmatrix} — & a_1^* & — \\ & \vdots & \\ — & a_k^* & — \end{pmatrix}.$$

Hence,

$$A^*A = \begin{pmatrix} a_1^*a_1 & a_1^*a_2 & \cdots & a_1^*a_k \\ a_2^*a_1 & a_2^*a_2 & \cdots & a_2^*a_k \\ \vdots & \vdots & \ddots & \vdots \\ a_k^*a_1 & a_k^*a_2 & \cdots & a_k^*a_k \end{pmatrix}$$
$$= \begin{pmatrix} ||a_1||^2 & \langle a_1, a_2\rangle & \cdots & \langle a_1, a_k\rangle \\ \langle a_2, a_1\rangle & ||a_2||^2 & \cdots & \langle a_2, a_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle a_k, a_1\rangle & \langle a_k, a_2\rangle & \cdots & ||a_k||^2 \end{pmatrix}.$$

On the other hand,

$$I_k = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Thus, $A^*A = I_k$ holds if and only if we have

$$\langle a_p, a_q\rangle = 0 \qquad \text{for all } p \neq q$$

and

$$\left|\left|a_p\right|\right|^2 = 1 \qquad \text{for each } p.$$

In other words, $A^*A = I_k$ holds if and only if we have

$$a_p \perp a_q \qquad \text{for all } p \neq q$$

and

$$||a_1|| = ||a_2|| = \cdots = ||a_k|| = 1.$$

In other words, A is an isometry if and only if $(a_1, a_2, \ldots, a_k)$ is orthonormal. This proves Proposition 1.4.2. □

Isometries are called isometries because they preserve lengths:

Proposition 1.4.3. Let $A \in \mathbb{C}^{n\times k}$ be an isometry. Then, each $x \in \mathbb{C}^k$ satisfies $||Ax|| = ||x||$.

Proof. We have $A^*A = I_k$ (since A is an isometry). Let $x \in \mathbb{C}^k$. Then, the definition of $||Ax||$ yields $||Ax|| = \sqrt{\langle Ax, Ax\rangle}$. Hence,

$$\begin{aligned}
||Ax||^2 &= \langle Ax, Ax\rangle \\
&= \underbrace{(Ax)^*}_{\substack{=x^*A^* \\ \text{(by Proposition 1.3.3 (c))}}} \qquad Ax \qquad \text{(by Proposition 1.1.5 \textbf{(a)})} \\
&= x^*A^*Ax \qquad (6) \\
&= x^*x \qquad (\text{since } A^*A = I_k) \\
&= \langle x, x\rangle \qquad (\text{by Proposition 1.1.5 \textbf{(a)}}) \\
&= ||x||^2
\end{aligned}$$

(since the definition of $||x||$ yields $||x|| = \sqrt{\langle x, x\rangle}$). In other words, we have $||Ax|| = ||x||$ (since $||Ax||$ and $||x||$ are nonnegative reals). This proves Proposition 1.4.3. □

Remark 1.4.4. Another warning on terminology: Some authors (e.g., Conrad in [Conrad, "Isometries"]) use the word "isometry" in a wider sense than we do. Namely, they use it for arbitrary maps from $\mathbb{C}^k$ to $\mathbb{C}^n$ that preserve distances. Our isometries can be viewed as **linear** isometries in this wider sense, because a matrix $A \in \mathbb{C}^{n\times k}$ corresponds to a linear map from $\mathbb{C}^k$ to $\mathbb{C}^n$. However, not all isometries in this wider sense are linear.

1.5. Unitary matrices

1.5.1. Definition, examples, basic properties

Definition 1.5.1. A matrix $U \in \mathbb{C}^{n\times k}$ is said to be *unitary* if and only if both U and U^* are isometries.

Lecture 2 starts here.

Example 1.5.2. (a) The matrix $A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ is unitary. Indeed, it is easy to see that $A^*A = I_2$, so that A is an isometry. Thus, A^* is an isometry as well, since $A^* = A$. Hence, A is unitary.

(b) A 1×1-matrix $\begin{pmatrix} \lambda \end{pmatrix} \in \mathbb{C}^{1\times 1}$ is unitary if and only if $|\lambda| = 1$.

(c) For any $n \in \mathbb{N}$, the identity matrix I_n is unitary.

(d) Let $n \in \mathbb{N}$, and let σ be a permutation of $[n]$ (that is, a bijective map from $[n]$ to $[n]$). Let P_σ be the *permutation matrix* of σ; this is the $n \times n$-matrix whose $(\sigma(j), j)$-th entry is 1 for each $j \in [n]$, and whose all other entries are 0. For instance, if $n = 3$ and if σ is the cyclic permutation sending $1, 2, 3$ to $2, 3, 1$ (respectively), then

$$P_\sigma = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The permutation matrix P_σ is always unitary (for any n and any permutation σ). Indeed, its conjugate transpose $(P_\sigma)^*$ is easily seen to be the permutation matrix $P_{\sigma^{-1}}$ of the inverse permutation σ^{-1}; but this latter permutation matrix $P_{\sigma^{-1}}$ is also the inverse of P_σ.

(e) A diagonal matrix $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{C}^{n\times n}$ is unitary if and only if its diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ lie on the unit circle (i.e., their absolute values $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|$ all equal 1).

Unitary matrices can be characterized in many other ways:

Theorem 1.5.3. Let $U \in \mathbb{C}^{n\times k}$ be a matrix. The following six statements are equivalent:

- $\mathcal{A}$: The matrix U is unitary.
- $\mathcal{B}$: The matrices U and U^* are isometries.
- $\mathcal{C}$: We have $UU^* = I_n$ and $U^*U = I_k$.
- $\mathcal{D}$: The matrix U is square (that is, $n = k$) and invertible and satisfies $U^{-1} = U^*$.
- $\mathcal{E}$: The columns of U form an orthonormal basis of $\mathbb{C}^n$.
- $\mathcal{F}$: The matrix U is square (that is, $n = k$) and is an isometry.

Proof. The equivalence $\mathcal{A} \iff \mathcal{B}$ follows immediately from Definition 1.5.1. The equivalence $\mathcal{B} \iff \mathcal{C}$ follows immediately from the definition of an isometry (since $(U^*)^* = U$). The implication $\mathcal{D} \implies \mathcal{C}$ is obvious. The implication $\mathcal{C} \implies \mathcal{D}$ follows from the known fact (see, e.g., [Treil15, Chapter 2, Corollary 3.7]) that every invertible matrix is square. Let us now prove some of the other implications:

- $\mathcal{D} \implies \mathcal{E}$*:* Assume that statement $\mathcal{D}$ holds. Then, $U^*U = I_k$ (since $U^{-1} = U^*$), and therefore U is an isometry. Hence, Proposition 1.4.2 shows that the tuple of columns of U is orthonormal. However, the columns of U form a basis of $\mathbb{C}^n$ (because U is invertible), and this basis is orthonormal (since we have just shown that the tuple of columns of U is orthonormal). Thus, statement $\mathcal{E}$ holds. We have thus proved the implication $\mathcal{D} \implies \mathcal{E}$.

- $\mathcal{E} \implies \mathcal{D}$*:* Assume that statement $\mathcal{E}$ holds. Then, the columns of U form an orthonormal basis, hence an orthonormal tuple. Thus, Proposition 1.4.2 shows that U is an isometry, so that $U^*U = I_k$. However, U is invertible because the columns of U form a basis of $\mathbb{C}^n$. Therefore, from $U^*U = I_k$, we obtain $U^{-1} = U^*$. Finally, the matrix U is square, since any invertible matrix is square. Thus, statement $\mathcal{D}$ holds. We have thus proved the implication $\mathcal{E} \implies \mathcal{D}$.

- $\mathcal{D} \implies \mathcal{F}$*:* The implication $\mathcal{D} \implies \mathcal{F}$ is easy (since $U^{-1} = U^*$ entails $U^*U = I_k$, which shows that U is an isometry).

- $\mathcal{F} \implies \mathcal{D}$*:* Assume that statement $\mathcal{F}$ holds. Thus, U is an isometry; that is, we have $U^*U = I_k = I_n$ (since $k = n$). However, it is known[4] that a square matrix A that has a left inverse (i.e., a further square matrix B satisfying $BA = I$) must be invertible. We can apply this to the square matrix U (which has a left inverse, since $U^*U = I_n$), and thus conclude that U is invertible. Hence, from $U^*U = I_n$, we obtain $U^{-1} = U^*$. Therefore, statement $\mathcal{D}$ holds. We have thus proved the implication $\mathcal{F} \implies \mathcal{D}$.

Altogether, we have thus proved that all six statements $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F}$ are equivalent. □

Note that Theorem 1.5.3 (specifically, the implication $\mathcal{A} \implies \mathcal{D}$) shows that any unitary matrix is square. In contrast, an isometry can be rectangular – but only tall, not wide, as the following exercise shows:

Exercise 1.5.1. [1] Let $A \in \mathbb{C}^{n \times k}$ be an isometry. Show that $n \geq k$.

[4] This is one part of the infamous "inverse matrix theorem" that lists many equivalent conditions for invertibility. See, for example, [Treil15, Chapter 2, Proposition 3.8].

Exercise 1.5.2. [3] **(a)** Prove that the product AB of two isometries $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{m\times k}$ is always an isometry.

(b) Prove that the product AB of two unitary matrices $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times n}$ is always unitary.

(c) Prove that the inverse of a unitary matrix $A \in \mathbb{C}^{n\times n}$ is always unitary.

Exercise 1.5.2 shows that the set of all unitary $n \times n$-matrices over $\mathbb{C}$ (for a given $n \in \mathbb{N}$) is a group under multiplication. This group is known as the *n-th unitary group*, and is denoted by $\mathrm{U}_n(\mathbb{C})$.

Exercise 1.5.3. [2] Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix.

(a) Prove that $|\det U| = 1$.

(b) Prove that any eigenvalue λ of U satisfies $|\lambda| = 1$.

1.5.2. Various constructions of unitary matrices

The next two exercises show some ways to generate unitary matrices:

Exercise 1.5.4. [3] Let $w \in \mathbb{C}^n$ be a nonzero vector. Then, $w^*w = \langle w, w\rangle > 0$ (by Proposition 1.1.6 **(b)**). Thus, we can define an $n \times n$-matrix

$$U_w := I_n - 2\left(w^*w\right)^{-1} ww^* \in \mathbb{C}^{n\times n}.$$

This is called a *Householder matrix*.

Show that this matrix U_w is unitary and satisfies $U_w^* = U_w$.

The next exercise uses the notion of a skew-Hermitian matrix:

Definition 1.5.4. A matrix $S \in \mathbb{C}^{n\times n}$ is said to be *skew-Hermitian* if and only if $S^* = -S$.

For instance, the matrix $\begin{pmatrix} i & 1 \\ -1 & 0 \end{pmatrix}$ is skew-Hermitian.

Exercise 1.5.5. [5] Let $S \in \mathbb{C}^{n\times n}$ be a skew-Hermitian matrix.

(a) Prove that the matrix $I_n - S$ is invertible.

[**Hint:** Show first that the matrix $I_n + S^*S$ is invertible, since each nonzero vector $v \in \mathbb{C}^n$ satisfies $v^*\left(I_n + S^*S\right)v = \underbrace{\langle v, v\rangle}_{>0} + \underbrace{\langle Sv, Sv\rangle}_{\geq 0} > 0$. Then, expand the product $\left(I_n - S^*\right)\left(I_n - S\right)$.]

(b) Prove that the matrices $I_n + S$ and $\left(I_n - S\right)^{-1}$ commute (i.e., satisfy $\left(I_n + S\right)\cdot\left(I_n - S\right)^{-1} = \left(I_n - S\right)^{-1}\cdot\left(I_n + S\right)$).

(c) Prove that the matrix $U := (I_n - S)^{-1} \cdot (I_n + S)$ is unitary.

(d) Prove that the matrix $U + I_n$ is invertible.

(e) Prove that $S = (U - I_n) \cdot (U + I_n)^{-1}$.

Exercise 1.5.5 constructs a map[5]

$$\left\{\text{skew-Hermitian matrices in } \mathbb{C}^{n\times n}\right\} \to \left\{U \in \mathrm{U}_n(\mathbb{C}) \mid U + I_n \text{ is invertible}\right\},$$
$$S \mapsto (I_n - S)^{-1} \cdot (I_n + S).$$

This map is known as the *Cayley parametrization* of the unitary matrices (and can be seen as an n-dimensional generalization of the stereographic projection from the imaginary axis to the unit circle – which is what it does for $n = 1$). Exercise 1.5.5 **(e)** shows that it is injective. It is not hard to check that it is surjective, too.

How close is the set $\{U \in \mathrm{U}_n(\mathbb{C}) \mid U + I_n \text{ is invertible}\}$ to the whole unitary group $\mathrm{U}_n(\mathbb{C})$? The answer is that it is almost the entire group $\mathrm{U}_n(\mathbb{C})$. Here is a rigorous way to state this:

Exercise 1.5.6. 3 Let $A \in \mathbb{C}^{n\times n}$ be a matrix. Prove the following:

(a) If A is unitary, then the matrix λA is unitary for each $\lambda \in \mathbb{C}$ satisfying $|\lambda| = 1$.

(b) The matrix $\lambda A + I_n$ is invertible for all but finitely many $\lambda \in \mathbb{C}$.

[**Hint:** The determinant $\det(\lambda A + I_n)$ is a polynomial function in λ.]

(c) The set $\{U \in \mathrm{U}_n(\mathbb{C}) \mid U + I_n \text{ is invertible}\}$ is dense in $\mathrm{U}_n(\mathbb{C})$. (That is, each unitary matrix in $\mathrm{U}_n(\mathbb{C})$ can be written as a limit $\lim_{k\to\infty} U_k$ of a sequence of unitary matrices U_k such that $U_k + I_n$ is invertible for each k.)

Thus, if the Cayley parametrization does not hit a unitary matrix, then at least it comes arbitrarily close.

Remark 1.5.5. A square matrix $A \in \mathbb{C}^{n\times n}$ satisfying $AA^T = A^TA = I_n$ is called *orthogonal*. Thus, unitary matrices differ from orthogonal matrices only in the use of the conjugate transpose A^* instead of the transpose A^T. In particular, a matrix $A \in \mathbb{R}^{n\times n}$ (with real entries) is orthogonal if and only if it is unitary.

Exercise 1.5.7. 5 A *Pythagorean triple* is a triple (p, q, r) of positive integers satisfying $p^2 + q^2 = r^2$. (In other words, it is a triple of positive integers that are the sides of a right-angled triangle.) Two famous Pythagorean triples are $(3, 4, 5)$ and $(5, 12, 13)$.

[5] Recall that $\mathrm{U}_n(\mathbb{C})$ denotes the n-th unitary group (i.e., the set of all unitary $n \times n$-matrices).

(a) Prove that a triple (p,q,r) of positive integers is Pythagorean if and only if the matrix $\begin{pmatrix} p/r & -q/r \\ q/r & p/r \end{pmatrix}$ is unitary.

(b) Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be any unitary matrix with rational entries. Assume that a and c are positive, and write a and c as p/r and q/r for some positive integers p,q,r. Show that (p,q,r) is a Pythagorean triple.

(c) Find infinitely many Pythagorean triples that are pairwise non-proportional (i.e., no two of them are obtained from one another just by multiplying all three entries by the same number).

[**Hint:** Use the $S \mapsto U$ construction from Exercise 1.5.5.]

We shall soon see one more way to construct unitary matrices from smaller ones, using the notion of block matrices, which we shall now introduce.

Incidentally, here is another simple but useful property of skew-Hermitian matrices:

Exercise 1.5.8. $\boxed{2}$ Let $A, B \in \mathbb{C}^{n \times n}$ be two skew-Hermitian matrices. Show that $AB - BA$ is again skew-Hermitian.

1.6. Block matrices

1.6.1. Definition

Definition 1.6.1. Let $\mathbb{F}$ be a field. Let $n, m, p, q \in \mathbb{N}$. Let $A \in \mathbb{F}^{n \times p}$, $B \in \mathbb{F}^{n \times q}$, $C \in \mathbb{F}^{m \times p}$ and $D \in \mathbb{F}^{m \times q}$ be four matrices. Then, $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ shall denote the $(n+m) \times (p+q)$-matrix obtained by "gluing" the four matrices A, B, C, D together in the manner suggested by the notation (i.e., we glue B to the right edge of A, we glue C to the bottom edge of A, and we glue D to the right edge of C and to the bottom edge of B). In other words, we set

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} := \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,p} & B_{1,1} & B_{1,2} & \cdots & B_{1,q} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,p} & B_{2,1} & B_{2,2} & \cdots & B_{2,q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,p} & B_{n,1} & B_{n,2} & \cdots & B_{n,q} \\ C_{1,1} & C_{1,2} & \cdots & C_{1,p} & D_{1,1} & D_{1,2} & \cdots & D_{1,q} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,p} & D_{2,1} & D_{2,2} & \cdots & D_{2,q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{m,1} & C_{m,2} & \cdots & C_{m,p} & D_{m,1} & D_{m,2} & \cdots & D_{m,q} \end{pmatrix}$$

(where, as we recall, the notation $M_{i,j}$ denotes the (i,j)-th entry of a matrix M).

Example 1.6.2. If $A = \begin{pmatrix} a & a' \\ a'' & a''' \end{pmatrix}$ and $B = \begin{pmatrix} b \\ b' \end{pmatrix}$ and $C = \begin{pmatrix} c & c' \end{pmatrix}$ and $D = \begin{pmatrix} d \end{pmatrix}$, then $\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} a & a' & b \\ a'' & a''' & b' \\ c & c' & d \end{pmatrix}$.

The notation introduced in Definition 1.6.1 is called *block matrix notation*, and can be generalized to more than four matrices:

Definition 1.6.3. Let $\mathbb{F}$ be a field. Let $u, v \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$ and $p_1, p_2, \ldots, p_v \in \mathbb{N}$. For each $i \in [u]$ and $j \in [v]$, let $A(i,j) \in \mathbb{F}^{n_i \times p_j}$ be a matrix. (We denote it by $A(i,j)$ instead of $A_{i,j}$ to avoid mistaking it for a single entry.) Then,

$$\begin{pmatrix} A(1,1) & A(1,2) & \cdots & A(1,v) \\ A(2,1) & A(2,2) & \cdots & A(2,v) \\ \vdots & \vdots & \ddots & \vdots \\ A(u,1) & A(u,2) & \cdots & A(u,v) \end{pmatrix} \tag{7}$$

shall denote the $(n_1 + n_2 + \cdots + n_u) \times (p_1 + p_2 + \cdots + p_v)$-matrix obtained by "gluing" the matrices $A(i,j)$ together in the manner suggested by the notation. In other words,

$$\begin{pmatrix} A(1,1) & A(1,2) & \cdots & A(1,v) \\ A(2,1) & A(2,2) & \cdots & A(2,v) \\ \vdots & \vdots & \ddots & \vdots \\ A(u,1) & A(u,2) & \cdots & A(u,v) \end{pmatrix}$$

shall denote the $(n_1 + n_2 + \cdots + n_u) \times (p_1 + p_2 + \cdots + p_v)$-matrix whose $\left(n_1 + n_2 + \cdots + n_{i-1} + k,\ p_1 + p_2 + \cdots + p_{j-1} + \ell\right)$-th entry is $\left(A(i,j)\right)_{k,\ell}$ for all $i \in [u]$ and $j \in [v]$ and $k \in [n_i]$ and $\ell \in [p_j]$.

Alternatively, this matrix can be defined abstractly using direct sums of vector spaces; see [Bourba74, Chapter II, §10, section 2] for this definition.

Example 1.6.4. Let $0_{2\times 2}$ denote the zero matrix of size 2×2. Then,

$$\begin{pmatrix} 0_{2\times 2} & I_2 & 0_{2\times 2} \\ I_2 & 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & -I_2 & I_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}.$$

In Definition 1.6.3, the big matrix (7) is called the *block matrix* formed out of the matrices $A(i,j)$; the single matrices $A(i,j)$ are called its *blocks*.

1.6.2. Multiplying block matrices

One of the most useful properties of block matrices is that they can be multiplied "as if the blocks were numbers" (i.e., by the same formula as for regular matrices), provided that the products make sense. Let us state this more precisely – first for the case of four blocks:

Proposition 1.6.5. Let $\mathbb{F}$ be a field. Let n, n', m, m', ℓ and ℓ' be six nonnegative integers. Let $A \in \mathbb{F}^{n\times m}$, $B \in \mathbb{F}^{n\times m'}$, $C \in \mathbb{F}^{n'\times m}$, $D \in \mathbb{F}^{n'\times m'}$, $A' \in \mathbb{F}^{m\times \ell}$, $B' \in \mathbb{F}^{m\times \ell'}$, $C' \in \mathbb{F}^{m'\times \ell}$ and $D' \in \mathbb{F}^{m'\times \ell'}$. Then,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} AA' + BC' & AB' + BD' \\ CA' + DC' & CB' + DD' \end{pmatrix}.$$

For comparison, here is the formula for the product of two 2×2-matrices (consisting of numbers, not blocks):

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{pmatrix}$$

(for any $a, b, c, d, a', b', c', d' \in \mathbb{F}$). Thus, Proposition 1.6.5 is saying that the same formula can be used to multiply block matrices made of appropriately sized blocks. Thus, roughly speaking, we can multiply block matrices "as if the blocks were numbers". To be fully honest, two caveats apply here:

- In the formula for $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$, we can write the right hand side in many different ways: e.g., we can replace aa' by $a'a$, because multiplication of numbers is commutative. In contrast, multiplication of matrices is not commutative, so that we cannot replace AA' by $A'A$ in Proposition 1.6.5. Thus, we can multiply block matrices "as if the blocks were numbers", but we have to keep the blocks in the correct order (viz., in the order in which they appear on the left hand side).
- We cannot use Proposition 1.6.5 to multiply two arbitrary block matrices; indeed, Proposition 1.6.5 requires the blocks to have "matching" dimensions. For example, A must have as many columns as A' has rows (this is enforced by the assumptions $A \in \mathbb{F}^{n\times m}$ and $A' \in \mathbb{F}^{m\times \ell}$). If this wasn't the case, then the product AA' on the right hand side wouldn't even make sense!

Proof of Proposition 1.6.5. Just check that each entry on the left hand side equals the corresponding entry on the right. This is a straightforward computation that is made painful by the notational load and the need to distinguish between four cases (depending on which block our entry lies in). Do one of the four cases to convince yourself that there is nothing difficult here. (See [Grinbe15] for all the gory details.) □

Unsurprisingly, Proposition 1.6.5 generalizes to the multi-block case:

Proposition 1.6.6. Let $\mathbb{F}$ be a field. Let $u, v, w \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$ and $p_1, p_2, \ldots, p_v \in \mathbb{N}$ and $q_1, q_2, \ldots, q_w \in \mathbb{N}$. For each $i \in [u]$ and $j \in [v]$, let $A(i,j) \in \mathbb{F}^{n_i \times p_j}$ be a matrix. For each $j \in [v]$ and $k \in [w]$, let $B(j,k) \in \mathbb{F}^{p_j \times q_k}$ be a matrix. Then,

$$\begin{pmatrix} A(1,1) & A(1,2) & \cdots & A(1,v) \\ A(2,1) & A(2,2) & \cdots & A(2,v) \\ \vdots & \vdots & \ddots & \vdots \\ A(u,1) & A(u,2) & \cdots & A(u,v) \end{pmatrix} \begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,w) \\ B(2,1) & B(2,2) & \cdots & B(2,w) \\ \vdots & \vdots & \ddots & \vdots \\ B(v,1) & B(v,2) & \cdots & B(v,w) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{v} A(1,j) B(j,1) & \sum_{j=1}^{v} A(1,j) B(j,2) & \cdots & \sum_{j=1}^{v} A(1,j) B(j,w) \\ \sum_{j=1}^{v} A(2,j) B(j,1) & \sum_{j=1}^{v} A(2,j) B(j,2) & \cdots & \sum_{j=1}^{v} A(2,j) B(j,w) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{v} A(u,j) B(j,1) & \sum_{j=1}^{v} A(u,j) B(j,2) & \cdots & \sum_{j=1}^{v} A(u,j) B(j,w) \end{pmatrix}.$$

Proof. Just like Proposition 1.6.5, but with more indices. In short, fun! □

1.6.3. Block-diagonal matrices

Definition 1.6.7. *Block-diagonal matrices* are block matrices of the form (7), where

- we have $u = v$,
- all matrices $A(i,i)$ are square (i.e., we have $n_i = p_i$ for all $i \in [u]$), and
- all $A(i,j)$ with $i \neq j$ are zero matrices.

In other words, block-diagonal matrices are block matrices of the form

$$\begin{pmatrix} A(1,1) & 0 & \cdots & 0 \\ 0 & A(2,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(u,u) \end{pmatrix},$$

where $A(1,1), A(2,2), \ldots, A(u,u)$ are arbitrary square matrices, and where each "0" means a zero matrix of appropriate dimensions.

As an easy consequence of Proposition 1.6.6, we obtain a multiplication rule for block-diagonal matrices that looks exactly like multiplication of usual diagonal matrices:

Corollary 1.6.8. Let $u \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$. For each $i \in [u]$, let $A(i,i)$ and $B(i,i)$ be two $n_i \times n_i$-matrices. Then,

$$\begin{pmatrix} A(1,1) & 0 & \cdots & 0 \\ 0 & A(2,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(u,u) \end{pmatrix} \begin{pmatrix} B(1,1) & 0 & \cdots & 0 \\ 0 & B(2,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B(u,u) \end{pmatrix}$$
$$= \begin{pmatrix} A(1,1)\,B(1,1) & 0 & \cdots & 0 \\ 0 & A(2,2)\,B(2,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(u,u)\,B(u,u) \end{pmatrix}.$$

(Here, each "0" means a zero matrix of appropriate dimensions.)

Example 1.6.9. Let $u = 2$ and $n_1 = 1$ and $n_2 = 2$. Let $A(1,1) = \begin{pmatrix} a \end{pmatrix}$ and $A(2,2) = \begin{pmatrix} b & c \\ d & e \end{pmatrix}$ and $B(1,1) = \begin{pmatrix} a' \end{pmatrix}$ and $B(2,2) = \begin{pmatrix} b' & c' \\ d' & e' \end{pmatrix}$. Then, Corollary 1.6.8 says that

$$\begin{pmatrix} A(1,1) & 0 \\ 0 & A(2,2) \end{pmatrix} \begin{pmatrix} B(1,1) & 0 \\ 0 & B(2,2) \end{pmatrix} = \begin{pmatrix} A(1,1)\,B(1,1) & 0 \\ 0 & A(2,2)\,B(2,2) \end{pmatrix},$$

i.e., that

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & c \\ 0 & d & e \end{pmatrix} \begin{pmatrix} a' & 0 & 0 \\ 0 & b' & c' \\ 0 & d' & e' \end{pmatrix} = \begin{pmatrix} aa' & 0 & 0 \\ 0 & bb'+cd' & bc'+ce' \\ 0 & db'+ed' & dc'+ee' \end{pmatrix}.$$

Corollary 1.6.8 can be stated (somewhat imprecisely) as follows: To multiply two block-diagonal matrices, we just multiply respective blocks with each other. The same applies to addition instead of multiplication. Thus, one can think of the diagonal blocks in a block-diagonal matrix as separate matrices, which are stuck together in a block-diagonal shape but don't interfere with each other.

Taking powers of block-diagonal matrices follows the same paradigm:

Corollary 1.6.10. Let $u \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$. For each $i \in [u]$, let $A(i,i)$

be an $n_i \times n_i$-matrix. Then,

$$\begin{pmatrix} A(1,1) & 0 & \cdots & 0 \\ 0 & A(2,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A(u,u) \end{pmatrix}^k$$
$$= \begin{pmatrix} (A(1,1))^k & 0 & \cdots & 0 \\ 0 & (A(2,2))^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (A(u,u))^k \end{pmatrix}$$

for any $k \in \mathbb{N}$. (Here, each "0" means a zero matrix of appropriate dimensions.)

Proof. Straightforward induction on k. The base case ($k = 0$) says that

$$I_{n_1+n_2+\cdots+n_u} = \begin{pmatrix} I_{n_1} & 0 & \cdots & 0 \\ 0 & I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_{n_u} \end{pmatrix},$$

which should be fairly clear. The induction step is an easy application of Corollary 1.6.8. □

Finally, the "diagonal blocks stuck together" philosophy for block-diagonal matrices holds for nullities as well. To wit, the nullity of a block-diagonal matrix is the sum of the nullities of its diagonal blocks. In other words:

Proposition 1.6.11. Let $u \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$. For each $i \in [u]$, let A_i be an $n_i \times n_i$-matrix. Then,

$$\dim\left(\operatorname{Ker}\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix}\right)$$
$$= \dim(\operatorname{Ker}(A_1)) + \dim(\operatorname{Ker}(A_2)) + \cdots + \dim(\operatorname{Ker}(A_u)). \tag{8}$$

Proof. Let $\mathbb{F}$ be the field that our matrices are defined over. If $v_{\langle 1\rangle}, v_{\langle 2\rangle}, \ldots, v_{\langle u\rangle}$ are u column vectors (of whatever sizes), then $\begin{pmatrix} v_{\langle 1\rangle} \\ v_{\langle 2\rangle} \\ \vdots \\ v_{\langle u\rangle} \end{pmatrix}$ shall mean the big column vector obtained by stacking these u column vectors $v_{\langle 1\rangle}, v_{\langle 2\rangle}, \ldots, v_{\langle u\rangle}$ atop one

another. (This is the particular case of the block matrix notation from Definition 1.6.3 for $v = 1$ and $p_1 = 1$.) It is easy to see (e.g., using Proposition 1.6.6) that if $v_{\langle 1\rangle}, v_{\langle 2\rangle}, \ldots, v_{\langle u\rangle}$ are u column vectors with $v_{\langle i\rangle} \in \mathbb{F}^{n_i}$ for each $i \in [u]$, then

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} \begin{pmatrix} v_{\langle 1\rangle} \\ v_{\langle 2\rangle} \\ \vdots \\ v_{\langle u\rangle} \end{pmatrix} = \begin{pmatrix} A_1 v_{\langle 1\rangle} \\ A_2 v_{\langle 2\rangle} \\ \vdots \\ A_u v_{\langle u\rangle} \end{pmatrix}. \tag{9}$$

Let $N := n_1 + n_2 + \cdots + n_u$. Any vector $v \in \mathbb{F}^N$ can be uniquely written in block-matrix notation as $\begin{pmatrix} v_{\langle 1\rangle} \\ v_{\langle 2\rangle} \\ \vdots \\ v_{\langle u\rangle} \end{pmatrix}$, where each $v_{\langle i\rangle}$ is a vector in $\mathbb{F}^{n_i}$. (To wit, we just subdivide v into blocks of sizes $n_1, n_2, \ldots, n_u$ from top to bottom; the topmost block will be $v_{\langle 1\rangle}$, the second-topmost will be $v_{\langle 2\rangle}$, and so on. Formally speaking, for each $i \in [u]$, we set $N_i := n_1 + n_2 + \cdots + n_{i-1}$, and we let $v_{\langle i\rangle}$ be the column vector in $\mathbb{F}^{n_i}$ whose entries are the (N_i+1)-st, (N_i+2)-nd, …, (N_i+n_i)-th entries of v.)

Now, consider a vector $v \in \mathbb{F}^N$ that is written in block-matrix notation $\begin{pmatrix} v_{\langle 1\rangle} \\ v_{\langle 2\rangle} \\ \vdots \\ v_{\langle u\rangle} \end{pmatrix}$, where each $v_{\langle i\rangle}$ is a vector in $\mathbb{F}^{n_i}$. Then,

$$\begin{aligned} \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} v &= \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} \begin{pmatrix} v_{\langle 1\rangle} \\ v_{\langle 2\rangle} \\ \vdots \\ v_{\langle u\rangle} \end{pmatrix} \\ &= \begin{pmatrix} A_1 v_{\langle 1\rangle} \\ A_2 v_{\langle 2\rangle} \\ \vdots \\ A_u v_{\langle u\rangle} \end{pmatrix} \qquad \text{(by (9))}. \end{aligned}$$

Hence, $\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} v = 0$ holds if and only if $A_i v_{\langle i\rangle} = 0$ holds for each

$i \in [u]$. In other words, $v \in \operatorname{Ker} \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix}$ holds if and only if $v_{\langle i \rangle} \in \operatorname{Ker}(A_i)$ holds for each $i \in [u]$. In other words, the vectors in $\operatorname{Ker} \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix}$ are precisely the vectors of the form $\begin{pmatrix} v_{\langle 1 \rangle} \\ v_{\langle 2 \rangle} \\ \vdots \\ v_{\langle u \rangle} \end{pmatrix}$, where $v_{\langle i \rangle} \in \operatorname{Ker}(A_i)$ for each $i \in [u]$. Thus,

$$\operatorname{Ker} \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} \cong \operatorname{Ker}(A_1) \oplus \operatorname{Ker}(A_2) \oplus \cdots \oplus \operatorname{Ker}(A_u)$$

as vector spaces. By taking dimensions on both sides, this yields (8). □

1.6.4. Unitarity

Now, we claim that a block-diagonal matrix is unitary if and only if its diagonal blocks are unitary:

Proposition 1.6.12. Let $u \in \mathbb{N}$. Let $n_1, n_2, \ldots, n_u \in \mathbb{N}$. For each $i \in [u]$, let $A_i \in \mathbb{C}^{n_i \times n_i}$ be a matrix. Then, the block-diagonal matrix $\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix}$ is unitary if and only if all u matrices $A_1, A_2, \ldots, A_u$ are unitary.

Proof. Let $N = n_1 + n_2 + \cdots + n_u$. Let

$$A = \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix}. \tag{10}$$

Thus, we must prove that A is unitary if and only if all u matrices $A_1, A_2, \ldots, A_u$ are unitary.

It is easy to see that

$$A^* = \begin{pmatrix} A_1^* & 0 & \cdots & 0 \\ 0 & A_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u^* \end{pmatrix}.$$

Multiplying this equality by (10), we obtain

$$\begin{aligned} A^*A &= \begin{pmatrix} A_1^* & 0 & \cdots & 0 \\ 0 & A_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u^* \end{pmatrix} \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u \end{pmatrix} \\ &= \begin{pmatrix} A_1^*A_1 & 0 & \cdots & 0 \\ 0 & A_2^*A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u^*A_u \end{pmatrix} \qquad \text{(by Corollary 1.6.8)}. \end{aligned}$$

On the other hand, it is again easy to see that

$$I_N = \begin{pmatrix} I_{n_1} & 0 & \cdots & 0 \\ 0 & I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_{n_u} \end{pmatrix}$$

(since $N = n_1 + n_2 + \cdots + n_u$). In light of these two equalities, we see that $A^*A = I_N$ holds if and only if

$$\begin{pmatrix} A_1^*A_1 & 0 & \cdots & 0 \\ 0 & A_2^*A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_u^*A_u \end{pmatrix} = \begin{pmatrix} I_{n_1} & 0 & \cdots & 0 \\ 0 & I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_{n_u} \end{pmatrix}$$

holds, i.e., if and only if we have $A_i^*A_i = I_{n_i}$ for each $i \in [u]$. Likewise, we can see that $AA^* = I_N$ holds if and only if we have $A_iA_i^* = I_{n_i}$ for each $i \in [u]$. Hence, we

have the following chain of equivalences:

$$\begin{aligned}
&(A \text{ is unitary}) \\
&\iff (AA^* = I_N \text{ and } A^*A = I_N) \\
&\qquad\qquad (\text{by the equivalence } \mathcal{A} \iff \mathcal{C} \text{ in Theorem 1.5.3}) \\
&\iff (\text{we have } A_i A_i^* = I_{n_i} \text{ and } A_i^* A_i = I_{n_i} \text{ for each } i \in [u]) \\
&\qquad\qquad \left(\begin{array}{c} \text{since we have shown that } AA^* = I_N \text{ holds if and only if} \\ \text{we have } A_i A_i^* = I_{n_i} \text{ for each } i \in [u] \text{, and since we have} \\ \text{shown that } A^*A = I_N \text{ holds if and only if} \\ \text{we have } A_i^* A_i = I_{n_i} \text{ for each } i \in [u] \end{array} \right) \\
&\iff (\text{the matrix } A_i \text{ is unitary for each } i \in [u]) \\
&\qquad\qquad (\text{by the equivalence } \mathcal{C} \iff \mathcal{A} \text{ in Theorem 1.5.3}) \\
&\iff (\text{all } u \text{ matrices } A_1, A_2, \ldots, A_u \text{ are unitary}) .
\end{aligned}$$

But this is precisely what we need to show. Thus, Proposition 1.6.12 is proven. □

1.7. The Gram–Schmidt process

Lecture 3 starts here.

We now come to one of the most crucial algorithms in linear algebra.

Theorem 1.7.1 (Gram–Schmidt process)**.** Let $(v_1, v_2, \ldots, v_m)$ be a linearly independent tuple of vectors in $\mathbb{C}^n$.

Then, there is an orthogonal tuple $(z_1, z_2, \ldots, z_m)$ of vectors in $\mathbb{C}^n$ that satisfies

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m] .$$

Furthermore, such a tuple $(z_1, z_2, \ldots, z_m)$ can be constructed by the following recursive process:

- For each $p \in [m]$, if the first $p-1$ entries $z_1, z_2, \ldots, z_{p-1}$ of this tuple have already been constructed, then we define the p-th entry z_p by the equality

$$z_p = v_p - \sum_{k=1}^{p-1} \frac{\left\langle v_p, z_k \right\rangle}{\left\langle z_k, z_k \right\rangle} z_k. \tag{11}$$

  (Note that the sum on the right hand side of (11) is an empty sum when $p = 1$; thus, (11) simplifies to $z_1 = v_1$ in this case.)

Roughly speaking, the claim of Theorem 1.7.1 is that if we start with any linearly independent tuple $(v_1, v_2, \ldots, v_m)$ of vectors in $\mathbb{C}^n$, then we can make this tuple orthogonal by tweaking it as follows:

- leave v_1 unchanged;
- modify v_2 by subtracting some scalar multiple of v_1;
- modify v_3 by subtracting some linear combination of v_1 and v_2;
- modify v_4 by subtracting some linear combination of v_1, v_2, v_3;
- and so on.

Specifically, the equation (11) tells us (recursively) the precise multiples (and linear combinations) that we need to subtract. This recursive tweaking process is known as *Gram–Schmidt orthogonalization* or the *Gram–Schmidt process*.

Example 1.7.2. Here is how the equalities (11) in Theorem 1.7.1 look like for $p \in \{1,2,3,4\}$:

$$\begin{aligned}
z_1 &= v_1; \\
z_2 &= v_2 - \frac{\langle v_2, z_1 \rangle}{\langle z_1, z_1 \rangle} z_1; \\
z_3 &= v_3 - \frac{\langle v_3, z_1 \rangle}{\langle z_1, z_1 \rangle} z_1 - \frac{\langle v_3, z_2 \rangle}{\langle z_2, z_2 \rangle} z_2; \\
z_4 &= v_4 - \frac{\langle v_4, z_1 \rangle}{\langle z_1, z_1 \rangle} z_1 - \frac{\langle v_4, z_2 \rangle}{\langle z_2, z_2 \rangle} z_2 - \frac{\langle v_4, z_3 \rangle}{\langle z_3, z_3 \rangle} z_3.
\end{aligned}$$

Example 1.7.3. Let us try out the recursive construction of $(z_1, z_2, \ldots, z_m)$ from Theorem 1.7.1 on an example. Let $n = 4$ and $m = 3$ and

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 0 \\ -2 \\ 0 \\ -2 \end{pmatrix}, \qquad v_3 = \begin{pmatrix} 2 \\ -2 \\ 0 \\ 0 \end{pmatrix}.$$

Then, (11) becomes

$$z_1 = v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix};$$

$$z_2 = v_2 - \frac{\langle v_2, z_1 \rangle}{\langle z_1, z_1 \rangle} z_1 = \begin{pmatrix} 0 \\ -2 \\ 0 \\ -2 \end{pmatrix} - \frac{-4}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix};$$

$$\begin{aligned} z_3 &= v_3 - \frac{\langle v_3, z_1 \rangle}{\langle z_1, z_1 \rangle} z_1 - \frac{\langle v_3, z_2 \rangle}{\langle z_2, z_2 \rangle} z_2 \\ &= \begin{pmatrix} 2 \\ -2 \\ 0 \\ 0 \end{pmatrix} - \frac{0}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \frac{4}{4} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}. \end{aligned}$$

So

$$(z_1, z_2, z_3) = \left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \right)$$

is an orthogonal tuple of vectors.

According to Proposition 1.2.5, we thus obtain an orthonormal tuple

$$\left(\frac{1}{||z_1||} z_1, \ \frac{1}{||z_2||} z_2, \ \frac{1}{||z_3||} z_3 \right) = \left(\begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 1/2 \\ -1/2 \\ 1/2 \\ -1/2 \end{pmatrix}, \begin{pmatrix} 1/2 \\ -1/2 \\ -1/2 \\ 1/2 \end{pmatrix} \right).$$

(We are in luck with this example; normally we would get square roots at this step.)

For more examples of the Gram–Schmidt process, see [Bartle14, Week 3, §4]. (These examples all use vectors in $\mathbb{R}^n$ rather than $\mathbb{C}^n$, which allows for visualization and saves one the trouble of complex conjugates.)

Our proof of Theorem 1.7.1 will require a simple lemma from elementary linear algebra:

Lemma 1.7.4. Let V be a vector space over some field. Let $v_1, v_2, \ldots, v_k$ be some vectors in V. Let x and y be two further vectors in V. Assume that $x - y \in \operatorname{span}\{v_1, v_2, \ldots, v_k\}$. Then,

$$\operatorname{span}\{v_1, v_2, \ldots, v_k, x\} = \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}.$$

Proof of Lemma 1.7.4. Set

$$S := \operatorname{span}\{v_1, v_2, \ldots, v_k\};$$
$$X := \operatorname{span}\{v_1, v_2, \ldots, v_k, x\};$$
$$Y := \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}.$$

These three sets S, X and Y are vector subspaces of V (since a span is always a vector subspace). By assumption, we have $x - y \in \operatorname{span}\{v_1, v_2, \ldots, v_k\} = S$. Therefore, $-(x-y) \in S$ as well (since S is a vector subspace of V). In other words, $y - x \in S$ (since $-(x-y) = y - x$). Hence, x and y play symmetric roles in our situation.

However, $x - y \in S = \operatorname{span}\{v_1, v_2, \ldots, v_k\}$ shows that $x - y$ is a linear combination of $v_1, v_2, \ldots, v_k$. In other words,

$$x - y = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k v_k, \tag{12}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are some scalars (i.e., elements of the base field). Consider these scalars. Solving the equality (12) for x, we obtain

$$x = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k v_k + y.$$

This shows that x is a linear combination of $v_1, v_2, \ldots, v_k, y$. In other words, $x \in \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}$. In other words, $x \in Y$ (since $Y = \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}$). On the other hand, each $i \in [k]$ satisfies

$$v_i \in \{v_1, v_2, \ldots, v_k, y\} \subseteq \operatorname{span}\{v_1, v_2, \ldots, v_k, y\} = Y.$$

In other words, the k vectors $v_1, v_2, \ldots, v_k$ belong to Y. Since we also know that $x \in Y$, we thus conclude that all $k+1$ vectors $v_1, v_2, \ldots, v_k, x$ belong to Y. Since Y is a vector subspace of V, this entails that any linear combination of $v_1, v_2, \ldots, v_k, x$ must belong to Y. In other words,

$$\operatorname{span}\{v_1, v_2, \ldots, v_k, x\} \subseteq Y$$

(since $\operatorname{span}\{v_1, v_2, \ldots, v_k, x\}$ is the set of all linear combinations of $v_1, v_2, \ldots, v_k, x$). In other words, $X \subseteq Y$ (since $X = \operatorname{span}\{v_1, v_2, \ldots, v_k, x\}$).

However, as we explained, x and y play symmetric roles in our situation. Swapping x with y results in the exchange of X with Y. Thus, just as we have proved $X \subseteq Y$, we can show that $Y \subseteq X$. Combining these two inclusions, we obtain $X = Y$. In view of $X = \operatorname{span}\{v_1, v_2, \ldots, v_k, x\}$ and $Y = \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}$, this rewrites as

$$\operatorname{span}\{v_1, v_2, \ldots, v_k, x\} = \operatorname{span}\{v_1, v_2, \ldots, v_k, y\}.$$

This proves Lemma 1.7.4. □

Proof of Theorem 1.7.1. We define a tuple $(z_1, z_2, \ldots, z_m)$ recursively by (11). First, we need to show that this tuple is actually well-defined – i.e., that the denominators

$\langle z_k, z_k \rangle$ in the equality (11) never become 0 in the process (which would render (11) meaningless and therefore prevent z_p from being well-defined). Second, we need to show that the resulting tuple does indeed satisfy

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m].$$

Finally, we need to show that the resulting tuple is orthogonal.

Let us prove the first two of these three claims in lockstep, by showing the following claim:

> *Claim 1:* For each $p \in \{0, 1, \ldots, m\}$, the vectors $z_1, z_2, \ldots, z_p$ are well-defined and satisfy
>
> $$\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}.$$

[*Proof of Claim 1:* We induct on p.

Induction base: Claim 1 is obviously true for $p = 0$ (since $\operatorname{span}\{\} = \operatorname{span}\{\}$).

Induction step: Fix some $p \in [m]$, and assume that the vectors $z_1, z_2, \ldots, z_{p-1}$ are well-defined and satisfy

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\}. \tag{13}$$

We now need to show that the vectors $z_1, z_2, \ldots, z_p$ are well-defined and satisfy

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}. \tag{14}$$

The tuple $(v_1, v_2, \ldots, v_p)$ is linearly independent (since the tuple $(v_1, v_2, \ldots, v_m)$ is linearly independent). Thus, the span $\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\}$ is $(p-1)$-dimensional and we have $v_p \notin \operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\}$. Hence,

$$v_p \notin \operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\} \qquad \text{(by (13))}.$$

Now, recall that the span $\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\}$ is $(p-1)$-dimensional. In view of (13), we can rewrite this as follows: The span $\operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\}$ is $(p-1)$-dimensional. In other words, the tuple $(z_1, z_2, \ldots, z_{p-1})$ is linearly independent. Hence, for each $k \in [p-1]$, we have $z_k \neq 0$ and therefore $\langle z_k, z_k \rangle > 0$ (by Proposition 1.1.6 **(b)**), so that $\langle z_k, z_k \rangle \neq 0$. Thus, the denominators on the right hand side of (11) are nonzero, so that z_p is well-defined. Hence, the vectors $z_1, z_2, \ldots, z_p$ are well-defined (since we already know that the vectors $z_1, z_2, \ldots, z_{p-1}$ are well-defined).

It remains to prove that

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}.$$

But this is easy: From (11), we obtain

$$v_p - z_p = \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k \in \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\}$$

(since $\sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k$ is clearly a linear combination of $z_1, z_2, \ldots, z_{p-1}$). Hence, Lemma 1.7.4 (applied to $k = p-1$ and $x = v_p$ and $y = z_p$) yields[6]

$$\begin{aligned}
\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}, v_p\right\} &= \operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}, z_p\right\} \\
&= \underbrace{\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\}}_{=\operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\}} + \operatorname{span}\left\{z_p\right\} \\
&\qquad \left(\begin{array}{c} \text{since } \operatorname{span}(A \cup B) = \operatorname{span} A + \operatorname{span} B \\ \text{for any two sets } A \text{ and } B \text{ of vectors} \end{array} \right) \\
&= \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\} + \operatorname{span}\left\{z_p\right\} \\
&= \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}, z_p\right\} \\
&\qquad \left(\begin{array}{c} \text{since } \operatorname{span} A + \operatorname{span} B = \operatorname{span}(A \cup B) \\ \text{for any two sets } A \text{ and } B \text{ of vectors} \end{array} \right).
\end{aligned}$$

In other words, $\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$. Thus, the induction step is complete, so that Claim 1 is proved by induction.]

Claim 1 (applied to $p = m$) shows that the vectors $z_1, z_2, \ldots, z_m$ are well-defined. In other words, the tuple $(z_1, z_2, \ldots, z_m)$ is well-defined. Furthermore, this tuple satisfies

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} = \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m]$$

(by Claim 1, applied to $p = j$). It now remains to show that this tuple is orthogonal. We shall achieve this by showing the following claim:

Claim 2: For any $j \in \{0, 1, \ldots, m\}$, the tuple $\left(z_1, z_2, \ldots, z_j\right)$ is orthogonal.

[*Proof of Claim 2:* We proceed by induction on j:

Induction base: Claim 2 clearly holds for $j = 0$, since the (empty) 0-tuple is vacuously orthogonal.

Induction step: Let $p \in [m]$. Assume (as the induction hypothesis) that Claim 2 holds for $j = p-1$. We must show that Claim 2 holds for $j = p$.

Our induction hypothesis says that Claim 2 holds for $j = p-1$. In other words, the tuple $\left(z_1, z_2, \ldots, z_{p-1}\right)$ is orthogonal. In other words, we have

$$z_a \perp z_b \qquad \text{whenever } a, b \in [p-1] \text{ satisfy } a \neq b. \tag{15}$$

[6] We are here using the following notion: If P and Q are two vector subspaces of a vector space V, then

$$P + Q := \{p + q \mid p \in P \text{ and } q \in Q\}.$$

This is again a vector subspace of V. (It is, in fact, the smallest subspace that contains both P and Q as subsets.)

In other words, we have

$$\langle z_a, z_b \rangle = 0 \qquad \text{whenever } a, b \in [p-1] \text{ satisfy } a \neq b. \tag{16}$$

We must show that Claim 2 holds for $j = p$. In other words, we must show that the tuple $(z_1, z_2, \ldots, z_p)$ is orthogonal. In other words, we must show that

$$z_a \perp z_b \qquad \text{whenever } a, b \in [p] \text{ satisfy } a \neq b. \tag{17}$$

It will clearly suffice to prove (17) in the case when one of a and b equals p (because in all other cases, we have $a, b \in [p-1]$, and thus $z_a \perp z_b$ follows from (15)).

Thus, let $a, b \in [p]$ satisfy $a \neq b$, and assume that one of a and b equals p. We must prove that $z_a \perp z_b$. Proposition 1.2.2 shows that $z_a \perp z_b$ is equivalent to $z_b \perp z_a$. Thus, a and b play symmetric roles in our claim. Hence, in our proof of $z_a \perp z_b$, we can WLOG assume that $a \leq b$ (since otherwise, we can swap a with b). Assume this. Hence, $a < b$ (since $a \neq b$). Thus, $a < b \leq p$, so that $a \neq p$. However, we assumed that one of a and b equals p; hence, $b = p$ (since $a \neq p$). Also, we have $a \in [p-1]$ (since $a < p$).

Now, (11) yields

$$\langle z_p, z_a \rangle = \left\langle v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k, z_a \right\rangle = \langle v_p, z_a \rangle - \left\langle \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k, z_a \right\rangle$$

(by Proposition 1.1.5 **(h)**). In view of

$$\begin{aligned}
&\left\langle \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k, z_a \right\rangle \\
&= \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} \langle z_k, z_a \rangle \qquad \text{(by Proposition 1.1.5 \textbf{(i)})} \\
&= \sum_{\substack{k \in [p-1]; \\ k \neq a}} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} \underbrace{\langle z_k, z_a \rangle}_{\substack{=0 \\ \text{(by (17), applied to } k \text{ and } a \\ \text{instead of } a \text{ and } b)}} + \underbrace{\frac{\langle v_p, z_a \rangle}{\langle z_a, z_a \rangle} \langle z_a, z_a \rangle}_{= \langle v_p, z_a \rangle} \\
&\qquad \left(\begin{array}{c} \text{here, we have split off the addend for } k = a \\ \text{from the sum, since } a \in [p-1] \end{array} \right) \\
&= \underbrace{\sum_{\substack{k \in [p-1]; \\ k \neq a}} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} 0}_{=0} + \langle v_p, z_a \rangle = \langle v_p, z_a \rangle ,
\end{aligned}$$

we can rewrite this as

$$\langle z_p, z_a \rangle = \langle v_p, z_a \rangle - \langle v_p, z_a \rangle = 0.$$

In view of $b = p$, this rewrites as $\langle z_b, z_a \rangle = 0$. Thus, $z_b \perp z_a$, so that $z_a \perp z_b$ (by Proposition 1.2.2).

As explained above, this completes our proof of the fact that Claim 2 holds for $j = p$. Thus, the induction step is complete, and Claim 2 is proven.]

Now, applying Claim 2 to $j = m$, we obtain that the tuple $(z_1, z_2, \ldots, z_m)$ is orthogonal. Thus, the proof of Theorem 1.7.1 is complete. □

One might wonder how the Gram–Schmidt process could be adapted to a tuple $(v_1, v_2, \ldots, v_m)$ of vectors that is **not** linearly independent. The equality (11) requires the vectors z_k to be nonzero, since the denominators in which they appear would be 0 otherwise. In Theorem 1.7.1, this requirement is indeed satisfied (as we have shown in the proof above). However, if we do not assume $(v_1, v_2, \ldots, v_m)$ to be linearly independent, then some of the z_k can be zero, and so the construction of the following z_p will fail. There are several ways to adapt the process to this complication. We will take the most stupid-sounding one: In the cases where the equality (11) would produce a zero vector z_p, we opt to instead pick some nonzero vector orthogonal to $z_1, z_2, \ldots, z_{p-1}$ (using Lemma 1.2.7) and declare it to be z_p. This works well as long as $m \leq n$; here is the result:

Theorem 1.7.5 (Gram–Schmidt process, take 2)**.** Let $(v_1, v_2, \ldots, v_m)$ be any tuple of vectors in $\mathbb{C}^n$ with $m \leq n$.

Then, there is an orthogonal tuple $(z_1, z_2, \ldots, z_m)$ of nonzero vectors in $\mathbb{C}^n$ that satisfies

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m].$$

Furthermore, such a tuple $(z_1, z_2, \ldots, z_m)$ can be constructed by the following recursive process:

- For each $p \in [m]$, if the first $p-1$ entries $z_1, z_2, \ldots, z_{p-1}$ of this tuple have already been constructed, then we define the p-th entry z_p as follows:
  - *If* $v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k \neq 0$, then we define z_p by the equality
  $$z_p = v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k. \tag{18}$$
  - *If* $v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k = 0$, then we pick an arbitrary nonzero vector $b \in \mathbb{C}^n$ that is orthogonal to each of $z_1, z_2, \ldots, z_{p-1}$ (indeed, such a vector b exists by Lemma 1.2.7, because $p - 1 < p \leq m \leq n$), and we set
  $$z_p = b. \tag{19}$$

Proof of Theorem 1.7.5. We define a tuple $(z_1, z_2, \ldots, z_m)$ by the recursive process described in Theorem 1.7.5. It is clear that this tuple is actually well-defined (indeed, the vectors z_p are nonzero by their construction, and thus the denominators $\langle z_k, z_k \rangle$ in (18) never become 0, because Proposition 1.1.6 **(b)** shows that any nonzero vector z satisfies $\langle z, z \rangle \neq 0$). We do, however, need to show that the resulting tuple does indeed satisfy

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m],$$

and that this tuple is orthogonal.

Let us prove the first of these two claims:

> *Claim 1:* For each $p \in \{0, 1, \ldots, m\}$, we have $\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$.

[*Proof of Claim 1:* We induct on p:
Induction base: Claim 1 obviously holds for $p = 0$.
Induction step: Fix some $p \in [m]$, and assume that

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\}. \tag{20}$$

We now need to show that

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}. \tag{21}$$

We shall first show that

$$v_p \in \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}. \tag{22}$$

Indeed, we recall our definition of z_p. This definition distinguishes between two cases, depending on whether the difference $v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k$ is $\neq 0$ or $= 0$. Let us analyze these two cases separately:

- *Case 1:* We have $v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k \neq 0$. In this case, z_p is defined by the equality (18). Solving this equality for v_p, we obtain

  $$v_p = z_p + \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k \in \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}.$$

  Thus, (22) is proved in Case 1.

- *Case 2:* We have $v_p - \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k = 0$. In this case, we have

  $$v_p = \sum_{k=1}^{p-1} \frac{\langle v_p, z_k \rangle}{\langle z_k, z_k \rangle} z_k \in \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}.$$

  Hence, (22) is proved in Case 2.

We have now proved (22) in both cases. However, for each $i \in [p-1]$, we have

$$\begin{aligned} v_i &\in \left\{v_1, v_2, \ldots, v_{p-1}\right\} \subseteq \operatorname{span}\left\{v_1, v_2, \ldots, v_{p-1}\right\} \\ &\subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_{p-1}\right\} \qquad \text{(by (20))} \\ &\subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}. \end{aligned}$$

In other words, all $p-1$ vectors $v_1, v_2, \ldots, v_{p-1}$ belong to $\operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$. Since the vector v_p also belongs to $\operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$ (by (22)), we thus conclude that all p vectors $v_1, v_2, \ldots, v_p$ belong to $\operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$. Therefore, each linear combination of these p vectors $v_1, v_2, \ldots, v_p$ must also belong to $\operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$ (because $\operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$ is a vector subspace of $\mathbb{C}^n$). In other words, $\operatorname{span}\left\{v_1, v_2, \ldots, v_p\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_p\right\}$. Thus, the induction step is complete, so that Claim 1 is proved by induction.]

It now remains to show that the tuple $(z_1, z_2, \ldots, z_m)$ is orthogonal. We shall achieve this by showing the following claim:

Claim 2: For any $j \in \{0, 1, \ldots, m\}$, the tuple $\left(z_1, z_2, \ldots, z_j\right)$ is orthogonal.

[*Proof of Claim 2:* We proceed by induction on j, similarly to the proof of Claim 2 in the proof of Theorem 1.7.1. Only one minor complication emerges in the induction step:

Induction step: Let $p \in [m]$. Assume (as the induction hypothesis) that Claim 2 holds for $j = p-1$. We must show that Claim 2 holds for $j = p$.

As in the proof of Theorem 1.7.1, we can convince ourselves that it suffices to show that

$$z_a \perp z_b \qquad \text{whenever } a, b \in [p] \text{ satisfy } a \neq b. \tag{23}$$

Moreover, we only need to show this in the case when one of a and b equals p (because in all other cases, it follows from the induction hypothesis). In other words, we only need to show that the vector z_p is orthogonal to each of $z_1, z_2, \ldots, z_{p-1}$.

Recall our definition of z_p. This definition distinguishes between two cases, depending on whether the difference $v_p - \sum_{k=1}^{p-1} \frac{\left\langle v_p, z_k \right\rangle}{\left\langle z_k, z_k \right\rangle} z_k$ is $\neq 0$ or $= 0$. In the first of these two cases, the proof proceeds exactly as in the proof of Theorem 1.7.1. Let us thus WLOG assume that we are in the second case. That is, we assume that $v_p - \sum_{k=1}^{p-1} \frac{\left\langle v_p, z_k \right\rangle}{\left\langle z_k, z_k \right\rangle} z_k = 0$. Hence, z_p is defined by (19), where b is a nonzero vector in $\mathbb{C}^n$ that is orthogonal to each of $z_1, z_2, \ldots, z_{p-1}$. This shows that z_p is orthogonal to each of $z_1, z_2, \ldots, z_{p-1}$. But as we explained above, this is exactly what we need to show. Thus, Claim 2 holds for $j = p$. The induction step is complete, and Claim 2 is proved.]

Now, applying Claim 2 to $j = m$, we obtain that the tuple $(z_1, z_2, \ldots, z_m)$ is orthogonal. Thus, the proof of Theorem 1.7.5 is complete. □

Corollary 1.7.6. Let $(v_1, v_2, \ldots, v_m)$ be any tuple of vectors in $\mathbb{C}^n$ with $m \leq n$.

Then, there is an orthonormal tuple $(q_1, q_2, \ldots, q_m)$ of vectors in $\mathbb{C}^n$ that satisfies

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} \subseteq \operatorname{span}\left\{q_1, q_2, \ldots, q_j\right\} \qquad \text{for all } j \in [m].$$

Proof of Corollary 1.7.6. We have $m \leq n$. Hence, Theorem 1.7.5 shows that there is an orthogonal tuple $(z_1, z_2, \ldots, z_m)$ of nonzero vectors in $\mathbb{C}^n$ that satisfies

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} \qquad \text{for all } j \in [m].$$

Consider this tuple $(z_1, z_2, \ldots, z_m)$. Proposition 1.2.5 (applied to $(z_1, z_2, \ldots, z_m)$ instead of $(u_1, u_2, \ldots, u_k)$) then shows that the tuple

$$\left(\frac{1}{||z_1||}z_1, \ \frac{1}{||z_2||}z_2, \ \ldots, \ \frac{1}{||z_m||}z_m\right)$$

is orthonormal. Moreover, we have

$$\operatorname{span}\left\{v_1, v_2, \ldots, v_j\right\} \subseteq \operatorname{span}\left\{z_1, z_2, \ldots, z_j\right\} = \operatorname{span}\left\{\frac{1}{||z_1||}z_1, \ \frac{1}{||z_2||}z_2, \ \ldots, \ \frac{1}{||z_j||}z_j\right\}$$

for all $j \in [m]$. Hence, Corollary 1.7.6 is proven (just take $q_i = \dfrac{1}{||z_i||}z_i$). □

1.8. QR factorization

Recall that an isometry is a matrix whose columns form an orthonormal tuple. (We saw this in Proposition 1.4.2.)

Theorem 1.8.1 (QR factorization, isometry version)**.** Let $A \in \mathbb{C}^{n \times m}$ satisfy $n \geq m$. Then, there exist an isometry $Q \in \mathbb{C}^{n \times m}$ and an upper-triangular matrix $R \in \mathbb{C}^{m \times m}$ such that $A = QR$.

The pair (Q, R) in Theorem 1.8.1 is called a *QR factorization* of A. (We are using the indefinite article, since it is usually not unique.)

Example 1.8.2. Let

$$A = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 1 & -2 & 0 & 2 \\ 1 & 0 & 1 & 0 \\ 1 & -2 & 0 & 0 \end{pmatrix} \in \mathbb{C}^{4 \times 4}.$$

Then, one QR factorization of A is given by

$$A = \underbrace{\begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & -1/2 & 1/2 \end{pmatrix}}_{=Q} \underbrace{\begin{pmatrix} 2 & -2 & 1 & 0 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}}_{=R}.$$

Another is given by

$$A = \underbrace{\begin{pmatrix} 1/2 & 1/2 & \sqrt{2}/2 & 0 \\ 1/2 & -1/2 & 0 & \sqrt{2}/2 \\ 1/2 & 1/2 & -\sqrt{2}/2 & 0 \\ 1/2 & -1/2 & 0 & -\sqrt{2}/2 \end{pmatrix}}_{=Q} \underbrace{\begin{pmatrix} 2 & -2 & 1 & 2 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & \sqrt{2} \\ 0 & 0 & 0 & \sqrt{2} \end{pmatrix}}_{=R}.$$

Proof of Theorem 1.8.1. Recall that $A_{\bullet,1}, A_{\bullet,2}, \ldots, A_{\bullet,m}$ denote the m columns of the matrix A. We have $m \leq n$ (since $n \geq m$). Hence, applying Corollary 1.7.6 to $(v_1, v_2, \ldots, v_m) = (A_{\bullet,1}, A_{\bullet,2}, \ldots, A_{\bullet,m})$, we conclude that there is an orthonormal tuple $(q_1, q_2, \ldots, q_m)$ of vectors in $\mathbb{C}^n$ that satisfies

$$\operatorname{span}\left\{A_{\bullet,1}, A_{\bullet,2}, \ldots, A_{\bullet,j}\right\} \subseteq \operatorname{span}\left\{q_1, q_2, \ldots, q_j\right\} \quad \text{for all } j \in [m]. \tag{24}$$

Consider this tuple $(q_1, q_2, \ldots, q_m)$. Let $Q \in \mathbb{C}^{n \times m}$ be the matrix whose columns are $q_1, q_2, \ldots, q_m$. Then, Q is an isometry (by Proposition 1.4.2, since its columns form an orthonormal tuple). The definition of Q shows that

$$Q_{\bullet,i} = q_i \qquad \text{for each } i \in [m]. \tag{25}$$

Now, let $j \in [m]$. Then,

$$A_{\bullet,j} \in \operatorname{span}\left\{A_{\bullet,1}, A_{\bullet,2}, \ldots, A_{\bullet,j}\right\} \subseteq \operatorname{span}\left\{q_1, q_2, \ldots, q_j\right\} \qquad \text{(by (24))}.$$

In other words, there exist scalars $r_{1,j}, r_{2,j}, \ldots, r_{j,j} \in \mathbb{C}$ such that $A_{\bullet,j} = \sum_{i=1}^{j} r_{i,j} q_i$. Consider these scalars $r_{1,j}, r_{2,j}, \ldots, r_{j,j}$. Also, set

$$r_{i,j} = 0 \qquad \text{for each integer } i > j. \tag{26}$$

Thus,

$$A_{\bullet,j} = \sum_{i=1}^{j} r_{i,j} q_i = \sum_{i=1}^{m} r_{i,j} q_i \tag{27}$$

(since $\sum_{i=1}^{m} r_{i,j} q_i = \sum_{i=1}^{j} r_{i,j} q_i + \sum_{i=j+1}^{m} \underbrace{r_{i,j}}_{\substack{=0 \\ \text{(by (26))}}} q_i = \sum_{i=1}^{j} r_{i,j} q_i$).

Forget that we fixed j. Thus, for each $j \in [m]$, we have defined scalars $r_{1,j}, r_{2,j}, r_{3,j}, \ldots \in \mathbb{C}$ that satisfy (26) and (27).

Now, let $R \in \mathbb{C}^{m \times m}$ be the $m \times m$-matrix whose (i,j)-th entry is $r_{i,j}$ for each $i,j \in [m]$. This matrix R is upper-triangular, because of (26). The definition of R yields[7]

$$R_{i,j} = r_{i,j} \qquad \text{for all } i,j \in [m]. \tag{28}$$

Furthermore, for each $j \in [m]$, we have

$$\begin{aligned} A_{\bullet,j} &= \sum_{i=1}^{m} \underbrace{r_{i,j}}_{\substack{=R_{i,j} \\ \text{(by (28))}}} \underbrace{q_i}_{\substack{=Q_{\bullet,i} \\ \text{(by (25))}}} \qquad \text{(by (27))} \\ &= \sum_{i=1}^{m} R_{i,j} Q_{\bullet,i} = (QR)_{\bullet,j} \end{aligned}$$

(by the definition of the product of two matrices[8]). In other words, $A = QR$.

Thus, we have found an isometry $Q \in \mathbb{C}^{n \times m}$ and an upper-triangular matrix $R \in \mathbb{C}^{m \times m}$ such that $A = QR$. This proves Theorem 1.8.1. □

Exercise 1.8.1. [4] Let $A \in \mathbb{C}^{n \times m}$ satisfy $n \geq m$ and $\operatorname{rank} A = m$. Prove that there exists exactly one QR factorization (Q,R) of A such that the diagonal entries of R are positive reals.

Note that there are other variants of QR factorization, such as the following one:

Theorem 1.8.3 (QR factorization, unitary version)**.** Let $A \in \mathbb{C}^{n \times m}$. Then, there exist a unitary matrix $Q \in \mathbb{C}^{n \times n}$ and an upper-triangular matrix $R \in \mathbb{C}^{n \times m}$ such that $A = QR$. Here, a rectangular matrix $R \in \mathbb{C}^{n \times m}$ is said to be *upper-triangular* if and only if it satisfies

$$R_{i,j} = 0 \qquad \text{for all } i > j.$$

Exercise 1.8.2. [5] Prove Theorem 1.8.3.
[**Hint:** Reduce both cases $n > m$ and $n < m$ to the case $n = m$.]

[7] Recall that $M_{i,j}$ is our general notation for the (i,j)-th entry of a matrix M.

[8] Actually, let's be a bit more explicit here: The standard definition of the product of two matrices yields

$$(QR)_{k,j} = \sum_{i=1}^{m} \underbrace{Q_{k,i} R_{i,j}}_{=R_{i,j} Q_{k,i}} = \sum_{i=1}^{m} R_{i,j} Q_{k,i} \qquad \text{for each } k \in [n].$$

In other words, $(QR)_{\bullet,j} = \sum_{i=1}^{m} R_{i,j} Q_{\bullet,i}$, which is precisely what we are claiming.

2. Schur triangularization ([HorJoh13, Chapter 2])

In this chapter, we will meet *Schur triangularization*: a way to transform an arbitrary $n \times n$-matrix with complex entries into an upper-triangular matrix by conjugating it (i.e., multiplying it by an invertible matrix W on the left and simultaneously by its inverse W^{-1} on the right). This is both of theoretical and of practical significance, but we will focus on the theoretical applications.

Before we get to Schur triangularization, we will have to set some groundwork.

2.0. Reminders on the characteristic polynomial and eigenvalues

First, let us recall some properties of the characteristic polynomial of an $n \times n$-matrix A, starting with its definition:

Definition 2.0.1. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix over $\mathbb{F}$.

The *characteristic polynomial* of A is defined to be the polynomial

$$\det\left(tI_n - A\right) \qquad \text{in the indeterminate } t \text{ with coefficients in } \mathbb{F}.$$

(Note that $tI_n - A$ is an $n \times n$-matrix whose entries are polynomials in t. Thus, its determinant $\det\left(tI_n - A\right)$ is itself a polynomial in t.)

The characteristic polynomial of A is denoted by p_A.

Example 2.0.2. Let $n = 2$ and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then,

$$\begin{aligned} tI_n - A = tI_2 - A &= t\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &= \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix} - \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} t-a & -b \\ -c & t-d \end{pmatrix}, \end{aligned}$$

so that

$$\begin{aligned} p_A &= \det\left(tI_n - A\right) = \det\begin{pmatrix} t-a & -b \\ -c & t-d \end{pmatrix} = (t-a)(t-d) - (-b)(-c) \\ &= t^2 - (a+d)\,t + (ad - bc). \end{aligned}$$

Example 2.0.3. Let $n = 3$ and $A = \begin{pmatrix} a & b & c \\ a' & b' & c' \\ a'' & b'' & c'' \end{pmatrix}$. Then,

$$tI_n - A = tI_3 - A = t\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} a & b & c \\ a' & b' & c' \\ a'' & b'' & c'' \end{pmatrix}$$
$$= \begin{pmatrix} t-a & -b & -c \\ -a' & t-b' & -c' \\ -a'' & -b'' & t-c'' \end{pmatrix},$$

so that

$$p_A = \det(tI_n - A) = \det\begin{pmatrix} t-a & -b & -c \\ -a' & t-b' & -c' \\ -a'' & -b'' & t-c'' \end{pmatrix}$$
$$= t^3 - \left(a + b' + c''\right)t^2 + \left(ab' - ba' + ac'' - ca'' + b'c'' - b''c'\right)t$$
$$- \left(ab'c'' - ab''c' - ba'c'' + ba''c' + ca'b'' - ca''b'\right).$$

Example 2.0.4. If $n = 1$ and $A = \begin{pmatrix} a \end{pmatrix}$, then $tI_n - A = \begin{pmatrix} t-a \end{pmatrix}$ and thus $p_A = t - a$.

Example 2.0.5. If $n = 0$ and $A = ()$, then $tI_n - A = ()$ and thus $p_A = 1$ (since the determinant of the 0×0-matrix $()$ is defined to be 1).

Remark 2.0.6. **(a)** Some authors define the characteristic polynomial p_A of an $n \times n$-matrix A to be $\det(A - tI_n)$ instead of $\det(tI_n - A)$. This differs from our definition only by a factor of $(-1)^n$, which is immaterial for most properties of the characteristic polynomial but still can cause the occasional confusion.

(b) Some other common notations for p_A are χ_A and c_A.

The patterns you might have spotted in Example 2.0.2 and in Example 2.0.3 are not accidental. Indeed, the coefficients of the characteristic polynomial of any square matrix can be expressed explicitly, if you consider sums of determinants to be explicit:

Proposition 2.0.7. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix over $\mathbb{F}$.

(a) The characteristic polynomial p_A is a monic polynomial in t of degree n. (That is, its leading term is t^n.)

(b) The constant term of the polynomial p_A is $(-1)^n \det A$.

(c) The t^{n-1}-coefficient of the polynomial p_A is $-\operatorname{Tr} A$. (Recall that $\operatorname{Tr} A$ is defined to be the sum of all diagonal entries of A; this sum is known as the *trace* of A.)

(d) More generally: Let $k \in \{0, 1, \ldots, n\}$. Then, the t^{n-k}-coefficient of the polynomial p_A is $(-1)^k$ times the sum of all principal $k \times k$-minors of A. (Recall that a $k \times k$-*minor* of A means the determinant of a $k \times k$-submatrix of A. This $k \times k$-minor is said to be *principal* if the $k \times k$-submatrix is obtained by removing some $n-k$ rows and the corresponding $n-k$ columns from A. For example, the principal 2×2-minors of a 3×3-matrix A are $\det \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$, $\det \begin{pmatrix} A_{1,1} & A_{1,3} \\ A_{3,1} & A_{3,3} \end{pmatrix}$ and $\det \begin{pmatrix} A_{2,2} & A_{2,3} \\ A_{3,2} & A_{3,3} \end{pmatrix}$.) In other words, the t^{n-k}-coefficient of p_A is

$$(-1)^k \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \det\left(\operatorname{sub}^{i_1, i_2, \ldots, i_k}_{i_1, i_2, \ldots, i_k} A\right),$$

where $\operatorname{sub}^{i_1, i_2, \ldots, i_k}_{i_1, i_2, \ldots, i_k} A$ denotes the $k \times k$-matrix whose (u, v)-th entry is A_{i_u, i_v} for all $u, v \in [k]$.

Proof sketch. We have $A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{pmatrix}$, so that

$$tI_n - A = t \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{pmatrix}$$

$$= \begin{pmatrix} t - A_{1,1} & -A_{1,2} & \cdots & -A_{1,n} \\ -A_{2,1} & t - A_{2,2} & \cdots & -A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -A_{n,1} & -A_{n,2} & \cdots & t - A_{n,n} \end{pmatrix}.$$

The determinant $\det(tI_n - A)$ of this matrix is a sum of certain products of its entries[9]. One of these products is

$$(t - A_{1,1})(t - A_{2,2}) \cdots (t - A_{n,n})$$
$$= t^n - (A_{1,1} + A_{2,2} + \cdots + A_{n,n}) t^{n-1} + (\text{terms with lower powers of } t).$$

[9]Here, we are using the *Leibniz formula* for the determinant of a matrix, which says that $\det B = \sum_{\sigma \in S_n} (-1)^\sigma B_{1,\sigma(1)} B_{2,\sigma(2)} \cdots B_{n,\sigma(n)}$ for each $n \times n$-matrix B. (We are applying this to $B = tI_n - A$.)

None of the other products appearing in this sum includes any power of t higher than t^{n-2} (because the product picks out at least two entries of A that lie outside of the main diagonal, and thus contain no t whatsoever; the remaining factors of the product contribute at most t^{n-2}). Hence, the entire determinant $\det(tI_n - A)$ can be written as

$$\det(tI_n - A) = t^n - (A_{1,1} + A_{2,2} + \cdots + A_{n,n})\, t^{n-1} + (\text{terms with lower powers of } t).$$

In other words,

$$p_A = t^n - (A_{1,1} + A_{2,2} + \cdots + A_{n,n})\, t^{n-1} + (\text{terms with lower powers of } t)$$

(since $p_A = \det(tI_n - A)$). This yields parts **(a)** and **(c)** of Proposition 2.0.7.

To prove Proposition 2.0.7 **(b)**, we substitute 0 for t in the polynomial identity $p_A = \det(tI_n - A)$. We obtain

$$p_A(0) = \det(0I_n - A) = \det(-A) = (-1)^n \det A.$$

Since $p_A(0)$ is the constant term of p_A (in fact, if f is any polynomial, then $f(0)$ is the constant term of f), we thus conclude that the constant term of p_A is $(-1)^n \det A$. This proves Proposition 2.0.7 **(b).**

Finally, Proposition 2.0.7 **(d)** can be established through a more accurate combinatorial analysis of the products that sum up to $\det(tI_n - A)$. See [Grinbe21, Proposition 6.4.29] for the details. (A combinatorially prepared reader might glean the idea from Example 2.0.3.)

We note that parts **(a)**, **(b)** and **(c)** of Proposition 2.0.7 can all be derived from part **(d)** as well. □

Next, we recall some basic notions around the eigenvalues of a matrix:

Definition 2.0.8. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ be an $n \times n$-matrix, and let $\lambda \in \mathbb{F}$.

(a) We say that λ is an *eigenvalue* of A if and only if $\det(\lambda I_n - A) = 0$. In other words, λ is an *eigenvalue* of A if and only if λ is a root of the characteristic polynomial $p_A = \det(tI_n - A)$.

(b) The *λ-eigenspace* of A is defined to be the set of all vectors $v \in \mathbb{F}^n$ satisfying $Av = \lambda v$. In other words, it is the kernel $\operatorname{Ker}(\lambda I_n - A) = \operatorname{Ker}(A - \lambda I_n)$. Thus, it is a vector subspace of $\mathbb{F}^n$. The elements of this λ-eigenspace are called the *λ-eigenvectors* of A (or the *eigenvectors* of A for eigenvalue λ). (Some authors exclude the zero vector 0 from being an eigenvector; we allow it. Thus, 0 is a λ-eigenvector for any λ, even if λ is not an eigenvalue.)

(c) The *algebraic multiplicity* of λ as an eigenvalue of A is defined to be the multiplicity of λ as a root of p_A. (If λ is not an eigenvalue of A, then this is 0.)

(d) The *geometric multiplicity* of λ as an eigenvalue of A is defined to be $\dim(\operatorname{Ker}(A - \lambda I_n))$. In other words, it is the dimension of the λ-eigenspace of A. In other words, it is the maximum number of linearly independent λ-eigenvectors. (If λ is not an eigenvalue of A, then this is 0.)

It can be shown that if $A \in \mathbb{F}^{n \times n}$ is a matrix and $\lambda \in \mathbb{F}$ is arbitrary, then the geometric multiplicity of λ as an eigenvalue of A is always $\leq$ to the algebraic multiplicity of λ as an eigenvalue of A. The two multiplicities can be equal, but don't have to be.

Example 2.0.9. Let $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{C}^{3 \times 3}$. Then, the only eigenvalue of A is 1. Its algebraic multiplicity is 3, while its geometric multiplicity is 2.

Theorem 2.0.10. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix with complex entries. Then:

(a) Its characteristic polynomial p_A factors into n linear terms:

$$p_A = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_n), \tag{29}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}$ are its roots (with their algebraic multiplicities).

(b) These roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A, appearing with their algebraic multiplicities.

(c) The sum of the algebraic multiplicities of all eigenvalues of A is n.

(d) The sum of all eigenvalues of A (with their algebraic multiplicities) is $\operatorname{Tr} A$ (that is, the trace of A).

(e) The product of all eigenvalues of A (with their algebraic multiplicities) is $\det A$.

(f) If $n > 0$, then the matrix A has at least one eigenvalue and at least one nonzero eigenvector.

Proof. The polynomial p_A is a monic polynomial of degree n (by Proposition 2.0.7 **(a)**), and therefore factors into linear terms (by the Fundamental Theorem of Algebra). This proves Theorem 2.0.10 **(a)**.

(b) This follows from the definition of eigenvalues and algebraic multiplicities.

(c) This follows from part **(b)**.

(d) Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the roots of p_A (with their multiplicities). Then, these roots are the eigenvalues of A, appearing with their algebraic multiplicities (by Theorem 2.0.10 **(b)**). Hence, their sum $\lambda_1 + \lambda_2 + \cdots + \lambda_n$ is the sum of the eigenvalues of A (with their algebraic multiplicities). On the other hand, we know from Theorem 2.0.10 **(a)** that the equality (29) holds. Comparing the coefficients of t^{n-1} on both sides of this equality, we obtain

$$\begin{aligned} \Big(\text{the coefficient of } t^{n-1} \text{ in } p_A\Big) &= \Big(\text{the coefficient of } t^{n-1} \text{ in } (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_n)\Big) \\ &= -(\lambda_1 + \lambda_2 + \cdots + \lambda_n). \end{aligned}$$

However, Proposition 2.0.7 **(c)** yields

$$\left(\text{the coefficient of } t^{n-1} \text{ in } p_A\right) = -\operatorname{Tr} A.$$

Comparing these two equalities, we obtain $-(\lambda_1 + \lambda_2 + \cdots + \lambda_n) = -\operatorname{Tr} A$. In other words, $\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{Tr} A$. This proves Theorem 2.0.10 **(d)** (since $\lambda_1 + \lambda_2 + \cdots + \lambda_n$ is the sum of the eigenvalues of A).

(e) This is similar to part **(d)**, except that we have to compare the coefficients of t^0 (instead of t^{n-1}) on both sides of (29), and we have to use Proposition 2.0.7 **(b)** (instead of Proposition 2.0.7 **(c)**).

(f) Assume that $n > 0$. Thus, $n \geq 1$. However, Theorem 2.0.10 **(b)** shows that A has exactly n eigenvalues, counted with algebraic multiplicities. Hence, A has at least one eigenvalue λ (since $n \geq 1$). Consider this λ. Since λ is an eigenvalue of A, we have $\det(\lambda I_n - A) = 0$. Hence, the $n \times n$-matrix $\lambda I_n - A$ is singular, so that its kernel $\operatorname{Ker}(\lambda I_n - A)$ is nonzero. In other words, there exists a nonzero vector $v \in \operatorname{Ker}(\lambda I_n - A)$. This vector v must be a λ-eigenvector of A (since $v \in \operatorname{Ker}(\lambda I_n - A)$). Hence, the matrix A has a nonzero eigenvector (namely, v). This completes the proof of Theorem 2.0.10 **(f)**. □

Exercise 2.0.1. [1] Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be any $n \times n$-matrix.

(a) Prove that $p_{A^T} = p_A$, where A^T denotes the transpose of the matrix A.

(b) Assume that $\mathbb{F} = \mathbb{C}$. Prove that $p_{A^*} = \overline{p_A}$, where $\overline{p_A}$ denotes the result of replacing all coefficients of the polynomial p_A by their complex conjugates.

For occasional future use, let us state some properties of traces as exercises:

Exercise 2.0.2. [1] Let $\mathbb{F}$ be a field. Let $n, m \in \mathbb{N}$. Let $A \in \mathbb{F}^{n \times m}$ and $B \in \mathbb{F}^{m \times n}$ be two matrices. Show that

$$\operatorname{Tr}(AB) = \operatorname{Tr}(BA).$$

Exercise 2.0.3. [1] Let $n, m \in \mathbb{N}$. Let $A \in \mathbb{C}^{n \times m}$ be any matrix.

(a) Show that

$$\operatorname{Tr}(A^* A) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left|A_{i,j}\right|^2.$$

(b) Show that $\operatorname{Tr}(A^* A) = 0$ if and only if $A = 0$.

2.1. Similarity of matrices

Next, let us recall the notion of similar matrices:

Definition 2.1.1. Let $\mathbb{F}$ be a field. Let A and B be two matrices in $\mathbb{F}^{n\times n}$. We say that A is *similar* to B if there exists an invertible matrix $W \in \mathbb{F}^{n\times n}$ such that $B = WAW^{-1}$.

We write "$A \sim B$" for "A is similar to B".

Example 2.1.2. The matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is similar to the matrix $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, since $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} = W \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} W^{-1}$ for the invertible matrix $W = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$. In other words, we have $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$.

The relation $\sim$ is easily seen to be an equivalence relation:[10]

Proposition 2.1.3. Let $\mathbb{F}$ be a field. Then:

(a) Any matrix $A \in \mathbb{F}^{n\times n}$ is similar to itself.

(b) If A and B are two matrices in $\mathbb{F}^{n\times n}$ such that A is similar to B, then B is similar to A.

(c) If A, B and C are three matrices in $\mathbb{F}^{n\times n}$ such that A is similar to B and such that B is similar to C, then A is similar to C.

Proof. **(a)** This follows from $A = I_n A I_n^{-1}$.

(b) Let A and B be two matrices in $\mathbb{F}^{n\times n}$ such that A is similar to B. Thus, there exists an invertible matrix $W \in \mathbb{F}^{n\times n}$ such that $B = WAW^{-1}$. Consider this W. From $B = WAW^{-1}$, we obtain $BW = WA$, so that $W^{-1}BW = A$. Thus, $A = W^{-1}B \underbrace{W}_{=\left(W^{-1}\right)^{-1}} = W^{-1}B\left(W^{-1}\right)^{-1}$. Since W^{-1} is invertible, this shows that B is similar to A. This proves Proposition 2.1.3 **(b)**.

(c) Let A, B and C be three matrices in $\mathbb{F}^{n\times n}$ such that A is similar to B and such that B is similar to C. Thus, there exists an invertible matrix $U \in \mathbb{F}^{n\times n}$ such that $B = UAU^{-1}$ (since A is similar to B), and there exists an invertible matrix $V \in \mathbb{F}^{n\times n}$ such that $C = VBV^{-1}$ (since B is similar to C). Consider these U and V.

The matrices V and U are invertible. Thus, so is their product VU, and its inverse is $(VU)^{-1} = U^{-1}V^{-1}$. (This is the famous "socks-and-shoes rule" for inverting products or compositions.) Now,

$$C = V \underbrace{B}_{=UAU^{-1}} V^{-1} = VUA \underbrace{U^{-1}V^{-1}}_{=(VU)^{-1}} = VUA\,(VU)^{-1}.$$

[10] Algebraists will recognize the relation $\sim$ (for matrices in $\mathbb{F}^{n\times n}$) as just being the conjugacy relation in the ring $\mathbb{F}^{n\times n}$ of all $n \times n$-matrices. (The meaning of the word "conjugacy" here has nothing to do with conjugates of complex numbers or with the conjugate transpose!)

In other words, $C = WAW^{-1}$ for the invertible matrix $W = VU$ (since we know that VU is invertible). This shows that A is similar to C. This proves Proposition 2.1.3 **(c)**. $\square$

Since the relation $\sim$ is symmetric (by Proposition 2.1.3 **(b)**), we can make the following definition:

Definition 2.1.4. Let $\mathbb{F}$ be a field. Let A and B be two matrices in $\mathbb{F}^{n\times n}$. We say that A and B are *similar* if A is similar to B (or, equivalently, B is similar to A).

Similar matrices have a lot in common. Here is a selection of invariants:

Proposition 2.1.5. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ and $B \in \mathbb{F}^{n\times n}$ be two similar matrices. Then:

(a) The matrices A and B have the same rank.

(b) The matrices A and B have the same nullity.

(c) The matrices A and B have the same determinant.

(d) The matrices A and B have the same characteristic polynomial.

(e) The matrices A and B have the same eigenvalues, with the same algebraic multiplicities and with the same geometric multiplicities.

(f) For any $k \in \mathbb{N}$, the matrix A^k is similar to B^k.

(g) For any $\lambda \in \mathbb{F}$, the matrix $\lambda I_n - A$ is similar to $\lambda I_n - B$.

(h) For any $\lambda \in \mathbb{F}$, the matrix $A - \lambda I_n$ is similar to $B - \lambda I_n$.

Proof. Since A is similar to B, there exists an invertible matrix $W \in \mathbb{F}^{n\times n}$ such that $B = WAW^{-1}$. Consider this W.

(b) Consider the kernels[11] $\operatorname{Ker} A$ and $\operatorname{Ker} B$ of A and B. For any $v \in \operatorname{Ker} A$, we have $Wv \in \operatorname{Ker} B$ (because $v \in \operatorname{Ker} A$ implies $Av = 0$, so that $\underbrace{B}_{=WAW^{-1}} Wv = WA\underbrace{W^{-1}W}_{=I_n}v = W\underbrace{Av}_{=0} = 0$ and therefore $Wv \in \operatorname{Ker} B$). Thus, we have found a linear map

$$\begin{aligned} \operatorname{Ker} A &\to \operatorname{Ker} B, \\ v &\mapsto Wv. \end{aligned}$$

This linear map is furthermore injective (because W is invertible, so that $Wu = Wv$ entails $u = v$). Hence, we obtain $\dim(\operatorname{Ker} A) \leq \dim(\operatorname{Ker} B)$. But A and B play symmetric roles in our situation (since the relation "similar" is symmetric), so that we can use the same reasoning to obtain $\dim(\operatorname{Ker} B) \leq \dim(\operatorname{Ker} A)$. Combining

[11] Recall that "kernel" is a synonym for "nullspace".

these two inequalities, we obtain $\dim(\operatorname{Ker} A) = \dim(\operatorname{Ker} B)$. In other words, A and B have the same nullity. This proves Proposition 2.1.5 **(b)**.

(a) The rank of an $n \times n$-matrix equals n minus its nullity (by the rank-nullity theorem). Hence, two $n \times n$-matrices that have the same nullity must also have the same rank. Thus, Proposition 2.1.5 **(a)** follows from Proposition 2.1.5 **(b)**.

(c) From $B = WAW^{-1}$, we obtain

$$\begin{aligned}\det B &= \det\left(WAW^{-1}\right) = \det W \cdot \det A \cdot \underbrace{\det\left(W^{-1}\right)}_{=(\det W)^{-1}} \\ &= \det W \cdot \det A \cdot (\det W)^{-1} = \det A.\end{aligned}$$

This proves Proposition 2.1.5 **(c)**.

(d) The characteristic polynomial of an $n \times n$-matrix M is defined to be $\det(tI_n - M)$ (where t is the indeterminate)[12]. Thus, we must show that $\det(tI_n - A) = \det(tI_n - B)$. However, we have

$$\begin{aligned}t \underbrace{I_n}_{\substack{=WW^{-1} \\ =WI_nW^{-1}}} - \underbrace{B}_{=WAW^{-1}} &= \underbrace{tWI_n}_{=W(tI_n)} W^{-1} - WAW^{-1} = W(tI_n)W^{-1} - WAW^{-1} \\ &= W(tI_n - A)W^{-1}.\end{aligned}$$

Thus,

$$\begin{aligned}\det(tI_n - B) &= \det\left(W(tI_n - A)W^{-1}\right) = \det W \cdot \det(tI_n - A) \cdot \underbrace{\det\left(W^{-1}\right)}_{=(\det W)^{-1}} \\ &= \det W \cdot \det(tI_n - A) \cdot (\det W)^{-1} = \det(tI_n - A).\end{aligned}$$

Thus, $\det(tI_n - A) = \det(tI_n - B)$, and Proposition 2.1.5 **(d)** is proven.

(e) The eigenvalues of a matrix, with their algebraic multiplicities, are the roots of the characteristic polynomial. Thus, from Proposition 2.1.5 **(d)**, we see that the matrices A and B have the same eigenvalues, with the same algebraic multiplicities. It remains to show that the geometric multiplicities are also the same.

Let λ be an eigenvalue of A (and therefore also of B, as we have just seen). The geometric multiplicity of λ as an eigenvalue of A is $\dim(\operatorname{Ker}(A - \lambda I_n))$. Likewise, the geometric multiplicity of λ as an eigenvalue of B is $\dim(\operatorname{Ker}(B - \lambda I_n))$. Hence, we must show that $\dim(\operatorname{Ker}(A - \lambda I_n)) = \dim(\operatorname{Ker}(B - \lambda I_n))$.

[12] At least this is our definition. As we already mentioned in Remark 2.0.6 **(a)**, another popular definition is $\det(M - tI_n)$. However, the two definitions differ only in a factor of $(-1)^n$, so they behave almost completely the same (and our argument works equally well for either of them).

We have

$$\underbrace{B}_{=WAW^{-1}} - \lambda \underbrace{I_n}_{\substack{=WW^{-1} \\ =WI_nW^{-1}}} = WAW^{-1} - \underbrace{\lambda W I_n}_{=W(\lambda I_n)} W^{-1} = WAW^{-1} - W(\lambda I_n) W^{-1}$$
$$= W(A - \lambda I_n) W^{-1}.$$

This shows that the matrices $A - \lambda I_n$ and $B - \lambda I_n$ are similar. Hence, Proposition 2.1.5 **(b)** shows that these two matrices $A - \lambda I_n$ and $B - \lambda I_n$ have the same nullity. In other words, $\dim(\operatorname{Ker}(A - \lambda I_n)) = \dim(\operatorname{Ker}(B - \lambda I_n))$. This is exactly what we needed to show; thus, Proposition 2.1.5 **(e)** is proven.

(f) Let $k \in \mathbb{N}$. We claim that

$$B^k = WA^kW^{-1}. \tag{30}$$

Once this is proved, it will clearly follow that A^k is similar to B^k.

One way to prove $B^k = WA^kW^{-1}$ is as follows: From $B = WAW^{-1}$, we obtain

$$B^k = \left(WAW^{-1}\right)^k = WA\underbrace{W^{-1} \cdot W}_{=I_n} A \underbrace{W^{-1} \cdot W}_{=I_n} AW^{-1} \cdot \cdots \cdot WA \underbrace{W^{-1} \cdot W}_{=I_n} AW^{-1}$$
$$= W \underbrace{AA \cdots A}_{k \text{ factors}} W^{-1} \qquad \left(\text{since all the } W^{-1} \cdot W\text{'s in the middle cancel out}\right)$$
$$= WA^kW^{-1}.$$

(To be precise, this works for $k \geq 1$; but the case $k = 0$ is trivial.)

A less handwavy proof of (30) would proceed by induction on k. As it is completely straightforward, I leave it to the reader.

(g) Let $\lambda \in \mathbb{F}$. Then,

$$\lambda \underbrace{I_n}_{\substack{=WW^{-1} \\ =WI_nW^{-1}}} - \underbrace{B}_{=WAW^{-1}} = \underbrace{\lambda W I_n}_{=W(\lambda I_n)} W^{-1} - WAW^{-1} = W(\lambda I_n) W^{-1} - WAW^{-1}$$
$$= W(\lambda I_n - A) W^{-1}.$$

This shows that the matrices $\lambda I_n - A$ and $\lambda I_n - B$ are similar. Thus, Proposition 2.1.5 **(g)** is proven.

(h) This differs from part **(g)** only in that the subtrahend and the minuend trade places. The proof is entirely analogous to part **(g)**. □

Note that neither part **(a)**, nor part **(b)**, nor part **(c)**, nor part **(d)**, nor part **(e)** of Proposition 2.1.5 is an "if and only if" statement: One can find two $n \times n$-matrices (for sufficiently large n) that have the same rank, nullity, determinant,

characteristic polynomial and eigenvalues but are not similar.[13] Thus, proving the similarity of two matrices is not as easy as comparing these data. We will later learn an algorithmic way to check whether two matrices are similar.

Exercise 2.1.1. [2] Prove that the two matrices $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ are not similar.

Exercise 2.1.2. [3] Let $A \in \mathbb{C}^{n \times n}$ be a matrix that is similar to some unitary matrix. Prove that A^{-1} is similar to A^*.

Remark 2.1.6. If you are used to thinking of matrices as linear maps, then similarity is a rather natural concept: Two $n \times n$-matrices $A \in \mathbb{F}^{n \times n}$ and $B \in \mathbb{F}^{n \times n}$ are similar if and only if they represent one and the same endomorphism $f : \mathbb{F}^n \to \mathbb{F}^n$ of $\mathbb{F}^n$ with respect to two (possibly different) bases of $\mathbb{F}^n$. To be more precise, A has to represent f with respect to some basis of $\mathbb{F}^n$, while B has to represent f with respect to a further basis of $\mathbb{F}^n$ (possibly the same, but usually not).

This fact is not hard to prove. Indeed, if A and B represent the same endomorphism f with respect to two bases of $\mathbb{F}^n$, then we have $B = WAW^{-1}$, where W is the change-of-basis matrix between these two bases. Conversely, if A and B are similar, then there exists some invertible matrix W satisfying $B = WAW^{-1}$, and then A and B represent the same endomorphism f with respect to two bases of $\mathbb{F}^n$ (namely, B represents the endomorphism

$$\begin{aligned} \mathbb{F}^n &\to \mathbb{F}^n, \\ v &\mapsto Bv \end{aligned}$$

with respect to the standard basis $(e_1, e_2, \ldots, e_n)$, whereas A represents the same endomorphism with respect to the basis $(We_1, We_2, \ldots, We_n)$).

Knowing this fact, many properties of similar matrices – including all parts of Proposition 2.1.5 – become essentially trivial: One just needs to recall that things like rank, nullity, determinant, eigenvalues etc. are properties of the endomorphism rather than properties of the matrix.

Two diagonal matrices are similar whenever they have the same diagonal entries up to order. In other words:

[13] Some of these examples are easy to find: For example, the matrices $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ have the same eigenvalues with the same algebraic multiplicities, but are not similar.

Proposition 2.1.7. Let $\mathbb{F}$ be a field. Let $n \in \mathbb{N}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{F}$. Let σ be a permutation of $[n]$ (that is, a bijective map from $[n]$ to $[n]$). Then,

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \sim \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right).$$

Example 2.1.8. For $n = 3$, Proposition 2.1.7 claims that $\operatorname{diag}(\lambda_1, \lambda_2, \lambda_3) \sim \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \lambda_{\sigma(3)}\right)$. For example, if σ is the permutation of $[3]$ that sends $1, 2, 3$ to $2, 3, 1$, respectively, then this is saying that $\operatorname{diag}(\lambda_1, \lambda_2, \lambda_3) \sim \operatorname{diag}(\lambda_2, \lambda_3, \lambda_1)$. In other words,

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \sim \begin{pmatrix} \lambda_3 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{pmatrix}.$$

Proof of Proposition 2.1.7. Let $P_\sigma \in \mathbb{F}^{n \times n}$ be the permutation matrix of σ (defined as in Example 1.5.2 **(d)**, but using the field $\mathbb{F}$ instead of $\mathbb{C}$). We recall that the (i,j)-th entry of this matrix P_σ is $\begin{cases} 1, & \text{if } i = \sigma(j); \\ 0, & \text{if } i \neq \sigma(j) \end{cases}$ for any $i, j \in [n]$.

Now, it is easy to see that

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \cdot P_\sigma = P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right). \tag{31}$$

[*Proof of (31):* It is straightforward to see that for any $i, j \in [n]$, both matrices $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \cdot P_\sigma$ and $P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$ have the same (i,j)-th entry, namely $\begin{cases} \lambda_i, & \text{if } i = \sigma(j); \\ 0, & \text{if } i \neq \sigma(j). \end{cases}$ Thus, these two matrices are equal. This proves (31).]

Since the permutation matrix P_σ is invertible, we can multiply both sides of (31) by P_σ^{-1} from the right, and thus we obtain

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right) \cdot P_\sigma^{-1}.$$

This shows that $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \sim \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$. Thus, Proposition 2.1.7 is proven. □

Proposition 2.1.7 actually has a converse: If two diagonal matrices are similar, then they have the same diagonal entries up to order. This follows easily from Proposition 2.1.5 **(e)**, because the diagonal entries of a diagonal matrix are its eigenvalues (with their algebraic multiplicities).

An analogue of Proposition 2.1.7 holds for block-diagonal matrices:

Proposition 2.1.9. Let $\mathbb{F}$ be a field. Let $n \in \mathbb{N}$. For each $i \in [n]$, let A_i be an $n_i \times n_i$-matrix (for some $n_i \in \mathbb{N}$). Let σ be a permutation of $[n]$ (that is, a bijective map from $[n]$ to $[n]$). Then,

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{pmatrix} \sim \begin{pmatrix} A_{\sigma(1)} & 0 & \cdots & 0 \\ 0 & A_{\sigma(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\sigma(n)} \end{pmatrix}.$$

Proof. This can be proved similarly to how we proved Proposition 2.1.7, except that now, instead of the permutation matrix P_σ, we need to use a "block permutation matrix" $\mathbf{P}_\sigma$. This matrix $\mathbf{P}_\sigma$ is defined to be the matrix that is written as

$$\begin{pmatrix} P(1,1) & P(1,2) & \cdots & P(1,n) \\ P(2,1) & P(2,2) & \cdots & P(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ P(n,1) & P(n,2) & \cdots & P(n,n) \end{pmatrix}$$

in block-matrix notation, where the (i,j)-th block $P(i,j)$ is defined by[14]

$$P(i,j) := \begin{cases} I_{n_i}, & \text{if } i = \sigma(j); \\ 0_{n_i \times n_{\sigma(j)}}, & \text{if } i \neq \sigma(j). \end{cases}$$

For example, if $n = 2$ and if σ is the permutation of $[2]$ that swaps 1 with 2, and if $n_1 = 1$ and $n_2 = 2$, then $\mathbf{P}_\sigma = \begin{pmatrix} 0 & I_1 \\ I_2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$. The formula analogous to (31) is

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{pmatrix} \cdot \mathbf{P}_\sigma = \mathbf{P}_\sigma \cdot \begin{pmatrix} A_{\sigma(1)} & 0 & \cdots & 0 \\ 0 & A_{\sigma(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\sigma(n)} \end{pmatrix}$$

this time; its proof is easy with the help of Proposition 1.6.6. □

Similarity of block-diagonal matrices can also come from similarity of the respective blocks:

[14] We let $0_{u \times v}$ denote the zero matrix of size $u \times v$.

Proposition 2.1.10. Let $\mathbb{F}$ be a field. Let $n \in \mathbb{N}$. For each $i \in [n]$, let A_i and B_i be two $n_i \times n_i$-matrices (for some $n_i \in \mathbb{N}$) satisfying $A_i \sim B_i$. Then,

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{pmatrix} \sim \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_n \end{pmatrix}.$$

Exercise 2.1.3. [2] Prove Proposition 2.1.10.

Exercise 2.1.4. [1] Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ and $B \in \mathbb{F}^{n\times n}$ be two matrices such that $A \sim B$.

(a) Prove that $A^T \sim B^T$. (Recall that C^T denotes the transpose of a matrix C.)

(b) Assume that $\mathbb{F} = \mathbb{C}$. Prove that $A^* \sim B^*$.

2.2. Unitary similarity

Unitary similarity is a more restrictive form of similarity, even though it is not immediately obvious from its definition:

Definition 2.2.1. Let A and B be two matrices in $\mathbb{C}^{n\times n}$. We say that A is *unitarily similar* to B if there exists a unitary matrix $W \in \mathrm{U}_n(\mathbb{C})$ such that $B = WAW^*$.

We write "$A \overset{\text{us}}{\sim} B$" for "$A$ is unitarily similar to B".

Example 2.2.2. The matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is unitarily similar to the matrix $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, since $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} = W \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} W^*$ for the unitary matrix $W = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$.

Exercise 2.2.1. [2] Prove that the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ is similar to the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, but not unitarily similar to it.

Just like the relation $\sim$, the relation $\overset{\text{us}}{\sim}$ is an equivalence relation:

Proposition 2.2.3. **(a)** Any matrix $A \in \mathbb{C}^{n\times n}$ is unitarily similar to itself.

(b) If A and B are two matrices in $\mathbb{C}^{n\times n}$ such that A is unitarily similar to B, then B is unitarily similar to A.

(c) If A, B and C are three matrices in $\mathbb{C}^{n\times n}$ such that A is unitarily similar to B and such that B is unitarily similar to C, then A is unitarily similar to C.

Proof. This is very similar to the proof of Proposition 2.1.3, and therefore left to the reader. (The only new idea is to use Exercise 1.5.2.) □

Definition 2.2.4. Let A and B be two matrices in $\mathbb{C}^{n\times n}$. We say that A and B are *unitarily similar* if A is unitarily similar to B (or, equivalently, B is unitarily similar to A).

As we promised, unitary similarity is a more restrictive version of similarity:

Proposition 2.2.5. Let A and B be two unitarily similar matrices in $\mathbb{C}^{n\times n}$. Then, A and B are similar.

Proof. There exists a unitary matrix $W \in \mathrm{U}_n(\mathbb{C})$ such that $B = WAW^*$ (since A is unitarily similar to B). Consider this W. The matrix W is unitary, and thus (by the implication $\mathcal{A} \Longrightarrow \mathcal{D}$ in Theorem 1.5.3) must be square and invertible and satisfy $W^{-1} = W^*$. Hence, $B = WA\underbrace{W^*}_{=W^{-1}} = WAW^{-1}$. But this shows that A is similar to B. Thus, Proposition 2.2.5 is proven. □

The following proposition is an analogue of Proposition 2.1.10 for unitary similarity:

Proposition 2.2.6. Let $n \in \mathbb{N}$. For each $i \in [n]$, let $A_i \in \mathbb{C}^{n_i\times n_i}$ and $B_i \in \mathbb{C}^{n_i\times n_i}$ be two $n_i \times n_i$-matrices (for some $n_i \in \mathbb{N}$) satisfying $A_i \overset{\mathrm{us}}{\sim} B_i$. Then,

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{pmatrix} \overset{\mathrm{us}}{\sim} \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_n \end{pmatrix}.$$

Exercise 2.2.2. [1] Prove Proposition 2.2.6.

We note further that the similarity in Proposition 2.1.7 can be upgraded to a unitary similarity if we work over the field $\mathbb{C}$:

Proposition 2.2.7. Let $n \in \mathbb{N}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}$. Let σ be a permutation of $[n]$ (that is, a bijective map from $[n]$ to $[n]$). Then,

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \overset{\mathrm{us}}{\sim} \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right).$$

Proof. Let $P_\sigma \in \mathbb{F}^{n\times n}$ be the permutation matrix of σ (defined in Example 1.5.2 **(d)**). In the proof of Proposition 2.1.7, we have shown that

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right) \cdot P_\sigma^{-1}.$$

However, the matrix P_σ is unitary (as we have already seen in Example 1.5.2 **(d)**). Hence, $P_\sigma^{-1} = P_\sigma^*$. Thus,

$$\begin{aligned}\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) &= P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right) \cdot \underbrace{P_\sigma^{-1}}_{=P_\sigma^*} \\ &= P_\sigma \cdot \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right) \cdot P_\sigma^*.\end{aligned}$$

This shows that $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \overset{\text{us}}{\sim} \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$. Thus, Proposition 2.2.7 is proven. □

Lecture 4 starts here.

2.3. Schur triangularization

2.3.1. The theorems

We are now ready for one more matrix decomposition: the so-called *Schur triangularization* (aka *Schur decomposition*):

Theorem 2.3.1 (Schur triangularization theorem)**.** Let $A \in \mathbb{C}^{n \times n}$. Then, there exist a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and an upper-triangular matrix $T \in \mathbb{C}^{n \times n}$ such that $A = UTU^*$. In other words, A is unitarily similar to some upper-triangular matrix.

The factorization $A = UTU^*$ in Theorem 2.3.1 (or, to be more precise, the pair (U, T)) is called a *Schur triangularization* of A. It is usually not unique.

Example 2.3.2. Let $A = \begin{pmatrix} 1 & 3 \\ -3 & 7 \end{pmatrix} \in \mathbb{C}^{2 \times 2}$. Then, a Schur triangularization of A is $A = UTU^*$, where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad T = \begin{pmatrix} 4 & 6 \\ 0 & 4 \end{pmatrix}.$$

(We chose A deliberately to obtain "nice" matrices U and T. The Schur triangularization of a typical $n \times n$-matrix will be more complicated, involving roots of n-th degree polynomials.)

We shall prove a slightly stronger form of Theorem 2.3.1:

Theorem 2.3.3 (Schur triangularization with prescribed diagonal)**.** Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be its eigenvalues (listed with their algebraic multiplicities). Then, there exists an upper-triangular matrix $T \in \mathbb{C}^{n \times n}$ such that $A \overset{\text{us}}{\sim} T$ (this means "A is unitarily similar to T", as we recall) and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_n$ in this order.

Example 2.3.4. Let $A = \begin{pmatrix} 1 & -2i \\ 0 & 3 \end{pmatrix} \in \mathbb{C}^{2\times 2}$. Theorem 2.3.1 then says that A is unitarily similar to some upper-triangular matrix. But this is trivial, since A is already upper-triangular (and clearly unitarily similar to itself). However, Theorem 2.3.3 can be used to draw a less obvious conclusion. In fact, the eigenvalues of A are 1 and 3. Let us list them in the order $3, 1$. Then, Theorem 2.3.3 (applied to $n = 2$ and $(\lambda_1, \lambda_2, \ldots, \lambda_n) = (3, 1)$) yields that there exists an upper-triangular matrix $T \in \mathbb{C}^{2\times 2}$ such that $A \overset{\text{us}}{\sim} T$ and such that the diagonal entries of T are 3 and 1 in this order. Finding such a T is not all that easy (in particular, A itself does not qualify, since its diagonal entries are 1 and 3 rather than 3 and 1). The answer is:

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad T = \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}.$$

Here, U is the unitary matrix satisfying $A = UTU^*$ (which confirms that $A \overset{\text{us}}{\sim} T$).

Actually, the form of the matrix T in this example is no accident; more generally, we have:

Exercise 2.3.1. [5] Let $a, b, c \in \mathbb{C}$. Prove that

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \overset{\text{us}}{\sim} \begin{pmatrix} c & -\overline{b} \\ 0 & a \end{pmatrix}.$$

2.3.2. The proofs

The following lemma about characteristic polynomials will help us in our proof of Theorem 2.3.3:

Lemma 2.3.5. Let $\mathbb{F}$ be a field. Let $n \in \mathbb{N}$. Let $p \in \mathbb{F}^{1\times n}$ be a row vector, and let $B \in \mathbb{F}^{n\times n}$ be an $n \times n$-matrix. Let $\lambda \in \mathbb{F}$ be a scalar. Let C be the $(n+1) \times (n+1)$-matrix $\begin{pmatrix} \lambda & p \\ 0 & B \end{pmatrix} \in \mathbb{F}^{(n+1)\times(n+1)}$ (written in block-matrix notation, where the "λ" stands for the 1×1-matrix $\begin{pmatrix} \lambda \end{pmatrix}$, and where the "0" stands for the zero vector in $\mathbb{F}^{n\times 1}$). Then,

$$p_C = (t - \lambda) \cdot p_B.$$

Proof of Lemma 2.3.5. The definition of a characteristic polynomial yields

$$p_B = \det(tI_n - B) \qquad \text{and} \qquad p_C = \det(tI_{n+1} - C).$$

However, from $I_{n+1} = \begin{pmatrix} 1 & 0 \\ 0 & I_n \end{pmatrix}$ and $C = \begin{pmatrix} \lambda & p \\ 0 & B \end{pmatrix}$ we obtain

$$tI_{n+1} - C = t\begin{pmatrix} 1 & 0 \\ 0 & I_n \end{pmatrix} - \begin{pmatrix} \lambda & p \\ 0 & B \end{pmatrix} = \begin{pmatrix} t-\lambda & -p \\ 0 & tI_n - B \end{pmatrix}.$$

Thus,

$$\det\left(tI_{n+1} - C\right) = \det\begin{pmatrix} t-\lambda & -p \\ 0 & tI_n - B \end{pmatrix} = (t-\lambda)\cdot\det\left(tI_n - B\right)$$

(here, we have applied Laplace expansion along the first column to compute the determinant, noticing that this column has only one nonzero entry). Thus,

$$p_C = \det\left(tI_{n+1} - C\right) = (t-\lambda)\cdot\underbrace{\det\left(tI_n - B\right)}_{=p_B} = (t-\lambda)\cdot p_B.$$

This proves Lemma 2.3.5. □

Let us now prove Theorem 2.3.3:

Proof of Theorem 2.3.3. We proceed by induction on n:

Induction base: For $n = 0$, Theorem 2.3.3 holds trivially[15].

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Theorem 2.3.3 is proved for $n = m-1$. We must prove that Theorem 2.3.3 holds for $n = m$.

So let $A \in \mathbb{C}^{m\times m}$ be an $m \times m$-matrix, and let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be its eigenvalues (listed with their algebraic multiplicities). We must show that there exists an upper-triangular matrix $T \in \mathbb{C}^{m\times m}$ such that $A \overset{\text{us}}{\sim} T$ and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order.

Since λ_1 is an eigenvalue of A, there exists at least one nonzero λ_1-eigenvector x of A. Pick any such x. Set $u_1 := \dfrac{1}{||x||}x$. Then, u_1 is still a λ_1-eigenvector of A, but additionally satisfies $||u_1|| = 1$. Hence, the 1-tuple (u_1) of vectors in $\mathbb{C}^m$ is orthonormal.

Thus, Corollary 1.2.9 (applied to $k = 1$ and $n = m$) shows that we can find $m-1$ vectors $u_2, u_3, \ldots, u_m \in \mathbb{C}^m$ such that $(u_1, u_2, \ldots, u_m)$ is an orthonormal basis of $\mathbb{C}^m$. Consider these $m-1$ vectors $u_2, u_3, \ldots, u_m$.

Let $U \in \mathbb{C}^{m\times m}$ be the $m \times m$-matrix whose columns are $u_1, u_2, \ldots, u_m$ in this order. Then, the columns of this matrix form an orthonormal basis of $\mathbb{C}^m$. Hence, Theorem 1.5.3 (specifically, the implication $\mathcal{E} \Longrightarrow \mathcal{A}$ in this theorem) yields that the matrix U is unitary. Therefore, $U^* = U^{-1}$. Moreover, since U is unitary, we have $UU^* = I_m$ and $U^*U = I_m$. Thus, U^* is unitary as well.

Define an $m \times m$-matrix

$$C := U^*AU \in \mathbb{C}^{m\times m}.$$

[15]There is only one 0×0-matrix, and we take T to be this matrix.

Since U^* is unitary, we have $A \overset{\text{us}}{\sim} U^* A \left(U^*\right)^*$. In other words, $A \overset{\text{us}}{\sim} C$ (since $U^* A \underbrace{\left(U^*\right)^*}_{=U} = U^* A U = C$). Hence, $A \sim C$ (by Proposition 2.2.5). Thus, Proposition 2.1.5 **(d)** shows that the matrices A and C have the same characteristic polynomial. In other words, $p_A = p_C$.

The definition of U yields $U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{pmatrix}$ and therefore

$$U^* = \begin{pmatrix} — & u_1^* & — \\ & \vdots & \\ — & u_m^* & — \end{pmatrix} \qquad \text{and} \qquad AU = \begin{pmatrix} | & & | \\ Au_1 & \cdots & Au_m \\ | & & | \end{pmatrix}.$$

Multiplying the latter two equalities, we obtain

$$U^* A U = \begin{pmatrix} u_1^* A u_1 & u_1^* A u_2 & \cdots & u_1^* A u_m \\ u_2^* A u_1 & u_2^* A u_2 & \cdots & u_2^* A u_m \\ \vdots & \vdots & \ddots & \vdots \\ u_m^* A u_1 & u_m^* A u_2 & \cdots & u_m^* A u_m \end{pmatrix}.$$

Since $C = U^* A U$, we can rewrite this as

$$C = \begin{pmatrix} u_1^* A u_1 & u_1^* A u_2 & \cdots & u_1^* A u_m \\ u_2^* A u_1 & u_2^* A u_2 & \cdots & u_2^* A u_m \\ \vdots & \vdots & \ddots & \vdots \\ u_m^* A u_1 & u_m^* A u_2 & \cdots & u_m^* A u_m \end{pmatrix}. \tag{32}$$

Hence,

$$\begin{aligned} C_{1,1} &= u_1^* \underbrace{A u_1}_{\substack{=\lambda_1 u_1 \\ \text{(since } u_1 \text{ is a } \lambda_1\text{-eigenvector of } A\text{)}}} = u_1^* \lambda_1 u_1 = \lambda_1 \underbrace{u_1^* u_1}_{\substack{=\langle u_1, u_1 \rangle = ||u_1||^2 = 1 \\ \text{(since } ||u_1|| = 1\text{)}}} \\ &= \lambda_1. \end{aligned} \tag{33}$$

Moreover, for each $i \in \{2, 3, \ldots, m\}$, we have

$$\begin{aligned} C_{i,1} &= u_i^* \underbrace{A u_1}_{\substack{=\lambda_1 u_1 \\ \text{(since } u_1 \text{ is a } \lambda_1\text{-eigenvector of } A\text{)}}} \qquad \text{(by (32))} \\ &= u_i^* \lambda_1 u_1 = \lambda_1 \underbrace{u_i^* u_1}_{\substack{=\langle u_1, u_i \rangle = 0 \\ \text{(since } (u_1, u_2, \ldots, u_m) \text{ is an orthonormal basis,} \\ \text{and thus } u_1 \perp u_i\text{)}}} = 0. \end{aligned}$$

Combining this with (33), we see that the 1-st column of the matrix C has entries $\lambda_1, 0, 0, \ldots, 0$ from top to bottom. In other words,

$$C = \begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{pmatrix},$$

where each asterisk ($*$) means an entry that we don't know or don't care about (in our case, both).

Let us write this in block-matrix notation:

$$C = \begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix}, \tag{34}$$

where $p \in \mathbb{C}^{1\times(m-1)}$ is a row vector and $B \in \mathbb{C}^{(m-1)\times(m-1)}$ is a matrix[16]. Consider this p and this B.

We shall now show that the eigenvalues of B are $\lambda_2, \lambda_3, \ldots, \lambda_m$. Indeed, Lemma 2.3.5 (applied to $\mathbb{F} = \mathbb{C}$ and $n = m-1$ and $\lambda = \lambda_1$) yields $p_C = (t-\lambda_1) \cdot p_B$ (because of (34)). Hence,

$$p_A = p_C = (t-\lambda_1) \cdot p_B.$$

On the other hand,

$$p_A = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m)$$

(since p_A is monic, and the roots of p_A are precisely the eigenvalues of A with algebraic multiplicities, which we know are $\lambda_1, \lambda_2, \ldots, \lambda_m$). Comparing these two equalities, we obtain

$$(t-\lambda_1) \cdot p_B = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m).$$

We can cancel the factor $t-\lambda_1$ from both sides of this equality (since the polynomial ring over $\mathbb{C}$ has no zero-divisors). Thus, we obtain

$$p_B = (t-\lambda_2)(t-\lambda_3)\cdots(t-\lambda_m).$$

In other words, the eigenvalues of the $(m-1)\times(m-1)$-matrix B are $\lambda_2, \lambda_3, \ldots, \lambda_m$ (listed with their algebraic multiplicities).

Hence, by the induction hypothesis, we can apply Theorem 2.3.3 to $m-1$, B and $\lambda_2, \lambda_3, \ldots, \lambda_m$ instead of n, A and $\lambda_1, \lambda_2, \ldots, \lambda_n$. As a result, we conclude that there exists an upper-triangular matrix $S \in \mathbb{C}^{(m-1)\times(m-1)}$ such that $B \overset{\text{us}}{\sim} S$ and such that the diagonal entries of S are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order. Consider this S.

We have $B \overset{\text{us}}{\sim} S$. In other words, there is a unitary matrix $V \in \mathbb{C}^{(m-1)\times(m-1)}$ such that $S = VBV^*$. Consider this V.

[16]The "0" here is actually the zero vector in $\mathbb{C}^{m-1}$.

Now, let

$$W := \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \in \mathbb{C}^{m \times m} \qquad \text{(in block-matrix notation)}.$$

This is a block-diagonal matrix, with $\begin{pmatrix} 1 \end{pmatrix}$ and V being its diagonal blocks. Hence, $W^* = \begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix}$. Moreover, W is a unitary matrix (since it is a block-diagonal matrix whose diagonal blocks are unitary[17]). Hence, $C \overset{\text{us}}{\sim} WCW^*$. Combining $A \overset{\text{us}}{\sim} C$ with $C \overset{\text{us}}{\sim} WCW^*$, we obtain $A \overset{\text{us}}{\sim} WCW^*$ (by Proposition 2.2.3 **(c)**).

However,

$$\begin{aligned}
&\underbrace{W}_{=\begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix}} \quad \underbrace{C}_{=\begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix}} \quad \underbrace{W^*}_{=\begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix}} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix} \\
&= \begin{pmatrix} 1 \cdot \lambda_1 \cdot 1 & 1 \cdot p \cdot V^* \\ V \cdot 0 \cdot 1 & V \cdot B \cdot V^* \end{pmatrix} \qquad \left(\begin{array}{c} \text{by using Proposition 1.6.5 twice} \\ \text{and simplifying the 0 addends away} \end{array} \right) \\
&= \begin{pmatrix} \lambda_1 & pV^* \\ 0 & VBV^* \end{pmatrix} = \begin{pmatrix} \lambda_1 & pV^* \\ 0 & S \end{pmatrix} \qquad (\text{since } VBV^* = S).
\end{aligned}$$

This matrix WCW^* is therefore upper-triangular (since the bottom-left block is a zero vector, and since the bottom-right block S is upper-triangular), and its diagonal entries are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order (because its first diagonal entry is visibly λ_1, whereas its remaining diagonal entries are the diagonal entries of S and therefore are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order[18]).

Thus, there exists an upper-triangular matrix $T \in \mathbb{C}^{m \times m}$ such that $A \overset{\text{us}}{\sim} T$ and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order (namely, WCW^* is such a matrix, because $A \overset{\text{us}}{\sim} WCW^*$ and because of what we just said).

Thus, we have shown that Theorem 2.3.3 holds for $n = m$. This completes the induction step. The proof of Theorem 2.3.3 is thus complete. □

Proof of Theorem 2.3.1. Theorem 2.3.1 follows from Theorem 2.3.3 (and the definition of unitary similarity). □

Exercise 2.3.2. [2] Find a Schur triangularization of the matrix $\begin{pmatrix} 1 & 0 \\ i & 1 \end{pmatrix}$.

[17] We are using Proposition 1.6.12 here.

[18] In fact, we have shown above that the diagonal entries of S are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order.

Exercise 2.3.3. [3] Find a Schur triangularization of the matrix $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$.

2.3.3. The diagonal entries of T

One remark is in order about the T in a Schur triangularization:

Proposition 2.3.6. Let $A \in \mathbb{C}^{n \times n}$. Let (U, T) be a Schur triangularization of A. Then, the diagonal entries of T are the eigenvalues of A (with their algebraic multiplicities).

Instead of proving this directly, let us show a more general result:

Proposition 2.3.7. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ and $T \in \mathbb{F}^{n \times n}$ be two similar matrices. Assume that T is upper-triangular. Then, the diagonal entries of T are the eigenvalues of A (with their algebraic multiplicities).

Proof of Proposition 2.3.7. We have assumed that A and T are similar. In other words, T is similar to A. Thus, Proposition 2.1.5 **(e)** shows that the matrices T and A have the same eigenvalues with the same algebraic multiplicities (and the same geometric multiplicities, but we don't need to know this). In other words, the eigenvalues of T are the eigenvalues of A (with the same algebraic multiplicities).

However, the matrix T is upper-triangular. Thus, the eigenvalues of T (with algebraic multiplicities) are the diagonal entries of T (this is a well-known fact[19]).

[19]Here is a *proof:* The matrix T is upper-triangular; thus, it has the form

$$T = \begin{pmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,n} \\ 0 & T_{2,2} & \cdots & T_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{n,n} \end{pmatrix}.$$

Thus, if t is an indeterminate, then

$$\begin{aligned} tI_n - T &= t \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,n} \\ 0 & T_{2,2} & \cdots & T_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{n,n} \end{pmatrix} \\ &= \begin{pmatrix} t - T_{1,1} & -T_{1,2} & \cdots & -T_{1,n} \\ 0 & t - T_{2,2} & \cdots & -T_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t - T_{n,n} \end{pmatrix}. \end{aligned}$$

This is still an upper-triangular matrix; thus, its determinant is the product of its diagonal entries. That is, we have

$$\det(tI_n - T) = (t - T_{1,1})(t - T_{2,2}) \cdots (t - T_{n,n}).$$

Since we know that the eigenvalues of T are the eigenvalues of A (with the same algebraic multiplicities), we thus conclude that the eigenvalues of A (with algebraic multiplicities) are the diagonal entries of T. This proves Proposition 2.3.7. □

Proof of Proposition 2.3.6. We assumed that (U, T) is a Schur triangularization of A. Hence, we have $A = UTU^*$, and the matrix U is unitary whereas the matrix T is upper-triangular. From $A = UTU^*$, we conclude that T is unitarily similar to A (by Definition 2.2.1, since U is unitary). Hence, T is similar to A (by Proposition 2.2.5). Therefore, Proposition 2.3.7 shows that the diagonal entries of T are the eigenvalues of A (with their algebraic multiplicities). This proves Proposition 2.3.6. □

We will see several applications of Theorem 2.3.1 in this chapter. First, however, let us see some variants of Schur triangularization.

2.3.4. Triangularization over an arbitrary field

The first variant gives a partial answer to the following natural question: What becomes of Theorem 2.3.1 if we replace $\mathbb{C}$ by an arbitrary field $\mathbb{F}$? Of course, Theorem 2.3.1 does not even make sense for an arbitrary field $\mathbb{F}$, since the notion of a unitary matrix is only defined over $\mathbb{C}$. Even if we take this loss and replace "unitary" by merely "invertible" (so U^* becomes U^{-1}, and "unitarily similar" just becomes "similar"), the theorem will still fail, because (for example) the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is not similar to any triangular matrix over the field $\mathbb{R}$. This is because its eigenvalues (which are i and $-i$) are not real.

However, apart from these two issues, things go well. So we can state the following weak version of Schur triangularization over an arbitrary field:

Theorem 2.3.8 (triangularization theorem)**.** Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$. Assume that the characteristic polynomial p_A of A factors as a product of n linear factors over $\mathbb{F}$ (so A has n eigenvalues in $\mathbb{F}$, counted with algebraic multiplicities). Then, there exist an invertible matrix $U \in \mathbb{F}^{n \times n}$ and an upper-triangular matrix $T \in \mathbb{F}^{n \times n}$ such that $A = UTU^{-1}$. In other words, A is similar to some upper-triangular matrix.

This is an analogue of Theorem 2.3.1; a corresponding analogue of Theorem 2.3.3 also exists:

Therefore, $T_{1,1}, T_{2,2}, \ldots, T_{n,n}$ are the roots of the polynomial $\det(tI_n - T)$ (with multiplicities).

However, $\det(tI_n - T)$ is the characteristic polynomial of T. Thus, the roots of this polynomial $\det(tI_n - T)$ are the eigenvalues of T (with algebraic multiplicities). Since we know that the roots of this polynomial are $T_{1,1}, T_{2,2}, \ldots, T_{n,n}$, we thus conclude that $T_{1,1}, T_{2,2}, \ldots, T_{n,n}$ are the eigenvalues of T (with algebraic multiplicities). In other words, the diagonal entries of T are the eigenvalues of T (with algebraic multiplicities). Qed.

Theorem 2.3.9 (triangularization with prescribed diagonal)**.** Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$. Assume that the characteristic polynomial p_A of A factors as a product of n linear factors over $\mathbb{F}$ (so A has n eigenvalues in $\mathbb{F}$, counted with algebraic multiplicities). Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A (listed with their algebraic multiplicities). Then, there exists an upper-triangular matrix $T \in \mathbb{F}^{n \times n}$ such that $A \sim T$ (this means "A is similar to T", as we recall) and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_n$ in this order.

The proofs of Theorem 2.3.8 and Theorem 2.3.9 are fairly similar to the above proofs of Theorem 2.3.1 and Theorem 2.3.3:

Proof of Theorem 2.3.9. We proceed by induction on n:

Induction base: For $n = 0$, Theorem 2.3.9 holds trivially[20].

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Theorem 2.3.9 is proved for $n = m - 1$. We must prove that Theorem 2.3.9 holds for $n = m$.

So let $A \in \mathbb{F}^{m \times m}$ be an $m \times m$-matrix whose characteristic polynomial p_A factors as a product of m linear factors, and let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be the eigenvalues of A (listed with their algebraic multiplicities). We must show that there exists an upper-triangular matrix $T \in \mathbb{F}^{m \times m}$ such that $A \sim T$ and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order.

Since λ_1 is an eigenvalue of A, there exists at least one nonzero λ_1-eigenvector x of A. Pick any such x. Set $u_1 := x$. The 1-tuple (u_1) of vectors in $\mathbb{F}^m$ is then linearly independent (since $u_1 = x$ is nonzero).

It is well-known that any linearly independent tuple of vectors in $\mathbb{F}^m$ can be extended to a basis of $\mathbb{F}^m$ (if you wish, this is an analogue of Corollary 1.2.9). Thus, in particular, the 1-tuple (u_1) of vectors in $\mathbb{F}^m$ can be extended to a basis of $\mathbb{F}^m$. In other words, we can find $m - 1$ vectors $u_2, u_3, \ldots, u_m \in \mathbb{F}^m$ such that $(u_1, u_2, \ldots, u_m)$ is a basis of $\mathbb{F}^m$. Consider these $m - 1$ vectors $u_2, u_3, \ldots, u_m$.

Let $U \in \mathbb{F}^{m \times m}$ be the $m \times m$-matrix whose columns are $u_1, u_2, \ldots, u_m$ in this order. Then, the columns of this matrix form a basis of $\mathbb{F}^m$. Hence, the matrix U is invertible.

Let $(e_1, e_2, \ldots, e_m)$ be the standard basis of the $\mathbb{F}$-vector space $\mathbb{F}^m$. It is known that if B is any $m \times m$-matrix, then

$$\text{the 1-st column of } B \text{ is } Be_1. \tag{35}$$

Applying this to $B = U$, we see that the 1-st column of U is Ue_1. Since we also know that the 1-st column of U is u_1 (by the definition of U), we thus conclude that $Ue_1 = u_1 = x$. However, $Ax = \lambda_1 x$ (since x is a λ_1-eigenvector of A). In view of $Ue_1 = x$, this rewrites as $AUe_1 = \lambda_1 Ue_1$.

Define an $m \times m$-matrix

$$C := U^{-1} A U \in \mathbb{F}^{m \times m}.$$

[20] There is only one 0×0-matrix, and we take T to be this matrix.

Thus, $A \sim C$. Hence, Proposition 2.1.5 **(d)** shows that the matrices A and C have the same characteristic polynomial. In other words, $p_A = p_C$.

Furthermore, (35) shows that the 1-st column of the matrix C is

$$Ce_1 = U^{-1} \underbrace{AUe_1}_{=\lambda_1 Ue_1} \qquad \left(\text{since } C = U^{-1}AU\right)$$

$$= U^{-1}\lambda_1 Ue_1 = \lambda_1 e_1 = \begin{pmatrix} \lambda_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In other words, the matrix C has the form

$$C = \begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{pmatrix},$$

where each asterisk ($*$) means an entry that we don't know or don't care about (in our case, both).

Let us write this in block-matrix notation:

$$C = \begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix}, \tag{36}$$

where $p \in \mathbb{F}^{1\times(m-1)}$ is a row vector and $B \in \mathbb{F}^{(m-1)\times(m-1)}$ is a matrix[21]. Consider this p and this B.

We shall now show that the eigenvalues of B are $\lambda_2, \lambda_3, \ldots, \lambda_m$. Indeed, Lemma 2.3.5 (applied to $n = m-1$ and $\lambda = \lambda_1$) yields $p_C = (t-\lambda_1) \cdot p_B$ (because of (36)). Hence,

$$p_A = p_C = (t-\lambda_1) \cdot p_B.$$

On the other hand,

$$p_A = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m)$$

(since p_A is monic, and the roots of p_A are precisely the eigenvalues of A with algebraic multiplicities, which we know are $\lambda_1, \lambda_2, \ldots, \lambda_m$). Comparing these two equalities, we obtain

$$(t-\lambda_1) \cdot p_B = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m).$$

[21] The "0" here is actually the zero vector in $\mathbb{F}^{m-1}$.

We can cancel the factor $t - \lambda_1$ from both sides of this equality (since the polynomial ring over $\mathbb{F}$ has no zero-divisors). Thus, we obtain

$$p_B = (t - \lambda_2)(t - \lambda_3) \cdots (t - \lambda_m).$$

In other words, the eigenvalues of the $(m-1) \times (m-1)$-matrix B are $\lambda_2, \lambda_3, \ldots, \lambda_m$ (listed with their algebraic multiplicities).

Hence, by the induction hypothesis, we can apply Theorem 2.3.9 to $m-1$, B and $\lambda_2, \lambda_3, \ldots, \lambda_m$ instead of n, A and $\lambda_1, \lambda_2, \ldots, \lambda_n$. As a result, we conclude that there exists an upper-triangular matrix $S \in \mathbb{F}^{(m-1)\times(m-1)}$ such that $B \sim S$ and such that the diagonal entries of S are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order. Consider this S.

We have $B \sim S$. In other words, there is a invertible matrix $V \in \mathbb{F}^{(m-1)\times(m-1)}$ such that $S = VBV^{-1}$. Consider this V.

Now, let

$$W := \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \in \mathbb{F}^{m\times m} \qquad \text{(in block-matrix notation)}.$$

This is a block-diagonal matrix, with $\begin{pmatrix} 1 \end{pmatrix}$ and V being its diagonal blocks. Hence, W is invertible with $W^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & V^{-1} \end{pmatrix}$. Therefore, $C \sim WCW^{-1}$. Combining $A \sim C$ with $C \sim WCW^{-1}$, we obtain $A \sim WCW^{-1}$ (by Proposition 2.1.3 **(c)**).

However,

$$\begin{aligned}
&\underbrace{W}\quad\quad\underbrace{C}\quad\quad\underbrace{W^{-1}} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & V^{-1} \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \lambda_1 & p \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V^{-1} \end{pmatrix} \\
&= \begin{pmatrix} 1 \cdot \lambda_1 \cdot 1 & 1 \cdot p \cdot V^{-1} \\ V \cdot 0 \cdot 1 & V \cdot B \cdot V^{-1} \end{pmatrix} \qquad \begin{pmatrix} \text{by using Proposition 1.6.5 twice} \\ \text{and simplifying the 0 addends away} \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 & pV^{-1} \\ 0 & VBV^{-1} \end{pmatrix} = \begin{pmatrix} \lambda_1 & pV^{-1} \\ 0 & S \end{pmatrix} \qquad \left(\text{since } VBV^{-1} = S\right).
\end{aligned}$$

This matrix WCW^{-1} is therefore upper-triangular (since the bottom-left block is a zero vector, and since the bottom-right block S is upper-triangular), and its diagonal entries are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order (because its first diagonal entry is visibly λ_1, whereas its remaining diagonal entries are the diagonal entries of S and therefore are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order[22]).

Thus, there exists an upper-triangular matrix $T \in \mathbb{F}^{m\times m}$ such that $A \sim T$ and such that the diagonal entries of T are $\lambda_1, \lambda_2, \ldots, \lambda_m$ in this order (namely, WCW^{-1} is such a matrix, because $A \sim WCW^{-1}$ and because of what we just said).

[22] In fact, we have shown above that the diagonal entries of S are $\lambda_2, \lambda_3, \ldots, \lambda_m$ in this order.

Thus, we have shown that Theorem 2.3.9 holds for $n = m$. This completes the induction step. The proof of Theorem 2.3.9 is thus complete. □

Proof of Theorem 2.3.8. Theorem 2.3.8 follows from Theorem 2.3.9 (and the definition of similarity). □

2.4. Commuting matrices

Next, we shall generalize Theorem 2.3.1 from the case of a single matrix to the case of several matrices that pairwise commute.

Definition 2.4.1. Two $n \times n$-matrices A and B are said to *commute* if $AB = BA$.

Examples of commuting matrices are easy to find; e.g., any two powers of a single matrix commute (i.e., we have $A^k A^\ell = A^\ell A^k$ for any $n \times n$-matrix A and any $k, \ell \in \mathbb{N}$). Also, any two diagonal matrices (of the same size) commute. But there are many more situations in which matrices commute. In this section, we shall extend Schur triangularization from a single matrix to a family of commuting matrices.

First, we need a lemma ([HorJoh13, Lemma 1.3.19]) which says that any family of pairwise commuting matrices in $\mathbb{C}^{n\times n}$ has a common eigenvector:

Lemma 2.4.2. Let $n > 0$. Let $\mathcal{F}$ be a subset of $\mathbb{C}^{n\times n}$ such that any two matrices in $\mathcal{F}$ commute (i.e., any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ satisfy $AB = BA$).

Then, there exists a nonzero vector $x \in \mathbb{C}^n$ such that x is an eigenvector of each $A \in \mathcal{F}$.

Proof. We shall use the following conventions:

- An *$\mathcal{F}$-eigenvector* will mean a vector $x \in \mathbb{C}^n$ such that x is an eigenvector of each $A \in \mathcal{F}$. Thus, our goal is to show that there exists a nonzero $\mathcal{F}$-eigenvector.
- A *subspace* shall mean a $\mathbb{C}$-vector subspace of $\mathbb{C}^n$.
- A subspace W of $\mathbb{C}^n$ is said to be *nontrivial* if it contains a nonzero vector (i.e., if its dimension is > 0).
- A subspace W of $\mathbb{C}^n$ is said to be *$\mathcal{F}$-invariant* if every $A \in \mathcal{F}$ and every $w \in W$ satisfy $Aw \in W$. (This means that applying a matrix $A \in \mathcal{F}$ to a vector $w \in W$ gives a vector in W again – i.e., that there is no way to "escape" W by applying matrices $A \in \mathcal{F}$.)

  [**Example:** If $n = 2$ and $\mathcal{F} = \left\{ \begin{pmatrix} 3 & a \\ 0 & 2 \end{pmatrix} \mid a \in \mathbb{R} \right\}$, then $\operatorname{span}(e_1) = \left\{ \begin{pmatrix} x \\ 0 \end{pmatrix} \mid x \in \mathbb{C} \right\}$ is an $\mathcal{F}$-invariant subspace, because every $A = \begin{pmatrix} 3 & a \\ 0 & 2 \end{pmatrix} \in \mathcal{F}$ and every

$w = \begin{pmatrix} x \\ 0 \end{pmatrix} \in \operatorname{span}(e_1)$ satisfy

$$Aw = \begin{pmatrix} 3 & a \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} 3x \\ 0 \end{pmatrix} \in \operatorname{span}(e_1).$$

More generally, if the set $\mathcal{F} \subseteq \mathbb{C}^{n \times n}$ consists entirely of upper-triangular matrices, then each of the subspaces

$$\operatorname{span}(e_1, e_2, \ldots, e_k) = \{x \in \mathbb{C}^n \mid \text{the last } n-k \text{ coordinates of } x \text{ are } 0\}$$

is $\mathcal{F}$-invariant, because the product Ax of an upper-triangular matrix $A \in \mathbb{C}^{n \times n}$ with a vector $x \in \operatorname{span}(e_1, e_2, \ldots, e_k)$ is always a vector in $\operatorname{span}(e_1, e_2, \ldots, e_k)$.]

It is clear that the trivial subspace $\{0\}$ is $\mathcal{F}$-invariant; so is the whole space $\mathbb{C}^n$ itself. The latter shows that there exists at least one nontrivial $\mathcal{F}$-invariant subspace (namely, $\mathbb{C}^n$).

Now, we shall show the following crucial claim:

Claim 1: Let W be a nontrivial $\mathcal{F}$-invariant subspace. Let $w \in W$. Assume that w is not an $\mathcal{F}$-eigenvector. Then, there exists a nontrivial $\mathcal{F}$-invariant subspace V such that V is a **proper subset** of W.

Roughly speaking, this claim is saying that if a nontrivial $\mathcal{F}$-invariant subspace contains a vector w that is not an $\mathcal{F}$-eigenvector, then there is a smaller nontrivial $\mathcal{F}$-invariant subspace inside it. This fact (once proved) allows us to start with any nontrivial $\mathcal{F}$-invariant subspace (for instance, $\mathbb{C}^n$ itself), and then successively replace it by smaller and smaller subspaces until we eventually find a nontrivial $\mathcal{F}$-invariant subspace that consists entirely of $\mathcal{F}$-eigenvectors. This will then yield the existence of a nonzero $\mathcal{F}$-eigenvector, and thus Lemma 2.4.2 will be proved. (We shall walk through this argument in more detail after proving Claim 1.)

So let us now prove Claim 1:

[*Proof of Claim 1:* We know that w is not an $\mathcal{F}$-eigenvector. In other words, there exists some $B \in \mathcal{F}$ such that w is not an eigenvector of B. Consider this B.

We have $Bv \in W$ for each $v \in W$ (since W is $\mathcal{F}$-invariant). Hence, the map

$$\begin{aligned} f : W &\to W, \\ v &\mapsto Bv \end{aligned}$$

is well-defined. This map f is furthermore $\mathbb{C}$-linear (for obvious reasons). Also, $\dim W > 0$ (because W is nontrivial).

However, it is well-known that any linear map from a finite-dimensional $\mathbb{C}$-vector space to itself has at least one nonzero eigenvector, provided that this vector space has dimension > 0. We thus conclude that the linear map $f : W \to W$ has at least one nonzero eigenvector[23]. Pick such a nonzero eigenvector x, and let $\lambda \in \mathbb{C}$ be

[23] since f is a $\mathbb{C}$-linear map from the finite-dimensional $\mathbb{C}$-vector space W to itself, and since $\dim W > 0$

the corresponding eigenvalue. Thus, $x \in W$ and $v \neq 0$ and $f(x) = \lambda x$. Since $f(x) = Bx$ (by the definition of f), we can rewrite the latter equality as $Bx = \lambda x$. In contrast, $Bw \neq \lambda w$ (since w is not an eigenvector of B).

Define a subset V of W by

$$V := \{v \in W \mid Bv = \lambda v\}.$$

It is easy to see that V is a subspace (since $B(u+v) = Bu + Bv$ and $B(\mu v) = \mu \cdot Bv$ for any $u, v \in W$ and $\mu \in \mathbb{C}$). Furthermore, this subspace V contains the nonzero vector x (since $Bx = \lambda x$), and thus is nontrivial. However, this subspace V does not contain w (because if it did, then we would have $Bw = \lambda w$ (by the definition of V), which would contradict $Bw \neq \lambda w$). Thus, $V \neq W$ (since W does contain w). Therefore, V is a **proper** subset of W (since V is a subset of W).

Let us now show that this subspace V is $\mathcal{F}$-invariant. Indeed, let $A \in \mathcal{F}$ and $v \in V$ be arbitrary. We shall show that $Av \in V$.

We have $v \in V \subseteq W$ and therefore $Av \in W$ (since W is $\mathcal{F}$-invariant). Moreover, $Bv = \lambda v$ (since $v \in V$). Furthermore, B and A commute (since any two matrices in $\mathcal{F}$ commute); thus, $BA = AB$. Hence, $BAv = A\underbrace{Bv}_{=\lambda v} = \lambda \cdot Av$. Thus, Av is a vector $z \in W$ that satisfies $Bz = \lambda z$ (since $Av \in W$). In other words, $Av \in V$ (by the definition of V).

Forget that we fixed A and v. We thus have shown that every $A \in \mathcal{F}$ and every $v \in V$ satisfy $Av \in V$. In other words, the subspace V is $\mathcal{F}$-invariant.

Thus, we have found a nontrivial $\mathcal{F}$-invariant subspace V such that V is a **proper subset** of W. This proves Claim 1.]

Now, we can complete the proof of Lemma 2.4.2 (using the strategy we outlined above):

We must prove that there exists a nonzero $\mathcal{F}$-eigenvector. Assume the contrary. Thus, there exists no nonzero $\mathcal{F}$-eigenvector.

We recursively construct a sequence $(V_0, V_1, V_2, V_3, \ldots)$ of nontrivial $\mathcal{F}$-invariant subspaces as follows:

- We begin by setting $V_0 := \mathbb{C}^n$. (This subspace is indeed $\mathcal{F}$-invariant, and furthermore is nontrivial because $n > 0$.)
- For each $i \in \mathbb{N}$, if the nontrivial $\mathcal{F}$-invariant subspace V_i has already been constructed, we define the next subspace V_{i+1} as follows: We pick an arbitrary nonzero vector $w \in V_i$. (Such a w exists, since V_i is nontrivial.) This w cannot be an $\mathcal{F}$-eigenvector (since there exists no nonzero $\mathcal{F}$-eigenvector). Hence, Claim 1 (applied to $W = V_i$) yields that there exists a nontrivial $\mathcal{F}$-invariant subspace V such that V is a **proper subset** of V_i. We choose such a V and define $V_{i+1} := V$.

Thus, we obtain a sequence $(V_0, V_1, V_2, V_3, \ldots)$ of subspaces such that each subspace V_{i+1} in this sequence is a **proper subset** of the preceding subspace V_i. Hence,

for each $i \in \mathbb{N}$, we have $\dim(V_{i+1}) < \dim(V_i)$ (since V_i and V_{i+1} are finite-dimensional vector spaces). Equivalently, for each $i \in \mathbb{N}$, we have $\dim(V_i) > \dim(V_{i+1})$. In other words,

$$\dim(V_0) > \dim(V_1) > \dim(V_2) > \cdots.$$

Hence, the sequence $(\dim(V_0), \dim(V_1), \dim(V_2), \ldots)$ is a strictly decreasing infinite sequence of nonnegative integers. However, this is absurd, since there exists no strictly decreasing infinite sequence of nonnegative integers[24]. Thus, we have obtained a contradiction. This proves Lemma 2.4.2. □

We can now generalize Theorem 2.3.1 to families of commuting matrices:

Theorem 2.4.3. Let $\mathcal{F}$ be a subset of $\mathbb{C}^{n\times n}$ such that any two matrices in $\mathcal{F}$ commute (i.e., any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ satisfy $AB = BA$).

Then, there exists a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ such that for each $A \in \mathcal{F}$, the matrix UAU^* is upper-triangular.

Proof. This can be proved by an induction on n, similarly to Theorem 2.3.3. But now, instead of picking an eigenvector x of a single matrix A, we pick a common eigenvector for all matrices in $\mathcal{F}$. The existence of such an eigenvector is guaranteed by Lemma 2.4.2.

Here is the argument in more detail:

We proceed by induction on n:

Induction base: For $n = 0$, Theorem 2.4.3 holds trivially[25].

Induction step: Let m be a positive integer. Assume (as the induction hypothesis) that Theorem 2.4.3 is proved for $n = m-1$. We must prove that Theorem 2.4.3 holds for $n = m$.

So let $\mathcal{F}$ be a subset of $\mathbb{C}^{m\times m}$ such that any two matrices in $\mathcal{F}$ commute (i.e., any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ satisfy $AB = BA$). We must show that there exists a unitary matrix $Q \in \mathrm{U}_m(\mathbb{C})$ such that for each $A \in \mathcal{F}$, the matrix QAQ^* is upper-triangular.

Lemma 2.4.2 (applied to $n = m$) shows that there exists a nonzero vector $x \in \mathbb{C}^m$ such that x is an eigenvector of each $A \in \mathcal{F}$. Pick any such x. Set $u_1 := \dfrac{1}{||x||}x$. Then, u_1 is still an eigenvector of each $A \in \mathcal{F}$, but additionally satisfies $||u_1|| = 1$. Hence, the 1-tuple (u_1) of vectors in $\mathbb{C}^m$ is orthonormal.

Thus, Corollary 1.2.9 (applied to $k = 1$ and $n = m$) shows that we can find $m-1$ vectors $u_2, u_3, \ldots, u_m \in \mathbb{C}^m$ such that $(u_1, u_2, \ldots, u_m)$ is an orthonormal basis of $\mathbb{C}^m$. Consider these $m-1$ vectors $u_2, u_3, \ldots, u_m$.

Let $U \in \mathbb{C}^{m\times m}$ be the $m \times m$-matrix whose columns are $u_1, u_2, \ldots, u_m$ in this order. Then, the columns of this matrix form an orthonormal basis of $\mathbb{C}^m$. Hence, Theorem 1.5.3 (specifically, the implication $\mathcal{E} \Longrightarrow \mathcal{A}$ in this theorem) yields that the

[24] This is an example of a "proof by infinite descent".

[25] because any 0×0-matrix is upper-triangular

matrix U is unitary. Therefore, $U^* = U^{-1}$. Moreover, since U is unitary, we have $UU^* = I_m$ and $U^*U = I_m$. Thus, the matrix U^* is unitary.

For each $A \in \mathcal{F}$, we now define an $m \times m$-matrix

$$C_A := U^* A U \in \mathbb{C}^{m \times m}.$$

Now, let $A \in \mathcal{F}$ be arbitrary. We know that u_1 is an eigenvector of A; let λ_A be the corresponding eigenvalue. Now, it is not hard to show[26] that the matrix C_A can be written in block-matrix notation as

$$C_A = \begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix}, \tag{37}$$

where $p_A \in \mathbb{C}^{1 \times (m-1)}$ is a row vector and $B_A \in \mathbb{C}^{(m-1) \times (m-1)}$ is a matrix[27]. Consider this p_A and this B_A.

Forget that we fixed A. Thus, for each $A \in \mathcal{F}$, we have defined a complex number λ_A, a row vector $p_A \in \mathbb{C}^{1 \times (m-1)}$ and a matrix $B_A \in \mathbb{C}^{(m-1) \times (m-1)}$ such that (37) holds.

Let $A \in \mathcal{F}$ and $A' \in \mathcal{F}$ be arbitrary. We are going to show that $B_{AA'} = B_A B_{A'}$.

Indeed, we first observe that the definitions of C_A and $C_{A'}$ yield

$$\begin{aligned} \underbrace{C_A}_{=U^*AU} \underbrace{C_{A'}}_{=U^*A'U} &= U^* A \underbrace{UU^*}_{=I_m} A' U = U^* A A' U \\ &= C_{AA'} \end{aligned} \tag{38}$$

(by the definition of $C_{AA'}$). However, (37) yields

$$C_A = \begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix}, \qquad C_{A'} = \begin{pmatrix} \lambda_{A'} & p_{A'} \\ 0 & B_{A'} \end{pmatrix},$$

$$\text{and} \qquad C_{AA'} = \begin{pmatrix} \lambda_{AA'} & p_{AA'} \\ 0 & B_{AA'} \end{pmatrix}.$$

In view of this, we can rewrite (38) as

$$\begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix} \begin{pmatrix} \lambda_{A'} & p_{A'} \\ 0 & B_{A'} \end{pmatrix} = \begin{pmatrix} \lambda_{AA'} & p_{AA'} \\ 0 & B_{AA'} \end{pmatrix}.$$

Hence,

$$\begin{pmatrix} \lambda_{AA'} & p_{AA'} \\ 0 & B_{AA'} \end{pmatrix} = \begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix} \begin{pmatrix} \lambda_{A'} & p_{A'} \\ 0 & B_{A'} \end{pmatrix} = \begin{pmatrix} \lambda_A \lambda_{A'} & \lambda_A p_{A'} + p_A B_{A'} \\ 0 & B_A B_{A'} \end{pmatrix}$$

[26] Indeed, this is essentially a carbon copy of the proof of (34) in our above proof of Theorem 2.3.3 (except that C, λ_1, p and B are now called C_A, λ_A, p_A and B_A); thus, we leave the details to the reader.

[27] The "0" here is actually the zero vector in $\mathbb{C}^{m-1}$.

(by applying Proposition 1.6.5 and simplifying). Comparing the bottom-right blocks of the block matrices on both sides, we obtain $B_{AA'} = B_A B_{A'}$.

Similarly, $B_{A'A} = B_{A'} B_A$. However, $AA' = A'A$ (since any two matrices in $\mathcal{F}$ commute). Thus, $B_{AA'} = B_{A'A}$. In view of $B_{AA'} = B_A B_{A'}$ and $B_{A'A} = B_{A'} B_A$, we can rewrite this as

$$B_A B_{A'} = B_{A'} B_A. \tag{39}$$

Forget that we fixed A and A'. We thus have proved (39) for all $A \in \mathcal{F}$ and $A' \in \mathcal{F}$.

Define a set

$$\mathcal{F}' := \{B_A \mid A \in \mathcal{F}\} \subseteq \mathbb{C}^{(m-1)\times(m-1)}.$$

Then, (39) shows that any two matrices in $\mathcal{F}'$ commute.

Hence, by the induction hypothesis, we can apply Theorem 2.4.3 to $m-1$ and $\mathcal{F}'$ instead of n and $\mathcal{F}$. As a result, we conclude that there exists a unitary matrix $V \in \mathrm{U}_{m-1}(\mathbb{C})$ such that for each $B \in \mathcal{F}'$, the matrix VBV^* is upper-triangular. Consider this V.

Now, let

$$W := \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \in \mathbb{C}^{m\times m} \qquad \text{(in block-matrix notation)}.$$

This is a block-diagonal matrix, with $\begin{pmatrix} 1 \end{pmatrix}$ and V being its diagonal blocks. Hence, $W^* = \begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix}$. Moreover, W is a unitary matrix (since it is a block-diagonal matrix whose diagonal blocks are unitary[28]).

Now, let Q be the matrix WU^*. Then, Q is unitary (by Exercise 1.5.2 **(b)**, since W and U^* are unitary). We shall show that for each $A \in \mathcal{F}$, the matrix QAQ^* is upper-triangular.

Indeed, let $A \in \mathcal{F}$. Then, $B_A \in \mathcal{F}'$ (by the definition of $\mathcal{F}'$). However, we know that for each $B \in \mathcal{F}'$, the matrix VBV^* is upper-triangular. Applying this to $B = B_A$, we conclude that the matrix VB_AV^* is upper-triangular. On the other

[28] We are using Proposition 1.6.12 here.

hand, from $Q = WU^*$, we obtain

$$\begin{aligned} QAQ^* &= WU^*A \underbrace{(WU^*)^*}_{=(U^*)^*W^*=UW^*} = W \underbrace{U^*AU}_{\substack{=C_A \\ \text{(by the definition of } C_A\text{)}}} W^* \\ &= \underbrace{W}_{=\begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix}} \underbrace{C_A}_{\substack{=\begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix} \\ \text{(by (37))}}} \underbrace{W^*}_{=\begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix}} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \lambda_A & p_A \\ 0 & B_A \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix} \\ &= \begin{pmatrix} 1 \cdot \lambda_A \cdot 1 & 1 \cdot p_A \cdot V^* \\ V \cdot 0 \cdot 1 & V \cdot B_A \cdot V^* \end{pmatrix} \qquad \begin{pmatrix} \text{by using Proposition 1.6.5 twice and} \\ \text{simplifying the 0 addends away} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_A & p_A V^* \\ 0 & V B_A V^* \end{pmatrix}. \end{aligned}$$

This matrix QAQ^* is therefore upper-triangular (since its bottom-right block VB_AV^* is upper-triangular).

Forget that we fixed A. We thus have shown that for each $A \in \mathcal{F}$, the matrix QAQ^* is upper-triangular. Since Q is unitary, we thus have found a unitary matrix $Q \in \mathrm{U}_m(\mathbb{C})$ such that for each $A \in \mathcal{F}$, the matrix QAQ^* is upper-triangular.

Thus, we have shown that Theorem 2.4.3 holds for $n = m$. This completes the induction step. The proof of Theorem 2.4.3 is thus complete. □

Lecture 5 starts here.

2.5. Normal matrices

We next define a fairly wide class of matrices with complex entries that contains several of our familiar classes as subsets:

Definition 2.5.1. A square matrix $A \in \mathbb{C}^{n \times n}$ is said to be *normal* if $AA^* = A^*A$.

In other words, a square matrix is normal if it commutes with its own conjugate transpose. This is not the most intuitive notion (nor is the word "normal" particularly expressive), so we shall give some examples:

Example 2.5.2. (a) Let $A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \in \mathbb{C}^{2 \times 2}$. Then, the matrix A is normal. Indeed, its conjugate transpose is $A^* = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, and it is easily seen that $AA^* = A^*A = 2I_2$.

(b) Let $B = \begin{pmatrix} 0 & i \\ 0 & 0 \end{pmatrix} \in \mathbb{C}^{2\times 2}$. Then, the matrix B is not normal. Indeed, $B^* = \begin{pmatrix} 0 & 0 \\ -i & 0 \end{pmatrix}$ and thus $BB^* \neq B^*B$, as can easily be verified.

(c) Let $a, b \in \mathbb{C}$ be arbitrary, and let $C = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \in \mathbb{C}^{2\times 2}$. Then, C is normal. Indeed, $C^* = \begin{pmatrix} \overline{a} & \overline{b} \\ \overline{b} & \overline{a} \end{pmatrix}$, so that it is easy to check that both CC^* and C^*C equal $\begin{pmatrix} a & b \\ b & a \end{pmatrix}\begin{pmatrix} \overline{a} & \overline{b} \\ \overline{b} & \overline{a} \end{pmatrix} = \begin{pmatrix} a\overline{a}+b\overline{b} & a\overline{b}+b\overline{a} \\ a\overline{b}+b\overline{a} & a\overline{a}+b\overline{b} \end{pmatrix}$.

As we promised, several familiar classes of matrices are normal. We recall a definition:

Definition 2.5.3. A square matrix $H \in \mathbb{C}^{n\times n}$ is said to be *Hermitian* if and only if $H^* = H$.

For example, the matrix $\begin{pmatrix} 1 & i \\ -i & 2 \end{pmatrix}$ is Hermitian. Any real symmetric matrix (i.e., any symmetric matrix with real entries) is Hermitian as well.

In contrast, a square matrix $S \in \mathbb{C}^{n\times n}$ is skew-Hermitian if and only if $S^* = -S$ (by Definition 1.5.4). Finally, a square matrix $U \in \mathbb{C}^{n\times n}$ is unitary if and only if $UU^* = U^*U = I_n$ (by Theorem 1.5.3, equivalence $\mathcal{A} \Longleftrightarrow \mathcal{C}$). Having recalled all these concepts, we can state the following:

Proposition 2.5.4. **(a)** Every Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is normal.

(b) Every skew-Hermitian matrix $S \in \mathbb{C}^{n\times n}$ is normal.

(c) Every unitary matrix $U \in \mathbb{C}^{n\times n}$ is normal.

(d) Every diagonal matrix $D \in \mathbb{C}^{n\times n}$ is normal.

Proof. **(a)** Let $H \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Then, $H^* = H$ (by the definition of "Hermitian"). Hence, $H\underbrace{H^*}_{=H} = \underbrace{H}_{=H^*}H = H^*H$. In other words, H is normal. This proves Proposition 2.5.4 **(a)**.

(b) This is analogous to part **(a)**, except for a minus sign that appears and disappears again.

(c) This is clear, since $UU^* = U^*U = I_n$ entails $UU^* = U^*U$.

(d) Let $D \in \mathbb{C}^{n\times n}$ be a diagonal matrix. Write D in the form

$$D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \qquad \text{for some } \lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}.$$

Then, $D^* = \operatorname{diag}\left(\overline{\lambda_1}, \overline{\lambda_2}, \ldots, \overline{\lambda_n}\right)$. Hence,

$$DD^* = \operatorname{diag}\left(\lambda_1\overline{\lambda_1}, \lambda_2\overline{\lambda_2}, \ldots, \lambda_n\overline{\lambda_n}\right) \qquad \text{and}$$
$$D^*D = \operatorname{diag}\left(\overline{\lambda_1}\lambda_1, \overline{\lambda_2}\lambda_2, \ldots, \overline{\lambda_n}\lambda_n\right).$$

The right hand sides of these two equalities are equal (since $\lambda_i\overline{\lambda_i} = \overline{\lambda_i}\lambda_i$ for each $i \in [n]$). Thus, the left hand sides must too be equal. In other words, $DD^* = D^*D$. This means that D is normal. This proves Proposition 2.5.4 **(d)**. □

Unlike the unitary matrices, the normal matrices are not closed under multiplication:

Exercise 2.5.1. $\boxed{2}$ Find two normal matrices $A, B \in \mathbb{C}^{2\times 2}$ such that neither $A + B$ nor AB is normal.

Here are three more ways to construct normal matrices out of existing normal matrices:

Proposition 2.5.5. Let $A \in \mathbb{C}^{n\times n}$ be a normal matrix.

(a) If $\lambda \in \mathbb{C}$ is arbitrary, then the matrix $\lambda I_n + A$ is normal.

(b) If $U \in \mathbb{C}^{n\times n}$ is a unitary matrix, then the matrix UAU^* is normal.

(c) The matrix A^* is normal.

Proof. We have $AA^* = A^*A$ (since A is normal).

(a) Let $\lambda \in \mathbb{C}$ be arbitrary. Then, Proposition 1.3.3 **(a)** yields

$$(\lambda I_n + A)^* = \underbrace{(\lambda I_n)^*}_{=\overline{\lambda} I_n} + A^* = \overline{\lambda} I_n + A^*.$$

(this is easily seen directly, or obtained from Proposition 1.3.3 **(b)**)

Hence,

$$(\lambda I_n + A)\underbrace{(\lambda I_n + A)^*}_{=\overline{\lambda} I_n + A^*} = (\lambda I_n + A)\left(\overline{\lambda} I_n + A^*\right)$$
$$= \lambda\overline{\lambda} I_n + \lambda A^* + \overline{\lambda} A + AA^*.$$

A similar computation shows that

$$(\lambda I_n + A)^*(\lambda I_n + A) = \overline{\lambda}\lambda I_n + \lambda A^* + \overline{\lambda} A + A^*A.$$

The right hand sides of these two equalities are equal (since $\lambda\overline{\lambda} = \overline{\lambda}\lambda$ and $AA^* = A^*A$). Hence, so are the left hand sides. In other words, $(\lambda I_n + A)(\lambda I_n + A)^* = (\lambda I_n + A)^*(\lambda I_n + A)$. In other words, the matrix $\lambda I_n + A$ is normal. This proves Proposition 2.5.5 **(a)**.

(b) Let $U \in \mathbb{C}^{n \times n}$ be a unitary matrix. Thus, $UU^* = U^*U = I_n$ (by the $\mathcal{A} \iff \mathcal{C}$ part of Theorem 1.5.3). Now, applying Proposition 1.3.3 **(c)** twice, we see that $(XYZ)^* = Z^*Y^*X^*$ for any three $n \times n$-matrices X, Y, Z. Hence,

$$(UAU^*)^* = \underbrace{(U^*)^*}_{=U} A^*U^* = UA^*U^*.$$

(by Proposition 1.3.3 **(d)**)

Hence,

$$(UAU^*) \underbrace{(UAU^*)^*}_{=UA^*U^*} = (UAU^*)(UA^*U^*) = UA \underbrace{U^*U}_{=I_n} A^*U^* = UAA^*U^*.$$

A similar computation shows that

$$(UAU^*)^* (UAU^*) = UA^*AU^*.$$

The right hand sides of these two equalities are equal (since $AA^* = A^*A$). Hence, so are the left hand sides. In other words, $(UAU^*)(UAU^*)^* = (UAU^*)^*(UAU^*)$. In other words, the matrix UAU^* is normal. This proves Proposition 2.5.5 **(b)**.

(c) This is left to the reader. □

Here is another normality-preserving way to transform matrices:

Definition 2.5.6. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be a square matrix. Let $p(x)$ be a polynomial in a single indeterminate x with coefficients in $\mathbb{F}$. Write $p(x)$ in the form $p(x) = a_0x^0 + a_1x^1 + \cdots + a_dx^d$, where $a_0, a_1, \ldots, a_d \in \mathbb{F}$.

Then, $p(A)$ denotes the matrix $a_0A^0 + a_1A^1 + \cdots + a_dA^d \in \mathbb{F}^{n \times n}$.

For instance, if $p(x) = x^3 - 2x^2 + 1$, then $p(A) = A^3 - 2A^2 + A^0 = A^3 - 2A^2 + I_n$.

Proposition 2.5.7. Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix. Let $p(x)$ be a polynomial in a single indeterminate x with coefficients in $\mathbb{C}$. Then, the matrix $p(A)$ is normal.

Exercise 2.5.2. [3] Prove Proposition 2.5.7.

Exercise 2.5.3. [2] Generalizing Proposition 2.5.5 **(b)**, we might claim the following:

Let $A \in \mathbb{C}^{k \times k}$ be a normal matrix. Let $U \in \mathbb{C}^{n \times k}$ be an isometry. Then, the matrix UAU^* is normal.

Is this generalization correct?

Exercise 2.5.4. [4] Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix. Prove the following:

(a) We have $||Ax|| = ||A^* x||$ for each $x \in \mathbb{C}^n$.

(b) We have $\operatorname{Ker} A = \operatorname{Ker}(A^*)$.

(c) Let $\lambda \in \mathbb{C}$. Then, the λ-eigenvectors of A are the $\overline{\lambda}$-eigenvectors of A^*.

Exercise 2.5.5. [4] **(a)** Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{C}^{m \times m}$ be two normal matrices, and $X \in \mathbb{C}^{n \times m}$. Prove that $AX = XB$ if and only if $A^* X = XB^*$. (This is known as the (finite) *Fuglede–Putnam theorem*.)

(b) Let $A \in \mathbb{C}^{n \times n}$ and $X \in \mathbb{C}^{n \times n}$ be two matrices such that A is normal. Prove that X commutes with A if and only if X commutes with A^*.

[**Hint:** For part **(a)**, set $C := AX - XB$ and $D := A^* X - XB^*$. Use Exercise 2.0.2 to show that $\operatorname{Tr}(C^* C) = \operatorname{Tr}(D^* D)$. Conclude using Exercise 2.0.3 **(b)**. Finally, observe that part **(b)** is a particular case of part **(a)**.]

Exercise 2.5.6. [5] Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{C}^{n \times n}$ be two normal matrices such that $AB = BA$. Prove that the matrices $A + B$ and AB are normal.

[**Hint:** Use Exercise 2.5.5 **(b)**.]

Exercise 2.5.7. [4] Let $A \in \mathbb{C}^{n \times n}$.

(a) Show that there is a **unique** pair (R, C) of Hermitian matrices R and C such that $A = R + iC$.

(b) Consider this pair (R, C). Show that A is normal if and only if R and C commute (that is, $RC = CR$).

[**Hint:** For part **(a)**, apply the "conjugate transpose" operation to $A = R + iC$ to obtain $A^* = R - iC$.]

We will now prove an innocent-looking property of normal matrices that will turn out crucial in characterizing them:

Lemma 2.5.8. Let $T \in \mathbb{C}^{n \times n}$ be a triangular matrix. Then, T is normal if and only if T is diagonal.

Proof. The "if" direction follows from Proposition 2.5.4 **(d)**. Thus, it remains to prove the "only if" direction.

So let us assume that T is normal. We shall show that T is diagonal.

The matrix T is normal; thus, T^* is normal as well (by Proposition 2.5.5 **(c)**). Since T is normal, we have $TT^* = T^* T$.

We have assumed that T is triangular. WLOG assume that T is upper-triangular (because otherwise, we can replace T by T^*). In other words,

$$T_{i,j} = 0 \qquad \text{for all } i, j \in [n] \text{ satisfying } i > j. \tag{40}$$

We must prove that the matrix T is diagonal. Assume the contrary. Thus, T has a nonzero off-diagonal entry[29]. Let i be the smallest element of $[n]$ such that the i-th row of T contains a nonzero off-diagonal entry. Hence, the i-th row of T contains a nonzero off-diagonal entry, but the 1-st, 2-nd, ..., $(i-1)$-st rows of T contain no such entries.

The definition of the product of two matrices yields that the (i,i)-th entry of T^*T is

$$\begin{aligned}
(T^*T)_{i,i} &= \sum_{k=1}^{n} \underbrace{(T^*)_{i,k}}_{=\overline{T_{k,i}}} T_{k,i} = \sum_{k=1}^{n} \overline{T_{k,i}} T_{k,i} \\
&\qquad \text{(by the definition of } T^*\text{)} \\
&= \sum_{k=1}^{i} \overline{T_{k,i}} T_{k,i} + \sum_{k=i+1}^{n} \overline{T_{k,i}} \underbrace{T_{k,i}}_{=0} \\
&\qquad \text{(by (40), applied to } k \text{ and } i \\
&\qquad \text{instead of } i \text{ and } j\text{)} \\
&= \sum_{k=1}^{i} \overline{T_{k,i}} T_{k,i} + \sum_{k=i+1}^{n} \overline{T_{k,i}} 0 = \sum_{k=1}^{i} \underbrace{\overline{T_{k,i}} T_{k,i}}_{=\left|T_{k,i}\right|^2} \\
&\qquad \text{(since } \overline{z}z = |z|^2 \text{ for each } z \in \mathbb{C}\text{)} \\
&= \sum_{k=1}^{i} |T_{k,i}|^2 .
\end{aligned}$$

A similar computation yields

$$(TT^*)_{i,i} = \sum_{k=i}^{n} |T_{i,k}|^2 .$$

The left hand sides of these two equalities are equal (since $T^*T = TT^*$). Thus, the right hand sides are equal as well. In other words, we have

$$\sum_{k=1}^{i} |T_{k,i}|^2 = \sum_{k=i}^{n} |T_{i,k}|^2 .$$

Both sides of this equality are sums with their $k = i$ addend equal to $|T_{i,i}|^2$. Thus, if we subtract $|T_{i,i}|^2$ from this equality, then both sums lose their $k = i$ addends, and we are left with

$$\sum_{k=1}^{i-1} |T_{k,i}|^2 = \sum_{k=i+1}^{n} |T_{i,k}|^2 . \tag{41}$$

Now, recall that the 1-st, 2-nd, ..., $(i-1)$-st rows of T contain no nonzero off-diagonal entries. In other words, if $k \in [i-1]$, then $T_{k,j} = 0$ for each $j \neq k$. Hence,

[29] An "off-diagonal entry" means an entry that does not lie on the diagonal.

in particular, if $k \in [i-1]$, then $T_{k,i} = 0$ (since $i \neq k$). Therefore, $\sum\limits_{k=1}^{i-1} \left| \underbrace{T_{k,i}}_{=0} \right|^2 = \sum\limits_{k=1}^{i-1} 0^2 = 0$. Comparing this with (41), we obtain

$$\sum_{k=i+1}^{n} |T_{i,k}|^2 = 0.$$

Therefore, all addends $|T_{i,i+1}|^2, |T_{i,i+2}|^2, \ldots, |T_{i,n}|^2$ in this sum are 0 (because a sum of nonnegative reals can only be 0 if all its addends are 0). In other words, all the numbers $T_{i,i+1}, T_{i,i+2}, \ldots, T_{i,n}$ are 0. Since all the numbers $T_{i,1}, T_{i,2}, \ldots, T_{i,i-1}$ are 0 as well (by (40)), we thus conclude that all the numbers $T_{i,1}, T_{i,2}, \ldots, T_{i,n}$ are 0 except for (possibly) $T_{i,i}$. In other words, the i-th row of T contains no nonzero off-diagonal entries. This contradicts the definition of i. Hence, we have obtained a contradiction, and our proof of Lemma 2.5.8 is complete. □

Exercise 2.5.8. [3] **(a)** Let $T \in \mathbb{C}^{n \times n}$ be an upper-triangular matrix. Prove that

$$\sum_{i=1}^{m} \left(TT^* - T^*T\right)_{i,i} = \sum_{i=1}^{m} \sum_{j=m+1}^{n} \left|T_{i,j}\right|^2$$

for each $m \in \{0, 1, \ldots, n\}$.

(b) Use this to give a direct proof (i.e., not a proof by contradiction) of Lemma 2.5.8.

For the next exercise, we recall the notion of a *nilpotent matrix*:

Definition 2.5.9. Let $\mathbb{F}$ be a field. A square matrix $A \in \mathbb{F}^{n \times n}$ is said to be *nilpotent* if there exists some nonnegative integer m such that $A^m = 0$.

For example, the matrix $\begin{pmatrix} 6 & 9 \\ -4 & -6 \end{pmatrix}$ is nilpotent, since $\begin{pmatrix} 6 & 9 \\ -4 & -6 \end{pmatrix}^2 = 0$. Also, every strictly upper-triangular matrix and every strictly lower-triangular matrix is nilpotent.

Exercise 2.5.9. [2] Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix that is nilpotent. Prove that $A = 0$.

Let us state another useful property of polynomials applied to matrices (Definition 2.5.6):

Exercise 2.5.10. [3] Let $A \in \mathbb{C}^{n \times n}$ be any matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A (listed with their algebraic multiplicities). Let $p(x)$ be a polynomial in a single indeterminate x with coefficients in $\mathbb{C}$.

Prove that the eigenvalues of the matrix $p(A)$ are $p(\lambda_1), p(\lambda_2), \ldots, p(\lambda_n)$ (listed with their algebraic multiplicities).

(This is known as the *spectral mapping theorem*.)

2.6. The spectral theorem

2.6.1. The spectral theorem for normal matrices

We are now ready to state the main theorem about normal matrices, the so-called *spectral theorem*:

Theorem 2.6.1 (spectral theorem for normal matrices)**.** Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix. Then:

(a) There exists a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and a diagonal matrix $D \in \mathbb{C}^{n \times n}$ such that

$$A = UDU^*.$$

In other words, A is unitarily similar to a diagonal matrix.

(b) Let $U \in \mathrm{U}_n(\mathbb{C})$ be a unitary matrix, and $D \in \mathbb{C}^{n \times n}$ be a diagonal matrix such that $A = UDU^*$. Then, the diagonal entries of D are the eigenvalues of A. Moreover, the columns of U are eigenvectors of A. Thus, there exists an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A.

Proof. **(a)** Theorem 2.3.1 yields that there exist a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and an upper-triangular matrix $T \in \mathbb{C}^{n \times n}$ such that $A = UTU^*$. Consider these U and T.

Since U is unitary, we have $U^*U = I_n$ and $UU^* = I_n$. The matrix U^* is unitary as well (since $U^* \underbrace{(U^*)^*}_{=U} = U^*U = I_n$ and $\underbrace{(U^*)^*}_{=U} U^* = UU^* = I_n$). Hence, Proposition 2.5.5 **(b)** (applied to U^* instead of U) yields that the matrix $U^* A (U^*)^*$ is normal. Since

$$U^* \underbrace{A}_{=UTU^*} \underbrace{(U^*)^*}_{=U} = \underbrace{U^*U}_{=I_n} T \underbrace{U^*U}_{=I_n} = I_n T I_n = T,$$

this rewrites as follows: The matrix T is normal. Since T is upper-triangular, we thus conclude by Lemma 2.5.8 that T is diagonal. Hence, if we set $D = T$, then we have constructed a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and a diagonal matrix $D \in \mathbb{C}^{n \times n}$ such that $A = UDU^*$. Hence, A is unitarily similar to a diagonal matrix. This proves Theorem 2.6.1 **(a)**.

(b) From $A = UDU^*$, we see that the matrix D is unitarily similar to A. Hence, D is similar to A (by Proposition 2.2.5). Moreover, the matrix D is upper-triangular (since it is diagonal). Thus, Proposition 2.3.7 (applied to $\mathbb{F} = \mathbb{C}$ and $T = D$)

yields that the diagonal entries of T are the eigenvalues of A (with their algebraic multiplicities).

Next, we shall show that the columns of U are eigenvectors of A. Indeed, from $A = UDU^*$, we obtain

$$AU = UD \qquad \underbrace{U^*U}_{\substack{=I_n \\ \text{(since } U \text{ is unitary)}}} \qquad = UD.$$

Now, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the diagonal entries of the diagonal matrix D. Let $i \in [n]$. Then, $De_i = \lambda_i e_i$ (where $(e_1, e_2, \ldots, e_n)$ denotes the standard basis of $\mathbb{C}^n$), since D is a diagonal matrix whose i-th diagonal entry is e_i. Therefore,

$$\underbrace{AU}_{=UD} e_i = U \underbrace{De_i}_{=\lambda_i e_i} = \lambda_i \cdot Ue_i.$$

This shows that Ue_i is an eigenvector of A (for eigenvalue λ_i). Since Ue_i is the i-th column of U, we can rewrite this as follows: The i-th column of U is an eigenvector of A.

Forget that we fixed i. We thus have shown that for each $i \in [n]$, the i-th column of U is an eigenvector of A. In other words, the columns of U are eigenvectors of A.

It remains to prove that there exists an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A. However, this is now easy: The matrix U is unitary. Thus, the columns of U form an orthonormal basis of $\mathbb{C}^n$ (by the implication $\mathcal{A} \Longrightarrow \mathcal{E}$ in Theorem 1.5.3). This basis consists of eigenvectors of A (since the columns of U are eigenvectors of A). Thus, there exists an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A (namely, this basis). This concludes the proof of Theorem 2.6.1 **(b)**. □

The decomposition $A = UDU^*$ in Theorem 2.4.3 (or, to be more precise, the pair (U, D)) is called a *spectral decomposition* of A. It is not unique (e.g., we can replace U by λU whenever $\lambda \in \mathbb{C}$ satisfies $|\lambda| = 1$; this does not change UDU^*). We can actually choose the order of the diagonal entries of D at will, as the following simple corollary shows:

Corollary 2.6.2. Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the n eigenvalues of A (listed with algebraic multiplicities, in an arbitrary order). Then, there exists a spectral decomposition (U, D) of A with $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Thus,

$$A \overset{\text{us}}{\sim} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{42}$$

Proof. Let $L = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. This is clearly a diagonal matrix.

Theorem 2.6.1 **(a)** yields that there exist a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and a diagonal matrix $D \in \mathbb{C}^{n \times n}$ such that $A = UDU^*$. Consider these U and D, and

denote them by W and F (since they are not yet the U and D that we are looking for). Thus, $W \in \mathrm{U}_n(\mathbb{C})$ is a unitary matrix and $F \in \mathbb{C}^{n \times n}$ is a diagonal matrix such that

$$A = WFW^*.$$

The definition of unitary similarity yields $A \overset{\text{us}}{\sim} F$ (since W is unitary and $A = WFW^*$). However, the diagonal entries of F are the eigenvalues of A (by Theorem 2.6.1 **(b)**, applied to $U = W$ and $D = F$). Since the eigenvalues of A are $\lambda_1, \lambda_2, \ldots, \lambda_n$, this shows that the diagonal entries of F are $\lambda_1, \lambda_2, \ldots, \lambda_n$ in some order. In other words, there exists a permutation σ of $[n]$ such that the diagonal entries of F are $\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}$. Consider this σ.

The matrix F is a diagonal matrix, and its diagonal entries are $\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}$. In other words, $F = \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$.

However, Proposition 2.2.7 yields $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \overset{\text{us}}{\sim} \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$. This rewrites as $L \overset{\text{us}}{\sim} F$ (since $L = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $F = \operatorname{diag}\left(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \ldots, \lambda_{\sigma(n)}\right)$). In other words, there exists a unitary matrix $Q \in \mathrm{U}_n(\mathbb{C})$ such that

$$F = QLQ^*.$$

Consider this Q. Now, the matrix WQ is unitary (by Exercise 1.5.2 **(b)**, since W and Q are unitary), and we have

$$A = W \underbrace{F}_{=QLQ^*} W^* = WQL \underbrace{Q^*W^*}_{=(WQ)^*} = WQL\,(WQ)^* = (WQ) \cdot L \cdot (WQ)^*.$$

This shows that (WQ, L) is a spectral decomposition of A (since WQ is unitary and L is diagonal). This spectral decomposition satisfies $L = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Thus, there exists a spectral decomposition (U, D) of A with $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ (namely, (WQ, L)). This furthermore shows that $A \overset{\text{us}}{\sim} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Corollary 2.6.2 is thus proven. □

Note that Theorem 2.6.1 **(b)** has a converse, which helps finding spectral decompositions in practice if one doesn't want to go through the trouble of Schur triangularization:

Proposition 2.6.3. Let $A \in \mathbb{C}^{n \times n}$. Let $U \in \mathrm{U}_n(\mathbb{C})$ be a unitary matrix and $D \in \mathbb{C}^{n \times n}$ a diagonal matrix. Assume that for each $i \in [n]$, we have $AU_{\bullet,i} = D_{i,i}U_{\bullet,i}$ (that is, the i-th column of U is an eigenvector of A for the eigenvalue $D_{i,i}$). Then, $A = UDU^*$, so that (U, D) is a spectral decomposition of A.

Exercise 2.6.1. [2] Prove Proposition 2.6.3.

Exercise 2.6.2. [5] **(a)** Find a spectral decomposition of the normal matrix $\begin{pmatrix} 1 & 1+i \\ 1+i & 1 \end{pmatrix}$.

(b) Find a spectral decomposition of the Hermitian matrix $\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$.

(c) Find a spectral decomposition of the skew-Hermitian matrix $\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$.

(d) Find a spectral decomposition of the unitary matrix $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

Exercise 2.6.3. [2] Describe all spectral decompositions of the $n \times n$ identity matrix I_n.

Only normal matrices can have a spectral decomposition. Indeed, if some $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ can be written as $A = UDU^*$ for some unitary U and some diagonal D, then D is normal (by Proposition 2.5.4 **(d)**), and therefore A is normal (by Proposition 2.5.5 **(b)**, applied to D instead of A). Thus, we obtain the following characterization of normal matrices:

Corollary 2.6.4. An $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ is normal if and only if it is unitarily similar to a diagonal matrix.

Proof. $\Longrightarrow$: Assume that A is normal. Then, Theorem 2.6.1 **(a)** shows that A is unitarily similar to a diagonal matrix. This proves the "$\Longrightarrow$" direction of Corollary 2.6.4.

$\Longleftarrow$: Assume that A is unitarily similar to a diagonal matrix. In other words, $A = UDU^*$ for some unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and some diagonal matrix $D \in \mathbb{C}^{n \times n}$. Consider these U and D. The matrix D is normal (by Proposition 2.5.4 **(d)**). Hence, the matrix UDU^* is normal (by Proposition 2.5.5 **(b)**, applied to D instead of A). In other words, the matrix A is normal (since $A = UDU^*$). This proves the "$\Longleftarrow$" direction of Corollary 2.6.4. □

Exercise 2.6.4. [3] Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{C}^{n \times n}$ be two normal matrices such that $A \sim B$. Prove that $A \overset{\mathrm{us}}{\sim} B$.

2.6.2. The spectral theorem for Hermitian matrices

The spectral decompositions of a Hermitian matrix have a special property:

Proposition 2.6.5. Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix, and let (U, D) be a spectral decomposition of A. Then, the diagonal entries of D are real.

Proof. The definition of a spectral decomposition yields that U is unitary and D is diagonal and $A = UDU^*$. However, since A is Hermitian, we have $A^* = A$. In view of $A = UDU^*$, this rewrites as $(UDU^*)^* = UDU^*$. Hence,

$$UDU^* = (UDU^*)^* = \underbrace{(U^*)^*}_{=U} D^* U^* = UD^*U^*.$$

Since the matrix U is unitary, we can cancel both U and U^* from this equality[30], and thus obtain $D = D^*$. However, if λ is a diagonal entry of D, then the corresponding diagonal entry of D^* must be $\overline{\lambda}$, and therefore we obtain $\lambda = \overline{\lambda}$ (since $D = D^*$ shows that these two entries are equal). Thus, each diagonal entry λ of D satisfies $\lambda = \overline{\lambda}$ and therefore $\lambda \in \mathbb{R}$ (because a complex number z satisfying $z = \overline{z}$ must automatically satisfy $z \in \mathbb{R}$). In other words, the diagonal entries of D are real. This proves Proposition 2.6.5. □

This allows us to characterize Hermitian matrices in a similar way as normal matrices were characterized by Corollary 2.6.4:

Corollary 2.6.6. An $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ is Hermitian if and only if it is unitarily similar to a diagonal matrix with real entries.

Proof. $\Longrightarrow$: Assume that A is Hermitian. Then, A is normal (by Proposition 2.5.4 **(a)**). Hence, Theorem 2.6.1 **(a)** shows that A is unitarily similar to a diagonal matrix. In other words, $A = UDU^*$ for some unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and some diagonal matrix $D \in \mathbb{C}^{n \times n}$. Consider these U and D. Clearly, A is unitarily similar to D. Moreover, (U, D) is a spectral decomposition of A (by the definition of a spectral decomposition). Hence, Proposition 2.6.5 yields that the diagonal entries of D are real. Thus, D is a diagonal matrix with real entries. Hence, A is unitarily similar to a diagonal matrix with real entries (since A is unitarily similar to D). This proves the "$\Longrightarrow$" direction of Corollary 2.6.6.

$\Longleftarrow$: Assume that A is unitarily similar to a diagonal matrix with real entries. In other words, $A = UDU^*$ for some unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and some diagonal matrix $D \in \mathbb{C}^{n \times n}$ that has real entries. Consider these U and D. The matrix D is a diagonal matrix with real entries; thus, it is easy to see that $D^* = D$ (since the diagonal entries of D are real and thus remain unchanged and unmoved in D^*, whereas all other entries of D are 0). Now, from $A = UDU^*$, we obtain

$$A^* = (UDU^*)^* = \underbrace{(U^*)^*}_{=U} \underbrace{D^*}_{=D} U^* = UDU^* = A.$$

In other words, the matrix A is Hermitian. This proves the "$\Longleftarrow$" direction of Corollary 2.6.6. □

[30] Indeed, the matrix U is unitary; therefore, U is invertible, and its inverse is $U^{-1} = U^*$. Hence, U^* is also invertible (being the inverse of U). Thus, we can multiply both sides of the equality $UDU^* = UD^*U^*$ from the left by U^{-1} and from the right by $(U^*)^{-1}$; as a result, we obtain $D = D^*$.

2.6.3. The spectral theorem for skew-Hermitian matrices

Similarly, we can handle skew-Hermitian matrices:

Proposition 2.6.7. Let $A \in \mathbb{C}^{n\times n}$ be a skew-Hermitian matrix, and let (U, D) be a spectral decomposition of A. Then, the diagonal entries of D are purely imaginary.

Corollary 2.6.8. An $n \times n$-matrix $A \in \mathbb{C}^{n\times n}$ is skew-Hermitian if and only if it is unitarily similar to a diagonal matrix with purely imaginary entries.

Exercise 2.6.5. [3] Prove Proposition 2.6.7 and Corollary 2.6.8.

2.6.4. The spectral theorem for unitary matrices

Likewise, we can handle unitary matrices:

Proposition 2.6.9. Let $A \in \mathbb{C}^{n\times n}$ be a unitary matrix, and let (U, D) be a spectral decomposition of A. Then, each of the diagonal entries of D has absolute value 1.

Corollary 2.6.10. An $n \times n$-matrix $A \in \mathbb{C}^{n\times n}$ is unitary if and only if it is unitarily similar to a diagonal matrix whose all diagonal entries have absolute value 1.

Exercise 2.6.6. [2] Prove Proposition 2.6.9 and Corollary 2.6.10.

Exercise 2.6.7. [2] Prove the following generalization of Theorem 2.6.1:

Let $\mathcal{F}$ be a subset of $\mathbb{C}^{n\times n}$ such that any matrix in $\mathcal{F}$ is normal, and such that any two matrices in $\mathcal{F}$ commute (i.e., any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ satisfy $AB = BA$).

Then, there exists a unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ such that for each $A \in \mathcal{F}$, the matrix UAU^* is diagonal.

Lecture 6 starts here.

2.7. The Cayley–Hamilton theorem

We will now state the famous Cayley–Hamilton theorem, and to prove it at least for matrices with complex entries. This will serve as a reminder of an important theorem (which will soon be used), and also as an illustration of how Schur triangularization can be applied.

In Definition 2.5.6, we have learnt how to substitute a square matrix into a polynomial. Something peculiar happens when a matrix is substituted into its own characteristic polynomial:

Theorem 2.7.1 (Cayley–Hamilton theorem)**.** Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. Then,

$$p_A(A) = 0.$$

(The "0" on the right hand side here means the zero matrix $0_{2\times 2}$.)

Example 2.7.2. Let $n = 2$ and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then, as we know from Example 2.0.2, we have

$$p_A = t^2 - (a+d)\,t + (ad - bc)\,.$$

Thus,

$$\begin{aligned} p_A(A) &= A^2 - (a+d)\,A + (ad-bc)\,I_2 \\ &= \begin{pmatrix} a & b \\ c & d \end{pmatrix}^2 - (a+d)\begin{pmatrix} a & b \\ c & d \end{pmatrix} + (ad-bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0. \end{aligned}$$

Thus, we have verified Theorem 2.7.1 for $n = 2$.

Remark 2.7.3. It is tempting to "prove" Theorem 2.7.1 by arguing that $p_A(A) = \det(AI_n - A)$ holds "by substituting A for t into $p_A = \det(tI_n - A)$". Unfortunately, such an argument is unjustified. Indeed, $tI_n - A$ is a matrix whose entries are polynomials in t. If you substitute A for t into it, it will become a matrix whose entries are matrices. This poses two problems: First, it is unclear how to take the determinant of such a matrix; second, this matrix is not $AI_n - A$. For example, for $n = 2$, substituting A for t in $tI_n - A$ gives

$$\begin{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} - a & -b \\ -c & \begin{pmatrix} a & b \\ c & d \end{pmatrix} - d \end{pmatrix},$$

which can be made sense of (if we treat the a, b, c, d as multiples of I_2), but which is certainly not the same as $AI_n - A$ (which is the zero matrix). There **is** a correct proof of the Cayley–Hamilton theorem along the lines of "substituting A for t", but it requires a lot of additional work (see `https://math.stackexchange.com/questions/1141648/` for some discussion of this).

Various proofs of Theorem 2.7.1 are found across the literature; see [Grinbe19, after Theorem 2.6] for a list of references (Theorem 2.7.1 is [Grinbe19, Theorem 2.5]). I can particularly recommend the algebraic proofs given in [Heffer20, Chapter Five, Section IV, Lemma 1.9], [Mate16, §4, Theorem 1] and [Shurma15], and the

combinatorial proof shown in [Straub83] and [Zeilbe85, §3]. Here, however, I will show a proof of Theorem 2.7.1 in the particular case when $\mathbb{F} = \mathbb{C}$.

This proof will rely on two lemmas. The first collects some useful properties of the application of polynomials to matrices:

Lemma 2.7.4. Let $n \in \mathbb{N}$. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. The word "polynomial" shall mean "polynomial in an indeterminate t with coefficients in $\mathbb{F}$". Then:

(a) If f and g are two polynomials, then $(fg)(A) = f(A) \cdot g(A)$.

(b) If $f_1, f_2, \ldots, f_k$ are several polynomials, then $(f_1 f_2 \cdots f_k)(A) = f_1(A) \cdot f_2(A) \cdot \cdots \cdot f_k(A)$.

(c) If f is a polynomial, and $W \in \mathbb{F}^{n \times n}$ is an invertible matrix, then $f\left(WAW^{-1}\right) = W \cdot f(A) \cdot W^{-1}$.

Proof of Lemma 2.7.4. **(a)** Let f and g be two polynomials. Write f in the form $f = \sum_{i=0}^{p} f_i t^i$ for some coefficients $f_0, f_1, \ldots, f_p \in \mathbb{F}$. Write g in the form $g = \sum_{j=0}^{q} g_j t^j$ for some coefficients $g_0, g_1, \ldots, g_q \in \mathbb{F}$. Definition 2.5.6 yields

$$f(A) = \sum_{i=0}^{p} f_i A^i \qquad \left(\text{since } f = \sum_{i=0}^{p} f_i t^i\right)$$

and

$$g(A) = \sum_{j=0}^{q} g_j A^j \qquad \left(\text{since } g = \sum_{j=0}^{q} g_j t^j\right).$$

Multiplying these two equalities, we obtain

$$\begin{aligned} f(A) \cdot g(A) &= \left(\sum_{i=0}^{p} f_i A^i\right) \cdot \left(\sum_{j=0}^{q} g_j A^j\right) = \sum_{i=0}^{p} \sum_{j=0}^{q} f_i g_j \underbrace{A^i A^j}_{=A^{i+j}} \\ &= \sum_{i=0}^{p} \sum_{j=0}^{q} f_i g_j A^{i+j}. \end{aligned} \tag{43}$$

Multiplying the equalities $f = \sum_{i=0}^{p} f_i t^i$ and $g = \sum_{j=0}^{q} g_j t^j$, we obtain

$$fg = \left(\sum_{i=0}^{p} f_i t^i\right)\left(\sum_{j=0}^{q} g_j t^j\right) = \sum_{k=0}^{p+q} \left(\sum_{\substack{i \in \{0,1,\ldots,k\}; \\ i \leq p;\ k-i \leq q}} f_i g_{k-i}\right) t^k$$

(by the definition of the product of two polynomials). Hence, Definition 2.5.6 yields

$$(fg)(A) = \sum_{k=0}^{p+q} \left(\sum_{\substack{i \in \{0,1,\ldots,k\}; \\ i \leq p;\ k-i \leq q}} f_i g_{k-i} \right) A^k = \underbrace{\sum_{k=0}^{p+q} \sum_{\substack{i \in \{0,1,\ldots,k\}; \\ i \leq p;\ k-i \leq q}}}_{= \sum_{i=0}^{p} \sum_{k=i}^{i+q}} f_i g_{k-i} A^k$$

(because both of these double sums are summing over all pairs (k,i) of nonnegative integers satisfying $i \leq k$ and $i \leq p$ and $k \leq i+q$)

$$= \sum_{i=0}^{p} \sum_{k=i}^{i+q} f_i g_{k-i} A^k = \sum_{i=0}^{p} \sum_{j=0}^{q} f_i \underbrace{g_{(i+j)-i}}_{=g_j} A^{i+j}$$

(here, we have substituted $i+j$ for k in the inner sum)

$$= \sum_{i=0}^{p} \sum_{j=0}^{q} f_i g_j A^{i+j} = f(A) \cdot g(A) \qquad \text{(by (43))}.$$

This proves Lemma 2.7.4 **(a)**.

(b) Lemma 2.7.4 **(b)** follows by induction on k. (The base case relies on $1(A) = I_n$, where 1 denotes the constant polynomial 1. The induction step uses Lemma 2.7.4 **(a)**.)

(c) Let f be a polynomial, and let $W \in \mathbb{F}^{n \times n}$ be an invertible matrix. Let $B := WAW^{-1}$. Write the polynomial f in the form $f = \sum_{k=0}^{p} f_k t^k$ for some coefficients $f_0, f_1, \ldots, f_p \in \mathbb{F}$. Thus, Definition 2.5.6 yields

$$f(A) = \sum_{k=0}^{p} f_k A^k \qquad \text{and} \tag{44}$$

$$f(B) = \sum_{k=0}^{p} f_k B^k. \tag{45}$$

However, for each $k \in \mathbb{N}$, we have

$$B^k = WA^kW^{-1} \tag{46}$$

(indeed, this is precisely the equality (30), which we have proved long ago). There-

fore, (45) becomes

$$f(B) = \sum_{k=0}^{p} f_k \underbrace{B^k}_{\substack{=WA^kW^{-1} \\ \text{(by (46))}}} = \sum_{k=0}^{p} f_k W A^k W^{-1}$$

$$= W \cdot \sum_{k=0}^{p} f_k A^k W^{-1} = W \cdot \underbrace{\left(\sum_{k=0}^{p} f_k A^k \right)}_{\substack{=f(A) \\ \text{(by (44))}}} \cdot W^{-1} = W \cdot f(A) \cdot W^{-1}.$$

This proves Lemma 2.7.4 **(c)**. □

The second lemma is an easy but neat property of triangular matrices:

Lemma 2.7.5. Let $n \in \mathbb{N}$. Let $\mathbb{F}$ be a field. Let $T_1, T_2, \ldots, T_n$ be n upper-triangular $n \times n$-matrices. Assume that for each $i \in [n]$, the i-th diagonal entry of the matrix T_i is 0 (that is, we have $(T_i)_{i,i} = 0$). Then,

$$T_1 T_2 \cdots T_n = 0.$$

(The 0 on the right hand side here is the zero matrix.)

Example 2.7.6. For $n = 3$, Theorem 2.7.5 is saying the following: If T_1, T_2, T_3 are three 3×3-matrices of the form

$$T_1 = \begin{pmatrix} 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{pmatrix}, \qquad T_2 = \begin{pmatrix} * & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{pmatrix}, \qquad T_3 = \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & 0 \end{pmatrix}$$

(where each asterisk "$*$" stands for an arbitrary entry – not necessarily equal to the other asterisk entries), then $T_1 T_2 T_3 = 0$.

Proof of Lemma 2.7.5. We claim that

$$\text{the first } k \text{ columns of the matrix } T_1 T_2 \cdots T_k \text{ are } 0 \tag{47}$$

for each $k \in \{0, 1, \ldots, n\}$.

[*Proof of (47):* We shall prove (47) by induction on k:

Base case: The first 0 columns of any matrix are 0 (indeed, this is vacuously true). Thus, (47) holds for $k = 0$.

Induction step: Let $p \in [n]$. Assume that (47) holds for $k = p - 1$. We must prove that (47) holds for $k = p$.

Let $A = T_1 T_2 \cdots T_{p-1}$ and $B = T_p$. Thus,

$$AB = \left(T_1 T_2 \cdots T_{p-1}\right) T_p = T_1 T_2 \cdots T_p. \tag{48}$$

We have assumed that (47) holds for $k = p-1$. In other words, the first $p-1$ columns of the matrix $T_1 T_2 \cdots T_{p-1}$ are 0. In other words, the first $p-1$ columns of the matrix A are 0 (since $A = T_1 T_2 \cdots T_{p-1}$). In other words,

$$A_{i,j} = 0 \qquad \text{for each } i \in [n] \text{ and } j \in [p-1]. \tag{49}$$

On the other hand, the assumption of Lemma 2.7.5 yields that the p-th diagonal entry of the matrix T_p is 0. In other words, $\left(T_p\right)_{p,p} = 0$. In other words, $B_{p,p} = 0$ (since $B = T_p$). Moreover, the matrix T_p is upper-triangular (by the assumption of Lemma 2.7.5). In other words, the matrix B is upper-triangular (since $B = T_p$). In other words,

$$B_{i,j} = 0 \qquad \text{for each } i, j \in [n] \text{ satisfying } i > j. \tag{50}$$

Now, let $i \in [n]$ and $j \in [p]$ be arbitrary. We shall show that $(AB)_{i,j} = 0$.

Indeed, the definition of the product of two matrices yields

$$\begin{aligned}
(AB)_{i,j} &= \sum_{k=1}^{n} A_{i,k} B_{k,j} = \sum_{k=1}^{p-1} \underbrace{A_{i,k}}_{\substack{=0 \\ \text{(by (49), applied to } k \text{ instead of } j \\ \text{(since } k \in [p-1]\text{))}}} B_{k,j} + \sum_{k=p}^{n} A_{i,k} B_{k,j} \\
&= \underbrace{\sum_{k=1}^{p-1} 0 B_{k,j}}_{=0} + \sum_{k=p}^{n} A_{i,k} B_{k,j} = \sum_{k=p}^{n} A_{i,k} B_{k,j}.
\end{aligned} \tag{51}$$

If $p > j$, then this becomes

$$(AB)_{i,j} = \sum_{k=p}^{n} A_{i,k} \underbrace{B_{k,j}}_{\substack{=0 \\ \text{(by (50), applied to } k \text{ instead of } i \\ \text{(since } k \geq p > j\text{))}}} = \sum_{k=p}^{n} A_{i,k} 0 = 0,$$

and therefore $(AB)_{i,j} = 0$ has been proved in this case. Hence, for the rest of the proof of $(AB)_{i,j} = 0$, we WLOG assume that we don't have $p > j$. Thus, $j \geq p$, so that $j = p$ (since $j \in [p]$). In other words, $p = j$. Now, (51) becomes

$$\begin{aligned}
(AB)_{i,j} &= \sum_{k=p}^{n} A_{i,k} B_{k,j} = A_{i,p} \underbrace{B_{p,j}}_{\substack{=B_{p,p} \\ \text{(since } j = p\text{)}}} + \sum_{k=p+1}^{n} A_{i,k} \underbrace{B_{k,j}}_{\substack{=0 \\ \text{(by (50), applied to } k \text{ instead of } i \\ \text{(since } k \geq p+1 > p = j\text{))}}} \\
&\qquad \text{(here, we have split off the addend for } k = p \text{ from the sum)} \\
&= A_{i,p} \underbrace{B_{p,p}}_{=0} + \underbrace{\sum_{k=p+1}^{n} A_{i,k} 0}_{=0} = 0.
\end{aligned}$$

Thus, we have proved that $(AB)_{i,j} = 0$.

Forget that we fixed i and j. We thus have shown that $(AB)_{i,j} = 0$ for all $i \in [n]$ and $j \in [p]$. In other words, the first p columns of the matrix AB are 0. In view of (48), we can rewrite this as follows: The first p columns of the matrix $T_1 T_2 \cdots T_p$ are 0. In other words, (47) holds for $k = p$. This completes the induction step. Thus, (47) is proven.]

Now, we can apply (47) to $k = n$, and conclude that the first n columns of the matrix $T_1 T_2 \cdots T_n$ is 0. Since this matrix $T_1 T_2 \cdots T_n$ has only n columns, this means that all of its columns are 0. In other words, the entire matrix $T_1 T_2 \cdots T_n$ is 0. This proves Lemma 2.7.5. □

Proof of Theorem 2.7.1 for $\mathbb{F} = \mathbb{C}$. Assume that $\mathbb{F} = \mathbb{C}$. The Schur triangularization theorem (Theorem 2.3.1) shows that A is unitarily similar to an upper-triangular matrix. Hence, A is similar to an upper-triangular matrix (because unitarily similar matrices always are similar). In other words, there exist an invertible matrix U and an upper-triangular matrix T such that $A = UTU^{-1}$. Consider these U and T.

From $A = UTU^{-1}$, we obtain

$$p_A(A) = p_A\left(UTU^{-1}\right) = U \cdot p_A(T) \cdot U^{-1}$$

(by Lemma 2.7.4 **(c)**, applied to p_A, U and T instead of f, W and A). Hence, in order to prove that $p_A(A) = 0$, it will suffice to show that $p_A(T) = 0$.

Now, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the diagonal entries of T. Then, by Proposition 2.3.7, these diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A (with algebraic multiplicities). Hence,

$$p_A = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_n)$$

(since p_A is monic, and the roots of p_A are precisely the eigenvalues of A with algebraic multiplicities). Therefore,

$$\begin{aligned} p_A(T) &= ((t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_n))(T) \\ &= (T - \lambda_1 I_n)(T - \lambda_2 I_n) \cdots (T - \lambda_n I_n) \end{aligned} \tag{52}$$

(by Lemma 2.7.4 **(b)**, applied to n, T and $t - \lambda_i$ instead of k, A and f_i).

We have

$$T_{i,i} = \lambda_i \qquad \text{for each } i \in [n] \tag{53}$$

(since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the diagonal entries of T).

However, the n matrices $T - \lambda_1 I_n, T - \lambda_2 I_n, \ldots, T - \lambda_n I_n$ are upper-triangular (since they are linear combinations of the upper-triangular matrices T and I_n). Moreover, for each $i \in [n]$, the i-th diagonal entry of the matrix $T - \lambda_i I_n$ is 0 (because this entry is $(T - \lambda_i I_n)_{i,i} = \underbrace{T_{i,i}}_{\substack{=\lambda_i \\ \text{(by (53))}}} - \lambda_i \underbrace{(I_n)_{i,i}}_{=1} = \lambda_i - \lambda_i = 0$). Thus, Lemma 2.7.5 (applied to $T_i = T - \lambda_i I_n$) yields

$$(T - \lambda_1 I_n)(T - \lambda_2 I_n) \cdots (T - \lambda_n I_n) = 0.$$

In view of (52), this rewrites as $p_A(T) = 0$. As explained above, this entails $p_A(A) = 0$. Thus, Theorem 2.7.1 is proved under the assumption that $\mathbb{F} = \mathbb{C}$. $\square$

The Cayley–Hamilton theorem has an interesting consequence: it yields that the inverse of an invertible matrix can be written as a polynomial applied to this matrix. (However, the specific polynomial that needs to be applied depends on this matrix.) In more detail:

Exercise 2.7.1. [3] Let $\mathbb{F}$ be a field. Let n be a positive integer. Let $A \in \mathbb{F}^{n \times n}$ be an invertible matrix with entries in $\mathbb{F}$. Prove that there exists a polynomial f of degree $n-1$ in the single indeterminate t over $\mathbb{F}$ such that $A^{-1} = f(A)$.

For example, for $n = 2$, we have $A^{-1} = uI_2 - vA$ with $u = \dfrac{\operatorname{Tr} A}{\det A}$ and $v = \dfrac{1}{\det A}$.

Another consequence of Cayley–Hamilton is that the powers of a given square matrix $A \in \mathbb{F}^{n \times n}$ span a vector space of dimension $\leq n$:

Exercise 2.7.2. [3] Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be a square matrix with entries in $\mathbb{F}$. Prove that for any nonnegative integer k, the power A^k can be written as an $\mathbb{F}$-linear combination of the first n powers $A^0, A^1, \ldots, A^{n-1}$.

Yet another rather curious consequence is an application to linearly recurrent sequences. We recall what these are:

Definition 2.7.7. Let $a_1, a_2, \ldots, a_k$ be k numbers. A sequence $(x_0, x_1, x_2, \ldots)$ of numbers is said to be $(a_1, a_2, \ldots, a_k)$-*recurrent* if each integer $i \geq k$ satisfies

$$x_i = a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_k x_{i-k}.$$

For instance, the famous Fibonacci sequence $(f_0, f_1, f_2, \ldots)$ (defined by the starting values $f_0 = 0$ and $f_1 = 1$ and the recurrence $f_i = f_{i-1} + f_{i-2}$) is $(1, 1)$-recurrent (by its very definition). Now, it is a simple exercise to check that the "even-indexed Fibonacci sequence" $(f_0, f_2, f_4, f_6, \ldots)$ and the "odd-indexed Fibonacci sequence" $(f_1, f_3, f_5, f_7, \ldots)$ themselves follow a simple recursion; to wit, they are both $(3, -1)$-recurrent (check this!). Likewise, the "multiples-of-3-indexed Fibonacci sequence" $(f_0, f_3, f_6, f_9, \ldots)$ as well as its companions $(f_1, f_4, f_7, f_{10}, \ldots)$ and $(f_2, f_5, f_8, f_{11}, \ldots)$ are $(4, 1)$-recurrent. This generalizes:

Exercise 2.7.3. [5] Let $a_1, a_2, \ldots, a_k$ be k numbers. Let $(x_0, x_1, x_2, \ldots)$ be any $(a_1, a_2, \ldots, a_k)$-recurrent sequence of numbers. Let d be a positive integer. Show that there exist k integers $b_1, b_2, \ldots, b_k$ such that each $i \geq kd$ satisfies

$$x_i = b_1 x_{i-d} + b_2 x_{i-2d} + \cdots + b_k x_{i-kd}.$$

(This means that the sequences $(x_{0+u}, x_{d+u}, x_{2d+u}, x_{3d+u}, \ldots)$ are $(b_1, b_2, \ldots, b_k)$-recurrent for all $u \geq 0$.)

[**Hint:** For each $j \geq 0$, define the column vector v_j by $v_j = \begin{pmatrix} x_j \\ x_{j+1} \\ \vdots \\ x_{j+k-1} \end{pmatrix} \in \mathbb{R}^k$. Let A be the $k \times k$-matrix $\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_k & a_{k-1} & a_{k-2} & \cdots & a_1 \end{pmatrix} \in \mathbb{R}^{k \times k}$. Start by showing that $Av_j = v_{j+1}$ for each $j \geq 0$.]

2.8. Sylvester's equation

We shall next see another application of the Cayley–Hamilton theorem. First, a notation:

Definition 2.8.1. Let $A \in \mathbb{C}^{n \times n}$. Then, the *spectrum* of A is defined to be the set of all eigenvalues of A. This spectrum is denoted by $\sigma(A)$.

Some authors write $\operatorname{spec} A$ instead of $\sigma(A)$. (Some also define it to be a multiset rather than a set; however, the set suffices for our purposes.)

We now claim the following:

Theorem 2.8.2. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix, and let $B \in \mathbb{C}^{m \times m}$ be an $m \times m$-matrix (both with complex entries). Let $C \in \mathbb{C}^{n \times m}$ be an $n \times m$-matrix. Then, the following two statements are equivalent:

- $\mathcal{U}$: There is a **unique** matrix $X \in \mathbb{C}^{n \times m}$ such that $AX - XB = C$.
- $\mathcal{V}$: We have $\sigma(A) \cap \sigma(B) = \varnothing$.

Example 2.8.3. Let us take $n = 1$ and $m = 1$, and see what Theorem 2.8.2 becomes. In this case, the matrices A, B and C are 1×1-matrices, so we can view them as scalars. Let us therefore write a, b and c for them. Then, Theorem 2.8.2 says that the following two statements are equivalent:

- $\mathcal{U}$: There is a **unique** complex number x such that $ax - xb = c$.
- $\mathcal{V}$: We have $\{a\} \cap \{b\} = \varnothing$ (that is, $a \neq b$).

This is not surprising, because the linear equation $ax - xb = c$ has a unique solution (namely, $x = \dfrac{c}{a-b}$) when $a \neq b$, and otherwise has either none or infinitely many solutions.

The equation $AX - XB = C$ in Theorem 2.8.2 is known as *Sylvester's equation*. It is much harder than the superficially similar equations $AX - BX = C$ and $XA - XB = C$ (see Exercise 2.8.1 for the first of these). In fact, since the X is on different sides in AX and in XB, it cannot be factored out from $AX - XB$ (matrices do not generally commute).

Exercise 2.8.1. $\boxed{2}$ Let $A \in \mathbb{C}^{n \times m}$, $B \in \mathbb{C}^{n \times m}$ and $C \in \mathbb{C}^{n \times p}$ be three complex matrices. Prove that there exists a matrix $X \in \mathbb{C}^{m \times p}$ such that $AX - BX = C$ if and only if each column of C belongs to the image (= column space) of $A - B$.

We shall prove only the $\mathcal{V} \Longrightarrow \mathcal{U}$ part of Theorem 2.8.2; the opposite direction will be left as an exercise (Exercise 2.8.2 **(b)**). Our proof of $\mathcal{V} \Longrightarrow \mathcal{U}$ will rely on the following lemma:

Lemma 2.8.4. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$, $B \in \mathbb{F}^{m \times m}$ and $X \in \mathbb{F}^{n \times m}$ be three matrices such that $AX = XB$. Then:

(a) We have $A^k X = XB^k$ for each $k \in \mathbb{N}$.

(b) Let p be a polynomial in a single indeterminate x with coefficients in $\mathbb{F}$. Then, $p(A)\, X = Xp(B)$.

Proof of Lemma 2.8.4. **(a)** Intuitively, this is easy: For instance, if $k = 4$, then this is saying that $A^4 X = XB^4$, but this follows from

$$A^4 B = AAA \underbrace{AB}_{=BA} = AA \underbrace{AB}_{=BA} A = A \underbrace{AB}_{=BA} AA = \underbrace{AB}_{=BA} AAA = BAAAA = BA^4.$$

Formally, Lemma 2.8.4 **(a)** is proved by induction on k:

Induction base: We have $A^0 X = XB^0$, since both sides of this equation equal X. Thus, Lemma 2.8.4 **(a)** holds for $k = 0$.

Induction step: Let $\ell \in \mathbb{N}$. Assume (as the induction hypothesis) that Lemma 2.8.4 **(a)** holds for $k = \ell$. We must prove that Lemma 2.8.4 **(a)** holds for $k = \ell + 1$.

We have assumed that Lemma 2.8.4 **(a)** holds for $k = \ell$. In other words, $A^\ell X = XB^\ell$. Thus,

$$\underbrace{A^{\ell+1}}_{=AA^\ell} X = A \underbrace{A^\ell X}_{=XB^\ell} = \underbrace{AX}_{=XB} B^\ell = X \underbrace{BB^\ell}_{=B^{\ell+1}} = XB^{\ell+1}.$$

In other words, Lemma 2.8.4 **(a)** holds for $k = \ell + 1$. This completes the induction step; thus, Lemma 2.8.4 **(a)** is proven.

(b) Write the polynomial p in the form $p(x) = \sum_{k=0}^{d} p_k x^k$ for some coefficients $p_0, p_1, \ldots, p_d \in \mathbb{F}$. Then, Definition 2.5.6 yields

$$p(A) = \sum_{k=0}^{d} p_k A^k \qquad \text{and} \qquad p(B) = \sum_{k=0}^{d} p_k B^k.$$

Hence,

$$p(A) X = \left(\sum_{k=0}^{d} p_k A^k \right) X = \sum_{k=0}^{d} p_k \underbrace{A^k X}_{\substack{= XB^k \\ \text{(by Lemma 2.8.4 (a))}}} = \sum_{k=0}^{d} p_k X B^k \qquad \text{and}$$

$$Xp(B) = X \sum_{k=0}^{d} p_k B^k = \sum_{k=0}^{d} p_k X B^k.$$

Comparing these two equalities, we find $p(A) X = Xp(B)$. Thus, Lemma 2.8.4 **(b)** is proven. □

Proof of the $\mathcal{V} \Longrightarrow \mathcal{U}$ part of Theorem 2.8.2. First, we observe that the matrix space $\mathbb{C}^{n \times m}$ is itself a $\mathbb{C}$-vector space of dimension nm.

Consider the map

$$\begin{aligned} L : \mathbb{C}^{n \times m} &\to \mathbb{C}^{n \times m}, \\ X &\mapsto AX - XB. \end{aligned}$$

This map L is linear, because for any $\alpha, \beta \in \mathbb{C}$ and any $X, Y \in \mathbb{C}^{n \times m}$, we have

$$\begin{aligned} L(\alpha X + \beta Y) &= A(\alpha X + \beta Y) - (\alpha X + \beta Y) B \\ &= \alpha AX + \beta AY - \alpha XB - \beta YB \\ &= \alpha \underbrace{(AX - XB)}_{= L(X)} + \beta \underbrace{(AY - YB)}_{= L(Y)} = \alpha L(X) + \beta L(Y). \end{aligned}$$

Now, assume that statement $\mathcal{V}$ holds. That is, we have $\sigma(A) \cap \sigma(B) = \varnothing$. We shall now show that $\operatorname{Ker} L = 0$. This will then yield that L is bijective.

Indeed, let $X \in \operatorname{Ker} L$. Thus, $X \in \mathbb{C}^{n \times m}$ and $L(X) = 0$. However, the definition of L yields $L(X) = AX - XB$. Therefore, $AX - XB = L(X) = 0$. In other words, $AX = XB$. Hence, we can apply Lemma 2.8.4.

Thus, Lemma 2.8.4 **(b)** (applied to $p = p_A$) yields $p_A(A) X = Xp_A(B)$. However, Theorem 2.7.1 **(a)** yields $p_A(A) = 0$, so that $p_A(A) X = 0X = 0$. Comparing this with $p_A(A) X = Xp_A(B)$, we obtain $Xp_A(B) = 0$.

We shall show that the matrix $p_A(B)$ is invertible. Indeed, Theorem 2.0.10 **(a)** shows that the polynomial p_A factors into n linear terms:

$$p_A = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_n), \tag{54}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}$ are its roots. Moreover, these roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A (by Theorem 2.0.10 **(b)**); thus, $\{\lambda_1, \lambda_2, \ldots, \lambda_n\} = \sigma(A)$.

Substituting the matrix B for t on both sides of the equality (54), we obtain

$$p_A(B) = (B - \lambda_1 I_n)(B - \lambda_2 I_n) \cdots (B - \lambda_n I_n) \tag{55}$$

(by Lemma 2.7.4 **(b)**, applied to B and n and $t - \lambda_i$ instead of A and k and f_i).

Now, let $i \in [n]$. Then, $\lambda_i \in \{\lambda_1, \lambda_2, \ldots, \lambda_n\} = \sigma(A)$. Therefore, $\lambda_i \notin \sigma(B)$ (since having $\lambda_i \in \sigma(B)$ would yield $\lambda_i \in \sigma(A) \cap \sigma(B)$, which would contradict $\sigma(A) \cap \sigma(B) = \varnothing$). In other words, λ_i is not an eigenvalue of B. In other words, $\det(\lambda_i I_n - B) \neq 0$ (by the definition of an eigenvalue). Hence, the matrix $\lambda_i I_n - B$ is invertible. In other words, the matrix $B - \lambda_i I_n$ is invertible (since $B - \lambda_i I_n = -(\lambda_i I_n - B)$).

Forget that we fixed i. We thus have shown that the matrix $B - \lambda_i I_n$ is invertible for each $i \in [n]$. In other words, the n matrices $B - \lambda_1 I_n, B - \lambda_2 I_n, \ldots, B - \lambda_n I_n$ are invertible. Hence, their product $(B - \lambda_1 I_n)(B - \lambda_2 I_n) \cdots (B - \lambda_n I_n)$ is invertible as well. In view of (55), this shows that $p_A(B)$ is invertible. Hence, from $X p_A(B) = 0$, we conclude that $X = 0$.

Now, forget that we fixed X. We thus have shown that $X = 0$ for each $X \in \operatorname{Ker} L$. In other words, $\operatorname{Ker} L = 0$. Hence, the linear map L is injective.

However, it is well-known that an injective linear map between two finite-dimensional vector spaces of the same dimension is necessarily bijective[31]. Hence, L is bijective (since L is an injective linear map between $\mathbb{C}^{n \times m}$ and $\mathbb{C}^{n \times m}$). Therefore, there exists a **unique** matrix $X \in \mathbb{C}^{n \times m}$ such that $L(X) = C$. In other words, there is a **unique** matrix $X \in \mathbb{C}^{n \times m}$ such that $AX - XB = C$ (since $L(X) = AX - XB$). In other words, statement $\mathcal{U}$ holds. Thus, the implication $\mathcal{V} \Longrightarrow \mathcal{U}$ is proven. □

Exercise 2.8.2. 5 Let A, B and C be as in Theorem 2.8.2.

(a) Let the linear map L be as in the above proof of the $\mathcal{V} \Longrightarrow \mathcal{U}$ part of Theorem 2.8.2. Prove that if $\lambda \in \sigma(A)$ and $\mu \in \sigma(B)$, then $\lambda - \mu$ is an eigenvalue of L (that is, there exists a nonzero matrix $X \in \mathbb{C}^{n \times m}$ satisfying $L(X) = (\lambda - \mu) X$).

(b) Prove the implication $\mathcal{U} \Longrightarrow \mathcal{V}$ in Theorem 2.8.2 (thus completing the proof of the theorem).

[31] *Proof.* Let $f : U \to V$ be an injective linear map between two finite-dimensional vector spaces of the same dimension. We must show that f is bijective. We have $\operatorname{Ker} f = 0$ (since f is injective) and thus $\dim(\operatorname{Ker} f) = 0$. The rank-nullity theorem yields

$$\dim U = \underbrace{\dim(\operatorname{Ker} f)}_{=0} + \dim(\operatorname{Im} f) = \dim(\operatorname{Im} f),$$

so that $\dim(\operatorname{Im} f) = \dim U = \dim V$ (since U and V have the same dimension), and therefore $\operatorname{Im} f = V$ (because $\operatorname{Im} f$ is a subspace of V). This shows that f is surjective. Since f is also injective, we thus conclude that f is bijective.

We note that more can be said: If A, B and C are as in Theorem 2.8.2, and if L is as in the above proof, then **all** eigenvalues of L have the form $\lambda - \mu$ for $\lambda \in \sigma(A)$ and $\mu \in \sigma(B)$. But this seems harder to prove at this point.

Lecture 7 starts here.

We shall next prove a somewhat surprising consequence of Theorem 2.8.2: a similarity criterion for certain block matrices:

Corollary 2.8.5. Let $A \in \mathbb{C}^{n \times n}$, $B \in \mathbb{C}^{m \times m}$ and $C \in \mathbb{C}^{n \times m}$ be three matrices such that $\sigma(A) \cap \sigma(B) = \varnothing$. Then, the two $(n+m) \times (n+m)$-matrices

$$\begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

(written in block matrix notation) are similar.

Example 2.8.6. Let $A = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 2 \end{pmatrix}$ and $C = \begin{pmatrix} 7 \\ 9 \end{pmatrix}$. Then, Corollary 2.8.5 says that the matrices

$$\begin{pmatrix} 1 & 3 & 7 \\ 0 & 1 & 9 \\ 0 & 0 & 2 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

are similar.

Proof of Corollary 2.8.5. Theorem 2.8.2 (specifically, its $\mathcal{V} \Longrightarrow \mathcal{U}$ direction) shows that there is a unique matrix $X \in \mathbb{C}^{n \times m}$ such that $AX - XB = C$. Consider this X.

Now, let $S = \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix}$. (This is an $(n+m) \times (n+m)$-matrix written in block matrix notation.) Now, I claim that this matrix S is invertible and that

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} = S \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} S^{-1}.$$

Once this claim is proved, the claim of Corollary 2.8.5 will follow (by the definition of "similar").

To see that S is invertible, we construct an inverse. Namely, we set $S' = \begin{pmatrix} I_n & -X \\ 0 & I_m \end{pmatrix}$

(again an $(n+m)\times(n+m)$-matrix). Then, the definitions of S and S' yield

$$\begin{aligned}
SS' &= \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix} \begin{pmatrix} I_n & -X \\ 0 & I_m \end{pmatrix} \\
&= \begin{pmatrix} I_n I_n + X\cdot 0 & I_n(-X) + XI_m \\ 0I_n + I_m\cdot 0 & 0(-X) + I_m I_m \end{pmatrix} \qquad \text{(by Proposition 1.6.5)} \\
&= \begin{pmatrix} I_n & 0 \\ 0 & I_m \end{pmatrix} \qquad \begin{pmatrix} \text{since } I_nI_n + X\cdot 0 = I_n \text{ and } I_n(-X) + XI_m = -X + X = 0 \\ \text{and } 0I_n + I_m\cdot 0 = 0 \text{ and } 0(-X) + I_mI_m = I_m \end{pmatrix} \\
&= I_{n+m}
\end{aligned}$$

and similarly $S'S = I_{n+m}$. Thus, the matrices S and S' are mutually inverse. Hence, S is invertible.

It remains to check that

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} = S \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} S^{-1}. \tag{56}$$

To do so, it suffices to check the equivalent identity

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} S = S \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \tag{57}$$

(indeed, these two identities are equivalent, since S is invertible). This we do by computing both sides and comparing: Namely, the definition of S yields

$$\begin{aligned}
\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} S &= \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix} \\
&= \begin{pmatrix} AI_n + 0\cdot 0 & A\cdot X + 0\cdot I_m \\ 0I_n + B\cdot 0 & 0X + BI_m \end{pmatrix} \qquad \text{(by Proposition 1.6.5)} \\
&= \begin{pmatrix} A & AX \\ 0 & B \end{pmatrix} \qquad \text{(by the obvious simplifications)}
\end{aligned}$$

and

$$\begin{aligned}
S \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} &= \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \\
&= \begin{pmatrix} I_nA + X\cdot 0 & I_nC + XB \\ 0A + I_m\cdot 0 & 0C + I_m\cdot B \end{pmatrix} \qquad \text{(by Proposition 1.6.5)} \\
&= \begin{pmatrix} A & C + XB \\ 0 & B \end{pmatrix} \qquad \text{(by the obvious simplifications)} \\
&= \begin{pmatrix} A & AX \\ 0 & B \end{pmatrix} \qquad \text{(since } AX - XB = C \text{ entails } C + XB = AX\text{)}.
\end{aligned}$$

Comparing these two equalities yields (57). Thus, we obtain (56) (by multiplying both sides of (57) with S^{-1} from the right). But this shows that the two matrices $\begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$ and $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ are similar. This proves Corollary 2.8.5. □

3. The Jordan canonical form ([HorJoh13, Chapter 3])

This chapter is devoted to the *Jordan canonical form* (and some of its variants), which is a normal form for $n \times n$-matrices over $\mathbb{C}$ with respect to similarity. This means that each $n \times n$-matrix is similar to a more-or-less unique matrix of a certain kind (namely, a block-diagonal matrix made of a specific type of blocks), called its "Jordan canonical form".

We recall that the notation "$A \sim B$" (where A and B are two $n \times n$-matrices) means that the matrices A and B are similar.

3.1. Jordan cells

The building blocks for the Jordan canonical form are the so-called Jordan cells. Let us define them:

Definition 3.1.1. Let $\mathbb{F}$ be a field. A *Jordan cell* is an $m \times m$-matrix of the form

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix} \qquad \text{for some } m > 0 \text{ and some } \lambda \in \mathbb{F}.$$

In other words, it is an $m \times m$-matrix

- whose diagonal entries are λ,
- whose entries directly above the diagonal (i.e., just one step upwards from a diagonal entry) are 1, and
- whose all remaining entries are 0.

In formal terms, it is the $m \times m$-matrix A whose entries are given by the rule

$$A_{i,j} = \begin{cases} \lambda, & \text{if } i = j; \\ 1, & \text{if } i = j-1; \\ 0, & \text{otherwise} \end{cases} \qquad \text{for all } i \in [m] \text{ and } j \in [m].$$

To be specific, this matrix is called the *Jordan cell of size m at eigenvalue λ*. It is denoted by $J_m(\lambda)$.

Example 3.1.2. **(a)** The Jordan cell of size 3 at eigenvalue -5 is

$$J_3(-5) = \begin{pmatrix} -5 & 1 & 0 \\ 0 & -5 & 1 \\ 0 & 0 & -5 \end{pmatrix}.$$

(b) The Jordan cell of size 2 at eigenvalue π is

$$J_2(\pi) = \begin{pmatrix} \pi & 1 \\ 0 & \pi \end{pmatrix}.$$

(c) For any $\lambda \in \mathbb{F}$, the Jordan cell of size 1 at eigenvalue λ is the 1×1-matrix $\begin{pmatrix} \lambda \end{pmatrix}$.

Remark 3.1.3. We will chiefly use Jordan cells as building blocks for the Jordan normal form. However, they are of some independent interest. In particular, they serve as matrix representations for several useful linear maps.

For example, fix $m \in \mathbb{N}$, and let P_m be the $\mathbb{C}$-vector space of all polynomials in a single variable t of degree $< m$ (with complex coefficients). Then, P_m has a basis $\left(t^0, t^1, \ldots, t^{m-1}\right)$. The derivative operator $\dfrac{d}{dt} : P_m \to P_m$ (which sends each polynomial $f \in P_m$ to its derivative f') is a $\mathbb{C}$-linear map that is represented by the matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 3 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

with respect to this basis. This matrix has the numbers $1, 2, \ldots, m-1$ in the cells directly above the main diagonal, and 0s everywhere else. It is not quite a Jordan cell. However, if we instead use the basis $\left(\dfrac{t^0}{0!}, \dfrac{t^1}{1!}, \ldots, \dfrac{t^{m-1}}{(m-1)!}\right)$, then the operator $\dfrac{d}{dt}$ is represented by the matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

which is precisely the Jordan cell $J_m(0)$. (This basis is just a rescaled version

of the basis $\left(t^0, t^1, \ldots, t^{m-1}\right)$, where the rescaling factors have been chosen to "normalize" the $1, 2, \ldots, m-1$ entries to be 1s.)

While the Jordan cell $J_m(\lambda)$ depends on both m and λ, its dependence on λ is not very substantial:

Proposition 3.1.4. Let $\mathbb{F}$ be a field. Let m be a positive integer, and let $\lambda \in \mathbb{F}$. Then,

$$J_m(\lambda) = J_m(0) + \lambda I_m.$$

Proof. Definition 3.1.1 yields

$$J_m(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix} \tag{58}$$

and

$$J_m(0) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{59}$$

On the other hand,

$$\lambda I_m = \begin{pmatrix} \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

Adding this equality to (59), we obtain

$$\begin{aligned} J_m(0) + \lambda I_m &= \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix} \\ &= \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix} = J_m(\lambda) \end{aligned}$$

(by (58)). This proves Proposition 3.1.4. □

Thanks to Proposition 3.1.4, we can reduce many questions about $J_m(\lambda)$ to the corresponding questions about $J_m(0)$. Let us compute the powers of $J_m(0)$:

Proposition 3.1.5. Let $\mathbb{F}$ be a field. Let m be a positive integer. Let $B = J_m(0)$. Let $p \in \mathbb{N}$. Then:

(a) The entries of the $m \times m$-matrix B^p are given by

$$\left(B^p\right)_{i,j} = \begin{cases} 1, & \text{if } i = j - p; \\ 0, & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in [m].$$

(b) We have $B^p = 0$ if $p \geq m$.

(c) We have $\dim\left(\operatorname{Ker}\left(B^p\right)\right) = p$ if $p \leq m$.

(d) We have $\dim\left(\operatorname{Ker}\left(B^p\right)\right) = m$ if $p \geq m$.

Example 3.1.6. For $m = 4$, the matrix $B = J_4(0)$ from Proposition 3.1.5 satisfies

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B^4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = 0.$$

Thus, as we go from B to B^2 to B^3 to B^4, the 1s in the cells directly above the main diagonal recede further and further upwards, until they eventually disappear beyond the borders of the matrix. (It is actually better to start this sequence with B^0 rather than B, so that the 1s start on the main diagonal.) Proposition 3.1.5 **(a)** is merely a formal way of stating this phenomenon. Parts **(b)**, **(c)** and **(d)** of Proposition 3.1.5 follow easily from part **(a)**.

Proof of Proposition 3.1.5. We have

$$B = J_m(0) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{60}$$

(by the definition of $J_m(0)$).

(a) We can prove Proposition 3.1.5 **(a)** by induction on p, using the definition of matrix multiplication (and the fact that $B^{q+1} = B^q B$ for each $q \in \mathbb{N}$). However, there is a more elegant proof using the action of B on basis vectors:

Forget that we fixed p. For each $i \in [m]$, let e_i be the column vector in $\mathbb{F}^m$ whose i-th entry is 1 while all its other entries are 0. That is,

$$e_i = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix}^T,$$

where the 1 is in the i-th position. The vectors $e_1, e_2, \ldots, e_m$ are the standard basis vectors of $\mathbb{F}^m$. It is well-known that every $m \times m$-matrix $C \in \mathbb{F}^{m \times m}$ satisfies

$$C_{\bullet,i} = Ce_i \qquad \text{for each } i \in [m]. \tag{61}$$

(Recall that $C_{\bullet,i}$ denotes the i-th column of C.)

We have so far defined the vectors e_i only for $i \in [m]$. Now, for each integer $i \notin [m]$, we define e_i to be the zero vector $0 \in \mathbb{F}^m$. Thus, we have defined a vector $e_i \in \mathbb{F}^m$ for each $i \in \mathbb{Z}$ (although it is nonzero only when $i \in [m]$). In particular, $e_0 = 0$ (since $0 \notin [m]$). Note that we have

$$(\text{the } k\text{-th entry of the column vector } e_i) = \begin{cases} 1, & \text{if } k = i; \\ 0, & \text{otherwise} \end{cases} \tag{62}$$

for each $i \in \mathbb{Z}$ and each $k \in [m]$ [32].

Now, from (60), we see that the columns of the matrix B are $0, e_1, e_2, \ldots, e_{m-1}$ in this order. In other words, the columns of B are $e_0, e_1, e_2, \ldots, e_{m-1}$ in this order (since $e_0 = 0$). In other words, we have

$$B_{\bullet,i} = e_{i-1} \qquad \text{for each } i \in [m]$$

(since $B_{\bullet,i}$ denotes the i-th column of B). However, (61) (applied to $C = B$) shows that we have

$$B_{\bullet,i} = Be_i \qquad \text{for each } i \in [m].$$

Comparing these two equalities, we obtain

$$Be_i = e_{i-1} \qquad \text{for each } i \in [m]. \tag{63}$$

However, this equality also holds for all $i \leq 0$ (because if $i \leq 0$, then both e_i and e_{i-1} equal the zero vector 0 (since $i \notin [m]$ and $i - 1 \notin [m]$), and therefore this equality boils down to $B \cdot 0 = 0$). Thus,

$$Be_i = e_{i-1} \qquad \text{for each integer } i \leq m. \tag{64}$$

Now, we claim that

$$B^p e_i = e_{i-p} \qquad \text{for each } p \in \mathbb{N} \text{ and } i \in [m]. \tag{65}$$

[32] *Proof.* If $i \in [m]$, then the equality (62) follows from the definition of e_i. On the other hand, if $i \notin [m]$, then $e_i = 0$ (by definition), so that both sides of the equality (62) are 0 (since we don't have $k = i$ (because $k \in [m]$ and $i \notin [m]$), and thus we have $\begin{cases} 1, & \text{if } k = i; \\ 0, & \text{otherwise} \end{cases} = 0$). Thus, the equality (62) holds in either case.

[*Proof of (65):* We induct on p:

Induction base: We have $B^0 = I_m$ and thus $B^0 e_i = I_m e_i = e_i = e_{i-0}$ for each $i \in [m]$. In other words, (65) holds for $p = 0$.

Induction step: Let $q \in \mathbb{N}$. Assume that (65) holds for $p = q$. We must prove that (65) holds for $p = q + 1$.

We have assumed that (65) holds for $p = q$. In other words, we have $B^q e_i = e_{i-q}$ for each $i \in [m]$. Now, let $i \in [m]$ be arbitrary. As we have just seen, we have $B^q e_i = e_{i-q}$. However, from $q \geq 0$, we obtain $i - q \leq i \leq m$ (since $i \in [m]$). Thus, (64) (applied to $i - q$ instead of i) yields $Be_{i-q} = e_{i-q-1}$. Now,

$$\underbrace{B^{q+1}}_{=BB^q} e_i = B \underbrace{B^q e_i}_{=e_{i-q}} = Be_{i-q} = e_{i-q-1} = e_{i-(q+1)}.$$

Forget that we fixed i. We thus have shown that $B^{q+1} e_i = e_{i-(q+1)}$ for each $i \in [m]$. In other words, (65) holds for $p = q + 1$. This completes the induction step. Thus, (65) is proved.]

Now, let $p \in \mathbb{N}$ and let $i, j \in [m]$. Then,

$$\begin{aligned} (B^p)_{\bullet,j} &= B^p e_j && \text{(by (61), applied to } B^p \text{ and } j \text{ instead of } C \text{ and } i) \\ &= e_{j-p} && \text{(by (65), applied to } j \text{ instead of } i)\,. \end{aligned}$$

Furthermore,

$$\begin{aligned} (B^p)_{i,j} &= \left(\text{the } i\text{-th entry of the column vector } \underbrace{(B^p)_{\bullet,j}}_{=e_{j-p}} \right) \\ &= \left(\text{the } i\text{-th entry of the column vector } e_{j-p} \right) \\ &= \begin{cases} 1, & \text{if } i = j - p; \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

(by (62), applied to i and $j - p$ instead of k and i). This proves Proposition 3.1.5 **(a)**.

(b) Assume that $p \geq m$. Let $i, j \in [m]$ be arbitrary. Then, $i \geq 1$ and $j \leq m$. Hence, $\underbrace{j}_{\leq m} - p \leq m - p \leq 0$ (since $p \geq m$), so that $0 \geq j - p$. On the other hand, $i \geq 1 > 0 \geq j - p$. Therefore, $i \neq j - p$. However, Proposition 3.1.5 **(a)** yields

$$(B^p)_{i,j} = \begin{cases} 1, & \text{if } i = j - p; \\ 0, & \text{otherwise} \end{cases} = 0 \qquad (\text{since } i \neq j - p)\,.$$

Forget that we fixed i and j. We thus have shown that $(B^p)_{i,j} = 0$ for all $i, j \in [m]$. In other words, all entries of the matrix B^p equal 0. Thus, $B^p = 0$. This proves Proposition 3.1.5 **(b)**.

(c) Assume that $p \leq m$. We shall use the column vectors $e_i \in \mathbb{F}^m$ that were defined for all $i \in \mathbb{Z}$ in our above proof of Proposition 3.1.5 **(a)**.

Let $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \in \mathbb{F}^m$ be a column vector. Thus,

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} = v_1 e_1 + v_2 e_2 + \cdots + v_m e_m = \sum_{i=1}^{m} v_i e_i.$$

Hence,

$$B^p v = B^p \cdot \sum_{i=1}^{m} v_i e_i = \sum_{i=1}^{m} v_i \underbrace{B^p e_i}_{\substack{=e_{i-p} \\ \text{(by (65))}}} = \sum_{i=1}^{m} v_i e_{i-p} = \sum_{j=1-p}^{m-p} v_{j+p} e_j$$

(here, we have substituted $j+p$ for i in the sum)

$$= \sum_{j=1-p}^{0} v_{j+p} \underbrace{e_j}_{\substack{=0 \\ \text{(since } j \notin [m] \\ \text{(because } j \leq 0))}} + \sum_{j=1}^{m-p} v_{j+p} e_j = \underbrace{\sum_{j=1-p}^{0} v_{j+p} 0}_{=0} + \sum_{j=1}^{m-p} v_{j+p} e_j = \sum_{j=1}^{m-p} v_{j+p} e_j$$

$$= v_{p+1} e_1 + v_{p+2} e_2 + \cdots + v_m e_{m-p} = \begin{pmatrix} v_{p+1} \\ v_{p+2} \\ \vdots \\ v_m \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Thus, $B^p v = 0$ holds if and only if $v_{p+1} = v_{p+2} = \cdots = v_m = 0$. In other words, $B^p v = 0$ holds if and only if $v \in \operatorname{span}\left(e_1, e_2, \ldots, e_p\right)$ (because $v_{p+1} = v_{p+2} = \cdots = v_m = 0$ is equivalent to $v \in \operatorname{span}\left(e_1, e_2, \ldots, e_p\right)$).

Now, forget that we fixed v. We thus have shown that a vector $v \in \mathbb{F}^m$ satisfies $B^p v = 0$ if and only if $v \in \operatorname{span}\left(e_1, e_2, \ldots, e_p\right)$. Thus,

$$\operatorname{Ker}\left(B^p\right) = \operatorname{span}\left(e_1, e_2, \ldots, e_p\right)$$

(since $\operatorname{Ker}\left(B^p\right)$ is defined as the set of all vectors $v \in \mathbb{F}^m$ satisfying $B^p v = 0$). Therefore,

$$\dim\left(\operatorname{Ker}\left(B^p\right)\right) = \dim\left(\operatorname{span}\left(e_1, e_2, \ldots, e_p\right)\right) = p$$

(since span $\left(e_1, e_2, \ldots, e_p\right)$ is clearly a p-dimensional subspace of $\mathbb{F}^m$). This proves Proposition 3.1.5 **(c)**.

(d) Assume that $p \geq m$. Then, Proposition 3.1.5 **(b)** yields $B^p = 0$. Hence, $\dim\left(\operatorname{Ker}\left(B^p\right)\right) = \dim\underbrace{\left(\operatorname{Ker} 0\right)}_{=\mathbb{F}^m} = \dim\left(\mathbb{F}^m\right) = 0$. Thus, Proposition 3.1.5 **(d)** is proven. □

Proposition 3.1.7. Let m be a positive integer, and let $\lambda \in \mathbb{C}$. The only eigenvalue of the matrix $J_m\left(\lambda\right)$ is λ. This eigenvalue has algebraic multiplicity m and geometric multiplicity 1.

Proof. The definition of $J_m\left(\lambda\right)$ yields

$$J_m\left(\lambda\right) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

This matrix is upper-triangular, so its characteristic polynomial is

$$p_{J_m(\lambda)} = \left(t-\lambda\right)\left(t-\lambda\right)\cdots\left(t-\lambda\right) = \left(t-\lambda\right)^m.$$

Thus, the only root of this polynomial is λ. In other words, the only eigenvalue of this matrix $J_m\left(\lambda\right)$ is λ. Its algebraic multiplicity is m (since this is its multiplicity as a root of $p_{J_m(\lambda)}$). It remains to show that its geometric multiplicity is 1.

Since the geometric multiplicity of λ is defined as $\dim\left(\operatorname{Ker}\left(J_m\left(\lambda\right) - \lambda I_m\right)\right)$, this means that it remains to show that $\dim\left(\operatorname{Ker}\left(J_m\left(\lambda\right) - \lambda I_m\right)\right) = 1$.

Let $B = J_m\left(0\right)$. Proposition 3.1.5 **(c)** (applied to $p = 1$) yields $\dim\left(\operatorname{Ker}\left(B^1\right)\right) = 1$ (since $1 \leq m$). In other words, $\dim\left(\operatorname{Ker} B\right) = 1$.

Proposition 3.1.4 yields

$$J_m\left(\lambda\right) = \underbrace{J_m\left(0\right)}_{=B} + \lambda I_m = B + \lambda I_m,$$

so that $J_m\left(\lambda\right) - \lambda I_m = B$. Hence,

$$\dim\left(\operatorname{Ker}\underbrace{\left(J_m\left(\lambda\right) - \lambda I_m\right)}_{=B}\right) = \dim\left(\operatorname{Ker} B\right) = 1.$$

This completes our proof of Proposition 3.1.7. □

3.2. Jordan canonical form: the theorem

Let us now build larger matrices out of Jordan cells:

Definition 3.2.1. Let $\mathbb{F}$ be a field. A *Jordan matrix* means a block-diagonal matrix whose diagonal blocks are Jordan cells. In other words, it is a matrix of the form

$$\begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) \end{pmatrix},$$

where $n_1, n_2, \ldots, n_k$ are positive integers and $\lambda_1, \lambda_2, \ldots, \lambda_k$ are scalars in $\mathbb{F}$ (not necessarily distinct, but not necessarily equal either).

We note that any Jordan matrix is upper-triangular.

We claim the following:[33]

Theorem 3.2.2 (Jordan canonical form theorem)**.** Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix over $\mathbb{C}$. Then:

(a) There exists a Jordan matrix J such that $A \sim J$.

(b) This Jordan matrix J is unique up to the order of the diagonal blocks.

This theorem is useful partly (but not only) because it allows to reduce questions about general square matrices to questions about Jordan matrices. And the latter can usually further be reduced to questions about Jordan cells, because a block-diagonal matrix "behaves like its diagonal blocks are separate" (see, e.g., the discussion before Proposition 1.6.11).

Note that we cannot hope for the matrix J in Theorem 3.2.2 to be fully unique, unless it has only one diagonal block (i.e., unless $k = 1$). Indeed, Proposition 1.6.6 shows that if we permute the diagonal blocks $J_{n_1}(\lambda_1), J_{n_2}(\lambda_2), \ldots, J_{n_k}(\lambda_k)$, then the matrix stays similar to A. Thus, the order of these diagonal blocks can be chosen arbitrary.

Definition 3.2.3. Let A be an $n \times n$-matrix over $\mathbb{C}$. Theorem 3.2.2 **(a)** says that there exists a Jordan matrix J such that $A \sim J$. Such a matrix J is called a *Jordan canonical form* of A (or a *Jordan normal form* of A).

We often use the definite article ("the Jordan canonical form of A"), because Theorem 3.2.2 **(b)** says that J is "more or less unique". (Strictly speaking, of course, it is not entirely appropriate.)

The diagonal blocks $J_{n_1}(\lambda_1), J_{n_2}(\lambda_2), \ldots, J_{n_k}(\lambda_k)$ of J are called the *Jordan blocks* (or *Jordan cells*) of A.

We often abbreviate "Jordan canonical form" as "JCF".

[33] Recall that "$A \sim B$" means that the matrix A is similar to the matrix B.

Example 3.2.4. A Jordan canonical form of $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 5 \\ 0 & 0 & 1 \end{pmatrix}$ is $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Indeed, $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ is a Jordan matrix (it can be written as $\begin{pmatrix} J_1(2) & 0 \\ 0 & J_2(1) \end{pmatrix}$) and it can be checked that

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

We can swap the two Jordan cells in this Jordan matrix, and obtain another Jordan canonical form of the same matrix: $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

Example 3.2.5. If $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, then a Jordan canonical form of A is $\begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = \begin{pmatrix} J_1(-i) & 0 \\ 0 & J_1(i) \end{pmatrix}$, where $i = \sqrt{-1} \in \mathbb{C}$. Note that this Jordan canonical form has imaginary entries, despite all entries of A being real. This is unavoidable when A has non-real eigenvalues.

Example 3.2.6. If D is a diagonal matrix, then D itself is a Jordan canonical form of D. Indeed, each diagonal entry λ of D can be viewed as a Jordan cell of size 1 (namely, $J_1(\lambda)$), so that D is a Jordan matrix.

3.3. Jordan canonical form: proof of uniqueness

We shall approach the proof of Theorem 3.2.2 slowly[34], making sure to record all auxiliary results obtained on the way (as they are themselves rather useful). We first try to explore how much of the structure of the Jordan normal form J can be read off the matrix A. This will lead us to the proof of the uniqueness part (i.e., part **(b)**) of Theorem 3.2.2.

We start with an example:

[34]See [Bourba03, Chapter VII, §5, section 4], [GalQua20, Theorem 31.17], [Heffer20, Chapter Five, Section IV, Theorem 2.8], [Loehr14, §8.10–§8.11], [OmClVi11, §4.6], [Prasol94, §12.2], [Shapir15, §4.3], [Taylor20, §2.4], [Treil15, Chapter 9, Theorem 5.1] and [Woerde16, Theorem 4.4.1] for other proofs (at least proofs of Theorem 3.2.2 **(a)**, which is the harder part of Theorem 3.2.2).

Example 3.3.1. Let $A \in \mathbb{C}^{7\times 7}$. Suppose that $A \sim J$ with

$$J = \begin{pmatrix} J_2(8) & 0 & 0 \\ 0 & J_3(8) & 0 \\ 0 & 0 & J_2(9) \end{pmatrix} = \begin{pmatrix} 8 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 9 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 9 \end{pmatrix}.$$

How are the structural elements of this matrix J (that is, the diagonal entries 8 and 9 and the sizes $2, 3, 2$ of the diagonal blocks) reflected in the matrix A ?

First, the matrix J is a Jordan matrix, and thus upper-triangular. Hence, Proposition 2.3.7 (applied to $T = J$) shows that the diagonal entries of J are the eigenvalues of A (with their algebraic multiplicities). Thus, the eigenvalues of A are 8 and 9, with respective algebraic multiplicities 5 and 2.

Now, what about the geometric multiplicities? That is, what are $\dim(\operatorname{Ker}(A - 8I_7))$ and $\dim(\operatorname{Ker}(A - 9I_7))$?

From $A \sim J$, we obtain $A - 8I_7 \sim J - 8I_7$ (by Proposition 2.1.5 **(h)**, applied to $B = J$ and $\lambda = 8$ and $n = 7$). Thus, Proposition 2.1.5 **(b)** (applied to $A - 8I_7$, $J - 8I_7$ and 7 instead of A, B and 8) yields that the matrices $A - 8I_7$ and $J - 8I_7$ have the same nullity. In other words,

$$\begin{aligned} \dim(\operatorname{Ker}(A - 8I_7)) &= \dim(\operatorname{Ker}(J - 8I_7)) \\ &= \dim\left(\operatorname{Ker}\begin{pmatrix} J_2(8) - 8I_2 & 0 & 0 \\ 0 & J_3(8) - 8I_3 & 0 \\ 0 & 0 & J_2(9) - 8I_2 \end{pmatrix}\right) \end{aligned}$$

(since

$$\begin{aligned} J - 8I_7 &= \begin{pmatrix} J_2(8) & 0 & 0 \\ 0 & J_3(8) & 0 \\ 0 & 0 & J_2(9) \end{pmatrix} - 8I_7 \\ &= \begin{pmatrix} J_2(8) - 8I_2 & 0 & 0 \\ 0 & J_3(8) - 8I_3 & 0 \\ 0 & 0 & J_2(9) - 8I_2 \end{pmatrix} \end{aligned}$$

). Thus,

$$\begin{aligned} &\dim(\operatorname{Ker}(A - 8I_7)) \\ &= \dim\left(\operatorname{Ker}\begin{pmatrix} J_2(8) - 8I_2 & 0 & 0 \\ 0 & J_3(8) - 8I_3 & 0 \\ 0 & 0 & J_2(9) - 8I_2 \end{pmatrix}\right) \\ &= \dim(\operatorname{Ker}(J_2(8) - 8I_2)) + \dim(\operatorname{Ker}(J_3(8) - 8I_3)) + \dim(\operatorname{Ker}(J_2(9) - 8I_2)) \end{aligned}$$

(by Proposition 1.6.11). Now, let us find the three dimensions on the right hand side.

Proposition 3.1.4 yields $J_2(8) = J_2(0) + 2I_2$, so that $J_2(8) - 8I_2 = J_2(0)$. Hence,

$$\dim\left(\operatorname{Ker}\left(J_2(8) - 8I_2\right)\right) = \dim\left(\operatorname{Ker}\left(J_2(0)\right)\right) = \dim\left(\operatorname{Ker}\left(\left(J_2(0)\right)^1\right)\right) = 1$$

(by Proposition 3.1.5 **(c)**, applied to $m = 2$ and $p = 1$, because the matrix $J_2(0)$ is what is called B in this proposition). Similarly, $\dim\left(\operatorname{Ker}\left(J_3(8) - 8I_3\right)\right) = 1$. On the other hand, the matrix $J_2(9) - 8I_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is an upper-triangular matrix with 1's on its main diagonal; thus, its determinant is $1 \cdot 1 = 1 \neq 0$, so that it is nonsingular. Hence, $\operatorname{Ker}\left(J_2(9) - 8I_2\right) = 0$, so that $\dim\left(\operatorname{Ker}\left(J_2(9) - 8I_2\right)\right) = 0$. Thus, our above computation of $\dim\left(\operatorname{Ker}\left(A - 8I_7\right)\right)$ becomes

$$\begin{aligned}
&\dim\left(\operatorname{Ker}\left(A - 8I_7\right)\right) \\
&= \underbrace{\dim\left(\operatorname{Ker}\left(J_2(8) - 8I_2\right)\right)}_{=1} + \underbrace{\dim\left(\operatorname{Ker}\left(J_3(8) - 8I_3\right)\right)}_{=1} + \underbrace{\dim\left(\operatorname{Ker}\left(J_2(9) - 8I_2\right)\right)}_{=0} \\
&= 1 + 1 + 0 = 2.
\end{aligned}$$

Looking back, we see that this comes from the fact that exactly 2 of the diagonal blocks in the Jordan canonical form J are Jordan cells at eigenvalue 8.

Generalizing this reasoning, we obtain the following:

Proposition 3.3.2. Let A be an $n \times n$-matrix, and let $J = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) \end{pmatrix}$ be its Jordan canonical form. Then:

(a) We have $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$.

(b) The geometric multiplicity of a number $\lambda \in \mathbb{C}$ as an eigenvalue of A is the number of Jordan cells of A at eigenvalue λ. In other words, it is the number of $i \in [k]$ satisfying $\lambda_i = \lambda$.

(c) The algebraic multiplicity of a number $\lambda \in \mathbb{C}$ as an eigenvalue of A is the **sum** of the sizes of all Jordan cells of A at eigenvalue λ. In other words, it is $\sum\limits_{\substack{i \in [k]; \\ \lambda_i = \lambda}} n_i$.

Proof. TODO: Scribe? □

With some more effort, we can obtain a more precise result:

Proposition 3.3.3. Let A be an $n \times n$-matrix, and let $J = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) \end{pmatrix}$ be its Jordan canonical form. Let $\lambda \in \mathbb{C}$. Let p be a positive integer. Then,

$$\begin{aligned} &(\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i \geq p) \\ &= \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right). \end{aligned}$$

Proof. We have $A \sim J$, so that $A - \lambda I_n \sim J - \lambda I_n$ (by Proposition 2.1.5 **(h)**, applied to $B = J$), and therefore $(A-\lambda I_n)^p \sim (J-\lambda I_n)^p$ (by Proposition 2.1.5 **(f)**, applied to $A - \lambda I_n$ and $B - \lambda I_n$ and p instead of A, B and k). Hence,

$$\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) = \dim\left(\operatorname{Ker}\left((J-\lambda I_n)^p\right)\right). \tag{66}$$

For each $i \in [k]$, we set

$$M_i := J_{n_i}(\lambda_i - \lambda). \tag{67}$$

However, for each $i \in [k]$, we have

$$J_{n_i}(\lambda_i) - \lambda I_{n_i} = J_{n_i}(\lambda_i - \lambda) \tag{68}$$

(because the two Jordan cells $J_{n_i}(\lambda_i)$ and $J_{n_i}(\lambda_i - \lambda)$ differ only in their diagonal entries, which are λ_i in the former matrix and $\lambda_i - \lambda$ in the latter). Comparing this with (67), we obtain

$$J_{n_i}(\lambda_i) - \lambda I_{n_i} = M_i \tag{69}$$

for each $i \in [k]$.

Now, we have

$$J = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) \end{pmatrix} \qquad \text{and}$$

$$\lambda I_n = \begin{pmatrix} \lambda I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda I_{n_k} \end{pmatrix}.$$

Subtracting these two equalities from one another, we obtain

$$J - \lambda I_n = \begin{pmatrix} J_{n_1}(\lambda_1) - \lambda I_{n_1} & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) - \lambda I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) - \lambda I_{n_k} \end{pmatrix}$$
$$= \begin{pmatrix} M_1 & 0 & \cdots & 0 \\ 0 & M_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_k \end{pmatrix}$$

(by (68)). Hence,

$$(J - \lambda I_n)^p = \begin{pmatrix} M_1 & 0 & \cdots & 0 \\ 0 & M_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_k \end{pmatrix}^p = \begin{pmatrix} M_1^p & 0 & \cdots & 0 \\ 0 & M_2^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_k^p \end{pmatrix}$$

(by Corollary 1.6.10). Thus,

$$\begin{aligned} &\dim\left(\operatorname{Ker}\left((J - \lambda I_n)^p\right)\right) \\ &= \dim\left(\operatorname{Ker}\begin{pmatrix} M_1^p & 0 & \cdots & 0 \\ 0 & M_2^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_k^p \end{pmatrix}\right) \\ &= \dim\left(\operatorname{Ker}\left(M_1^p\right)\right) + \dim\left(\operatorname{Ker}\left(M_2^p\right)\right) + \cdots + \dim\left(\operatorname{Ker}\left(M_k^p\right)\right) \\ &\qquad \text{(by Proposition 1.6.11)} \\ &= \sum_{i=1}^{k} \dim\left(\operatorname{Ker}\left(M_i^p\right)\right) \end{aligned} \tag{70}$$

Now, fix an $i \in [k]$ satisfying $\lambda_i \neq \lambda$. Thus, $\lambda_i - \lambda \neq 0$. The matrix $J_{n_i}(\lambda_i - \lambda)$ is upper-triangular, and its diagonal entries are all $\lambda_i - \lambda$. Hence, its determinant is $\det\left(J_{n_i}(\lambda_i - \lambda)\right) = (\lambda_i - \lambda)^{n_i} \neq 0$ (since $\lambda_i - \lambda \neq 0$). Therefore, this matrix $J_{n_i}(\lambda_i - \lambda)$ is invertible. In other words, the matrix M_i is invertible (since $M_i = J_{n_i}(\lambda_i - \lambda)$). Hence, its p-th power M_i^p is also invertible, and therefore has nullity 0. In other words, $\dim\left(\operatorname{Ker}\left(M_i^p\right)\right) = 0$.

Forget that we fixed i. We thus have shown that if $i \in [k]$ satisfies $\lambda_i \neq \lambda$, then

$$\dim\left(\operatorname{Ker}\left(M_i^p\right)\right) = 0. \tag{71}$$

Hence, (70) becomes

$$\begin{aligned}
&\dim\left(\operatorname{Ker}\left(\left(J-\lambda I_n\right)^p\right)\right) \\
&= \sum_{i=1}^{k} \dim\left(\operatorname{Ker}\left(M_i^p\right)\right) \\
&= \sum_{\substack{i\in[k];\\ \lambda_i\neq\lambda}} \underbrace{\dim\left(\operatorname{Ker}\left(M_i^p\right)\right)}_{\substack{=0\\ \text{(by (71))}}} + \sum_{\substack{i\in[k];\\ \lambda_i=\lambda}} \dim\left(\operatorname{Ker}\left(M_i^p\right)\right) \\
&\qquad\qquad \left(\text{since each } i\in[k] \text{ satisfies either } \lambda_i\neq\lambda \text{ or } \lambda_i=\lambda\right) \\
&= \sum_{\substack{i\in[k];\\ \lambda_i=\lambda}} \dim\left(\operatorname{Ker}\left(M_i^p\right)\right). \qquad (72)
\end{aligned}$$

Now, let us fix some $i\in[k]$ satisfying $\lambda_i=\lambda$. Then, $\lambda_i-\lambda=0$. Hence, $J_{n_i}\left(\lambda_i-\lambda\right)=J_{n_i}\left(0\right)=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$. Let us denote this matrix $J_{n_i}\left(0\right)$ by B. Then, Proposition 3.1.5 **(c)** (applied to $m=n_i$) shows that we have

$$\dim\left(\operatorname{Ker}\left(B^p\right)\right)=p \qquad \text{if } p\leq n_i.$$

On the other hand, Proposition 3.1.5 **(c)** (applied to $m=n_i$) shows that we have

$$\dim\left(\operatorname{Ker}\left(B^p\right)\right)=n_i \qquad \text{if } p>n_i.$$

Combining these two equalities, we obtain

$$\dim\left(\operatorname{Ker}\left(B^p\right)\right)=\begin{cases} p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p\geq n_i. \end{cases}$$

[35] In other words,

$$\dim\left(\operatorname{Ker}\left(M_i^p\right)\right)=\begin{cases} p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p\geq n_i \end{cases} \qquad (73)$$

(since $B=J_{n_i}\left(0\right)=J_{n_i}\left(\lambda_i-\lambda\right)=M_i$ (by (67))).

Now, forget that we fixed i. We thus have proved (71) for each $i\in[k]$ satisfying $\lambda_i=\lambda$. Therefore, (72) becomes

$$\dim\left(\operatorname{Ker}\left(\left(J-\lambda I_n\right)^p\right)\right)=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}} \underbrace{\dim\left(\operatorname{Ker}\left(M_i^p\right)\right)}_{=\begin{cases} p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p\geq n_i \end{cases}\ \text{(by (73))}} = \sum_{\substack{i\in[k];\\ \lambda_i=\lambda}} \begin{cases} p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p\geq n_i. \end{cases}$$

[35] Note that the two cases $p\leq n_i$ and $p\geq n_i$ are not mutually exclusive: They overlap when $p=n_i$. (But the answers in this case are identical.)

Thus, (66) becomes

$$\dim\left(\operatorname{Ker}\left(\left(A-\lambda I_n\right)^p\right)\right)=\dim\left(\operatorname{Ker}\left(\left(J-\lambda I_n\right)^p\right)\right)=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p>n_i.\end{cases}$$

However, we can also apply the same argument to $p-1$ instead of p (since $p-1\in\mathbb{N}$). Thus, we obtain

$$\dim\left(\operatorname{Ker}\left(\left(A-\lambda I_n\right)^{p-1}\right)\right)=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}p-1, & \text{if } p-1\leq n_i;\\ n_i, & \text{if } p-1\geq n_i.\end{cases}$$

Subtracting these two equalities, we obtain

$$\begin{aligned}
&\dim\left(\operatorname{Ker}\left(\left(A-\lambda I_n\right)^p\right)\right)-\dim\left(\operatorname{Ker}\left(\left(A-\lambda I_n\right)^{p-1}\right)\right)\\
&=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p>n_i\end{cases}-\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}p-1, & \text{if } p-1\leq n_i;\\ n_i, & \text{if } p-1\geq n_i\end{cases}\\
&=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\underbrace{\left(\begin{cases}p, & \text{if } p\leq n_i;\\ n_i, & \text{if } p>n_i\end{cases}-\begin{cases}p-1, & \text{if } p-1\leq n_i;\\ n_i, & \text{if } p-1\geq n_i\end{cases}\right)}_{=\begin{cases}1, & \text{if } p\leq n_i;\\ 0, & \text{if } p>n_i\end{cases}}\\
&\qquad\qquad\begin{array}{c}\text{(this can be directly checked in each}\\ \text{of the two cases } p\leq n_i \text{ and } p>n_i\text{)}\end{array}\\
&=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}1, & \text{if } p\leq n_i;\\ 0, & \text{if } p>n_i\end{cases}=\sum_{\substack{i\in[k];\\ \lambda_i=\lambda}}\begin{cases}1, & \text{if } n_i\geq p;\\ 0, & \text{if } n_i<p\end{cases}\\
&\qquad\qquad\left(\begin{array}{c}\text{since the condition } p\leq n_i \text{ is equivalent to } n_i\geq p,\\ \text{and since the condition } p>n_i \text{ is equivalent to } n_i<p\end{array}\right)\\
&=\left(\text{the number of } i\in[k] \text{ such that } \lambda_i=\lambda \text{ and } n_i\geq p\right)
\end{aligned}$$

(because the sum has an addend equal to 1 for each $i\in[k]$ satisfying $\lambda_i=\lambda$ and $n_i\geq p$, whereas all remaining addends of this sum are 0). Thus, Proposition 3.3.3 is proved. □

Corollary 3.3.4. Let A be an $n\times n$-matrix, and let $J=\begin{pmatrix}J_{n_1}(\lambda_1) & 0 & \cdots & 0\\ 0 & J_{n_2}(\lambda_2) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & J_{n_k}(\lambda_k)\end{pmatrix}$ be its Jordan canonical form. Let $\lambda\in\mathbb{C}$.

Let p be a positive integer. Then,

$$\begin{aligned}
&(\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i = p) \\
&= 2\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) \\
&\qquad - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p+1}\right)\right).
\end{aligned}$$

Proof. An integer z equals p if and only if it satisfies $z \geq p$ but does not satisfy $z \geq p+1$. Hence, an $i \in [k]$ satisfies $n_i = p$ if and only if it satisfies $n_i \geq p$ but does not satisfy $n_i \geq p+1$. Thus,

$$\begin{aligned}
&(\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i = p) \\
&= \underbrace{(\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i \geq p)}_{\substack{=\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right)-\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right) \\ \text{(by Proposition 3.3.3)}}} \\
&\qquad - \underbrace{(\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i \geq p+1)}_{\substack{=\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p+1}\right)\right)-\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p}\right)\right) \\ \text{(by Proposition 3.3.3,} \\ \text{applied to } p+1 \text{ instead of } p\text{)}}} \\
&\qquad\qquad (\text{because any } i \in [k] \text{ satisfying } n_i \geq p+1 \text{ must also satisfy } n_i \geq p) \\
&= \left(\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right)\right) \\
&\qquad - \left(\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p+1}\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right)\right) \\
&= 2\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p+1}\right)\right).
\end{aligned}$$

This proves Corollary 3.3.4. □

Now, we can easily prove Theorem 3.2.2 **(b)**:

*Proof of Theorem 3.2.2 **(b)**.* Let $A \in \mathbb{C}^{n\times n}$ be an $n \times n$-matrix. Let J be a Jordan matrix such that $A \sim J$. Write J as $J = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{n_k}(\lambda_k) \end{pmatrix}$ as in Corollary 3.3.4. Then, the Jordan blocks of J are $J_{n_1}(\lambda_1), J_{n_2}(\lambda_2), \ldots, J_{n_k}(\lambda_k)$. Hence, for any $\lambda \in \mathbb{C}$ and any positive integer p, we have

$$\begin{aligned}
&(\text{the number of Jordan blocks of } J \text{ of size } p \text{ at eigenvalue } \lambda) \\
&= (\text{the number of } i \in [k] \text{ such that } \lambda_i = \lambda \text{ and } n_i = p) \\
&= 2\dim\left(\operatorname{Ker}\left((A-\lambda I_n)^p\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p-1}\right)\right) - \dim\left(\operatorname{Ker}\left((A-\lambda I_n)^{p+1}\right)\right).
\end{aligned}$$

Therefore, this number is uniquely determined by A, λ and p. Hence, the whole structure of J is determined uniquely by A, up to the order of the Jordan blocks. This proves Theorem 3.2.2 **(b)**. □

Example 3.3.5. Let A be an 8×8-matrix. Assume that we know that the Jordan canonical form of A has

- 1 Jordan block of size 1 at eigenvalue 17;
- 2 Jordan blocks of size 1 at eigenvalue 35;
- 1 Jordan block of size 2 at eigenvalue 35;
- 1 Jordan block of size 3 at eigenvalue 59;
- no Jordan blocks of other sizes or at other eigenvalues.

Then, the Jordan canonical form of A must be the block-diagonal matrix

$$\begin{pmatrix} J_1(17) & 0 & 0 & 0 & 0 \\ 0 & J_2(35) & 0 & 0 & 0 \\ 0 & 0 & J_1(35) & 0 & 0 \\ 0 & 0 & 0 & J_1(35) & 0 \\ 0 & 0 & 0 & 0 & J_3(59) \end{pmatrix}$$

or one that is obtained from it by permuting the diagonal blocks.

Let us now approach the existence of the Jordan canonical form (Theorem 3.2.2 **(a)**).

3.4. Jordan canonical form: proof of existence

We will prove the existence of the Jordan canonical form in several steps, each of which will bring our matrix A "closer" to a Jordan matrix. Along the way, we will obtain several results of independent interest.

3.4.1. Step 1: Schur triangularization

Our first step will be an application of Schur triangularization. As we recall, the "weak" Schur triangularization theorem (Theorem 2.3.1) tells us that if $A \in \mathbb{C}^{n \times n}$ is an $n \times n$-matrix, then A is unitarily similar to an upper-triangular matrix T. The diagonal entries of the latter matrix T will be the eigenvalues of A in some order (by Proposition 2.3.6). However, let us now be a bit pickier. To wit, we now want the triangular matrix T to have the property that equal eigenvalues come in contiguous runs on the main diagonal. In other words, we want T to have the property that if two diagonal entries of T are equal, then all the diagonal entries between them

are also equal to them. For instance, if the eigenvalues of A are $1,1,2,2$, we don't want[36]

$$T = \begin{pmatrix} 1 & * & * & * \\ & 2 & * & * \\ & & 1 & * \\ & & & 2 \end{pmatrix},$$

but instead we want

$$T = \begin{pmatrix} 1 & * & * & * \\ & 1 & * & * \\ & & 2 & * \\ & & & 2 \end{pmatrix}.$$

Fortunately, we can achieve this using Theorem 2.3.3: Indeed, if we list the eigenvalues of A as $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ in such a way that equal eigenvalues come in contiguous runs in this list, then Theorem 2.3.3 shows that we can find an upper-triangular matrix T that is unitarily similar to A and that has diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ in this order. This matrix T is what we want.

3.4.2. Step 2: Separating distinct eigenvalues

Theorem 2.3.3 brings any $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ to a certain simplified form (upper-triangular with eigenvalues placed contiguously on the diagonal) that is not yet a Jordan canonical form, but already has some of its aspects. We will now transform it further to get a bit closer to a Jordan canonical form. To wit, we will get rid of some of the entries above the diagonal (or, to be more precise, we will turn them into 0). Let us demonstrate this on an example:

Example 3.4.1. Let $a, b, c, \ldots, p \in \mathbb{C}$ be any numbers. We shall now show that

$$\begin{pmatrix} 1 & a & b & c & d & e \\ & 1 & f & g & h & i \\ & & 2 & j & k & \ell \\ & & & 2 & m & n \\ & & & & 2 & p \\ & & & & & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & a & & & & \\ & 1 & & & & \\ & & 2 & j & k & \\ & & & 2 & m & \\ & & & & 2 & \\ & & & & & 3 \end{pmatrix} \tag{74}$$

(where the entries in the empty cells are understood to be 0s).

Indeed, the triangular matrices $\begin{pmatrix} 1 & a \\ & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & j & k & \ell \\ & 2 & m & n \\ & & 2 & p \\ & & & 3 \end{pmatrix}$ have disjoint spectra (i.e., they have no eigenvalues in common), because their diagonals have

[36] In the following equation, an empty cell of the matrix must be filled with a 0, whereas a "$*$" in a cell means that any arbitrary value can go into that cell.

no entries in common. So, by Corollary 2.8.5, we have

$$\begin{pmatrix} 1 & a & b & c & d & e \\ & 1 & f & g & h & i \\ & & 2 & j & k & \ell \\ & & & 2 & m & n \\ & & & & 2 & p \\ & & & & & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & a & & & & \\ & 1 & & & & \\ & & 2 & j & k & \ell \\ & & & 2 & m & n \\ & & & & 2 & p \\ & & & & & 3 \end{pmatrix}. \tag{75}$$

Furthermore, the triangular matrices $\begin{pmatrix} 1 & a & & \\ & 1 & & \\ & & 2 & j & k \\ & & & 2 & m \\ & & & & 2 \end{pmatrix}$ and $\begin{pmatrix} 3 \end{pmatrix}$ have disjoint spectra, so Corollary 2.8.5 yields

$$\begin{pmatrix} 1 & a & & & & \\ & 1 & & & & \\ & & 2 & j & k & \ell \\ & & & 2 & m & n \\ & & & & 2 & p \\ & & & & & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & a & & & & \\ & 1 & & & & \\ & & 2 & j & k & \\ & & & 2 & m & \\ & & & & 2 & \\ & & & & & 3 \end{pmatrix}. \tag{76}$$

Since $\sim$ is an equivalence relation, we can combine the two similarities (75) and (76), and we conclude that the claim (74) holds.

This example generalizes:

Theorem 3.4.2. Let $T \in \mathbb{C}^{n \times n}$ be an upper-triangular matrix. Assume that the diagonal entries of T come in contiguous runs (i.e., if $i, j \in [n]$ satisfy $i < j$ and $T_{i,i} = T_{j,j}$, then $T_{i,i} = T_{i+1,i+1} = T_{i+2,i+2} = \cdots = T_{j,j}$). Let S be the matrix obtained from T by setting all entries $T_{i,j}$ with $T_{i,i} \neq T_{j,j}$ to 0. In other words, let $S \in \mathbb{C}^{n \times n}$ be the $n \times n$-matrix defined by setting

$$S_{i,j} = \begin{cases} T_{i,j}, & \text{if } T_{i,i} = T_{j,j}; \\ 0, & \text{otherwise} \end{cases} \qquad \text{for all } i, j \in [n].$$

Then, $T \sim S$.

Proof. TODO: Scribe! □

Roughly speaking, Theorem 3.4.2 says that whenever we have an upper-triangular matrix T whose diagonal has no interlaced values (i.e., there is never a μ between two λ's on the diagonal when $\mu \neq \lambda$), we can "clean out" all the above-diagonal entries that correspond to different diagonal entries (i.e., all above-diagonal entries

$T_{i,j}$ with $T_{i,i} \neq T_{j,j}$) by a similarity (i.e., if we set all these entries to 0, the resulting matrix will be similar to T).

Now, combining this with Theorem 2.3.3, we obtain the following:

Proposition 3.4.3. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Then, A is similar to a block-diagonal matrix of the form

$$\begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_k \end{pmatrix},$$

where each B_i is an upper-triangular matrix such that all entries on the diagonal of B_i are equal. (Here, the cells that we left empty are understood to be filled with zero matrices.)

Proof. TODO: Scribe! □

Note that we have given up unitary similarity at this point: The word "similar" in Proposition 3.4.3 cannot be replaced by "unitarily similar". (A counterexample is easily obtained from Exercise 2.2.1.)

3.4.3. Step 3: Strictly upper-triangular matrices

The block-diagonal matrix in Proposition 3.4.3 is not yet a Jordan canonical form, but it is already somewhat close. At least, we have separated out all the distinct eigenvalues of A and "cleaned out the space between them". We now can work with the matrices $B_1, B_2, \ldots, B_k$ separately; each of these matrices has one distinct eigenvalue. Our next goal is to show that each of these matrices $B_1, B_2, \ldots, B_k$ is similar to a Jordan matrix. (This will easily yield that the total block-diagonal matrix $\begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_k \end{pmatrix}$ is similar to a Jordan matrix, and therefore the same holds for A.)

For each $i \in [k]$, the matrix B_i has all its diagonal entries equal. Let us say these diagonal entries all equal μ_i. Thus, $B_i - \mu_i I$ is a strictly upper-triangular matrix. (Recall: a *strictly upper-triangular* matrix is an upper-triangular matrix whose diagonal entries are 0.) We want to show that B_i is similar to a Jordan matrix. Because of Proposition 2.1.5 **(g)**, it will suffice to show that the strictly upper-triangular matrix $B_i - \mu_i I$ is similar to a Jordan matrix (because adding $\mu_i I$ to a Jordan matrix always gives a Jordan matrix again).

Thus, our goal is now to show that every strictly upper-triangular matrix A is similar to a Jordan matrix. Before we approach this goal in general, let us convince ourselves that it is achievable for 2×2-matrices.

Example 3.4.4. A strictly upper-triangular 2×2-matrix $A \in \mathbb{C}^{2\times 2}$ must have the form $\begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}$ for some $a \in \mathbb{C}$.

- If $a = 0$, then A is the Jordan matrix $\begin{pmatrix} J_1(0) & \\ & J_1(0) \end{pmatrix}$.
- If $a \neq 0$, then A is similar to the Jordan matrix $J_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Indeed,

$$A = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}^{-1}.$$

Now, we return to the general case. Let $\mathbb{F}$ be any field. Let $A \in \mathbb{F}^{n\times n}$ be any strictly upper-triangular $n \times n$-matrix. (We don't need to restrict ourselves to the case $\mathbb{F} = \mathbb{C}$ here.) We want to prove that A is similar to a Jordan matrix.

The key to this proof will be to restate the question in terms of certain bases of $\mathbb{F}^n$, and then to construct these bases by an iterative process. We begin with a few notions:

Convention 3.4.5. We fix a nonnegative integer $n \in \mathbb{N}$, a field $\mathbb{F}$ and a strictly upper-triangular matrix $A \in \mathbb{F}^{n\times n}$ for the rest of Subsection 3.4.3.

We observe that

$$A^n = 0. \tag{77}$$

(This is a well-known property of strictly upper-triangular $n \times n$-matrices. It can be obtained by applying Lemma 2.7.5 to $T_i = A$, because A is an upper-triangular matrix whose all diagonal entries are 0. An alternative proof can be found, e.g., in [Grinbe19, Corollary 3.78][37].)

Definition 3.4.6. (a) An *orbit* shall mean a tuple of the form $\left(A^0 v,\ A^1 v,\ \ldots,\ A^k v\right)$, where $v \in \mathbb{F}^n$ is a vector and $k \in \mathbb{N}$ is an integer satisfying $A^{k+1} v = 0$. (We can also write this tuple as $\left(v,\ Av,\ A^2 v,\ \ldots,\ A^k v\right)$.)

(b) The *concatenation* of two tuples $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \ldots, b_\ell)$ is defined to be the tuple $(a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_\ell)$. Thus, concatenation is a binary operation on the set of tuples. Since this operation is associative, we thus obtain the notion of concatenation of several tuples. For example, the concatenation of three tuples $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \ldots, b_\ell)$ and $(c_1, c_2, \ldots, c_m)$ is $(a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_\ell, c_1, c_2, \ldots, c_m)$.

[37] To be precise, [Grinbe19, Corollary 3.78] proves the analogous property for strictly **lower**-triangular matrices. But the case of strictly upper-triangular matrices is analogous (the roles of rows and columns are swapped).

(c) A tuple $(v_1, v_2, \ldots, v_m)$ of vectors in $\mathbb{F}^n$ will be called *forwarded* if each $i \in [m]$ satisfies $Av_i = v_{i+1}$ or $Av_i = 0$. (Here, v_{m+1} is understood to be 0.)

(d) A tuple $(v_1, v_2, \ldots, v_m)$ of vectors in $\mathbb{F}^n$ will be called *backwarded* if each $i \in [m]$ satisfies $Av_i = v_{i-1}$ or $Av_i = 0$. (Here, v_0 is understood to be 0.)

Note that the notions of "orbit", "forwarded" and "backwarded" depend on A, but we do not mention A since A is fixed.

Example 3.4.7. Let p, q, r be three vectors in $\mathbb{F}^n$ satisfying $A^3 p = 0$ and $A^2 q = 0$ and $A^4 r = 0$. Then, the 9-tuple

$$\left(p,\ Ap,\ A^2 p,\ q,\ Aq,\ r,\ Ar,\ A^2 r,\ A^3 r\right)$$

is forwarded. Indeed, if we rename this tuple as $(v_1, v_2, \ldots, v_9)$, then each $i \in \{1, 2, 4, 6, 7, 8\}$ satisfies $Av_i = v_{i+1}$, whereas each $i \in \{3, 5\}$ satisfies $Av_i = 0$. This 9-tuple is furthermore the concatenation of the orbits $\left(p,\ Ap,\ A^2 p\right)$, $\left(q,\ Aq\right)$ and $\left(r,\ Ar,\ A^2 r,\ A^3 r\right)$. Reversing this 9-tuple yields a new 9-tuple

$$\left(A^3 r,\ A^2 r,\ Ar,\ r,\ Aq,\ q,\ A^2 p,\ Ap,\ p\right),$$

which is backwarded.

What we have seen in this example can be generalized:

Proposition 3.4.8. (a) A tuple $(v_1, v_2, \ldots, v_m)$ of vectors in $\mathbb{F}^n$ is forwarded if and only if it is a concatenation of finitely many orbits.

(b) A tuple $(v_1, v_2, \ldots, v_m)$ of vectors in $\mathbb{F}^n$ is backwarded if and only if the tuple $(v_m, v_{m-1}, \ldots, v_1)$ is forwarded.

Proof. TODO: Scribe? □

More importantly, backwarded tuples are closely related to Jordan forms. To wit:

Proposition 3.4.9. Let $(s_1, s_2, \ldots, s_n)$ be a basis of $\mathbb{F}^n$. Let $S \in \mathbb{F}^{n \times n}$ be the $n \times n$-matrix with columns $s_1, s_2, \ldots, s_n$. Then, $S^{-1}AS$ is a Jordan matrix if and only if the n-tuple $(s_1, s_2, \ldots, s_n)$ is backwarded.

Proof. We shall only prove the "if" part, since this is the only part that we will use; however, the proof of the "only if" part can essentially be obtained from the proof of the "if" part by reading it in reverse.

So let us prove the "if" part. Thus, we assume that the n-tuple $(s_1, s_2, \ldots, s_n)$ is backwarded. In other words, each $i \in [n]$ satisfies

$$As_i = s_{i-1} \text{ or } As_i = 0 \tag{78}$$

(where s_0 means 0). Our goal is to show that $S^{-1}AS$ is a Jordan matrix.

The matrix S is invertible, since its columns $s_1, s_2, \ldots, s_n$ form a basis of $\mathbb{F}^n$.

We call an $i \in [n]$

- *red* if it satisfies $As_i = s_{i-1} \neq 0$, and
- *blue* if it satisfies $As_i = 0$.

Thus, (78) shows that each $i \in [n]$ is either red or blue. Note that 1 is always blue, since $s_{1-1} = s_0 = 0$.

Let J be the $n \times n$-matrix whose (i,j)-th entry is

$$\begin{cases} 1, & \text{if } j = i+1 \text{ and } j \text{ is red;} \\ 0, & \text{otherwise.} \end{cases}$$

For instance, if $n = 8$ and if $2, 3, 5, 7, 8$ are red whereas $1, 4, 6$ are blue, then

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

Thus, all entries of J are 0 except for some 1s placed in cells directly above the main diagonal. This shows that J is a Jordan matrix, with each Jordan block covering the rows and columns between one blue $i \in [n]$ and the next. Explicitly, if $i_1, i_2, \ldots, i_k$ are the blue i's listed from smallest to largest (i.e., with $i_1 < i_2 < \cdots < i_k$), then J is the block-diagonal matrix $\begin{pmatrix} J_{i_2 - i_1}(0) & 0 & \cdots & 0 \\ 0 & J_{i_3 - i_2}(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{i_{k+1} - i_k}(0) \end{pmatrix}$, where we set $i_{k+1} := n+1$.

We shall now show that $AS = SJ$. This will entail that $S^{-1}AS = J$, which as we know is a Jordan matrix.

For each $i \in [n]$, we have

$$\begin{aligned} &(\text{the } i\text{-th column of the matrix } AS) \\ &= A \cdot \underbrace{(\text{the } i\text{-th column of the matrix } S)}_{\substack{= s_i \\ (\text{by the definition of } S)}} \qquad (\text{by the rules for multiplying matrices}) \\ &= As_i = \begin{cases} s_{i-1}, & \text{if } i \text{ is red;} \\ 0, & \text{if } i \text{ is blue} \end{cases} \qquad (\text{by the definition of "red" and "blue"}). \end{aligned}$$

On the other hand, for each $i \in [n]$, we have

$$\begin{aligned}
&(\text{the } i\text{-th column of the matrix } SJ) \\
&= S \cdot \underbrace{(\text{the } i\text{-th column of the matrix } J)}_{\substack{= \begin{cases} e_{i-1}, & \text{if } i \text{ is red;} \\ 0, & \text{if } i \text{ is blue} \end{cases} \\ \text{(by the definition of } J\text{)}}} \qquad \text{(by the rules for multiplying matrices)} \\
&= S \cdot \begin{cases} e_{i-1}, & \text{if } i \text{ is red;} \\ 0, & \text{if } i \text{ is blue} \end{cases} = \begin{cases} Se_{i-1}, & \text{if } i \text{ is red;} \\ 0, & \text{if } i \text{ is blue} \end{cases} \\
&= \begin{cases} s_{i-1}, & \text{if } i \text{ is red;} \\ 0, & \text{if } i \text{ is blue.} \end{cases}
\end{aligned}$$

(since $Se_{i-1} = (\text{the } (i-1)\text{-th column of the matrix } S) = s_{i-1}$). Comparing these two equalities, we obtain

$$(\text{the } i\text{-th column of the matrix } AS) = (\text{the } i\text{-th column of the matrix } SJ)$$

for each $i \in [n]$. Hence, $AS = SJ$. Therefore, $S^{-1}AS = J$. Hence, $S^{-1}AS$ is a Jordan matrix (since we know that J is a Jordan matrix). This proves the "if" direction of Proposition 3.4.9. □

Recall that our goal is to show that A is similar to a Jordan matrix. Proposition 3.4.9 shows us a way to this goal: We just need to find a basis for $\mathbb{F}^n$ that is forwarded. In view of Proposition 3.4.8 **(b)**, this is tantamount to finding a basis for $\mathbb{F}^n$ that is backwarded. Let us first see how to do so on examples:

[...]

TODO: Polish from here!

TODO: Empty cells = 0 entries.

Example 3.4.10. Let $n = 4$ and $A =$... (where the cells we leave empty are understood to contain zeroes). Then, ... find some interesting orbits and bases

TODO: Scribe?

We begin by finding forwarded bases in some examples:

Example 3.4.11. Let $n = 2$. Then, $A = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}$ for some $a \in \mathbb{F}$.

We are looking for an invertible matrix $S \in \mathbb{F}^{2\times 2}$ such that $S^{-1}AS$ is a Jordan matrix.

If $a = 0$, then this is obvious (just take $S = I_2$), since $A = \begin{pmatrix} J_1(0) & \\ & J_1(0) \end{pmatrix}$ is already a Jordan matrix.

Now assume $a \neq 0$.

Consider our unknown invertible matrix S. Let s_1 and s_2 be its columns. Then, s_1 and s_2 are linearly independent (since S is invertible). Moreover, we want $S^{-1}AS = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. In other words, we want $AS = S\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. However, $S = \begin{pmatrix} s_1 & s_2 \end{pmatrix}$ (in block-matrix notation), so $S\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & s_1 \end{pmatrix}$. Thus our equation $AS = S\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is equivalent to

$$\begin{pmatrix} As_1 & As_2 \end{pmatrix} = \begin{pmatrix} 0 & s_1 \end{pmatrix}.$$

In other words, $As_1 = 0$ and $As_2 = s_1$.

So we are looking for two linearly independent vectors $s_1, s_2 \in \mathbb{F}^2$ such that $As_1 = 0$ and $As_2 = s_1$.

One way to do so is to pick some nonzero vector $s_1 \in \operatorname{Ker} A$, and then define s_2 to be some preimage of s_1 under A. (It can be shown that such preimage exists.) This way, however, does not generalize to higher n.

Another (better) way is to start by picking $s_2 \in \mathbb{F}^2 \setminus \operatorname{Ker} A$ and then setting $s_1 = As_2$. We claim that s_1 and s_2 are linearly independent, and that $As_1 = 0$.

To show that $As_1 = 0$, we just observe that $As_1 = \underbrace{AA}_{=A^2=0} s_2 = 0$.

To show that s_1 and s_2 are linearly independent, we argue as follows: Let $\lambda_1, \lambda_2 \in \mathbb{F}$ be such that $\lambda_1 s_1 + \lambda_2 s_2 = 0$. Applying A to this, we obtain $A \cdot (\lambda_1 s_1 + \lambda_2 s_2) = A \cdot 0 = 0$. However,

$$A \cdot (\lambda_1 s_1 + \lambda_2 s_2) = \lambda_1 \underbrace{As_1}_{=0} + \lambda_2 \underbrace{As_2}_{=s_1} = \lambda_2 s_1,$$

so this becomes $\lambda_2 s_1 = 0$. However, $s_1 \neq 0$ (because $s_1 = As_2$ but $s_2 \notin \operatorname{Ker} A$). Hence, $\lambda_2 = 0$. Now, $\lambda_1 s_1 + \lambda_2 s_2 = 0$ becomes $\lambda_1 s_1 = 0$. Since $s_1 \neq 0$, this yields $\lambda_1 = 0$. Now both λ_is are 0, qed.

Example 3.4.12. Let $n = 3$ and $A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

Our first method above doesn't work, because most vectors in $\operatorname{Ker} A$ do not have preimages under A.

However, our second method can be made to work:

We pick a vector $s_3 \notin \operatorname{Ker} A$. To wit, we pick $s_3 = e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Then, $As_3 = e_1$. Set $s_2 = As_3 = e_1$. Note that $s_2 \in \operatorname{Ker} A$. Let s_1 be another nonzero vector in $\operatorname{Ker} A$, namely $e_2 - e_3$. These three vectors s_1, s_2, s_3 are linearly independent and satisfy $As_1 = 0$ and $As_2 = 0$ and $As_3 = s_2$.

So $S = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}$. And indeed, $S^{-1}AS = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ is a Jordan matrix.

So what is the general algorithm here? Can we always find n linearly independent vectors $s_1, s_2, \ldots, s_n$ such that each As_i is either 0 or s_{i-1} ?

Now, we return to the general case: How do we find a backwarded basis $(s_1, s_2, \ldots, s_n)$ of $\mathbb{F}^n$?

(The following proof is due to Terence Tao [Tao07].)

We recall that an *orbit* was defined to be a tuple of the form $(v, Av, A^2v, \ldots, A^kv)$, where $v \in \mathbb{F}^n$ satisfies $A^{k+1}v = 0$. Note that for each $v \in \mathbb{F}^n$, there is an orbit that starts with v, since $A^n = 0$.

Now, we claim the existence of a forwarded basis:

Lemma 3.4.13 (orbit basis lemma)**.** There exists a basis of $\mathbb{F}^n$ that is a concatenation of orbits.

Once Lemma 3.4.13 is proved, we will be done, because such a basis will be a forwarded basis (by Proposition 3.4.8 **(a)**), and therefore reading it backwards will gives us a backwarded basis (by Proposition 3.4.8 **(b)**), which is precisely what we wish. For example, if the basis that Lemma 3.4.13 gives us is

$$\left(u, Au, A^2u, \quad v, Av, A^2v, A^3v, \quad w, Aw\right)$$

(with $A^3u = 0$ and $A^4v = 0$ and $A^2w = 0$), then reading it backwards gives

$$\left(Aw, w, \quad A^3v, A^2v, Av, v, \quad A^2u, Au, u\right),$$

which is a backwarded basis of $\mathbb{F}^n$.

Proof of Lemma 3.4.13. It is easy to find a finite **spanning set** of $\mathbb{F}^n$ that is a concatenation of orbits. Indeed, we can start with the standard basis $(e_1, e_2, \ldots, e_n)$, and extend it to the list

$$\begin{aligned}(&e_1, Ae_1, A^2e_1, \ldots, A^{n-1}e_1, \\ &e_2, Ae_2, A^2e_2, \ldots, A^{n-1}e_2, \\ &\ldots, \\ &e_n, Ae_n, A^2e_n, \ldots, A^{n-1}e_n).\end{aligned}$$

This is clearly a spanning set of $\mathbb{F}^n$ (since $e_1, e_2, \ldots, e_n$ already span $\mathbb{F}^n$), and also a concatenation of orbits (since $A^n = 0$).

Now, we will gradually shorten this spanning set (i.e., replace it by smaller ones) until we get a basis. We have to do this in such a way that it remains a spanning set

throughout the process, and that it remains a concatenation of orbits throughout the process.

For the sake of concreteness, let us assume that our spanning set is

$$\left(x, Ax, \ \ y, Ay, A^2y, \ \ z, Az, A^2z, A^3z, \ \ w\right),$$

with $A^2x = 0$ and $A^3y = 0$ and $A^4z = 0$ and $Aw = 0$. If this spanning set is linearly independent, then it is already a basis, and we are done. So assume that it is not. Thus, there exists some linear dependence relation – say,

$$3x + 4Ax + 5Ay + 6A^2y + 7A^2z + 8w = 0.$$

Applying A to this relation, we obtain

$$\begin{aligned} 3Ax + 4A^2x + 5A^2y + 6A^3y + 7A^3z + 8Aw = 0, \qquad &\text{i.e.} \\ 3Ax + 5A^2y + 7A^3z = 0 \end{aligned}$$

(since $A^2x = 0$ and $A^3y = 0$ and $Aw = 0$). Applying A to this relation again, we obtain

$$\begin{aligned} 3A^2x + 5A^3y + 7A^4z = 0, \qquad &\text{i.e.} \\ 0 = 0. \end{aligned}$$

We have gone too far, so let us revert to the previous equation:

$$3Ax + 5A^2y + 7A^3z = 0.$$

So this is a linear dependence relation between the **final** vectors of the orbits in our spanning set. ("*Final*" means the last vector in the orbit.) Factoring out an A in this relation, we obtain

$$A\left(3x + 5Ay + 7A^2z\right) = 0.$$

Thus, the 1-tuple $\left(3x + 5Ay + 7A^2z\right)$ is an orbit.

Now, let us replace the orbit (x, Ax) in our spanning set $\left(x, Ax, \ \ y, Ay, A^2y, \ \ z, Az, A^2z, A^3z, \ \ w\right)$ by the orbit $\left(3x + 5Ay + 7A^2z\right)$. We get

$$\left(3x + 5Ay + 7A^2z, \ \ y, Ay, A^2y, \ \ z, Az, A^2z, A^3z, \ \ w\right).$$

This is still a concatenation of orbits, since the 1-tuple $\left(3x + 5Ay + 7A^2z\right)$ is an orbit. Furthermore, this is still a spanning set of $\mathbb{F}^n$; why? Because we removed the vector Ax, which was unnecessary for spanning $\mathbb{F}^n$ (because the equality $3Ax + 5A^2y + 7A^3z = 0$ reveals that it is a linear combination of the other vectors in our spanning set), and we replaced x by $3x + 5Ay + 7A^2z$ (which does not change the span, because Ay and A^2z are still in the spanning set).

This example generalizes. In the general case, you have a spanning set $\mathbf{s}$ that is a concatenation of orbits:

$$\mathbf{s} = \left(v_1, Av_1, \ldots, A^{m_1}v_1,\ v_2, Av_2, \ldots, A^{m_2}v_2,\ \ldots,\ v_k, Av_k, \ldots, A^{m_k}v_k\right),$$

where $A^{m_1+1}v_1 = 0$ and $A^{m_2+1}v_2 = 0$ and $\ldots$ and $A^{m_k+1}v_k = 0$. If this spanning set $\mathbf{s}$ is a basis, you are done. If not, you pick a linear dependence relation:

$$\sum_{i,j} \lambda_{i,j} A^j v_i = 0.$$

By multiplying this by A an appropriate amount of times (namely, you keep multiplying until it becomes $0 = 0$, and then you take a step back), you obtain a linear dependence relation that involves only the **final** vectors of the orbits (i.e., the vectors $A^{m_1}v_1,\ A^{m_2}v_2,\ \ldots,\ A^{m_k}v_k$). Thus, it will look like this:

$$\mu_1 A^{m_1}v_1 + \mu_2 A^{m_2}v_2 + \cdots + \mu_k A^{m_k}v_k = 0$$

(with at least one of $\mu_1, \mu_2, \ldots, \mu_k$ being nonzero). Assume WLOG that the first p of the scalars $\mu_1, \mu_2, \ldots, \mu_k$ are nonzero, while the remaining $k - p$ are 0 (this can always be achieved by permuting the orbits, which of course does not change anything about the spanning set being a spanning set). So the relation becomes

$$\mu_1 A^{m_1}v_1 + \mu_2 A^{m_2}v_2 + \cdots + \mu_p A^{m_p}v_p = 0,$$

with $\mu_1, \mu_2, \ldots, \mu_p$ being nonzero. Note that $p > 0$ (since at least one of $\mu_1, \mu_2, \ldots, \mu_k$ is nonzero), so that $\mu_1 \neq 0$. Assume WLOG that $m_1 = \min\left\{m_1, m_2, \ldots, m_p\right\}$, and factor out A^{m_1} from this relation. This yields

$$A^{m_1}\left(\mu_1 v_1 + \mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p\right) = 0.$$

Now, set $w_1 = \mu_1 v_1 + \mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p$. Thus, $A^{m_1}w_1 = 0$. Hence, $\left(w_1, Aw_1, A^2w_1, \ldots, A^{m_1-1}w_1\right)$ is an orbit of length m_1. Now, replace the orbit $\left(v_1, Av_1, \ldots, A^{m_1}v_1\right)$ in the spanning set $\mathbf{s}$ by the shorter orbit $\left(w_1, Aw_1, A^2w_1, \ldots, A^{m_1-1}w_1\right)$. The resulting list

$$\left(w_1, Aw_1, A^2w_1, \ldots, A^{m_1-1}w_1,\ v_2, Av_2, \ldots, A^{m_2}v_2,\ \ldots,\ v_k, Av_k, \ldots, A^{m_k}v_k\right)$$

is still a concatenation of orbits (since $A^{m_1}w_1 = 0$). Also, it still spans $\mathbb{F}^n$, because the $m_1 + 1$ vectors $v_1, Av_1, \ldots, A^{m_1}v_1$ that we have removed from $\mathbf{s}$ can be recovered

as linear combinations of the vectors in our new list as follows:

$$\begin{aligned}
v_1 &= \frac{1}{\mu_1}\left(w_1 - \left(\mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p\right)\right) \\
&\qquad\qquad \left(\text{since } w_1 = \mu_1 v_1 + \mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p\right); \\
Av_1 &= A\cdot\frac{1}{\mu_1}\left(w_1 - \left(\mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p\right)\right) \\
&= \frac{1}{\mu_1}\left(Aw_1 - \left(\mu_2 A^{m_2-m_1+1}v_2 + \cdots + \mu_p A^{m_p-m_1+1}v_p\right)\right); \\
A^2v_1 &= A^2\cdot\frac{1}{\mu_1}\left(w_1 - \left(\mu_2 A^{m_2-m_1}v_2 + \cdots + \mu_p A^{m_p-m_1}v_p\right)\right) \\
&= \frac{1}{\mu_1}\left(A^2w_1 - \left(\mu_2 A^{m_2-m_1+2}v_2 + \cdots + \mu_p A^{m_p-m_1+2}v_p\right)\right); \\
&\cdots; \\
A^{m_1}v_1 &= \frac{1}{\mu_1}\left(\underbrace{A^{m_1}w_1}_{=0} - \left(\mu_2 A^{m_2}v_2 + \cdots + \mu_p A^{m_p}v_p\right)\right) \\
&= \frac{1}{\mu_1}\left(-\left(\mu_2 A^{m_2}v_2 + \cdots + \mu_p A^{m_p}v_p\right)\right).
\end{aligned}$$

So we have found a new spanning set of $\mathbb{F}^n$ that is still a concatenation of orbits, but is shorter than $\mathbf{s}$ (namely, it has one less vector than $\mathbf{s}$). In other words, we have found a way to replace a spanning set of $\mathbb{F}^n$ that is a concatenation of orbits by a smaller such set as long as it is linearly independent. Performing this process repeatedly, we will eventually obtain a basis (since we cannot keep making a finite list shorter and shorter indefinitely). This proves Lemma 3.4.13. □

As we said, Lemma 3.4.13 gives us a basis of $\mathbb{F}^n$ that is a concatenation of orbits. In other words, it gives us a forwarded basis (by Proposition 3.4.8 **(a)**), and therefore reading it backwards will gives us a backwarded basis (by Proposition 3.4.8 **(b)**). In view of Proposition 3.4.9, this lets us find an invertible matrix $S \in \mathbb{F}^{n\times n}$ such that $S^{-1}AS$ is a Jordan matrix. This completes the proof of Theorem 3.2.2 **(a)** (the existence part of the Jordan canonical form).

Example 3.4.14. Let $\mathbb{F} = \mathbb{C}$ and

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

This matrix A is not strictly upper-triangular, but it is nilpotent, with $A^3 = 0$, so the above argument goes equally well with this A.

[TODO: Replace this by a better example, with an actual strictly upper-triangular A.]

Let us try to find a basis of $\mathbb{F}^6$ that is a concatenation of orbits.

We begin with the spanning set

$$\left(e_1, Ae_1, A^2e_1,\ \ e_2, Ae_2, A^2e_2,\ \ \ldots,\ \ e_6, Ae_6, A^2e_6\right).$$

It has lots of linear dependencies. For one, $Ae_1 = 0$. Multiplying it by A gives $A^2e_1 = 0$, so we can replace $\left(e_1, Ae_1, A^2e_1\right)$ by (e_1, Ae_1). So our spanning set becomes

$$\left(e_1, Ae_1,\ \ e_2, Ae_2, A^2e_2,\ \ \ldots,\ \ e_6, Ae_6, A^2e_6\right).$$

One more step of the same form gives

$$\left(e_1,\ \ e_2, Ae_2, A^2e_2,\ \ \ldots,\ \ e_6, Ae_6, A^2e_6\right).$$

Now, observe that $Ae_3 = e_2$. That is, $e_2 - Ae_3 = 0$. Multiplying it by A^2, we obtain $A^2e_2 = 0$ (since $A^2 \cdot Ae_3 = A^3e_3 = 0$). So we replace the orbit $\left(e_2, Ae_2, A^2e_2\right)$ by (e_2, Ae_2). So we get the spanning set

$$\left(e_1,\ \ e_2, Ae_2,\ \ e_3, Ae_3, A^2e_3,\ \ e_4, Ae_4, A^2e_4,\ \ e_5, Ae_5, A^2e_5,\ \ e_6, Ae_6, A^2e_6\right).$$

We observe that

$$Ae_2 = e_1 + e_2 + e_3 + e_4 + e_5 + e_6.$$

In other words,

$$Ae_2 - e_1 - e_2 - e_3 - e_4 - e_5 - e_6 = 0.$$

Multiplying this by A^2, we obtain

$$-A^2e_3 - A^2e_4 - A^2e_5 - A^2e_6 = 0.$$

In other words,

$$A^2\left(-e_3 - e_4 - e_5 - e_6\right) = 0.$$

Thus, we set $w_1 := -e_3 - e_4 - e_5 - e_6$, and we replace $\left(e_3, Ae_3, A^2e_3\right)$ by (w_1, Aw_1). So we get the spanning set

$$\left(e_1,\ \ e_2, Ae_2,\ \ w_1, Aw_1,\ \ e_4, Ae_4, A^2e_4,\ \ e_5, Ae_5, A^2e_5,\ \ e_6, Ae_6, A^2e_6\right).$$

Keep making these steps. Eventually, there will be no more linear dependencies, so we will have a basis.

Exercise 3.4.1. [3] Compute the Jordan canonical form of the matrix $\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$.

Exercise 3.4.2. [4] Let $A \in \mathbb{C}^{n \times n}$ be a matrix. Prove that the following three statements are equivalent:

- $\mathcal{A}$: The matrix A is nilpotent.
- $\mathcal{B}$: We have $A^n = 0$.
- $\mathcal{C}$: The only eigenvalue of A is 0 (that is, we have $\sigma(A) = \{0\}$).

Exercise 3.4.3. [4] Let $A \in \mathbb{C}^{n \times n}$ be a matrix. Let $\lambda \in \mathbb{C}$ be nonzero. Prove the following:

(a) If A is nilpotent, then $A \sim \lambda A$.

(b) If $A \sim \lambda A$ and if all n numbers $\lambda^1, \lambda^2, \ldots, \lambda^n$ are distinct from 1, then A is nilpotent.

3.5. Powers and the Jordan canonical form

Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n \times n}$ for some field $\mathbb{F}$. Assume that we know the JCF J of A (this always exists when $\mathbb{F} = \mathbb{C}$, but sometimes exists for other fields as well) and an invertible matrix $S \in \mathbb{F}^{n \times n}$ such that

$$A = SJS^{-1}.$$

Then, it is fairly easy to compute all powers A^m of A. Indeed, recall that

- $\left(SJS^{-1}\right)^m = SJ^mS^{-1}$ for any $m \in \mathbb{N}$.
- $$\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}^m = \begin{pmatrix} A_1^m & & & \\ & A_2^m & & \\ & & \ddots & \\ & & & A_k^m \end{pmatrix} \text{ for any } m \in \mathbb{N}.$$

Thus, it suffices to compute the m-th power of any Jordan cell $J_k(\lambda)$.
So let us consider a Jordan cell

$$C := J_5(\lambda) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & 1 & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}.$$

Then,

$$C^2 = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & & \\ & \lambda^2 & 2\lambda & 1 & \\ & & \lambda^2 & 2\lambda & 1 \\ & & & \lambda^2 & 2\lambda \\ & & & & \lambda^2 \end{pmatrix}; \qquad C^3 = \begin{pmatrix} \lambda^3 & 3\lambda^2 & 3\lambda & 1 & \\ & \lambda^3 & 3\lambda^2 & 3\lambda & 1 \\ & & \lambda^3 & 3\lambda^2 & 3\lambda \\ & & & \lambda^3 & 3\lambda^2 \\ & & & & \lambda^3 \end{pmatrix};$$

$$C^4 = \begin{pmatrix} \lambda^4 & 4\lambda^3 & 6\lambda^2 & 4\lambda & 1 \\ & \lambda^4 & 4\lambda^3 & 6\lambda^2 & 4\lambda \\ & & \lambda^4 & 4\lambda^3 & 6\lambda^2 \\ & & & \lambda^4 & 4\lambda^3 \\ & & & & \lambda^4 \end{pmatrix}; \qquad C^5 = \begin{pmatrix} \lambda^5 & 5\lambda^4 & 10\lambda^3 & 10\lambda^2 & 5\lambda \\ & \lambda^5 & 5\lambda^4 & 10\lambda^3 & 10\lambda^2 \\ & & \lambda^5 & 5\lambda^4 & 10\lambda^3 \\ & & & \lambda^5 & 5\lambda^4 \\ & & & & \lambda^5 \end{pmatrix}.$$

In general, we have the following:

Theorem 3.5.1. Let $\mathbb{F}$ be a field. Let $k > 0$ and $\lambda \in \mathbb{F}$. Let $C = J_k(\lambda)$. Let $m \in \mathbb{N}$. Then, C^m is the upper-triangular $k \times k$-matrix whose (i,j)-th entry is $\binom{m}{j-i}\lambda^{m-j+i}$ for all $i, j \in [k]$. (Here, we follow the convention that $\binom{m}{\ell}\lambda^{m-\ell} := 0$ when $\ell \notin \mathbb{N}$. Also, recall that $\binom{n}{\ell} = 0$ when $n \in \mathbb{N}$ and $\ell > n$.)

First proof of Theorem 3.5.1. Induct on m and use $C^m = CC^{m-1}$ as well as Pascal's recursion

$$\binom{n}{\ell} = \binom{n-1}{\ell} + \binom{n-1}{\ell-1}.$$

□

Second proof of Theorem 3.5.1. Set $B := J_k(0) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$. Proposition 3.1.5 **(a)** tells us what the powers of B are: Namely, B^i has 1s i steps above the main diagonal, and 0s everywhere else.

However, $C = B + \lambda I_k$. The matrices λI_k and B commute (i.e., we have $B \cdot \lambda I_k = \lambda I_k \cdot B$). It is a general fact that if X and Y are two commuting $n \times n$-matrices, then the binomial formula

$$(X+Y)^m = \sum_{i=0}^{m} \binom{m}{i} X^i Y^{m-i} \qquad \text{holds.}$$

(This can be proved in the same way as for numbers, because the commutativity of X and Y lets you move any Xes past any Ys.) Applying this formula to $X = B$ and

$Y = \lambda I_k$, we obtain

$$(B+\lambda I_k)^m = \sum_{i=0}^{m} \binom{m}{i} B^i \underbrace{(\lambda I_k)^{m-i}}_{=\lambda^{m-i} I_k} = \sum_{i=0}^{m} \binom{m}{i} \lambda^{m-i} B^i$$

$$= \begin{pmatrix} \lambda^m & \binom{m}{1}\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots & \cdots & \cdots & \cdots \\ & \lambda^m & \binom{m}{1}\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots & \cdots & \cdots \\ & & \lambda^m & \binom{m}{1}\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots & \cdots \\ & & & \lambda^m & \binom{m}{1}\lambda^{m-1} & \cdots & \cdots \\ & & & & \lambda^m & \cdots & \vdots \\ & & & & & \ddots & \vdots \\ & & & & & & \lambda^m \end{pmatrix},$$

which is precisely the matrix claimed in the theorem. □

Now we know how to take powers of Jordan cells, and therefore how to take powers of any matrix that we know how to bring to a Jordan canonical form.

Corollary 3.5.2. Let $A \in \mathbb{C}^{n \times n}$. Then, $\lim\limits_{m \to \infty} A^m = 0$ if and only if all eigenvalues of A have absolute value < 1.

Proof. $\Longrightarrow$: Suppose that $\lim\limits_{m \to \infty} A^m = 0$, and let λ be an eigenvalue of A. We must show that $|\lambda| < 1$.

Consider a nonzero eigenvector x for eigenvalue λ. Thus, $Ax = \lambda x$. Then, $A^2 x = \lambda^2 x$ (since $A^2 x = A \underbrace{Ax}_{=\lambda x} = \lambda \underbrace{Ax}_{=\lambda x} = \lambda\lambda x = \lambda^2 x$) and similarly $A^3 x = \lambda^3 x$ and $A^4 x = \lambda^4 x$ and so on. Thus,

$$A^m x = \lambda^m x \qquad \text{for each } m \in \mathbb{N}.$$

Now, as $m \to \infty$, the vector $A^m x$ goes to 0 (since $A^m \to 0$). Thus, the vector $\lambda^m x$ goes to 0 as well (since $A^m x = \lambda^m x$ for each $m \in \mathbb{N}$). Since $x \neq 0$, this entails that the scalar λ^m goes to 0 as well. Hence, $|\lambda| < 1$ (because if $|\lambda|$ was ≥ 1, then λ^m would either oscillate along the unit circle, or move further and further away from the origin).

$\Longleftarrow$: Suppose that all eigenvalues of A have absolute value < 1.

Let $A = SJS^{-1}$ be the Jordan canonical form of A. Write J as $\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{pmatrix}$, where $J_1, J_2, \ldots, J_p$ are Jordan cells.

It suffices to show that $\lim\limits_{m\to\infty} J_h^m = 0$ for each $h \in [p]$ (because this will yield $\lim\limits_{m\to\infty} J^m = 0$, and therefore

$$\lim_{m\to\infty} A^m = \lim_{m\to\infty} \underbrace{\left(SJS^{-1}\right)^m}_{=SJ^mS^{-1}} = \lim_{m\to\infty} SJ^mS^{-1} = S\underbrace{\left(\lim_{m\to\infty} J^m\right)}_{=0}S^{-1} = 0,$$

which is what we want to show).

Fix some $h \in [p]$. Write J_h as $J_k(\lambda)$, with $|\lambda| < 1$. Theorem 3.5.1 thus yields that the powers J_h^m of this matrix J_h have a very specific form; in particular, each J_h^m is an upper-triangular $k \times k$-matrix whose (i,j)-th entry is $\binom{m}{j-i}\lambda^{m-j+i}$. Thus, we need to show that for each $i, j \in [k]$, we have

$$\lim_{m\to\infty} \binom{m}{j-i}\lambda^{m-j+i} = 0.$$

However, this follows from a standard asymptotics argument:

$$\underbrace{\binom{m}{j-i}}_{=\frac{m(m-1)(m-2)\cdots(m-j+i)}{(j-i)!}} \underbrace{\lambda^{m-j+i}}_{\substack{\text{exponential in } m,\\ \text{with quotient } \lambda \text{ having absolute value } |\lambda|<1}} \to 0$$

(for $i \leq j$; otherwise the claim is trivial)

because exponential functions with a quotient of absolute value < 1 converge to 0 faster than polynomials can go to ∞. □

Exercise 3.5.1. [6] Let $\lambda \in \mathbb{C}$. Let n and k be two positive integers. Prove the following:

(a) If a $k \times k$-matrix C has eigenvalue λ with algebraic multiplicity k and geometric multiplicity 1, then $C \sim J_k(\lambda)$.

(b) We have $(J_k(\lambda))^n \sim J_k(\lambda^n)$ if $\lambda \neq 0$.

(c) If $A \in \mathbb{C}^{k\times k}$ is an invertible matrix such that A^n is diagonalizable, then A is diagonalizable.

Exercise 3.5.2. [4] Let $\mathbb{F}$ be a field. Compute $(J_k(\lambda))^{-1}$ for any nonzero $\lambda \in \mathbb{F}$ and any $k > 0$.

3.6. The minimal polynomial

Recall: A polynomial $p(t) \in \mathbb{F}[t]$ (where $\mathbb{F}$ is any field, and t is an indeterminate) is said to be *monic* if its leading coefficient is 1 – that is, if it can be written in the

form

$$p(t) = t^m + p_{m-1}t^{m-1} + p_{m-2}t^{m-2} + \cdots + p_0 t^0$$
$$\text{for some } m \in \mathbb{N} \text{ and } p_0, p_1, \ldots, p_{m-1} \in \mathbb{F}.$$

Definition 3.6.1. Given a matrix $A \in \mathbb{F}^{n\times n}$ and a polynomial $p(t) \in \mathbb{F}[t]$, we say that $p(t)$ *annihilates* A if $p(A) = 0$.

The Cayley–Hamilton theorem says that the characteristic polynomial p_A of a square matrix A always annihilates A. However, often there are matrices that are annihilated by other – sometimes simpler – polynomials.

Example 3.6.2. The identity matrix I_n is annihilated by the polynomial $p(t) := t - 1$, because

$$p(I_n) = I_n - I_n = 0.$$

Example 3.6.3. The matrix $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ is annihilated by the polynomial $p(t) := t^2$, since its square is 0.

Example 3.6.4. The diagonal matrix $\begin{pmatrix} 2 & & \\ & 2 & \\ & & 3 \end{pmatrix}$ (where empty cells are understood to be filled with zeroes) is annihilated by the polynomial $p(t) := (t-2)(t-3)$, since

$$p\left(\begin{pmatrix} 2 & & \\ & 2 & \\ & & 3 \end{pmatrix}\right) = \left(\begin{pmatrix} 2 & & \\ & 2 & \\ & & 3 \end{pmatrix} - 2I_3\right)\left(\begin{pmatrix} 2 & & \\ & 2 & \\ & & 3 \end{pmatrix} - 3I_3\right)$$
$$= \begin{pmatrix} 0 & & \\ & 0 & \\ & & 1 \end{pmatrix}\begin{pmatrix} -1 & & \\ & -1 & \\ & & 0 \end{pmatrix} = 0.$$

Theorem 3.6.5. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ be an $n \times n$-matrix. Then, there is a **unique** monic polynomial $q_A(t)$ of minimum degree that annihilates A.

Proof. The Cayley–Hamilton theorem (Theorem 2.7.1) shows that p_A annihilates A. Since p_A is monic, we thus conclude that there exists **some** monic polynomial that annihilates A. Hence, there exists such a polynomial of minimum degree.

It remains to show that it is unique. To do so, we let q_A and $\widetilde{q}_A$ be two monic polynomials of minimum degree that annihilate A. Our goal then is to show that $q_A = \widetilde{q}_A$.

Assume the contrary. Thus, $q_A - \widetilde{q}_A \neq 0$. However, the two polynomials q_A and $\widetilde{q}_A$ have the same degree (since both have minimum degree) and the same leading coefficients (because they are both monic). Thus, their difference $q_A - \widetilde{q}_A$ is a polynomial of smaller degree than q_A and $\widetilde{q}_A$; furthermore, this difference $q_a - \widetilde{q}_A$ annihilates A (because $\left(q_A - \widetilde{q}_A\right)(A) = q_A(A) - \widetilde{q}_A(A) = 0 - 0 = 0$). Thus, by scaling this difference by an appropriate scalar in $\mathbb{F}$, we can make it monic (since it is nonzero), and of course it will still annihilate A. Therefore, we obtain a monic polynomial of smaller degree than q_A that annihilates A. This contradicts the minimality of q_A's degree. This concludes the proof of Theorem 3.6.5. □

Definition 3.6.6. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. Theorem 3.6.5 shows that there is a **unique** monic polynomial $q_A(t)$ of minimum degree that annihilates A. This unique polynomial will be denoted $q_A(t)$ and will be called the *minimal polynomial* of A.

Example 3.6.7. Let $\mathbb{F} = \mathbb{C}$. Let A be the diagonal matrix $\begin{pmatrix} 2 & & \\ & 2 & \\ & & 3 \end{pmatrix}$ (where empty cells are supposed to contain 0 entries). Then,

$$q_A(t) = (t-2)(t-3).$$

Indeed, we already know that the monic polynomial $(t-2)(t-3)$ annihilates A. If there was any monic polynomial of smaller degree that would annihilate A, then it would have the form $t - \lambda$ for some $\lambda \in \mathbb{F}$, but λ cannot be 2 and 3 at the same time.

For comparison: The characteristic polynomial of A is $p_A(t) = (t-2)^2(t-3)$.

Exercise 3.6.1. [2] Find the minimal polynomial of a diagonal matrix whose **distinct** diagonal entries are $\lambda_1, \lambda_2, \ldots, \lambda_k$. (Each of these $\lambda_1, \lambda_2, \ldots, \lambda_k$ can appear on the diagonal any positive number of times.)

Exercise 3.6.2. [3] Let $\mathbb{F}$ be a field. Let $n \geq 2$ be an integer. Let $x_1, x_2, \ldots, x_n \in \mathbb{F}$ and $y_1, y_2, \ldots, y_n \in \mathbb{F}$. Let A be the $n \times n$-matrix $\begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_n \end{pmatrix} \in \mathbb{F}^{n \times n}$.

(a) Find the minimal polynomial of A under the assumption that $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are nonzero.

(b) What changes if we drop this assumption?

[**Hint:** Compute A^2.]

Theorem 3.6.8. Let $A \in \mathbb{F}^{n\times n}$ be an $n \times n$-matrix. Let $f(t) \in \mathbb{F}[t]$ be any polynomial. Then, f annihilates A if and only if f is a multiple of q_A (that is, $f(t) = q_A(t) \cdot g(t)$ for some polynomial $g(t) \in \mathbb{F}[t]$).

Proof. $\Longrightarrow$: Assume that f annihilates A. Thus, $f(A) = 0$. WLOG, assume that $f \neq 0$. Thus, we can make f monic by scaling it. Thus, $\deg f \geq \deg q_A$ (since q_A had minimum degree). Hence, we can divide f by q_A with remainder, obtaining

$$f(t) = q_A(t) \cdot g(t) + r(t), \tag{79}$$

where $g(t)$ and $r(t)$ are two polynomials with $\deg r < \deg q_A$. (Note that $r(t)$ is allowed to be the zero polynomial.) Consider these $g(t)$ and $r(t)$.

Substituting A for t in the equality (79) (and using Lemma 2.7.4 **(a)**), we obtain

$$f(A) = \underbrace{q_A(A)}_{\substack{=0 \\ \text{(since } q_A \text{ annihilates } A\text{)}}} \cdot g(A) + r(A) = r(A),$$

so that $r(A) = f(A) = 0$. In other words, r annihilates A. Since $\deg r < \deg q_A$, this entails that $r = 0$ (since otherwise, we could scale the polynomial $r(t)$ to make it monic, and then we would obtain a monic polynomial of degree $\deg r < \deg q_A$ that annihilates A; but this would contradict the minimality of $\deg q_A$). Thus,

$$f(t) = q_A(t) \cdot g(t) + \underbrace{r(t)}_{=0} = q_A(t) \cdot g(t).$$

Thus, f is a multiple of q_A.

$\Longleftarrow$: Easy and LTTR. □

Corollary 3.6.9. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ be a matrix. Then, $q_A(t) \mid p_A(t)$.

Proof. Apply the previous theorem to $f = p_A$, recalling that p_A annihilates A. □

The corollary yields that any root of q_A must be a root of p_A, that is, an eigenvalue of A. Conversely, we can show that any eigenvalue of A is a root of q_A (but we don't know with which multiplicity):

Proposition 3.6.10. Let $A \in \mathbb{C}^{n\times n}$ be an $n \times n$-matrix. If $\lambda \in \sigma(A)$, then $q_A(\lambda) = 0$.

Proof. Let $\lambda \in \sigma(A)$. Thus, there exists a nonzero eigenvector x for λ.

Then, $Ax = \lambda x$. As we have seen above, this entails $A^m x = \lambda^m x$ for each $m \in \mathbb{N}$. Therefore, $f(A)\,x = f(\lambda)\,x$ for each polynomial $f(t) \in \mathbb{C}[t]$ (because you can write $f(t)$ as $f_0 t^0 + f_1 t^1 + \cdots + f_p t^p$, and then apply $A^m x = \lambda^m x$ to each of $m = 0, 1, \ldots, p$). Hence, $q_A(A)\,x = q_A(\lambda)\,x$, so that

$$q_A(\lambda)\,x = \underbrace{q_A(A)}_{\substack{=0 \\ \text{(since } q_A \text{ annihilates } A\text{)}}} x = 0.$$

Since $x \neq 0$, this entails $q_A(\lambda) = 0$, qed. □

Combining Corollary 3.6.9 with Proposition 3.6.10, we see that the roots of $q_A(t)$ are precisely the eigenvalues of A (when $A \in \mathbb{C}^{n\times n}$); we just don't know yet with which multiplicities they appear as roots. In other words, we have

$$q_A(t) = (t-\lambda_1)^{k_1}(t-\lambda_2)^{k_2}\cdots\left(t-\lambda_p\right)^{k_p},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the distinct eigenvalues of A, and the $k_1, k_2, \ldots, k_p$ are positive integers; but we don't know these $k_1, k_2, \ldots, k_p$ yet. So let us find them. We will use some lemmas for this.

Lemma 3.6.11. Let $\mathbb{F}$ be a field. Let A and B be two similar $n \times n$-matrices in $\mathbb{F}^{n\times n}$. Then, $q_A(t) = q_B(t)$.

Proof. This is obvious from the viewpoint of endomorphisms (see Remark 2.1.6). For a pedestrian proof, you can just argue that a polynomial f annihilates A if and only if it annihilates B. But this is easy: We have $A = SBS^{-1}$ for some invertible S (since A and B are similar), and therefore every polynomial f satisfies

$$f(A) = f\left(SBS^{-1}\right) = Sf(B)S^{-1}$$

and therefore $f(A) = 0$ holds if and only if $f(B) = 0$. □

We recall the notion of the lcm (= least common multiple) of several polynomials. It is defined as one would expect: If $p_1, p_2, \ldots, p_m$ are m nonzero polynomials (in a single indeterminate t), then $\operatorname{lcm}(p_1, p_2, \ldots, p_m)$ is the monic polynomial of smallest degree that is a common multiple of $p_1, p_2, \ldots, p_m$. For example,

$$\begin{aligned}\operatorname{lcm}\left(t^2-1,\ t^3-1\right) &= \operatorname{lcm}\left((t-1)(t+1),\ (t-1)\left(t^2+t+1\right)\right)\\ &= (t-1)(t+1)\left(t+t^2+1\right) = t^4+t^3-t-1.\end{aligned}$$

(Again, the lcm of several polynomials is unique. This can be shown in the same way that we used to prove uniqueness of the minimal polynomial.)

Our next lemma tells us what the minimal polynomial of a block-diagonal matrix is:

Lemma 3.6.12. Let $A_1, A_2, \ldots, A_m$ be m square matrices. Let

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix}.$$

Then,

$$q_A = \operatorname{lcm}\left(q_{A_1},\ q_{A_2},\ \ldots,\ q_{A_m}\right).$$

Proof. For any polynomial $f \in \mathbb{F}[t]$, we have

$$f(A) = f\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix} = \begin{pmatrix} f(A_1) & & & \\ & f(A_2) & & \\ & & \ddots & \\ & & & f(A_m) \end{pmatrix}$$

(indeed, the last equality follows from

$$\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix}^k = \begin{pmatrix} A_1^k & & & \\ & A_2^k & & \\ & & \ddots & \\ & & & A_m^k \end{pmatrix}$$

and from the fact that a polynomial f is just a $\mathbb{F}$-linear combination of t^ks). Thus, $f(A) = 0$ holds if and only if

$$f(A_1) = 0 \text{ and } f(A_2) = 0 \text{ and } \cdots \text{ and } f(A_m) = 0.$$

However, $f(A) = 0$ holds if and only if f is a multiple of q_A, whereas $f(A_i) = 0$ holds if and only if f is a multiple of q_{A_i}. Thus, the previous sentence says that f is a multiple of q_A if and only if f is a multiple of all of the q_{A_i}s. In other words, the multiples of q_A are precisely the common multiples of all the q_{A_i}s. Therefore, q_A is the lcm of the q_{A_i}s (because the universal property of an lcm characterizes the lcm of the q_{A_i}s as the unique monic polynomial whose multiples are the common multiples of all the q_{A_i}s). □

Lemma 3.6.13. Let $\mathbb{F}$ be a field. Let $k > 0$ and $\lambda \in \mathbb{F}$. Let $A = J_k(\lambda)$. Then,

$$q_A = (t - \lambda)^k.$$

Proof. It is easy to see that $q_A = q_{A - \lambda I_k}(t - \lambda)$, because for a polynomial $f \in \mathbb{F}[t]$ to annihilate $A - \lambda I_k$ is the same as for the polynomial $f(t - \lambda)$ to annihilate A. So we need to find $q_{A-\lambda I_k}$. Recall (from Proposition 3.1.4) that

$$A - \lambda I_k = J_k(0) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & \end{pmatrix}.$$

Therefore, for any polynomial $f = f_0 t^0 + f_1 t^1 + f_2 t^2 + \cdots$, we have

$$f\left(A - \lambda I_k\right) = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{k-1} \\ & f_0 & f_1 & \cdots & f_{k-2} \\ & & f_0 & \cdots & f_{k-3} \\ & & & \ddots & \vdots \\ & & & & f_0 \end{pmatrix}.$$

So $f\left(A - \lambda I_k\right) = 0$ if and only if $f_0 = f_1 = \cdots = f_{k-1} = 0$, i.e., if and only if the first k coefficients of f are 0. Now, the monic polynomial of smallest degree whose first k coefficients are 0 is the polynomial t^k. So the monic polynomial f of smallest degree that satisfies $f\left(A - \lambda I_k\right) = 0$ is t^k. In other words, $q_{A-\lambda I_k} = t^k$.

Now, recall that $q_A = q_{A-\lambda I_k}\left(t - \lambda\right) = \left(t - \lambda\right)^k$ (since $q_{A-\lambda I_k} = t^k$). This proves Lemma 3.6.13. □

Theorem 3.6.14. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Let J be the Jordan canonical form of A. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the distinct eigenvalues of A. Then,

$$q_A = \left(t - \lambda_1\right)^{k_1} \left(t - \lambda_2\right)^{k_2} \cdots \left(t - \lambda_p\right)^{k_p},$$

where k_i is the size of the largest Jordan cell at eigenvalue λ_i in J.

Example 3.6.15. Let A have Jordan canonical form

$$J = \begin{pmatrix} 5 & 1 & & & & & & \\ & 5 & 1 & & & & & \\ & & 5 & & & & & \\ & & & 5 & 1 & & & \\ & & & & 5 & & & \\ & & & & & 8 & & \\ & & & & & & 8 & 1 \\ & & & & & & & 8 \end{pmatrix}.$$

Then,

$$q_A = \left(t - 5\right)^3 \left(t - 8\right)^2.$$

Proof of Theorem 3.6.14. We have $A \sim J$, so that $q_A = q_J$ (by Lemma 3.6.11).

Recall that J is a Jordan matrix, i.e., a block-diagonal matrix whose diagonal blocks are Jordan cells $J_1, J_2, \ldots, J_m$. Thus, by Lemma 3.6.12, we have

$$\begin{aligned} q_J &= \operatorname{lcm}\left(q_{J_1},\ q_{J_2},\ \ldots,\ q_{J_m}\right) \\ &= \operatorname{lcm}\left(\left(t - \lambda_{J_1}\right)^{k_{J_1}},\ \left(t - \lambda_{J_2}\right)^{k_{J_2}},\ \ldots,\ \left(t - \lambda_{J_m}\right)^{k_{J_m}}\right), \end{aligned}$$

where each J_i has eigenvalue λ_{J_i} and size k_{J_i} (by Lemma 3.6.13). This lcm must be divisible by each $t - \lambda$ at least as often as each of the $\left(t - \lambda_{J_i}\right)^{k_{J_i}}$s is; i.e., it must be

divisible by $(t-\lambda)^k$, where k is the largest size of a Jordan cell of J at eigenvalue λ. So the lcm is the product of these $(t-\lambda)^k$s. But this is precisely our claim. $\square$

Exercise 3.6.3. $\boxed{1}$ Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times n}$ be any $n \times n$-matrix.

(a) Prove that $q_{A^T} = q_A$, where A^T denotes the transpose of the matrix A.

(b) Assume that $\mathbb{F} = \mathbb{C}$. Prove that $q_{A^*} = \overline{q_A}$, where $\overline{q_A}$ denotes the result of replacing all coefficients of the polynomial q_A by their complex conjugates.

Exercise 3.6.4. $\boxed{5}$ **(a)** A matrix $A \in \mathbb{C}^{3\times 3}$ has characteristic polynomial $t(t-1)(t-2)$. What can its JCF be?

(b) A matrix $A \in \mathbb{C}^{3\times 3}$ has characteristic polynomial $t^2(t-2)$. What can its JCF be?

(c) A matrix $A \in \mathbb{C}^{3\times 3}$ has minimal polynomial $t^2(t-2)$. What can its JCF be?

(d) A matrix $A \in \mathbb{C}^{3\times 3}$ has minimal polynomial $t(t-2)$. What can its JCF be?

Exercise 3.6.5. $\boxed{5}$ Let $A \in \mathbb{C}^{n\times n}$ be a matrix. Prove that $A^T \sim A$, where A^T denotes the transpose of the matrix A.

[**Hint:** Reduce to the case of a Jordan cell.]

3.7. Application of functions to matrices

Consider a square matrix $A \in \mathbb{C}^{n\times n}$. We have already defined what it means to apply a polynomial f to A: We just write f as $\sum_i f_i t^i$, and substitute A for t.

Can we do the same with non-polynomial functions f ? For example, can we define $\exp A$ or $\sin A$?

One option to do so is to follow the same rule as for polynomials, but using the Taylor series for f. For example, since exp has Taylor series $\exp t = \sum_{i\in\mathbb{N}} \frac{t^i}{i!}$, we can set

$$\exp A = \sum_{i\in\mathbb{N}} \frac{A^i}{i!}.$$

This indeed works for exp and for sin, as the sums you get always converge. But it doesn't generally work, e.g., for $f = \tan$, since its Taylor series only converges in a certain neighborhood of 0. Is this the best we can do?

There is a different approach that gives a more general definition. We begin with a lemma about Jordan cells:

Lemma 3.7.1. Let $k > 0$ and $\lambda \in \mathbb{C}$. Let $A = J_k(\lambda)$. Then, for any polynomial $f \in \mathbb{C}[t]$, we have

$$f(A) = \begin{pmatrix} \frac{f(\lambda)}{0!} & \frac{f'(\lambda)}{1!} & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & \frac{f(\lambda)}{0!} & \frac{f'(\lambda)}{1!} & \cdots & \frac{f^{(k-2)}(\lambda)}{(k-2)!} \\ & & \frac{f(\lambda)}{0!} & \cdots & \frac{f^{(k-3)}(\lambda)}{(k-3)!} \\ & & & \ddots & \vdots \\ & & & & \frac{f(\lambda)}{0!} \end{pmatrix}$$

Exercise 3.7.1. [2] Prove Lemma 3.7.1.

Now, we aim to define $f(A)$ by the above formula, at least when A is a Jordan cell. This only requires f to be $(k-1)$-times differentiable at λ.

Definition 3.7.2. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix that has minimal polynomial

$$q_A(t) = (t - \lambda_1)^{k_1} (t - \lambda_2)^{k_2} \cdots (t - \lambda_p)^{k_p},$$

where the $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the distinct eigenvalues of A.

Let f be a function from $\mathbb{C}$ to $\mathbb{C}$ that is defined at each of the numbers $\lambda_1, \lambda_2, \ldots, \lambda_p$ and is holomorphic at each of them, or at least $(k_i - 1)$-times differentiable at each λ_i if λ_i is real. Then, we can define an $n \times n$-matrix $f(A) \in \mathbb{C}^{n \times n}$ as follows: Write $A = SJS^{-1}$, where J is a Jordan matrix and S is invertible. Write

J as $\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{pmatrix}$, where the $J_1, J_2, \ldots, J_m$ are Jordan cells. Then, we set

$$f(A) := Sf(J)S^{-1}, \qquad \text{where}$$

$$f(J) := \begin{pmatrix} f(J_1) & & & \\ & f(J_2) & & \\ & & \ddots & \\ & & & f(J_m) \end{pmatrix}, \qquad \text{where}$$

$$f(J_k(\lambda)) := \begin{pmatrix} \frac{f(\lambda)}{0!} & \frac{f'(\lambda)}{1!} & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & \frac{f(\lambda)}{0!} & \frac{f'(\lambda)}{1!} & \cdots & \frac{f^{(k-2)}(\lambda)}{(k-2)!} \\ & & \frac{f(\lambda)}{0!} & \cdots & \frac{f^{(k-3)}(\lambda)}{(k-3)!} \\ & & & \ddots & \vdots \\ & & & & \frac{f(\lambda)}{0!} \end{pmatrix}.$$

Theorem 3.7.3. This definition is actually well-defined. That is, the value $f(A)$ does not depend on the choice of S and J.

Exercise 3.7.2. [5] Prove this.

[Hint: Use Hermite interpolation to find a polynomial $g \in \mathbb{C}[t]$ such that $g^{(m)}(\lambda) = f^{(m)}(\lambda)$ for each $\lambda \in \sigma(A)$ and each $m \in \{0, 1, \ldots, m_\lambda - 1\}$, where m_λ is the algebraic multiplicity of λ as an eigenvalue of A.]

3.8. The companion matrix

For each $n \times n$-matrix A, we have defined its characteristic polynomial p_A and its minimal polynomial q_A. What variety of polynomials do we get this way? Do all characteristic polynomials share some property, or can any monic polynomial be a characteristic polynomial?

The latter turns out to be true (and moreover, the same holds for the minimal polynomial). We shall prove this by explicitly constructing a matrix with a given polynomial as its characteristic polynomial.

Definition 3.8.1. Let $\mathbb{F}$ be a field, and let $n \in \mathbb{N}$.

Let $f(t) = t^n + f_{n-1}t^{n-1} + f_{n-2}t^{n-2} + \cdots + f_1 t^1 + f_0 t^0$ be a monic polynomial of degree n with coefficients in $\mathbb{F}$. Then, the *companion matrix* of $f(t)$ is defined to be the matrix

$$C_f := \begin{pmatrix} 0 & & & & -f_0 \\ 1 & 0 & & & -f_1 \\ & 1 & 0 & & -f_2 \\ & & 1 & \ddots & \vdots \\ & & & \ddots & 0 & -f_{n-2} \\ & & & & 1 & -f_{n-1} \end{pmatrix} \in \mathbb{F}^{n \times n}$$

(where each cell that is left empty is supposed to be filled with a 0). This is the $n \times n$-matrix whose first $n-1$ columns are the standard basis vectors $e_2, e_3, \ldots, e_n$, and whose last column is $(-f_0, -f_1, \ldots, -f_{n-1})^T$.

Proposition 3.8.2. For any monic polynomial $f(t)$, we have

$$p_{C_f}(t) = q_{C_f}(t) = f(t).$$

Proof. Let us first show that $p_{C_f}(t) = f(t)$.

To do so, we induct on n. The *base case* (that is, the case $n = 0$) is obvious (since the determinant of the 0×0-matrix is 1 by definition). Let us thus proceed to the *induction step*: Let n be a positive integer. Let $f(t) = t^n + f_{n-1}t^{n-1} + f_{n-2}t^{n-2} + \cdots + f_1 t^1 + f_0 t^0$ be a monic polynomial of degree n with coefficients in $\mathbb{F}$. We must show that $p_{C_f}(t) = f(t)$. We assume (as the induction hypothesis) that the same holds for all monic polynomials of degree $n-1$.

Let $g(t)$ be the polynomial $t^{n-1} + f_{n-1}t^{n-2} + \cdots + f_2 t^1 + f_1 t^0$. This is a monic polynomial of degree $n-1$; thus, we can apply the induction hypothesis to it. Thus, we conclude that $p_{C_g}(t) = g(t)$.

The definition of the characteristic polynomial yields

$$p_{C_f}(t) = \det\left(tI_n - C_f\right) = \det \begin{pmatrix} t & & & & & f_0 \\ -1 & t & & & & f_1 \\ & -1 & t & & & f_2 \\ & & & -1 & \ddots & \vdots \\ & & & & \ddots & t & f_{n-2} \\ & & & & & -1 & t + f_{n-1} \end{pmatrix}.$$

We compute this determinant by Laplace expansion along the first row (exploiting

the fact that only two entries of this first row are nonzero):

$$\det \begin{pmatrix} t & & & & & f_0 \\ -1 & t & & & & f_1 \\ & -1 & t & & & f_2 \\ & & -1 & \ddots & & \vdots \\ & & & \ddots & t & f_{n-2} \\ & & & & -1 & t+f_{n-1} \end{pmatrix}$$

$$= t \det \underbrace{\begin{pmatrix} t & & & & & f_1 \\ -1 & t & & & & f_2 \\ & -1 & t & & & f_3 \\ & & -1 & \ddots & & \vdots \\ & & & \ddots & t & f_{n-2} \\ & & & & -1 & t+f_{n-1} \end{pmatrix}}_{\substack{=tI_{n-1}-C_g \\ \text{(by the definition of } C_g\text{)}}} + (-1)^{n+1} f_0 \det \underbrace{\begin{pmatrix} -1 & t & & & \\ & -1 & t & & \\ & & -1 & \ddots & \\ & & & \ddots & t \\ & & & & -1 \end{pmatrix}}_{\substack{=(-1)^{n-1} \\ \text{(since this matrix is upper-triangular of} \\ \text{size } n-1\text{, and its diagonal entries are } -1\text{'s)}}}$$

$$= t \underbrace{\det\left(tI_{n-1}-C_g\right)}_{\substack{=p_{C_g}(t) \\ \text{(by the definition of the characteristic polynomial)}}} + (-1)^{n+1} \underbrace{f_0 (-1)^{n-1}}_{=(-1)^{n-1} f_0}$$

$$= t \underbrace{p_{C_g}(t)}_{\substack{=g(t) \\ =t^{n-1}+f_{n-1}t^{n-2}+\cdots+f_2t^1+f_1t^0}} + \underbrace{(-1)^{n+1}(-1)^{n-1}}_{=1} f_0$$

$$= t\left(t^{n-1}+f_{n-1}t^{n-2}+\cdots+f_2t^1+f_1t^0\right)+f_0$$

$$= t^n + f_{n-1}t^{n-1} + f_2t^2 + f_1t^1 + f_0 = f(t).$$

Thus, $p_{C_f}(t) = f(t)$ is proved. This completes the induction step. Hence, we have proved the $p_{C_f}(t) = f(t)$ part of Proposition 3.8.2.

Now, let us show that $q_{C_f}(t) = f(t)$. Indeed, both $q_{C_f}(t)$ and $f(t)$ are monic polynomials, and we know from Corollary 3.6.9 that $q_{C_f}(t) \mid p_{C_f}(t) = f(t)$. Hence, if $q_{C_f}(t) \neq f(t)$, then $q_{C_f}(t)$ is a proper divisor of $f(t)$, thus has degree $< n$ (since $f(t)$ has degree n). So we just need to rule out the possibility that $q_{C_f}(t)$ has degree $< n$.

Indeed, assume (for the sake of contradiction) that $q_{C_f}(t)$ has degree $< n$. Thus, $q_{C_f}(t) = a_k t^k + a_{k-1}t^{k-1} + \cdots + a_0 t^0$ with $k < n$ and $a_k = 1$ (since q_{C_f} is monic of degree $< n$). However, the definition of q_{C_f} yields $q_{C_f}\left(C_f\right) = 0$. In other words,

$$a_k C_f^k + a_{k-1} C_f^{k-1} + \cdots + a_0 C_f^0 = 0.$$

However, let us look at what C_f does to the standard basis vector $e_1 = (1,0,0,0,\ldots,0)^T$.

We have

$$\begin{aligned} C_f^0 e_1 &= e_1; \\ C_f^1 e_1 &= C_f e_1 = e_2; \\ C_f^2 e_1 &= C_f e_2 = e_3; \\ &\ldots; \\ C_f^{n-1} e_1 &= e_n. \end{aligned}$$

Thus, applying our equality

$$a_k C_f^k + a_{k-1} C_f^{k-1} + \cdots + a_0 C_f^0 = 0$$

to e_1, we obtain

$$a_k e_{k+1} + a_{k-1} e_k + \cdots + a_0 e_1 = 0 \qquad (\text{since } k < n).$$

But this is absurd, since $e_1, e_2, \ldots, e_n$ are linearly independent. So we found a contradiction, and thus we conclude that $q_{C_f}(t)$ has degree $\geq n$. So, by the above, we obtain $q_{C_f}(t) = f(t)$. □

Remark 3.8.3. For algebraists: The companion matrix C_f has a natural meaning. To wit, consider the quotient ring $\mathbb{F}[t] / (f(t))$ as an n-dimensional $\mathbb{F}$-vector space with basis $\left(\overline{t^0}, \overline{t^1}, \ldots, \overline{t^{n-1}}\right)$. Then, the companion matrix C_f represents the endomorphism "multiply by t" (that is, the endomorphism that sends each residue class $\overline{g(t)}$ to $\overline{t \cdot g(t)}$) in this basis.

Exercise 3.8.1. [3] Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix that has n **distinct** eigenvalues. Prove that $A \sim C_{p_A}$.

Exercise 3.8.2. [4] Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. Prove that $A \sim C_{p_A}$ if and only if there exists a vector $v \in \mathbb{F}^n$ such that

$$\left(v,\ Av,\ A^2 v,\ \ldots,\ A^{n-1} v\right) = \left(A^0 v,\ A^1 v,\ \ldots,\ A^{n-1} v\right)$$

is a basis of $\mathbb{F}^n$.

3.9. The Jordan–Chevalley decomposition

Recall that:

- A matrix $A \in \mathbb{C}^{n \times n}$ is said to be *diagonalizable* if it is similar to a diagonal matrix.

- A matrix $A \in \mathbb{C}^{n \times n}$ is said to be *nilpotent* if some power of it is the zero matrix (i.e., if $A^k = 0$ for some $k \in \mathbb{N}$). As we know from Exercise 3.4.2, for an $n \times n$-matrix A to be nilpotent, it is necessary and sufficient that $A^n = 0$.

Theorem 3.9.1 (Jordan–Chevalley decomposition)**.** Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix.

(a) Then, there exists a unique pair (D, N) consisting of

- a diagonalizable matrix $D \in \mathbb{C}^{n \times n}$ and
- a nilpotent matrix $N \in \mathbb{C}^{n \times n}$

such that $DN = ND$ and $A = D + N$.

(b) Both matrices D and N in this pair can be written as polynomials in A. In other words, there exist two polynomials $f, g \in \mathbb{C}[t]$ such that $D = f(A)$ and $N = g(A)$.

The pair (D, N) in this theorem is known as the *Jordan–Chevalley decomposition* (or the *Dunford decomposition*) of A.

For a complete proof of Theorem 3.9.1, see [Bourba03, Chapter VII, §5, section 9, Theorem 1]. An outline can also be found on the Wikipedia page for "Jordan–Chevalley decomposition".

Partial proof of Theorem 3.9.1. We will only show the following claim:

Claim 1: There exists a Jordan–Chevalley decomposition of A.

To prove Claim 1, we can WLOG assume that A is a Jordan matrix. Indeed, if $A = SJS^{-1}$ for some invertible $S \in \mathbb{C}^{n \times n}$ and some Jordan matrix $J \in \mathbb{C}^{n \times n}$, and if (D', N') is a Jordan–Chevalley decomposition of J, then $\left(SD'S^{-1}, SN'S^{-1}\right)$ is a Jordan–Chevalley decomposition of A.

So we WLOG assume that A is a Jordan matrix. Thus,

$$A = \begin{pmatrix} J_{k_1}(\lambda_1) & & & \\ & J_{k_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{k_p}(\lambda_p) \end{pmatrix}$$

for some $\lambda_1, \lambda_2, \ldots, \lambda_p$ and some $k_1, k_2, \ldots, k_p$. (Here, empty cells are understood to be filled with zero matrices.)

We want to find a Jordan–Chevalley decomposition of A. In other words, we want to find a pair (D, N), where $D \in \mathbb{C}^{n \times n}$ is a diagonalizable matrix and $N \in$

$\mathbb{C}^{n\times n}$ is a nilpotent matrix satisfying $DN = ND$ and $A = D + N$. We do this by setting

$$D := \begin{pmatrix} \lambda_1 I_{k_1} & & & \\ & \lambda_2 I_{k_2} & & \\ & & \ddots & \\ & & & \lambda_p I_{k_p} \end{pmatrix} \quad \text{and} \quad N := \begin{pmatrix} J_{k_1}(0) & & & \\ & J_{k_2}(0) & & \\ & & \ddots & \\ & & & J_{k_p}(0) \end{pmatrix}.$$

It is easy to check that $A = D + N$ (since $J_k(\lambda) = \lambda I_k + J_k(0)$ for each $k > 0$ and $\lambda \in \mathbb{C}$) and $DN = ND$ (since block-diagonal matrices can be multiplied block by block, and since matrices of the form λI_k for $k > 0$ and $\lambda \in \mathbb{C}$ commute with every $k \times k$-matrix). Clearly, the matrix D is diagonalizable (since D is diagonal) and the matrix N is nilpotent (since N is strictly upper-triangular). Thus, Claim 1 is proved.

The rest of the proof of Theorem 3.9.1 is omitted for now. □

3.10. The real Jordan canonical form

Given a matrix $A \in \mathbb{R}^{n\times n}$ with real entries, its Jordan canonical form doesn't necessarily have real entries. Indeed, the eigenvalues of A don't have to be real. Sometimes, we want to find a "simple" form for A that does have real entries. What follows is a way to tweak the Jordan canonical form to this use case.

We observe the following:

Lemma 3.10.1. Let $A \in \mathbb{R}^{n\times n}$ and $\lambda \in \mathbb{C}$. Then, the "Jordan structure of A at λ" (meaning the multiset of the sizes of the Jordan blocks of A at λ) equals the Jordan structure of A at $\overline{\lambda}$. In other words, for each $p > 0$, we have

$$\begin{aligned} &(\text{the number of Jordan blocks of } A \text{ at } \lambda \text{ having size } p) \\ &= \left(\text{the number of Jordan blocks of } A \text{ at } \overline{\lambda} \text{ having size } p\right). \end{aligned}$$

In other words, Jordan blocks at λ and Jordan blocks at $\overline{\lambda}$ come in pairs of equal sizes (when $\lambda \neq \overline{\lambda}$).

Exercise 3.10.1. [2] Prove Lemma 3.10.1.

So we can try to combine each Jordan block at λ with an equally sized Jordan block at $\overline{\lambda}$ (when $\lambda \notin \mathbb{R}$) and hope that something real comes out somehow, in the same way as multiplying the complex polynomials $t - \lambda$ and $t - \overline{\lambda}$ yields the real polynomial $(t - \lambda)\left(t - \overline{\lambda}\right) = t^2 - 2(\operatorname{Re}\lambda)\, t + |\lambda|^2 \in \mathbb{R}[t]$.

How to do this? For Jordan blocks of size 1, this is easy:

Lemma 3.10.2. Let $\lambda \in \mathbb{C}$. Let L be the 2×2-matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix}$. Let $a = \operatorname{Re} \lambda$ and $b = \operatorname{Im} \lambda$ (so that $\lambda = a + bi$). Then,

$$L \sim \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

Exercise 3.10.2. $\boxed{2}$ Prove Lemma 3.10.2.

Now, let us see how to combine a Jordan block at λ with an equally sized Jordan block at $\overline{\lambda}$ when the size is arbitrary. We can WLOG assume that these two Jordan blocks are adjacent (since we can permute the Jordan blocks at will). Thus, they form the following matrix together:[38]

$$\begin{pmatrix} J_p(\lambda) & \\ & J_p(\overline{\lambda}) \end{pmatrix} = \begin{pmatrix} \lambda & 1 & & & & & & \\ & \lambda & \ddots & & & & & \\ & & \ddots & 1 & & & & \\ & & & \lambda & & & & \\ & & & & \overline{\lambda} & 1 & & \\ & & & & & \overline{\lambda} & \ddots & \\ & & & & & & \ddots & 1 \\ & & & & & & & \overline{\lambda} \end{pmatrix}.$$

This matrix is similar to the $2p \times 2p$-matrix

$$L_p := \begin{pmatrix} \lambda & & 1 & & & & & \\ & \overline{\lambda} & & 1 & & & & \\ & & \lambda & & 1 & & & \\ & & & \overline{\lambda} & & 1 & & \\ & & & & \ddots & & \ddots & \\ & & & & & \ddots & & \ddots \\ & & & & & & \lambda & \\ & & & & & & & \overline{\lambda} \end{pmatrix} = \begin{pmatrix} L & I_2 & & & \\ & L & I_2 & & \\ & & \ddots & \ddots & \\ & & & L & I_2 \\ & & & & L \end{pmatrix},$$

where L is the 2×2-matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix}$. (In fact, we can easily see that $\begin{pmatrix} J_p(\lambda) & \\ & J_p(\overline{\lambda}) \end{pmatrix} = P_\sigma^{-1} L_p P_\sigma$, where P_σ is the permutation matrix of the permutation σ of $[2p]$ that

[38] Again, empty cells in matrices signify 0s (or zero matrices).

sends $1, 2, 3, \ldots, p, p+1, p+2, p+3, \ldots, 2p$ to $1, 3, 5, \ldots, 2p-1, 2, 4, 6, \ldots, 2p$.) However, Lemma 3.10.2 yields

$$L \sim \begin{pmatrix} a & b \\ -b & a \end{pmatrix},$$

where $a = \operatorname{Re}\lambda$ and $b = \operatorname{Im}\lambda$ (so that $\lambda = a + bi$). So our matrix L_p is similar to

$$\begin{pmatrix} a & b & 1 & & & & & \\ -b & a & & 1 & & & & \\ & & a & b & 1 & & & \\ & & -b & a & & 1 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & a & b \\ & & & & & & -b & a \end{pmatrix}$$

(why?). Hence, altogether, we obtain

$$\begin{pmatrix} J_p(\lambda) & \\ & J_p(\overline{\lambda}) \end{pmatrix} \sim \begin{pmatrix} a & b & 1 & & & & & \\ -b & a & & 1 & & & & \\ & & a & b & 1 & & & \\ & & -b & a & & 1 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & a & b \\ & & & & & & -b & a \end{pmatrix}.$$

The matrix on the right is a real matrix. Thus, we can replace our two Jordan blocks (at λ and $\overline{\lambda}$, of equal sizes) by a real $2p \times 2p$-matrix, obtaining a similar matrix. Performing the same procedure with all Jordan blocks at non-real eigenvalues, we thus obtain a "normal form" that has real entries. See [HorJoh13, §3.4.1] for details and for further results in this direction.

3.11. The centralizer of a matrix

Here is a fairly natural question: Which matrices commute with a given square matrix A ?

Proposition 3.11.1. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. Let f and g be two polynomials in a single variable t over $\mathbb{F}$. Then, $f(A)$ commutes with $g(A)$.

Proof. Write $f(t)$ as $f(t) = \sum_{i=0}^{n} f_i t^i$, and write $g(t)$ as $g(t) = \sum_{j=0}^{m} g_j t^j$. Then,

$$f(A) = \sum_{i=0}^{n} f_i A^i \qquad \text{and} \qquad g(A) = \sum_{j=0}^{m} g_j A^j.$$

Thus,

$$f(A) \cdot g(A) = \left(\sum_{i=0}^{n} f_i A^i \right) \cdot \left(\sum_{j=0}^{m} g_j A^j \right) = \sum_{i=0}^{n} \sum_{j=0}^{m} f_i g_j \underbrace{A^i A^j}_{=A^{i+j}} = \sum_{i=0}^{n} \sum_{j=0}^{m} f_i g_j A^{i+j}.$$

A similar computation shows that

$$g(A) \cdot f(A) = \sum_{i=0}^{n} \sum_{j=0}^{m} f_i g_j A^{i+j}.$$

Comparing these two, we obtain $f(A) \cdot g(A) = g(A) \cdot f(A)$, qed. □

Thus, in particular, $f(A)$ commutes with A for any polynomial f (because $A = g(A)$ for $g(t) = t$).

But are there other matrices that commute with A ?

There certainly can be. For instance, if $A = \lambda I_n$ for some $\lambda \in \mathbb{F}$, then **every** $n \times n$-matrix commutes with A (but very few matrices are of the form $f(A)$ for some polynomial f). This is, in a sense, the "best case scenario". Only for $A = \lambda I_n$ is it true that every $n \times n$-matrix commutes with A.

Let us study the general case now.

Definition 3.11.2. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. The *centralizer* of A is defined to be the set of all $n \times n$-matrices $B \in \mathbb{F}^{n \times n}$ such that $AB = BA$. We denote this set by $\operatorname{Cent} A$.

We thus want to know what $\operatorname{Cent} A$ is.

We begin with some general properties:

Proposition 3.11.3. Let $A \in \mathbb{F}^{n \times n}$ be an $n \times n$-matrix. Then, $\operatorname{Cent} A$ is a subset of $\mathbb{F}^{n \times n}$ that is closed under addition, scaling and multiplication and contains λI_n for all $\lambda \in \mathbb{F}$. In other words:

(a) For any $B, C \in \operatorname{Cent} A$, we have $B + C \in \operatorname{Cent} A$.

(b) For any $B \in \operatorname{Cent} A$ and $\lambda \in \mathbb{F}$, we have $\lambda B \in \operatorname{Cent} A$.

(c) For any $B, C \in \operatorname{Cent} A$, we have $BC \in \operatorname{Cent} A$.

(d) For any $\lambda \in \mathbb{F}$, we have $\lambda I_n \in \operatorname{Cent} A$.

This implies, in particular, that $\operatorname{Cent} A$ is a vector subspace of $\mathbb{F}^{n \times n}$. Furthermore, it shows that $\operatorname{Cent} A$ is an $\mathbb{F}$-subalgebra of $\mathbb{F}^{n \times n}$ (in particular, a subring of $\mathbb{F}^{n \times n}$).

Proof of Proposition 3.11.3. Let me just show part **(c)**; the other parts are even easier.

(c) Let $B, C \in \operatorname{Cent} A$. Thus, $AB = BA$ and $AC = CA$. Now,

$$\underbrace{AB}_{=BA} C = B \underbrace{AC}_{=CA} = BCA.$$

This shows that $BC \in \operatorname{Cent} A$. Thus, part **(c)** is proved. □

Now, as an example, let us compute $\operatorname{Cent} A$ in the case when A is a single Jordan cell $J_n(0)$. So we fix an $n > 0$, and we set

$$A := J_n(0) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Let $B \in \mathbb{F}^{n \times n}$ be arbitrary. We want to know when $B \in \operatorname{Cent} A$. In other words, we want to know when $AB = BA$.

We have

$$AB = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} B_{1,1} & B_{1,2} & B_{1,3} & \cdots & B_{1,n} \\ B_{2,1} & B_{2,2} & B_{2,3} & \cdots & B_{2,n} \\ B_{3,1} & B_{3,2} & B_{3,3} & \cdots & B_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_{n,1} & B_{n,2} & B_{n,3} & \cdots & B_{n,n} \end{pmatrix}$$

$$= \begin{pmatrix} B_{2,1} & B_{2,2} & B_{2,3} & \cdots & B_{2,n} \\ B_{3,1} & B_{3,2} & B_{3,3} & \cdots & B_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_{n,1} & B_{n,2} & B_{n,3} & \cdots & B_{n,n} \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$BA = \begin{pmatrix} B_{1,1} & B_{1,2} & B_{1,3} & \cdots & B_{1,n} \\ B_{2,1} & B_{2,2} & B_{2,3} & \cdots & B_{2,n} \\ B_{3,1} & B_{3,2} & B_{3,3} & \cdots & B_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_{n,1} & B_{n,2} & B_{n,3} & \cdots & B_{n,n} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & B_{1,1} & B_{1,2} & \cdots & B_{1,n-1} \\ 0 & B_{2,1} & B_{2,2} & \cdots & B_{2,n-1} \\ 0 & B_{3,1} & B_{3,2} & \cdots & B_{3,n-1} \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ 0 & B_{n,1} & B_{n,2} & \cdots & B_{n,n-1} \end{pmatrix}.$$

Thus, $AB = BA$ holds if and only if

$$\begin{pmatrix} B_{2,1} & B_{2,2} & B_{2,3} & \cdots & B_{2,n} \\ B_{3,1} & B_{3,2} & B_{3,3} & \cdots & B_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_{n,1} & B_{n,2} & B_{n,3} & \cdots & B_{n,n} \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 0 & B_{1,1} & B_{1,2} & \cdots & B_{1,n-1} \\ 0 & B_{2,1} & B_{2,2} & \cdots & B_{2,n-1} \\ 0 & B_{3,1} & B_{3,2} & \cdots & B_{3,n-1} \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ 0 & B_{n,1} & B_{n,2} & \cdots & B_{n,n-1} \end{pmatrix},$$

i.e., if

$$\begin{aligned} B_{2,j} &= B_{1,j-1} \qquad \text{for all } j \in [n] \text{ (where } B_{1,0} := 0); \\ B_{3,j} &= B_{2,j-1} \qquad \text{for all } j \in [n] \text{ (where } B_{2,0} := 0); \\ B_{4,j} &= B_{3,j-1} \qquad \text{for all } j \in [n] \text{ (where } B_{3,0} := 0); \\ &\ldots; \\ B_{n,j} &= B_{n-1,j-1} \qquad \text{for all } j \in [n] \text{ (where } B_{n-1,0} := 0); \\ 0 &= B_{n,j} \qquad \text{for all } j \in [n-1]. \end{aligned}$$

The latter system of equations can be restated as follows:

$$\begin{aligned} &\ldots; \\ B_{n,n-2} &= B_{n-1,n-3} = B_{n-2,n-4} = \cdots = B_{3,1} = 0; \\ B_{n,n-1} &= B_{n-1,n-2} = B_{n-2,n-3} = \cdots = B_{2,1} = 0; \\ B_{n,n} &= B_{n-1,n-1} = B_{n-2,n-2} = \cdots = B_{1,1}; \\ B_{n-1,n} &= B_{n-2,n-1} = B_{n-3,n-2} = \cdots = B_{1,2}; \\ B_{n-2,n} &= B_{n-3,n-1} = B_{n-4,n-2} = \cdots = B_{1,3}; \\ &\ldots. \end{aligned}$$

In other words, it means that the matrix B looks as follows:

$$B = \begin{pmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \\ & b_0 & b_1 & \cdots & b_{n-2} \\ & & b_0 & \cdots & b_{n-3} \\ & & & \ddots & \vdots \\ & & & & b_0 \end{pmatrix}$$

(where the empty cells have entries equal to 0). This is called an *upper-triangular Toeplitz matrix*. We can also rewrite it as

$$B = b_0 I_n + b_1 A + b_2 A^2 + \cdots + b_{n-1} A^{n-1}.$$

So we have proved the following:

Theorem 3.11.4. Let $n > 0$. Let $A = J_n(0)$. Then,

$$\begin{aligned}
\operatorname{Cent} A &= \left\{ \begin{pmatrix} b_0 & b_1 & b_2 & \cdots & b_{n-1} \\ & b_0 & b_1 & \cdots & b_{n-2} \\ & & b_0 & \cdots & b_{n-3} \\ & & & \ddots & \vdots \\ & & & & b_0 \end{pmatrix} \;\mid\; b_0, b_1, \ldots, b_{n-1} \in \mathbb{F} \right\} \\
&= \left\{ b_0 I_n + b_1 A + b_2 A^2 + \cdots + b_{n-1} A^{n-1} \;\mid\; b_0, b_1, \ldots, b_{n-1} \in \mathbb{F} \right\} \\
&= \left\{ f(A) \;\mid\; f \in \mathbb{F}[t] \text{ is a polynomial of degree } \leq n-1 \right\}.
\end{aligned}$$

So this is the worst-case scenario: The only matrices commuting with A are the matrices of the form $f(A)$ (which, as we recall, must always commute with A).

What happens for an arbitrary A ? Is the answer closer to the best-case scenario or to the worst-case scenario? The answer is that the worst-case scenario holds for a randomly chosen matrix, but we can actually answer the question "what is Cent A exactly" if we know the Jordan canonical form of A.

We start with simple propositions:

Proposition 3.11.5. Let $A \in \mathbb{F}^{n \times n}$ and $\lambda \in \mathbb{F}$. Then, $\operatorname{Cent}(A - \lambda I_n) = \operatorname{Cent} A$.

Exercise 3.11.1. [1] Prove this.

Proposition 3.11.6. Let A, B and S be three $n \times n$-matrices such that S is invertible. Then,

$$(B \in \operatorname{Cent} A) \iff \left(SBS^{-1} \in \operatorname{Cent}\left(SAS^{-1} \right) \right).$$

Exercise 3.11.2. [1] Prove this.

Thus, if A is a matrix with complex entries, and if we want to compute Cent A, it suffices to compute Cent J, where J is the JCF of A.

Therefore, we now focus on centralizers of Jordan matrices. One further simplification stems from the following proposition:

Proposition 3.11.7. Let $A_1, A_2, \ldots, A_k$ be square matrices with complex entries. Assume that the spectra of these matrices are disjoint – i.e., if $i \neq j$, then $\sigma(A_i) \cap \sigma\left(A_j\right) = \varnothing$.

Then,

$$\operatorname{Cent}\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix} = \left\{ \begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_k \end{pmatrix} \mid B_i \in \operatorname{Cent}(A_i) \text{ for each } i \in [k] \right\}.$$

Proof. The $\supseteq$ inclusion is obvious. We thus need to prove the $\subseteq$ inclusion only.

Let A_i be an $n_i \times n_i$-matrix for each $i \in [k]$.

Let $B \in \operatorname{Cent}\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}$. We want to show that B has the form $\begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_k \end{pmatrix}$ where $B_i \in \operatorname{Cent}(A_i)$ for each $i \in [k]$.

Write B as a block matrix

$$B = \begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix},$$

where each $B(i,j)$ is an $n_i \times n_j$-matrix. Then, by the rule for multiplying block matrices, we have

$$\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix} \begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix}$$
$$= \begin{pmatrix} A_1B(1,1) & A_1B(1,2) & \cdots & A_1B(1,k) \\ A_2B(2,1) & A_2B(2,2) & \cdots & A_2B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ A_kB(k,1) & A_kB(k,2) & \cdots & A_kB(k,k) \end{pmatrix}$$

and

$$\begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix} \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}$$
$$= \begin{pmatrix} B(1,1)A_1 & B(1,2)A_2 & \cdots & B(1,k)A_k \\ B(2,1)A_1 & B(2,2)A_2 & \cdots & B(2,k)A_k \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1)A_1 & B(k,2)A_2 & \cdots & B(k,k)A_k \end{pmatrix}.$$

However, these two matrices must be equal, since $\begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix} \in$ Cent $\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}$. Thus, we have

$$\begin{pmatrix} A_1B(1,1) & A_1B(1,2) & \cdots & A_1B(1,k) \\ A_2B(2,1) & A_2B(2,2) & \cdots & A_2B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ A_kB(k,1) & A_kB(k,2) & \cdots & A_kB(k,k) \end{pmatrix} = \begin{pmatrix} B(1,1)A_1 & B(1,2)A_2 & \cdots & B(1,k)A_k \\ B(2,1)A_1 & B(2,2)A_2 & \cdots & B(2,k)A_k \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1)A_1 & B(k,2)A_2 & \cdots & B(k,k)A_k \end{pmatrix}.$$

Comparing blocks, we can rewrite this as

$$A_iB(i,j) = B(i,j)A_j \qquad \text{for all } i,j \in [k].$$

Now, let $i,j \in [k]$ be distinct. Consider this equality $A_iB(i,j) = B(i,j)A_j$. We can rewrite it as $A_iB(i,j) - B(i,j)A_j = 0$. Thus, $B(i,j)$ is an $n_i \times n_j$-matrix X satisfying $A_iX - XA_j = 0$. However, because $\sigma(A_i) \cap \sigma(A_j) = \varnothing$, a theorem we proved before (the $\mathcal{V} \Longrightarrow \mathcal{U}$ direction of Theorem 2.8.2) tells us that there is a **unique** $n_i \times n_j$-matrix X satisfying $A_iX - XA_j = 0$. Clearly, this unique matrix X must be the 0 matrix (since the 0 matrix satisfies $A_i0 - 0A_j = 0$). So we conclude that $B(i,j)$ is the 0 matrix. In other words, $B(i,j) = 0$.

So we have shown that $B(i,j) = 0$ whenever i and j are distinct. Thus,

$$B = \begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix} = \begin{pmatrix} B(1,1) & & & \\ & B(2,2) & & \\ & & \ddots & \\ & & & B(k,k) \end{pmatrix}.$$

This shows that B is block-diagonal. Now, applying the equation

$$A_i B(i,j) = B(i,j) A_j \qquad \text{for all } i,j \in [k]$$

to $j = i$, we obtain $A_i B(i,i) = B(i,i) A_i$, which of course means that $B(i,i) \in \operatorname{Cent}(A_i)$. Thus, B has the form $\begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_k \end{pmatrix}$ where $B_i \in \operatorname{Cent}(A_i)$ for each $i \in [k]$. This completes the proof of Proposition 3.11.7. □

So we only need to compute $\operatorname{Cent} J$ when J is a Jordan matrix with only one eigenvalue.

We can WLOG assume that this eigenvalue is 0, since we know that $\operatorname{Cent}(A - \lambda I_n) = \operatorname{Cent} A$.

So we only need to compute $\operatorname{Cent} J$ when J is a Jordan matrix with zeroes on its diagonal.

If J is just a single Jordan cell, we already know the result (by Theorem 3.11.4). In the general case, we have the following:

Proposition 3.11.8. Let $J \in \mathbb{C}^{n \times n}$ be a Jordan matrix whose Jordan blocks are

$$J_{n_1}(0), \ J_{n_2}(0), \ \ldots, \ J_{n_k}(0).$$

Let B be an $n \times n$-matrix, written as a block matrix

$$B = \begin{pmatrix} B(1,1) & B(1,2) & \cdots & B(1,k) \\ B(2,1) & B(2,2) & \cdots & B(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ B(k,1) & B(k,2) & \cdots & B(k,k) \end{pmatrix},$$

where each $B(i,j)$ is an $n_i \times n_j$-matrix. Then, $B \in \operatorname{Cent} J$ if and only if each of the k^2 blocks $B(i,j)$ is an upper-triangular Toeplitz matrix in the wide sense.

Here, we say that a matrix is an *upper-triangular Toeplitz matrix in the wide sense* if it

- has the form $\begin{pmatrix} 0 & U \end{pmatrix}$, where U is an upper-triangular Toeplitz (square) matrix and 0 is a zero matrix, or
- has the form $\begin{pmatrix} U \\ 0 \end{pmatrix}$, where U is an upper-triangular Toeplitz (square) matrix and 0 is a zero matrix.

(The zero matrices are allowed to be empty.)

Proof. Essentially the same argument that we used to prove Theorem 3.11.4, just with a lot more bookkeeping involved. See [OmClVi11, Proposition 3.1.2] for details. □

We can summarize our results into a single theorem:

Theorem 3.11.9. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix with Jordan canonical form J. Then, Cent A is a vector subspace of $\mathbb{C}^{n \times n}$ with dimension

$$\sum_{\lambda \in \sigma(A)} g_\lambda(A).$$

Here, for each eigenvalue λ of A, the number $g_\lambda(A)$ is a nonnegative integer defined as follows: Let $n_1, n_2, \ldots, n_k$ be the sizes of the Jordan blocks at eigenvalue λ that appear in J; then, we set

$$g_\lambda(A) := \sum_{i=1}^{k} \sum_{j=1}^{k} \min\left\{n_i, n_j\right\}.$$

Proof. Combine Proposition 3.11.6, Proposition 3.11.7, Proposition 3.11.5, and Proposition 3.11.8, and count the degrees of freedom. □

Now, let us return to the worst-case scenario: When is Cent $A = \{f(A) \mid f \in \mathbb{C}[t]\}$? We can answer this, too, although the proof takes longer.

Definition 3.11.10. An $n \times n$-matrix $A \in \mathbb{F}^{n \times n}$ is said to be *nonderogatory* if $q_A = p_A$ (that is, the minimal polynomial of A equals the characteristic polynomial of A).

"Most" matrices are nonderogatory (in the sense that a "randomly chosen" matrix with complex entries will be nonderogatory with probability 1); but there are exceptions. It is easy to see that if a matrix A has n distinct eigenvalues, then A is nonderogatory, but this is not an "if and only if"; a single Jordan cell is also nonderogatory. Here is a necessary and sufficient criterion:

Proposition 3.11.11. An $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ is nonderogatory if and only if its Jordan canonical form has exactly one Jordan block for each eigenvalue.

Exercise 3.11.3. [2] Prove this.

Theorem 3.11.12. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Then,

$$\text{Cent } A = \{f(A) \mid f \in \mathbb{C}[t]\}$$

if and only if A is nonderogatory. Moreover, in this case,

$$\text{Cent } A = \{f(A) \mid f \in \mathbb{C}[t] \text{ is a polynomial of degree } \leq n-1\}.$$

Exercise 3.11.4. [8] Prove this.

4. Hermitian matrices ([HorJoh13, Chapter 4])

Recall: A *Hermitian matrix* is an $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ such that $A^* = A$.

Note that this is the complex analogue of real symmetric matrices (i.e., matrices $A \in \mathbb{R}^{n \times n}$ such that $A^T = A$).

If A is a Hermitian matrix, then $A_{i,i} \in \mathbb{R}$ and $A_{i,j} = \overline{A_{j,i}}$.

For instance, the matrix $\begin{pmatrix} -1 & i & 2 \\ -i & 5 & 1-i \\ 2 & 1+i & 0 \end{pmatrix}$ is Hermitian.

4.1. Basics

Theorem 4.1.1. Let $A \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Then, the following are equivalent:

- $\mathcal{A}$: The matrix A is Hermitian (i.e., we have $A^* = A$).
- $\mathcal{B}$: We have $A = UDU^*$ for some unitary matrix $U \in \mathbb{C}^{n \times n}$ and some real diagonal matrix $D \in \mathbb{C}^{n \times n}$ (that is, D is a diagonal matrix with real entries).
- $\mathcal{C}$: The matrix A is normal and its eigenvalues are real.
- $\mathcal{D}$: We have $\langle Ax, x \rangle \in \mathbb{R}$ for each $x \in \mathbb{C}^n$.
- $\mathcal{E}$: The matrix S^*AS is Hermitian for all $S \in \mathbb{C}^{n \times k}$ (for all $k \in \mathbb{N}$).

To prove this, we will need two lemmas:

Lemma 4.1.2. Let $M \in \mathbb{C}^{n \times n}$ be an $n \times n$-matrix. Let $u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$ be two vectors in $\mathbb{C}^n$. Then,

$$\langle Mu, v \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j} u_j \overline{v_i}.$$

Proof. For each $i \in [n]$, let w_i denote the i-th entry of the column vector w. According to the definition of matrix multiplication, this entry is given by

$$w_i = M_{i,1}u_1 + M_{i,2}u_2 + \cdots + M_{i,n}u_n = \sum_{j=1}^{n} M_{i,j}u_j. \tag{80}$$

However, $Mu = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$; therefore, the definition of the inner product yields

$$\langle Mu, v\rangle = w_1\overline{v_1} + w_2\overline{v_2} + \cdots + w_n\overline{v_n} = \sum_{i=1}^{n} \underbrace{w_i}_{\substack{=\sum_{j=1}^{n} M_{i,j}u_j \\ \text{(by (80))}}} \overline{v_i} = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j}u_j\overline{v_i}.$$

This proves Lemma 4.1.2. □

Lemma 4.1.3. Let $M \in \mathbb{C}^{n\times n}$ be an $n \times n$-matrix. Assume that $\langle Mx, x\rangle = 0$ for each $x \in \mathbb{C}^n$. Then, $M = 0$.

Proof of Lemma 4.1.3. For every $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{C}^n$, we have

$$\begin{aligned} \langle Mx, x\rangle &= \sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j}x_j\overline{x_i} \qquad \text{(by Lemma 4.1.2)} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j}\overline{x_i}x_j \end{aligned}$$

and therefore

$$\sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j}\overline{x_i}x_j = \langle Mx, x\rangle = 0 \tag{81}$$

(since we assumed that $\langle Mx, x\rangle = 0$ for each $x \in \mathbb{C}^n$). In particular:

- We can apply (81) to $x = e_1 = (1, 0, 0, \ldots, 0)^T$, and we obtain $M_{1,1} \cdot \overline{1} \cdot 1 = 0$, which means $M_{1,1} = 0$. Similarly, we can find that

$$M_{i,i} = 0 \qquad \text{for all } i \in [n]. \tag{82}$$

- We can apply (81) to $x = e_1 + e_2 = (1, 1, 0, 0, \ldots, 0)^T$, and we obtain

$$M_{1,1} \cdot \overline{1} \cdot 1 + M_{1,2} \cdot \overline{1} \cdot 1 + M_{2,1} \cdot \overline{1} \cdot 1 \cdot M_{2,2} \cdot \overline{1} \cdot 1 = 0.$$

  This simplifies to

$$M_{1,1} + M_{1,2} + M_{2,1} + M_{2,2} = 0.$$

  However, the previous bullet point yields $M_{1,1} = 0$ and $M_{2,2} = 0$, so this simplifies further to

$$M_{1,2} + M_{2,1} = 0.$$

- We can apply (81) to $x = e_1 + ie_2 = (1, i, 0, 0, \ldots, 0)^T$ (where $i = \sqrt{-1}$), and we obtain

$$M_{1,1} \cdot \overline{1} \cdot 1 + M_{1,2} \cdot \overline{1} \cdot i + M_{2,1} \cdot \overline{i} \cdot 1 \cdot M_{2,2} \cdot \overline{i} \cdot i = 0.$$

  This simplifies to

$$M_{1,1} + iM_{1,2} - iM_{2,1} + M_{2,2} = 0.$$

  However, we know that $M_{1,1} = 0$ and $M_{2,2} = 0$, so this simplifies further to

$$iM_{1,2} - iM_{2,1} = 0.$$

  Thus,

$$M_{1,2} - M_{2,1} = 0.$$

  Adding this to

$$M_{1,2} + M_{2,1} = 0,$$

  we obtain $2M_{1,2} = 0$. In other words, $M_{1,2} = 0$. Similarly, we can show that

$$M_{i,j} = 0 \qquad \text{for all } i \neq j. \tag{83}$$

Combining (82) with (83), we conclude that all entries of M are 0. In other words, $M = 0$. This proves Lemma 4.1.3. □

Now we can prove the theorem:

Proof of Theorem 4.1.1. The equivalence $\mathcal{A} \iff \mathcal{B}$ has already been proved (it is Corollary 2.6.6). The implication $\mathcal{A} \implies \mathcal{C}$ follows from Proposition 2.5.4 **(a)** and Proposition 2.6.5 and Theorem 2.6.1 **(b)**. The implication $\mathcal{C} \implies \mathcal{B}$ follows from Theorem 2.6.1. Combining these facts, we obtain the equivalence $\mathcal{A} \iff \mathcal{B} \iff \mathcal{C}$. So we only need to prove the equivalence $\mathcal{A} \iff \mathcal{D} \iff \mathcal{E}$.

- *Proof of $\mathcal{A} \implies \mathcal{D}$:* Assume that $\mathcal{A}$ holds. Thus, $A = A^*$. Now, let $x \in \mathbb{C}^n$. Then, $\langle x, Ax \rangle = \overline{\langle Ax, x \rangle}$ (by Proposition 1.1.5 **(b)**). However, by Proposition 1.1.5 **(a)**, we have

$$\langle Ax, x \rangle = x^* Ax \qquad \text{and} \qquad \langle x, Ax \rangle = \underbrace{(Ax)^*}_{=x^*A^*} x = x^* \underbrace{A^*}_{=A} x = x^* Ax.$$

Comparing these two equalities, we see that $\langle Ax, x\rangle = \langle x, Ax\rangle = \overline{\langle Ax, x\rangle}$. This entails $\langle Ax, x\rangle \in \mathbb{R}$ (since the only complex numbers $z \in \mathbb{C}$ that satisfy $z = \overline{z}$ are the real numbers). Thus, statement $\mathcal{D}$ is proved.

- *Proof of $\mathcal{D} \Longrightarrow \mathcal{A}$:* Assume that statement $\mathcal{D}$ holds. Thus, $\langle Ax, x\rangle \in \mathbb{R}$ for each $x \in \mathbb{C}^n$. Again, Proposition 1.1.5 **(a)** shows that each $x \in \mathbb{C}^n$ satisfies

$$\langle Ax, x\rangle = x^*Ax \qquad \text{and} \qquad \langle x, Ax\rangle = \underbrace{(Ax)^*}_{=x^*A^*} x = x^*A^*x.$$

Thus, each $x \in \mathbb{C}^n$ satisfies

$$\begin{aligned} x^*Ax &= \langle Ax, x\rangle = \overline{\langle Ax, x\rangle} \qquad (\text{since } \langle Ax, x\rangle \in \mathbb{R}) \\ &= \langle x, Ax\rangle \qquad (\text{by Proposition 1.1.5 } \textbf{(b)}) \\ &= x^*A^*x \end{aligned}$$

and thus

$$x^*\left(A^* - A\right)x = x^*A^*x - \underbrace{x^*Ax}_{=x^*A^*x} = x^*A^*x - x^*A^*x = 0.$$

Applying Lemma 4.1.3 to $M = A^* - A$, we thus conclude that $A^* - A = 0$. In other words, $A^* = A$. This proves statement $\mathcal{A}$.

- *Proof of $\mathcal{A} \Longrightarrow \mathcal{E}$:* If A is Hermitian, then $A^* = A$, so that every matrix $S \in \mathbb{C}^{n\times k}$ satisfies

$$(S^*AS)^* = S^* \underbrace{A^*}_{=A} \underbrace{(S^*)^*}_{=S} = S^*AS,$$

and therefore S^*AS is again Hermitian. This proves the implication $\mathcal{A} \Longrightarrow \mathcal{E}$.

- *Proof of $\mathcal{E} \Longrightarrow \mathcal{A}$:* If statement $\mathcal{E}$ holds, then we can apply it to $S = I_n$ (and $k = n$), and conclude that $I_n^*AI_n$ is Hermitian; but this is simply saying that A is Hermitian. So the implication $\mathcal{E} \Longrightarrow \mathcal{A}$ follows.

Theorem 4.1.1 is thus proved. □

Exercise 4.1.1. $\boxed{1}$ **(a)** Prove the converse of Proposition 1.4.3: If a matrix $A \in \mathbb{C}^{n\times k}$ satisfies $||Ax|| = ||x||$ for each $x \in \mathbb{C}^k$, then A is an isometry.

(b) Prove the converse of Exercise 2.5.4 **(a)**: If a matrix $A \in \mathbb{C}^{n\times n}$ satisfies $||Ax|| = ||A^*x||$ for each $x \in \mathbb{C}^n$, then A is normal.

Let us recall again that sums of Hermitian matrices are Hermitian, but products are not (in general).

4.2. Definiteness and semidefiniteness

Definition 4.2.1. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix.

(a) We say that A is *positive semidefinite* if it satisfies

$$\langle Ax, x\rangle \geq 0 \qquad \text{for all } x \in \mathbb{C}^n.$$

(b) We say that A is *positive definite* if it satisfies

$$\langle Ax, x\rangle > 0 \qquad \text{for all nonzero } x \in \mathbb{C}^n.$$

(c) We say that A is *negative semidefinite* if it satisfies

$$\langle Ax, x\rangle \leq 0 \qquad \text{for all } x \in \mathbb{C}^n.$$

(d) We say that A is *negative definite* if it satisfies

$$\langle Ax, x\rangle < 0 \qquad \text{for all nonzero } x \in \mathbb{C}^n.$$

(e) We say that A is *indefinite* if it is neither positive semidefinite nor negative semidefinite, i.e., if there exist vectors $x, y \in \mathbb{C}^n$ such that

$$\langle Ax, x\rangle < 0 < \langle Ay, y\rangle .$$

Here are some examples of matrices that are definite, semidefinite or neither:

Example 4.2.2. Let $n \in \mathbb{N}$. Let $J = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$. This matrix J is real symmetric, thus Hermitian. Is it positive definite? Positive semidefinite? Let us see.

Let $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$. Then, Lemma 4.1.2 yields

$$\begin{aligned}\langle Jx, x\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{x_j} x_i = \left(\sum_{j=1}^{n} \overline{x_j}\right)\left(\sum_{i=1}^{n} x_i\right) = \left(\sum_{i=1}^{n} \overline{x_i}\right)\left(\sum_{i=1}^{n} x_i\right) \\ &= \overline{\left(\sum_{i=1}^{n} x_i\right)}\left(\sum_{i=1}^{n} x_i\right) = \left|\sum_{i=1}^{n} x_i\right|^2 \geq 0.\end{aligned}$$

So J is positive semidefinite.

Is J positive definite? Again, let $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$. We just have shown that $\langle Jx, x\rangle = \left|\sum_{i=1}^{n} x_i\right|^2$. Therefore, to have $\langle Jx, x\rangle = 0$ is equivalent to having $\sum_{i=1}^{n} x_i = 0$. When $n = 1$ (or $n = 0$), this is equivalent to having $x = 0$, so

we can conclude that J is positive definite in this case. However, if $n > 1$, then this is not equivalent to having $x = 0$, and in fact the vector $e_1 - e_2$ is an example of a nonzero vector $x \in \mathbb{C}^n$ such that $\langle Jx, x\rangle = 0$. So J is not positive definite unless $n \leq 1$.

Example 4.2.3. Consider a diagonal matrix

$$D := \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

with $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$. When is D positive semidefinite?

We want $\langle Dx, x\rangle \geq 0$ for all $x \in \mathbb{C}^n$. Let $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$. Then,

$$\langle Dx, x\rangle = \sum_{i=1}^{n} \lambda_i \underbrace{\overline{x_i} x_i}_{=|x_i|^2} = \sum_{i=1}^{n} \lambda_i |x_i|^2 .$$

If $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$, then we therefore conclude that $\langle Dx, x\rangle \geq 0$, so that D is positive semidefinite. Otherwise, D is not positive semidefinite, since we can pick an $x = e_j$ where j satisfies $\lambda_j < 0$. So D is positive semidefinite if and only if $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$. A similar argument shows that D is positive definite if and only if $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$.

Example 4.2.4. The Hilbert matrix

$$\begin{pmatrix} \frac{1}{1} & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n} \end{pmatrix}$$

(i.e., the $n \times n$-matrix whose (i,j)-th entry is $\dfrac{1}{i+j-1}$) is positive definite. In other words, for any $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\overline{x_i} x_j}{i+j-1} \geq 0.$$

This is not obvious at all, and the proof will be the content of the next exercise. More generally, if $a_1, a_2, \ldots, a_n$ are positive reals, then the $n \times n$-matrix whose (i,j)-th entry is $\dfrac{1}{a_i + a_j}$ is positive definite.

Exercise 4.2.1. [4] Let $a_1, a_2, \ldots, a_n$ be positive reals. Let A be the $n \times n$-matrix whose (i,j)-th entry is $\dfrac{1}{a_i + a_j}$. Prove that A is positive definite.

[**Hint:** Recall that $\dfrac{1}{m} = \int_0^1 t^{m-1} dt$ for each $m > 0$. Also, integrating an $\mathbb{R}_{\geq 0}$-valued function over $[0,1]$ yields a nonnegative real.]

Exercise 4.2.2. [5] Let $a_1, a_2, \ldots, a_n$ be reals. Let $A \in \mathbb{R}^{n \times n}$ be the $n \times n$-matrix

$$\begin{pmatrix} a_1 & a_1 & \cdots & a_1 & a_1 \\ a_1 & a_2 & \cdots & a_2 & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1 & a_2 & \cdots & a_{n-1} & a_{n-1} \\ a_1 & a_2 & \cdots & a_{n-1} & a_n \end{pmatrix}$$

(that is, the $n \times n$-matrix whose (i,j)-th entry is $a_{\min\{i,j\}}$). This matrix A is real symmetric and thus Hermitian.

(a) Set $a_0 := 0$, and let $d_i := a_i - a_{i-1}$ for each $i \in [n]$. Let D be the diagonal matrix $\operatorname{diag}(d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n \times n}$. Let U be the upper-triangular matrix

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n},$$

all of whose entries on and above the main diagonal are 1. Prove that $A = U^* D U$.

(b) Prove that A is positive definite if and only if $0 < a_1 < a_2 < \cdots < a_n$.

Note that a Hermitian matrix A is negative definite if and only if $-A$ is positive definite. Similarly, a Hermitian matrix A is negative semidefinite if and only if $-A$ is positive semidefinite. (These claims follow easily from the definitions.)

As an application of positive semidefiniteness, the Schoenberg theorem generalizes the triangle inequality. Recall that the triangle inequality says that three nonnegative real numbers x, y, z are the mutual distances of 3 points in the plane if and only if $x \leq y + z$ and $y \leq z + x$ and $z \leq x + y$. More generally, Schoenberg's theorem gives a criterion for when a bunch of nonnegative reals can be realized as mutual distances of n points in an r-dimensional real vector space:

Theorem 4.2.5 (Schoenberg's theorem)**.** Let $n \in \mathbb{N}$ and $r \in \mathbb{N}$. Let $d_{i,j}$ be a nonnegative real for each $i, j \in [n]$. Assume that $d_{i,i} = 0$ for all $i \in [n]$, and

furthermore $d_{i,j} = d_{j,i}$ for all $i, j \in [n]$. Then, there exist n points $P_1, P_2, \ldots, P_n \in \mathbb{R}^r$ satisfying

$$\left|P_i - P_j\right| = d_{i,j} \qquad \text{for all } i, j \in [n]$$

if and only if the $(n-1) \times (n-1)$-matrix whose (i,j)-th entry is

$$d_{i,n}^2 + d_{j,n}^2 - d_{i,j}^2 \qquad \text{for all } i, j \in [n-1]$$

is positive semidefinite and has rank $\leq r$.

We will not prove this here. (See [LibLav15, Theorem 7.1] for a proof.)

Remark 4.2.6. If $A \in \mathbb{R}^{n \times n}$ and $\langle Ax, x \rangle \geq 0$ for all $x \in \mathbb{R}^n$, then we **cannot** conclude that A is positive semidefinite. The reason is that it does not follow that A is symmetric. For example, $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ satisfies

$$\langle Ax, x \rangle = 2x_1^2 + x_1 x_2 + 2x_2^2 = \frac{1}{2}\left(x_1 + x_2\right)^2 + \frac{3}{2}\left(x_1^2 + x_2^2\right) \geq 0$$

for each $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, but it is not symmetric.

Theorem 4.2.7. Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Then:

(a) The matrix A is positive semidefinite if and only if all eigenvalues of A are nonnegative.

(b) The matrix A is positive definite if and only if all eigenvalues of A are positive.

(Recall that the eigenvalues of A are real by the spectral theorem.)

Proof. By the spectral theorem (Corollary 2.6.6), the matrix A is unitarily similar to a diagonal matrix with real entries. In other words, $A = UDU^*$ for some unitary matrix $U \in \mathrm{U}_n(\mathbb{C})$ and some diagonal matrix $D \in \mathbb{C}^{n \times n}$ that has real entries. Consider these U and D. From Theorem 2.6.1 **(b)**, we know that the diagonal entries of D are the eigenvalues of A. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the diagonal entries of D, so that $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Then, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A (since the diagonal entries of D are the eigenvalues of A).

(a) $\Longrightarrow$: Assume that A is positive semidefinite. Let λ be an eigenvalue of A. Let $x \neq 0$ be a corresponding eigenvector. Then, $Ax = \lambda x$. However, $\langle Ax, x \rangle \geq 0$ since A is positive semidefinite. However, from $Ax = \lambda x$, we obtain $\langle Ax, x \rangle = \langle \lambda x, x \rangle = \lambda \langle x, x \rangle$. Thus, $\lambda \langle x, x \rangle = \langle Ax, x \rangle \geq 0$. We can cancel $\langle x, x \rangle$ from this inequality (since $\langle x, x \rangle > 0$). Thus, we get $\lambda \geq 0$. Therefore, all eigenvalues of A are ≥ 0.

$\Longleftarrow$: Assume that all eigenvalues of A are ≥ 0. In other words, $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ (since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A). Thus, the square roots $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$

are well-defined nonnegative reals. Set $E := \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right) \in \mathbb{C}^{n \times n}$. Then,

$$\begin{aligned} E^2 &= \left(\operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right)\right)^2 = \operatorname{diag}\left(\left(\sqrt{\lambda_1}\right)^2, \left(\sqrt{\lambda_2}\right)^2, \ldots, \left(\sqrt{\lambda_n}\right)^2\right) \\ &= \operatorname{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right) = D, \end{aligned}$$

so that $D = E^2 = EE$. Moreover, $E^* = E$, since E is a diagonal matrix with real entries. Now,

$$A = U \underbrace{D}_{=EE} U^* = UE \underbrace{E}_{=E^*} U^* = \underbrace{UE}_{=\left((UE)^*\right)^*} \underbrace{E^* U^*}_{=(UE)^*} = \left((UE)^*\right)^* (UE)^*.$$

Hence, for each $x \in \mathbb{C}^n$, we have

$$\begin{aligned} \langle Ax, x \rangle &= x^* A x \qquad \text{(by the formula } \langle u, v \rangle = v^* u\text{)} \\ &= x^* \left((UE)^*\right)^* (UE)^* x \qquad \left(\text{since } A = \left((UE)^*\right)^* (UE)^*\right) \\ &= \left\langle (UE)^* x,\ (UE)^* x \right\rangle \qquad \text{(by the formula } \langle u, v \rangle = v^* u\text{)} \\ &\geq 0 \qquad \text{(since } \langle u, u \rangle \geq 0 \text{ for each } u \in \mathbb{C}^n\text{)}. \end{aligned}$$

Thus, A is positive semidefinite.

We have thus proved Proposition 4.2.7 **(a)**. The proof of Proposition 4.2.7 **(b)** is similar, except that we need to also observe that $x \neq 0$ entails $(UE)^* x \neq 0$ (because U and E are invertible, thus UE is invertible, thus $(UE)^*$ is invertible). □

Exercise 4.2.3. [4] Let $A, B \in \mathbb{C}^{n \times n}$ be two positive definite Hermitian matrices.

(a) Prove that $A + B$ is positive definite.

(b) Find a counterexample showing that AB is not necessarily Hermitian.

(c) Now assume that $AB = BA$. Prove that AB is Hermitian and positive definite.

[**Hint:** In part **(c)**, use Exercise 2.6.7.]

4.3. The Cholesky decomposition

Theorem 4.3.1 (Cholesky decomposition for positive definite matrices)**.** Let $A \in \mathbb{C}^{n \times n}$ be a positive definite Hermitian matrix. Then, A has a unique factorization of the form

$$A = LL^*,$$

where $L \in \mathbb{C}^{n \times n}$ is a lower-triangular matrix whose diagonal entries are positive reals.

Example 4.3.2. For $n = 1$, the theorem is trivial: In this case, $A = \begin{pmatrix} a \end{pmatrix}$ for some $a \in \mathbb{R}$, and this a is > 0 because A is positive definite. Thus, setting $L = \begin{pmatrix} \sqrt{a} \end{pmatrix}$, we obtain $A = LL^*$. Moreover, this is clearly the only choice for L.

Example 4.3.3. Let us manually verify Theorem 4.3.1 for $n = 2$. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a positive definite Hermitian matrix. We are looking for a lower-triangular matrix $L = \begin{pmatrix} \lambda & 0 \\ x & \delta \end{pmatrix}$ whose diagonal entries λ and δ are positive reals that satisfies $A = LL^*$.

So we need

$$\begin{aligned} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A = LL^* &= \begin{pmatrix} \lambda & 0 \\ x & \delta \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ x & \delta \end{pmatrix}^* \\ &= \begin{pmatrix} \lambda & 0 \\ x & \delta \end{pmatrix} \begin{pmatrix} \lambda & \overline{x} \\ 0 & \delta \end{pmatrix} \qquad (\text{since } \lambda, \delta \text{ are real}) \\ &= \begin{pmatrix} \lambda^2 & \lambda\overline{x} \\ \lambda x & x\overline{x} + \delta^2 \end{pmatrix} = \begin{pmatrix} \lambda^2 & \lambda\overline{x} \\ \lambda x & |x|^2 + \delta^2 \end{pmatrix}. \end{aligned}$$

So we need to solve the system of equations

$$\begin{cases} a = \lambda^2; \\ b = \lambda\overline{x}; \\ c = \lambda x; \\ d = |x|^2 + \delta^2. \end{cases}$$

First, we solve the equation $a = \lambda^2$ by setting $\lambda = \sqrt{a}$. Since A is positive definite, we have $a = \langle Ae_1, e_1 \rangle > 0$, so that $\sqrt{a}$ is well-defined, and we get a positive real λ. Next, we solve the equation $c = \lambda x$ by setting $x = \dfrac{c}{\lambda}$. Next, the equation $b = \lambda\overline{x}$ is automatically satisfied, since the Hermitianness of A entails $b = \overline{c} = \overline{\lambda x} = \lambda\overline{x}$ (since λ is real). Finally, we solve the equation $d = |x|^2 + \delta^2$ by setting $\delta = \sqrt{d - |x|^2}$. Here, we need to convince ourselves that $d - |x|^2$ is a positive real, i.e., that $d > |x|^2$. Why is this the case?

I claim that this follows from applying $\langle Az, z \rangle \geq 0$ to the vector $z = \begin{pmatrix} b \\ -a \end{pmatrix}$. Indeed, setting $z = \begin{pmatrix} b \\ -a \end{pmatrix}$, we obtain

$$Az = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} b \\ -a \end{pmatrix} = \begin{pmatrix} 0 \\ bc - ad \end{pmatrix},$$

so that

$$\langle Az, z \rangle = \left\langle \begin{pmatrix} 0 \\ bc - ad \end{pmatrix}, \begin{pmatrix} b \\ -a \end{pmatrix} \right\rangle = (bc - ad)\,\overline{-a} = a\,(ad - bc)$$

and thus $a(ad-bc) = \langle Az, z\rangle > 0$ (by the positive definiteness of A, since $z \neq 0$). We can divide this inequality by a (since $a > 0$), and obtain $ad - bc > 0$. Now, recall that $x = \dfrac{c}{\lambda}$ and $\lambda = \sqrt{a}$. Hence,

$$\begin{aligned} d - |x|^2 &= d - \left|\frac{c}{\lambda}\right|^2 = d - \frac{c\overline{c}}{\lambda^2} = d - \frac{cb}{a} \qquad \left(\text{since } \overline{c} = b \text{ and } \lambda^2 = a\right) \\ &= \frac{ad-bc}{a} > 0 \qquad (\text{since } ad - bc > 0 \text{ and } a > 0)\,. \end{aligned}$$

This is what we need. So Theorem 4.3.1 is proved for $n = 2$.

To prove Theorem 4.3.1 in general, we need a lemma that essentially generalizes our above argument for $d - |x|^2 > 0$:

Lemma 4.3.4. Let $Q \in \mathbb{C}^{n\times n}$ be a invertible matrix. Let $x \in \mathbb{C}^n$ be some column vector. Let $d \in \mathbb{R}$. Let

$$A := \begin{pmatrix} QQ^* & Qx \\ (Qx)^* & d \end{pmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}.$$

Assume that A is positive definite. Then, $||x||^2 < d$.

Proof of Lemma 4.3.4. Set $Q^{-*} := \left(Q^{-1}\right)^* = \left(Q^*\right)^{-1}$. (This is well-defined, since Q is invertible.) Set $u = \begin{pmatrix} Q^{-*}x \\ -1 \end{pmatrix} \in \mathbb{C}^{n+1}$. (This is in block-matrix notation. Explicitly, this is the column vector obtained by appending the extra entry -1 at the bottom of $Q^{-*}x$.)

The definitions of A and u yield

$$\begin{aligned} Au &= \begin{pmatrix} QQ^* & Qx \\ (Qx)^* & d \end{pmatrix} \begin{pmatrix} Q^{-*}x \\ -1 \end{pmatrix} = \begin{pmatrix} QQ^*Q^{-*}x + Qx(-1) \\ (Qx)^* Q^{-*}x + d(-1) \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ x^*Q^*Q^{-*}x - d \end{pmatrix} = \begin{pmatrix} 0 \\ x^*x - d \end{pmatrix} = \begin{pmatrix} 0 \\ ||x||^2 - d \end{pmatrix} \end{aligned}$$

(since $x^*x = \langle x, x\rangle = ||x||^2$). Hence,

$$\langle Au, u\rangle = \left\langle \begin{pmatrix} 0 \\ ||x||^2 - d \end{pmatrix}, \begin{pmatrix} Q^{-*}x \\ -1 \end{pmatrix} \right\rangle = \left(||x||^2 - d\right)\left(\overline{-1}\right) = d - ||x||^2\,.$$

However, the vector u is nonzero (since its last entry is -1), and the matrix A is positive definite (by assumption). Thus, $\langle Au, u\rangle > 0$. Since $\langle Au, u\rangle = d - ||x||^2$, we thus obtain $d - ||x||^2 > 0$. In other words, $d > ||x||^2$. This proves Lemma 4.3.4. □

Now, let us prove the Cholesky factorization theorem:

Proof of Theorem 4.3.1. We proceed by induction on n.

The *base cases* $n = 0$ and $n = 1$ are essentially obvious ($n = 1$ was done in Example 4.3.2).

Induction step: Assume that Theorem 4.3.1 holds for some n. We must prove that it holds for $n+1$ as well.

Let $A \in \mathbb{C}^{(n+1)\times(n+1)}$ be a positive definite Hermitian matrix. Write A in the block-matrix form

$$A = \begin{pmatrix} B & b \\ b^* & d \end{pmatrix},$$

where $B \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$ and $d \in \mathbb{C}$. Note that the b^* on the bottom of the right hand side is because A is Hermitian, so all entries in the last row of A are the complex conjugates of the corresponding entries in the last column of A. Also, $d = d^*$ for the same reason, so $d \in \mathbb{R}$. Moreover, B is Hermitian (since A is Hermitian).

Next, we claim that B is positive definite. Indeed, for any nonzero vector $x \in \mathbb{C}^n$, we have $\langle Bx, x\rangle = \langle Ax', x'\rangle$, where x' is the nonzero vector $\begin{pmatrix} x \\ 0 \end{pmatrix} \in \mathbb{C}^{n+1}$. Thus, positive definiteness of B follows from positive definiteness of A. (More generally, any principal submatrix of a positive definite matrix is positive definite.)

Therefore, by the induction hypothesis, we can apply Theorem 4.3.1 to the $n \times n$-matrix B instead of A. We conclude that B can be uniquely written as a product $B = QQ^*$, where $Q \in \mathbb{C}^{n\times n}$ is a lower-triangular matrix whose diagonal entries are positive reals. Consider this Q. Note that the matrix Q is invertible (since it is lower-triangular and its diagonal entries are positive).

Now, we want to find a vector $x \in \mathbb{C}^n$ and a positive real δ such that if we set

$$L := \begin{pmatrix} Q & 0 \\ x^* & \delta \end{pmatrix},$$

then $A = LL^*$. If we can find such x and δ, then at least the existence part of Theorem 4.3.1 will be settled.

So let us set $L := \begin{pmatrix} Q & 0 \\ x^* & \delta \end{pmatrix}$, and see what conditions $A = LL^*$ places on x and δ. We want

$$\begin{aligned} \begin{pmatrix} B & b \\ b^* & d \end{pmatrix} = A = LL^* &= \begin{pmatrix} Q & 0 \\ x^* & \delta \end{pmatrix} \begin{pmatrix} Q^* & (x^*)^* \\ 0 & \overline{\delta} \end{pmatrix} \\ &= \begin{pmatrix} Q & 0 \\ x^* & \delta \end{pmatrix} \begin{pmatrix} Q^* & x \\ 0 & \delta \end{pmatrix} \qquad \left(\text{since } \delta \in \mathbb{R} \text{ entails } \overline{\delta} = \delta\right) \\ &= \begin{pmatrix} QQ^* & Qx \\ x^*Q^* & x^*x + \delta^2 \end{pmatrix}. \end{aligned}$$

In other words, we want

$$\begin{cases} B = QQ^*; \\ b = Qx; \\ b^* = x^*Q^*; \\ d = x^*x + \delta^2. \end{cases}$$

The first of these four equations is already satisfied (we know that $B = QQ^*$). The second equation will be satisfied if we set $x = Q^{-1}b$. We can indeed set $x = Q^{-1}b$, since the matrix Q is invertible. The third equation follows automatically from the second (indeed, $b = Qx$ entails $b^* = (Qx)^* = x^*Q^*$). Finally, the fourth equation rewrites as $d = ||x||^2 + \delta^2$. We can satisfy it by setting $\delta = \sqrt{d - ||x||^2}$, as long as we can show that $d - ||x||^2 > 0$. Fortunately, we can indeed show this, because Lemma 4.3.4 yields that $||x||^2 < d$. Thus, we have found x and δ, and constructed a lower-triangular matrix L whose diagonal entries are positive reals and which satisfies $A = LL^*$.

It remains to show that this L is unique. Indeed, we can basically read our argument above backwards. **If** $L \in \mathbb{C}^{(n+1)\times(n+1)}$ is a lower-triangular matrix whose diagonal entries are positive reals and which satisfies $A = LL^*$, then we can write A in the form $A = \begin{pmatrix} Q & 0 \\ x^* & \delta \end{pmatrix}$ for some $Q \in \mathbb{C}^{n\times n}$ and $x \in \mathbb{C}^n$ and some positive real δ, where Q is lower-triangular with its diagonal entries being real. The equation $A = LL^*$ then rewrites as

$$\begin{pmatrix} B & b \\ b^* & d \end{pmatrix} = \begin{pmatrix} QQ^* & Qx \\ x^*Q^* & x^*x + \delta^2 \end{pmatrix}. \tag{84}$$

Thus, in particular, $B = QQ^*$. By the induction hypothesis, the lower-triangular matrix $Q \in \mathbb{C}^{n\times n}$ with real diagonal entries that satisfies $B = QQ^*$ is unique. Hence, our new Q is exactly the Q that was constructed above. Furthermore, (84) shows that $b = Qx$, so that $x = Q^{-1}b$, so again our new x is our old x. Finally, (84) yields $d = x^*x + \delta^2$, whence $\delta^2 = d - x^*x = d - ||x||^2$. Thus, $\delta = \sqrt{d - ||x||^2}$, because δ has to be positive. So our δ is our old δ. Thus, our L is the L that we constructed above. This proves the uniqueness of the L. Theorem 4.3.1 is proved. □

Theorem 4.3.1 can be used to prove several facts about positive definite matrices:

Exercise 4.3.1. [2] Let $A \in \mathbb{C}^{n\times n}$ be a positive definite Hermitian matrix. Prove that $\det A$ is a positive real.

Exercise 4.3.2. [4] Let $n > 0$. Let A and B be two positive definite Hermitian matrices in $\mathbb{C}^{n\times n}$. Prove that $\operatorname{Tr}(AB)$ is real and $\operatorname{Tr}(AB) > 0$.

Exercise 4.3.3. [4] Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with real entries. Assume that every nonzero vector $x \in \mathbb{R}^n$ (with **real** entries) satisfies $\langle Ax, x \rangle > 0$. Prove that A is positive definite.

Some properties of positive semidefinite matrices can be deduced from corresponding properties of positive definite matrices:

Exercise 4.3.4. [4] Let $A \in \mathbb{C}^{n \times n}$ be a positive semidefinite Hermitian matrix.

(a) Prove that $A + \varepsilon I_n$ is positive definite whenever ε is a positive real number.

(b) Prove that $\det A$ is a nonnegative real.

(c) Let $B \in \mathbb{C}^{n \times n}$ be a further positive semidefinite Hermitian matrix. Prove that $\operatorname{Tr}(AB)$ is real and $\operatorname{Tr}(AB) \geq 0$.

There is a version of Cholesky decomposition for positive semidefinite matrices, but we omit it for now.

4.4. Rayleigh quotients

4.4.1. Definition and basic properties

Definition 4.4.1. Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix, and $x \in \mathbb{C}^n$ be a nonzero vector. Then, the *Rayleigh quotient* for A and x is defined to be the real number

$$R(A, x) := \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \frac{x^* A x}{x^* x} = \frac{x^* A x}{||x||^2}.$$

Proposition 4.4.2. Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix, and $x \in \mathbb{C}^n$ be a nonzero vector. Let $y = \dfrac{x}{||x||}$. Then,

$$R(A, x) = R(A, y) = y^* A y.$$

Proof. Let $\lambda = ||x||$. Thus, $y = \dfrac{x}{\lambda}$, so that $x = \lambda y$. Hence,

$$R(A, x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \frac{\langle A \cdot \lambda y, \lambda y \rangle}{\langle \lambda y, \lambda y \rangle} = \frac{\overline{\lambda} \lambda \langle Ay, y \rangle}{\overline{\lambda} \lambda \langle y, y \rangle} = \frac{\langle Ay, y \rangle}{\langle y, y \rangle} = R(A, y)$$

(by the definition of $R(A, y)$). Moreover, the definition of y yields $||y|| = \left|\left| \dfrac{x}{||x||} \right|\right| = \dfrac{||x||}{||x||} = 1$. Now, the definition of $R(A, y)$ yields

$$R(A, y) = \frac{y^* A y}{||y||^2} = y^* A y \qquad (\text{since } ||y|| = 1).$$

The proof of Proposition 4.4.2 is thus complete. □

4.4.2. The Courant–Fisher theorem: statement

Let us explore what Rayleigh quotients can tell us about the eigenvalues of a Hermitian matrix.

Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix with $n > 0$. By the spectral theorem, we have $A = UDU^*$ for some unitary $U \in \mathbb{C}^{n \times n}$ and some real diagonal matrix $D \in \mathbb{R}^{n \times n}$. Consider these U and D. We have $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A. We WLOG assume that

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

(indeed, we can always achieve this by permuting rows/columns of D and integrating the permutation matrices into U [39]). We set

$$\lambda_{\min}(A) := \lambda_1 \qquad \text{and} \qquad \lambda_{\max}(A) := \lambda_n.$$

Let us now pick some vector $x \in \mathbb{C}^n$ of length 1 (that is, $||x|| = 1$). Set $z = U^*x$. Then, $z^* = (U^*x)^* = x^*(U^*)^* = x^*U$. Also, the matrix U^* is an isometry (since U is unitary), and thus we have $||U^*x|| = ||x|| = 1$. In other words, $||z|| = 1$ (since $z = U^*x$). In other words, $\sum_{k=1}^{n} |z_k|^2 = 1$ (since $||z|| = \sqrt{\sum_{k=1}^{n} |z_k|^2}$). Furthermore, writing z as $\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$, we have

$$x^* \underbrace{A}_{=UDU^*} x = \underbrace{x^*U}_{=z^*} D \underbrace{U^*x}_{=z} = z^*Dz = \sum_{k=1}^{n} \lambda_k \overline{z_k} z_k = \sum_{k=1}^{n} \underbrace{\lambda_k}_{\leq \lambda_n} |z_k|^2$$
$$\leq \sum_{k=1}^{n} \lambda_n |z_k|^2 = \lambda_n \underbrace{\sum_{k=1}^{n} |z_k|^2}_{=1} = \lambda_n.$$

Thus, we have shown that each vector $x \in \mathbb{C}^n$ of length 1 satisfies $x^*Ax \leq \lambda_n$. This inequality becomes an equality at least for one vector x of length 1: namely, for the vector $x = Ue_n$ (because for this vector, we have $z = \underbrace{U^*U}_{=I_n} e_n = e_n$, so that $z_k = 0$ for all $k < n$, and therefore the inequality $\sum_{k=1}^{n} \lambda_k |z_k|^2 \leq \sum_{k=1}^{n} \lambda_n |z_k|^2$ becomes

[39] Alternatively, we can achieve $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ right away by applying Theorem 2.3.3.

an equality[40]). Thus,

$$\begin{aligned}\lambda_n &= \max\left\{x^*Ax \;\mid\; x \in \mathbb{C}^n \text{ is a vector of length } 1\right\} \\ &= \max\left\{\frac{x^*Ax}{x^*x} \;\mid\; x \in \mathbb{C}^n \text{ is nonzero}\right\} \\ &= \max\left\{R\left(A,x\right) \;\mid\; x \in \mathbb{C}^n \text{ is nonzero}\right\}.\end{aligned}$$

Since $\lambda_n = \lambda_{\max}\left(A\right)$, we thus have proved the following fact:

Proposition 4.4.3. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix with $n > 0$. Then, the largest eigenvalue of A is

$$\begin{aligned}\lambda_{\max}\left(A\right) &= \max\left\{x^*Ax \;\mid\; x \in \mathbb{C}^n \text{ is a vector of length } 1\right\} \\ &= \max\left\{R\left(A,x\right) \;\mid\; x \in \mathbb{C}^n \text{ is nonzero}\right\}.\end{aligned}$$

Similarly, we can prove the following:

Proposition 4.4.4. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix with $n > 0$. Then, the smallest eigenvalue of A is

$$\begin{aligned}\lambda_{\min}\left(A\right) &= \min\left\{x^*Ax \;\mid\; x \in \mathbb{C}^n \text{ is a vector of length } 1\right\} \\ &= \min\left\{R\left(A,x\right) \;\mid\; x \in \mathbb{C}^n \text{ is nonzero}\right\}.\end{aligned}$$

What about the other eigenvalues? Can we characterize λ_2 (for example) in terms of Rayleigh quotients? Yes, but the characterization is more complicated:

Theorem 4.4.5 (Courant–Fisher theorem)**.** Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A, with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then, for each $k \in [n]$, we have

$$\lambda_k = \min_{\substack{S \subseteq \mathbb{C}^n \text{ is a subspace;} \\ \dim S = k}} \; \max_{\substack{x \in S; \\ x \neq 0}} R\left(A,x\right) \tag{85}$$

and

$$\lambda_k = \max_{\substack{S \subseteq \mathbb{C}^n \text{ is a subspace;} \\ \dim S = n-k+1}} \; \min_{\substack{x \in S; \\ x \neq 0}} R\left(A,x\right). \tag{86}$$

(The notation "$\min_{x\in\Omega} f\left(x\right)$" we are using here is a synonym for $\min\left\{f\left(x\right) \mid x \in \Omega\right\}$ whenever Ω is a set and f is a function defined on this set. The same applies to maxima. For instance, $\max_{\substack{x\in S; \\ x\neq 0}} R\left(A,x\right)$ means $\max\left\{R\left(A,x\right) \mid x \in S \text{ and } x \neq 0\right\}$.)

[40] and because the vector Ue_n does have length 1 (since U is an isometry, so that $||Ue_n|| = ||e_n|| = 1$)

To prove this theorem, we will use some elementary facts about subspaces of finite-dimensional vector spaces. We begin by recalling a classical definition:

4.4.3. The Courant–Fisher theorem: lemmas

Definition 4.4.6. Let S_1 and S_2 be two subspaces of a vector space V. Then,

$$S_1 + S_2 := \{s_1 + s_2 \mid s_1 \in S_1 \text{ and } s_2 \in S_2\}.$$

This is again a subspace of V. (This is the smallest subspace of V that contains both S_1 and S_2 as subspaces.)

Proposition 4.4.7. Let $\mathbb{F}$ be a field. Let V be a finite-dimensional $\mathbb{F}$-vector space. Let S_1 and S_2 be two subspaces of V. Then,

$$\dim(S_1 \cap S_2) + \dim(S_1 + S_2) = \dim S_1 + \dim S_2.$$

Proof. Pick any basis $(x_1, x_2, \ldots, x_k)$ of the vector space $S_1 \cap S_2$.

Then, $(x_1, x_2, \ldots, x_k)$ is a linearly independent list of vectors in S_1. Thus, we can extend it to a basis of S_1 by inserting some new vectors $y_1, y_2, \ldots, y_p$. Hence,

$$(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_p) \text{ is a basis of } S_1.$$

On the other hand, $(x_1, x_2, \ldots, x_k)$ is a linearly independent list of vectors in S_2. Thus, we can extend it to a basis of S_2 by inserting some new vectors $z_1, z_2, \ldots, z_q$. Hence,

$$(x_1, x_2, \ldots, x_k, z_1, z_2, \ldots, z_q) \text{ is a basis of } S_2.$$

The above three bases yield $\dim(S_1 \cap S_2) = k$ and $\dim S_1 = k + p$ and $\dim S_2 = k + q$.

Now, we claim that

$$\mathbf{w} := (x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_p, z_1, z_2, \ldots, z_q) \text{ is a basis of } S_1 + S_2.$$

Once this is proved, we will conclude that $\dim(S_1 + S_2) = k + p + q$, and then Proposition 4.4.7 will follow by a simple computation (namely, $k + (k + p + q) = (k + p) + (k + q)$).

So let us prove our claim. To prove that $\mathbf{w}$ is a basis of $S_1 + S_2$, we need to check the following two statements:

1. The list $\mathbf{w}$ is linearly independent.
2. The list $\mathbf{w}$ spans $S_1 + S_2$.

Proving statement 2 is easy: Any element of $S_1 + S_2$ is an element of S_1 plus an element of S_2, and thus can be written as

$$\begin{aligned}
&\big(\text{a linear combination of } x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_p\big) \\
&\qquad + \big(\text{a linear combination of } x_1, x_2, \ldots, x_k, z_1, z_2, \ldots, z_q\big) \\
&= \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p \\
&\qquad + \mu_1 x_1 + \mu_2 x_2 + \cdots + \mu_k x_k + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q \\
&\qquad\qquad \big(\text{for some scalars } \lambda_i, \alpha_j, \mu_i, \beta_v \in \mathbb{F}\big) \\
&= (\lambda_1 + \mu_1)\, x_1 + (\lambda_2 + \mu_2)\, x_2 + \cdots + (\lambda_k + \mu_k)\, x_k \\
&\qquad + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q \\
&= \big(\text{a linear combination of } x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_p, z_1, z_2, \ldots, z_q\big);
\end{aligned}$$

thus it belongs to the span of $\mathbf{w}$.

Let us now prove statement 1. We need to show that $\mathbf{w}$ is linearly independent. So let us assume that

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q = 0$$

for some coefficients $\lambda_m, \alpha_i, \beta_j$ that are not all equal to 0. We want a contradiction.

Let

$$v := \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p.$$

Then,

$$\begin{aligned}
v &= -\left(\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q\right) \qquad \text{(by the above equation)} \\
&\in S_2 \qquad \big(\text{since the } z_j\text{'s lie in } S_2\big).
\end{aligned}$$

On the other hand, the definition of v yields $v \in S_1$ (since the x_m's and the y_i's lie in S_1). Thus, v lies in both S_1 and S_2. This entails that $v \in S_1 \cap S_2$. Since $(x_1, x_2, \ldots, x_k)$ is a basis of $S_1 \cap S_2$, this entails that

$$v = \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_k x_k \qquad \text{for some } \xi_1, \xi_2, \ldots, \xi_k \in \mathbb{F}.$$

Comparing this with

$$v = -\left(\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q\right),$$

we obtain

$$\xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_k x_k = -\left(\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q\right).$$

In other words,

$$\xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_k x_k + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q = 0.$$

Since the list $(x_1, x_2, \ldots, x_k, z_1, z_2, \ldots, z_q)$ is linearly independent (being a basis of S_2), this entails that all coefficients ξ_m and β_j are 0. Thus, $v = 0$ (since $v = \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_k x_k$). In view of

$$v = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p,$$

this rewrites as

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p = 0.$$

Since the list $(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_p)$ is linearly independent (being a basis of S_1), this entails that all coefficients λ_m and α_i are 0.

Now we know that all λ_m and α_i and β_j are 0, which contradicts our assumption that some of them are nonzero. This completes the proof of Statement 1.

As we said, we now conclude that the list $\mathbf{w}$ is a basis of the vector space $S_1 + S_2$. Since this list $\mathbf{w}$ contains $k + p + q$ vectors, we thus obtain $\dim(S_1 + S_2) = k + p + q$, so that

$$\begin{aligned}\underbrace{\dim(S_1 \cap S_2)}_{=k} + \underbrace{\dim(S_1 + S_2)}_{=k+p+q} &= k + (k + p + q) \\ &= \underbrace{(k+p)}_{=\dim S_1} + \underbrace{(k+q)}_{=\dim S_2} \\ &= \dim S_1 + \dim S_2.\end{aligned}$$

This proves Proposition 4.4.7. □

Remark 4.4.8. A well-known fact in elementary set theory says that if A_1 and A_2 are two finite sets, then

$$|A_1 \cap A_2| + |A_1 \cup A_2| = |A_1| + |A_2|.$$

Proposition 4.4.7 is an analogue of this fact for vector spaces (noticing that the sum $S_1 + S_2$ is a vector-space analogue of the union).

Note, however, that the "next level" of the above formula has no vector space analogue. We do have

$$\begin{aligned}&|A_1 \cup A_2 \cup A_3| + |A_1 \cap A_2| + |A_1 \cap A_3| + |A_2 \cap A_3| \\ &= |A_1| + |A_2| + |A_3| + |A_1 \cap A_2 \cap A_3|\end{aligned}$$

for any three finite sets A_1, A_2, A_3, but no such relation holds for three subspaces of a vector space.

Corollary 4.4.9. Let $\mathbb{F}$ be a field, and let $n \in \mathbb{N}$. Let V be an n-dimensional $\mathbb{F}$-vector space. Let $S_1, S_2, \ldots, S_k$ be subspaces of V (with $k \geq 1$). Let

$$\delta := \dim(S_1) + \dim(S_2) + \cdots + \dim(S_k) - (k-1)\, n.$$

(a) Then, $\dim(S_1 \cap S_2 \cap \cdots \cap S_k) \geq \delta$.

(b) If $\mathbb{F} = \mathbb{C}$ and $V = \mathbb{C}^n$ and $\delta > 0$, then there exists a vector $x \in S_1 \cap S_2 \cap \cdots \cap S_k$ with $||x|| = 1$.

Proof. **(a)** We induct on k. The *base case* ($k = 1$) is obvious (since $\dim(S_1 \cap S_2 \cap \cdots \cap S_k) = \dim(S_1) = \delta$ in this case).

Induction step: Suppose the statement holds for some k. Now consider $k+1$ subspaces $S_1, S_2, \ldots, S_{k+1}$ of V, and let

$$\delta_{k+1} := \dim(S_1) + \dim(S_2) + \cdots + \dim(S_{k+1}) - kn.$$

We want to prove that $\dim(S_1 \cap S_2 \cap \cdots \cap S_{k+1}) \geq \delta_{k+1}$.

Then,

$$\begin{aligned}&\dim(S_1 \cap S_2 \cap \cdots \cap S_{k+1}) \\ &= \dim(S_1 \cap S_2 \cap \cdots \cap S_{k-1} \cap (S_k \cap S_{k+1})).\end{aligned}$$

Now, set

$$\delta_k := \dim(S_1) + \dim(S_2) + \cdots + \dim(S_{k-1}) + \dim(S_k \cap S_{k+1}) - (k-1)\,n.$$

By the induction hypothesis, we can apply Corollary 4.4.9 **(a)** to $S_k \cap S_{k+1}$ and δ_k instead of S_k and δ. Thus, we obtain

$$\dim(S_1 \cap S_2 \cap \cdots \cap S_{k-1} \cap (S_k \cap S_{k+1})) \geq \delta_k.$$

It remains to show that $\delta_k \geq \delta_{k+1}$. Equivalently, we need to show that

$$\dim(S_k \cap S_{k+1}) - (k-1)\,n \geq \dim(S_k) + \dim(S_{k+1}) - kn.$$

In other words, we need to show that

$$\dim(S_k \cap S_{k+1}) + n \geq \dim(S_k) + \dim(S_{k+1}).$$

However, $S_k + S_{k+1}$ is a subspace of V, so its dimension is $\dim(S_k + S_{k+1}) \leq \dim V = n$. Therefore,

$$\dim(S_k \cap S_{k+1}) + \underbrace{n}_{\geq \dim(S_k + S_{k+1})} \geq \dim(S_k \cap S_{k+1}) + \dim(S_k + S_{k+1}) = \dim(S_k) + \dim(S_{k+1})$$

(by Proposition 4.4.7). So the induction step is complete, and Corollary 4.4.9 **(a)** is proved.

(b) Assume that $\mathbb{F} = \mathbb{C}$ and $V = \mathbb{C}^n$ and $\delta > 0$. Then, part **(a)** yields

$$\dim(S_1 \cap S_2 \cap \cdots \cap S_k) \geq \delta > 0.$$

Thus, the subspace $S_1 \cap S_2 \cap \cdots \cap S_k$ is not just $\{0\}$. Therefore, it contains a nonzero vector. Scaling this vector by the reciprocal of its length, we obtain a vector of length 1. This proves Corollary 4.4.9 **(b)**. □

Lemma 4.4.10. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Let $(v_1, v_2, \ldots, v_k)$ be an orthonormal tuple of eigenvectors of A, and let $\mu_1, \mu_2, \ldots, \mu_k$ be the corresponding eigenvalues (so that $A_i v_i = \mu_i v_i$ for each $i \in [k]$). Assume that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$. Then, each $v \in \operatorname{span}(v_1, v_2, \ldots, v_k)$ satisfies

$$\langle Av, v\rangle \geq \mu_1 \langle v, v\rangle \tag{87}$$

and

$$\langle Av, v\rangle \leq \mu_k \langle v, v\rangle . \tag{88}$$

Proof. Let $v \in \operatorname{span}(v_1, v_2, \ldots, v_k)$. Thus, we can write v in the form

$$v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k \tag{89}$$

for some $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{C}$. Consider these $\alpha_1, \alpha_2, \ldots, \alpha_k$. From (89), we obtain

$$v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = \sum_{i=1}^{k} \alpha_i v_i$$

and thus

$$Av = A \sum_{i=1}^{k} \alpha_i v_i = \sum_{i=1}^{k} \alpha_i \underbrace{Av_i}_{=\mu_i v_i} = \sum_{i=1}^{k} \alpha_i \mu_i v_i.$$

Combining this with $v = \sum_{i=1}^{k} \alpha_i v_i = \sum_{j=1}^{k} \alpha_j v_j$, we obtain

$$\langle Av, v\rangle = \left\langle \sum_{i=1}^{k} \alpha_i \mu_i v_i, \ \sum_{j=1}^{k} \alpha_j v_j \right\rangle = \sum_{i=1}^{k} \ \sum_{j=1}^{k} \alpha_i \mu_i \overline{\alpha_j} \langle v_i, v_j\rangle . \tag{90}$$

However, the inner products $\langle v_i, v_j\rangle$ in this sum are 0 whenever $i \neq j$ (since the tuple $(v_1, v_2, \ldots, v_k)$ is orthonormal). Thus, we can simplify this sum as follows:

$$\sum_{i=1}^{k} \ \sum_{j=1}^{k} \alpha_i \mu_i \overline{\alpha_j} \langle v_i, v_j\rangle = \sum_{i=1}^{k} \alpha_i \mu_i \overline{\alpha_i} \underbrace{\langle v_i, v_i\rangle}_{\substack{=\|v_i\|^2=1 \\ \text{(since the tuple } (v_1,v_2,\ldots,v_k) \\ \text{is orthonormal)}}} = \sum_{i=1}^{k} \underbrace{\alpha_i \mu_i \overline{\alpha_i}}_{=\alpha_i\overline{\alpha_i}\cdot\mu_i}$$

$$= \sum_{i=1}^{k} \underbrace{\alpha_i \overline{\alpha_i}}_{=|\alpha_i|^2} \cdot \mu_i = \sum_{i=1}^{k} |\alpha_i|^2 \mu_i.$$

Thus, (90) rewrites as

$$\langle Av, v\rangle = \sum_{i=1}^{k} |\alpha_i|^2 \mu_i. \tag{91}$$

On the other hand, from $v = \sum_{i=1}^{k} \alpha_i v_i$ and $v = \sum_{j=1}^{k} \alpha_j v_j$, we obtain

$$\begin{aligned}
\langle v, v \rangle &= \left\langle \sum_{i=1}^{k} \alpha_i v_i, \ \sum_{j=1}^{k} \alpha_j v_j \right\rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \overline{\alpha_j} \left\langle v_i, v_j \right\rangle \\
&= \sum_{i=1}^{k} \underbrace{\alpha_i \overline{\alpha_i}}_{=|\alpha_i|^2} \underbrace{\left\langle v_i, v_i \right\rangle}_{\substack{=||v_i||^2=1 \\ \text{(since the tuple } (v_1,v_2,\ldots,v_k) \\ \text{is orthonormal)}}} \\
&\qquad \left(\text{since the inner products } \left\langle v_i, v_j \right\rangle \text{ are 0 whenever } i \neq j\right) \\
&= \sum_{i=1}^{k} |\alpha_i|^2 . \qquad (92)
\end{aligned}$$

Now, (91) becomes

$$\begin{aligned}
\langle Av, v \rangle &= \sum_{i=1}^{k} |\alpha_i|^2 \underbrace{\mu_i}_{\substack{\geq \mu_1 \\ \text{(since } \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k\text{)}}} \geq \sum_{i=1}^{k} |\alpha_i|^2 \mu_1 = \mu_1 \underbrace{\sum_{i=1}^{k} |\alpha_i|^2}_{\substack{=\langle v,v \rangle \\ \text{(by (92))}}} \\
&= \mu_1 \langle v, v \rangle .
\end{aligned}$$

This proves (87). Likewise, (91) becomes

$$\begin{aligned}
\langle Av, v \rangle &= \sum_{i=1}^{k} |\alpha_i|^2 \underbrace{\mu_i}_{\substack{\leq \mu_k \\ \text{(since } \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k\text{)}}} \leq \sum_{i=1}^{k} |\alpha_i|^2 \mu_k = \mu_k \underbrace{\sum_{i=1}^{k} |\alpha_i|^2}_{\substack{=\langle v,v \rangle \\ \text{(by (92))}}} \\
&= \mu_k \langle v, v \rangle .
\end{aligned}$$

This proves (88). □

4.4.4. The Courant–Fisher theorem: proof

Now, we are ready for the proof of the Courant–Fisher theorem:

Proof of Theorem 4.4.5. The spectral theorem (Theorem 2.6.1 **(a)**) says that $A = UDU^*$ for some unitary U and some diagonal matrix D. Consider these U and D. Proposition 2.6.5 shows that the diagonal entries of D are real.

The columns of U form an orthonormal basis of $\mathbb{C}^n$ (since U is unitary); let $(u_1, u_2, \ldots, u_n)$ be this basis. Then, $u_1, u_2, \ldots, u_n$ are eigenvectors of A (by Theorem 2.6.1 **(b)**). We WLOG assume that the corresponding eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$

(otherwise, permute the diagonal entries of D and correspondingly permute the columns of U). Thus, $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ (by the proof of Theorem 2.6.1 **(b)**).

Let $k \in [n]$.

Let S be a vector subspace of $\mathbb{C}^n$ with $\dim S = k$. Let $S' = \operatorname{span}(u_k, u_{k+1}, \ldots, u_n)$. Then, by Proposition 4.4.7, we have

$$\dim\left(S \cap S'\right) + \dim\left(S + S'\right) = \underbrace{\dim S}_{=k} + \underbrace{\dim S'}_{=n-k+1} = n + 1 > n \geq \dim\left(S + S'\right)$$

(since $S + S'$ is a subspace of $\mathbb{C}^n$). Subtracting $\dim(S + S')$ from this inequality, we obtain $\dim(S \cap S') > 0$. Thus, $S \cap S'$ contains a nonzero vector. Thus, $\sup\limits_{\substack{x \in S \cap S'; \\ x \neq 0}} R(A, x)$ and $\inf\limits_{\substack{x \in S \cap S'; \\ x \neq 0}} R(A, x)$ are well-defined.

Now,

$$\begin{aligned} \sup_{\substack{x \in S; \\ x \neq 0}} R(A, x) &\geq \sup_{\substack{x \in S \cap S'; \\ x \neq 0}} R(A, x) \geq \inf_{\substack{x \in S \cap S'; \\ x \neq 0}} R(A, x) \\ &\geq \inf_{\substack{x \in S'; \\ x \neq 0}} R(A, x). \end{aligned} \tag{93}$$

However, I claim that $\inf\limits_{\substack{x \in S'; \\ x \neq 0}} R(A, x) \geq \lambda_k$ (actually, this is an equality, but we will not need this). Indeed, $(u_k, u_{k+1}, \ldots, u_n)$ is an orthonormal tuple of eigenvalues of A (since $(u_1, u_2, \ldots, u_n)$ is an orthonormal tuple of eigenvalues of A) with corresponding eigenvalues $\lambda_k, \lambda_{k+1}, \ldots, \lambda_n$ satisfying $\lambda_k \leq \lambda_{k+1} \leq \cdots \leq \lambda_n$ (since $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$). Thus, any $x \in \operatorname{span}(u_k, u_{k+1}, \ldots, u_n)$ satisfies

$$\langle Ax, x \rangle \geq \lambda_k \langle x, x \rangle$$

(by (87), applied to $(u_k, u_{k+1}, \ldots, u_n)$ and $(\lambda_k, \lambda_{k+1}, \ldots, \lambda_n)$ and x instead of $(v_1, v_2, \ldots, v_k)$ and $(\mu_1, \mu_2, \ldots, \mu_k)$ and v) and thus $R(A, x) = \dfrac{\langle Ax, x \rangle}{\langle x, x \rangle} \geq \lambda_k$. In other words, any $x \in S'$ satisfies $R(A, x) \geq \lambda_k$ (since $S' = \operatorname{span}(u_k, u_{k+1}, \ldots, u_n)$). Hence, we have

$$\inf_{\substack{x \in S'; \\ x \neq 0}} R(A, x) \geq \lambda_k.$$

Combining this with (93), we obtain

$$\sup_{\substack{x \in S; \\ x \neq 0}} R(A, x) \geq \lambda_k.$$

Furthermore, this supremum is a maximum, because

$$\sup_{\substack{x\in S;\\ x\neq 0}} R(A,x) = \sup_{\substack{y\in S;\\ ||y||=1}} R(A,y) \qquad \left(\text{since } R(A,x) = R(A,y) \text{ where } y = \frac{x}{||x||}\right)$$

$$= \max_{\substack{y\in S;\\ ||y||=1}} R(A,y) \qquad \left(\begin{array}{c}\text{since the set of all } y \in S \text{ satisfying } ||y|| = 1\\ \text{is compact, and since a continuous function}\\ \text{on a compact set always has a maximum}\end{array}\right)$$

$$= \max_{\substack{x\in S;\\ x\neq 0}} R(A,x).$$

So we conclude that

$$\max_{\substack{x\in S;\\ x\neq 0}} R(A,x) = \sup_{\substack{x\in S;\\ x\neq 0}} R(A,x) \geq \lambda_k.$$

Forget that we fixed S. We thus have shown that if S is any k-dimensional subspace of $\mathbb{C}^n$, then $\max\limits_{\substack{x\in S;\\ x\neq 0}} R(A,x)$ exists and satisfies

$$\max_{\substack{x\in S;\\ x\neq 0}} R(A,x) \geq \lambda_k.$$

However, by choosing S appropriately, we can achieve equality here; indeed, we have to choose $S = \operatorname{span}(u_1, u_2, \ldots, u_k)$ for this. (Why? Because each $x \in \operatorname{span}(u_1, u_2, \ldots, u_k)$ can easily be seen to satisfy $\langle Ax, x\rangle \leq \lambda_k \langle x, x\rangle$ by a similar argument to the one we used above[41].)

Thus, we have shown that the value $\max\limits_{\substack{x\in S;\\ x\neq 0}} R(A,x)$ is $\geq \lambda_k$ for each S, but is $= \lambda_k$ for a certain S. Therefore, λ_k is the smallest possible value of $\max\limits_{\substack{x\in S;\\ x\neq 0}} R(A,x)$. In other words,

$$\lambda_k = \min_{\substack{S\subseteq\mathbb{C}^n \text{ is a subspace;}\\ \dim S = k}} \max_{\substack{x\in S;\\ x\neq 0}} R(A,x).$$

Thus, we have proved (85). It remains to prove the other part of the theorem – i.e., the equality (86).

One way to prove this is by arguing similarly to the above proof. Alternatively, we can simply apply the already proved equality (85) to $-A$ instead of A, after noticing that $-A$ is a Hermitian matrix with eigenvalues

$$-\lambda_n \leq -\lambda_{n-1} \leq \cdots \leq -\lambda_1.$$

[41] but using (88) instead of (87)

Keep in mind that $-\lambda_k$ is not the k-th smallest eigenvalue of $-A$, but rather is the k-th largest eigenvalue of $-A$, and thus the $(n-k+1)$-st smallest eigenvalue of $-A$. Thus, we have to apply the equality (85) to $-A$ and $n-k+1$ instead of A and k. Taking negatives turns minima into maxima and vice versa (i.e., we have $\min\limits_{x\in\Omega}\left(-f\left(x\right)\right) = -\max\limits_{x\in\Omega} f\left(x\right)$ and $\max\limits_{x\in\Omega}\left(-f\left(x\right)\right) = -\min\limits_{x\in\Omega} f\left(x\right)$). Finally, it is helpful to know that $R\left(-A,x\right) = -R\left(A,x\right)$ for any vector $x \in \mathbb{C}^n$. Armed with these observations, we can easily derive (86) from (85). The proof of Theorem 4.4.5 is now complete. □

4.4.5. The Weyl inequalities

The Courant–Fisher theorem can be used to connect the eigenvalues of $A+B$ with the eigenvalues of A and B.

Theorem 4.4.11 (Weyl's inequalities)**.** Let A and B be two Hermitian matrices in $\mathbb{C}^{n\times n}$. Let $i \in [n]$ and $j \in \{0,1,\ldots,n-i\}$.

(a) Then,

$$\lambda_i\left(A+B\right) \leq \lambda_{i+j}\left(A\right) + \lambda_{n-j}\left(B\right).$$

Here, $\lambda_k\left(C\right)$ means the k-th smallest eigenvalue of a Hermitian matrix C.

Moreover, this inequality becomes an equality if and only if there exists a nonzero vector $x \in \mathbb{C}^n$ satisfying

$$Ax = \lambda_{i+j}\left(A\right)x, \qquad Bx = \lambda_{n-j}\left(B\right)x, \qquad \left(A+B\right)x = \lambda_i\left(A+B\right)x$$

(at the same time).

(b) Furthermore,

$$\lambda_{i-k+1}\left(A\right) + \lambda_k\left(B\right) \leq \lambda_i\left(A+B\right) \qquad \text{for any } k \in [i].$$

Theorem 4.4.11 lets us bound the eigenvalues of $A+B$ in terms of those of A and B under the assumption that A and B are Hermitian matrices. (In contrast, if A and B are arbitrary – not Hermitian – matrices, then no such bounds are possible for $n > 1$.)

Example 4.4.12. Applying Theorem 4.4.11 **(a)** to $i = n$ and $j = 0$, we obtain

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

Applying Theorem 4.4.11 **(b)** to $i = 1$ and $k = 1$, we obtain

$$\lambda_1\left(A\right) + \lambda_1\left(B\right) \leq \lambda_1\left(A+B\right).$$

Proof of Theorem 4.4.11. Let $(x_1, x_2, \ldots, x_n)$, $(y_1, y_2, \ldots, y_n)$ and $(z_1, z_2, \ldots, z_n)$ be three orthonormal bases of $\mathbb{C}^n$ with

$$Ax_k = \lambda_k(A)\, x_k, \qquad By_k = \lambda_k(B)\, y_k, \qquad (A+B)\, z_k = \lambda_k(A+B)\, z_k$$

for all $k \in [n]$. (As above, we can find such bases by using the spectral decompositions of A, B and $A+B$.)

Let

$$\begin{aligned} S_1 &= \operatorname{span}\left(x_1, x_2, \ldots, x_{i+j}\right); \\ S_2 &= \operatorname{span}\left(y_1, y_2, \ldots, y_{n-j}\right); \\ S_3 &= \operatorname{span}\left(z_i, z_{i+1}, \ldots, z_n\right). \end{aligned}$$

Then,

$$\delta := \underbrace{\dim(S_1)}_{=i+j} + \underbrace{\dim(S_2)}_{=n-j} + \underbrace{\dim(S_3)}_{=n-i+1} - 2n = 1 > 0.$$

Hence, Corollary 4.4.9 **(b)** yields that there is a length-1 vector v in $S_1 \cap S_2 \cap S_3$. This v satisfies

$$\begin{aligned} \lambda_i(A+B) &\leq \langle (A+B)\, v,\ v\rangle \qquad \left(\text{by (87), since } v \in S_3 = \operatorname{span}(z_i, z_{i+1}, \ldots, z_n)\right) \\ &= \langle Av + Bv,\ v\rangle = \underbrace{\langle Av,\ v\rangle}_{\substack{\leq \lambda_{i+j}(A) \\ \text{(by (88),} \\ \text{since } v \in S_1 = \operatorname{span}(x_1, x_2, \ldots, x_{i+j}))}} + \underbrace{\langle Bv,\ v\rangle}_{\substack{\leq \lambda_{n-j}(B) \\ \text{(by (88),} \\ \text{since } v \in S_2 = \operatorname{span}(y_1, y_2, \ldots, y_{n-j}))}} \\ &\leq \lambda_{i+j}(A) + \lambda_{n-j}(B). \end{aligned}$$

This proves the inequality part of Theorem 4.4.11 **(a)**. The equality case is not hard to analyze following the above argument; we leave this to the reader.

The proof of Theorem 4.4.11 **(b)** is left to the reader as well. □

4.5. ([Missing lecture]) The interlacing theorem

4.6. Consequences of the interlacing theorem

Recall: If $A \in \mathbb{C}^{n \times n}$ is a Hermitian matrix (i.e., a square matrix satisfying $A^* = A$), then we denote its eigenvalues by $\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)$ in weakly increasing order (with multiplicities). This makes sense, since we know that these eigenvalues are reals.

Last time, Hugo proved:

Theorem 4.6.1 (Cauchy's interlacing theorem, aka eigenvalue interlacing theorem)**.** Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Let $j \in [n]$. Let $B \in \mathbb{C}^{(n-1)\times(n-1)}$ be the matrix obtained from A by removing the j-th row and the j-th column. Then,

$$\lambda_1(A) \leq \lambda_1(B) \leq \lambda_2(A) \leq \lambda_2(B) \leq \cdots \leq \lambda_{n-1}(A) \leq \lambda_{n-1}(B) \leq \lambda_n(A).$$

In other words,

$$\lambda_i(A) \leq \lambda_i(B) \leq \lambda_{i+1}(A) \qquad \text{for each } i \in [n-1].$$

A converse of this theorem also holds:

Proposition 4.6.2. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mu_1, \mu_2, \ldots, \mu_{n-1}$ be real numbers satisfying

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \lambda_{n-1} \leq \mu_{n-1} \leq \lambda_n.$$

Then, there exist $n-1$ reals $y_1, y_2, \ldots, y_{n-1} \in \mathbb{R}$ and a real $a \in \mathbb{R}$ such that the matrix

$$A := \begin{pmatrix} \mu_1 & & & & y_1 \\ & \mu_2 & & & y_2 \\ & & \ddots & & \vdots \\ & & & \mu_{n-1} & y_{n-1} \\ y_1 & y_2 & \cdots & y_{n-1} & a \end{pmatrix}$$

(where all empty cells are supposed to be filled with 0s) has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. (This matrix is, of course, Hermitian, since it is real symmetric.)

Proof. Omitted. (Exercise?) □

Now, let us derive some consequences from Cauchy's interlacing theorem. We begin with a straightforward generalization:

Corollary 4.6.3 (Cauchy's interlacing theorem for multiple deletions)**.** Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Let $r \in \{0, 1, \ldots, n\}$. Let $C \in \mathbb{C}^{r \times r}$ be the result of removing $n-r$ rows and the corresponding $n-r$ columns from A. (That is, we pick some $j_1 < j_2 < \cdots < j_{n-r}$, and we remove the j_1-st, j_2-nd, …, j_{n-r}-th rows from A, and we remove the j_1-st, j_2-nd, …, j_{n-r}-th columns from A.) Then, for each $j \in [r]$, we have

$$\lambda_j(A) \leq \lambda_j(C) \leq \lambda_{j+n-r}(A).$$

Proof. Induction on $n-r$.

The *base case* ($n-r=0$) is trivial, since $C = A$ in this case.

In the *induction step*, we obtain C from B by removing a single row and the corresponding column. Thus, by the original Cauchy interlacing theorem, we get $\lambda_j(B) \leq \lambda_j(C)$. However, by the induction hypothesis, we get $\lambda_j(A) \leq \lambda_j(B)$. Combining these inequalities, we get $\lambda_j(A) \leq \lambda_j(C)$. The remaining inequality $\lambda_j(C) \leq \lambda_{j+n-r}(A)$ is proved similarly: By the original Cauchy interlacing theorem, we get $\lambda_j(C) \leq \lambda_{j+1}(B)$. However, by the induction hypothesis, we get $\lambda_{j+1}(B) \leq \lambda_{j+1+(n-r-1)}(A) = \lambda_{j+n-r}(A)$. □

The next corollary provides a minimum/maximum description of the sum of the first m smallest/largest eigenvalues of a Hermitian matrix:

Corollary 4.6.4. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Let $m \in \{0,1,\ldots,n\}$. Then:

(a) We have

$$\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) = \min_{\text{isometries } V\in\mathbb{C}^{n\times m}} \operatorname{Tr}(V^*AV).$$

(b) We have

$$\lambda_{n-m+1}(A) + \lambda_{n-m+2}(A) + \cdots + \lambda_n(A) = \max_{\text{isometries } V\in\mathbb{C}^{n\times m}} \operatorname{Tr}(V^*AV).$$

Proof. First of all, it suffices to show the first equality, because the second follows by applying the first to $-A$ instead of A.

First, we shall show that

$$\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) \leq \operatorname{Tr}(V^*AV) \qquad \text{for every isometry } V \in \mathbb{C}^{n\times m}.$$

Indeed, let $V \in \mathbb{C}^{n\times m}$ be an isometry. Thus, V is an $n \times m$-matrix whose columns are orthonormal. As we know from Corollary 1.2.9, we can extend each orthonormal tuple of vectors to an orthonormal basis. Doing this to the columns of V, we thus obtain an orthonormal basis of $\mathbb{C}^n$ whose first m entries are the columns of V. Let U be the matrix whose columns are the entries of this basis. Then,

$$U = \begin{pmatrix} V & \widetilde{V} \end{pmatrix} \qquad \text{(in block-matrix notation)}$$

by construction of this basis, and furthermore the matrix U is unitary since its columns form an orthonormal basis.

Since U is unitary, we have $U^*AU \sim A$ and therefore

$$\lambda_j(U^*AU) = \lambda_j(A) \qquad \text{for all } j \in [n].$$

However, $U = \begin{pmatrix} V & \widetilde{V} \end{pmatrix}$ entails

$$U^*AU = \begin{pmatrix} V & \widetilde{V} \end{pmatrix}^* A \begin{pmatrix} V & \widetilde{V} \end{pmatrix} = \begin{pmatrix} V^* \\ \widetilde{V}^* \end{pmatrix} A \begin{pmatrix} V & \widetilde{V} \end{pmatrix} = \begin{pmatrix} V^*AV & * \\ * & * \end{pmatrix},$$

where the three $*$s mean blocks that we don't care about. So the matrix V^*AV is obtained from U^*AU by removing a bunch of rows and the corresponding columns. Hence, Corollary 4.6.3 yields

$$\lambda_j(U^*AU) \leq \lambda_j(V^*AV) \qquad \text{for all } j \in [m]$$

(since U^*AU is Hermitian (because A is Hermitian)). In other words,

$$\lambda_j(A) \leq \lambda_j(V^*AV) \qquad \text{for all } j \in [m]$$

(since $\lambda_j(U^*AU) = \lambda_j(A)$). Adding these inequalities together, we obtain

$$\begin{aligned}
&\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) \\
&\leq \lambda_1(V^*AV) + \lambda_2(V^*AV) + \cdots + \lambda_m(V^*AV) \\
&= (\text{the sum of all eigenvalues of } V^*AV) \\
&\qquad\qquad \left(\begin{array}{c} \text{since } V^*AV \text{ is an } m \times m\text{-matrix} \\ \text{and thus has } m \text{ eigenvalues} \end{array} \right) \\
&= \operatorname{Tr}(V^*AV)
\end{aligned}$$

(since the sum of all eigenvalues of a matrix is the trace of this matrix).

Now, we need to show that there exists a unitary matrix $V \in \mathbb{C}^{n\times m}$ such that

$$\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) = \operatorname{Tr}(V^*AV).$$

To do this, we construct V as follows: We pick an eigenvector x_i of A at eigenvalue $\lambda_i(A)$ for each $i \in [n]$ in such a way that $(x_1, x_2, \ldots, x_n)$ is an orthonormal basis of $\mathbb{C}^n$. (This is possible because of Theorem 2.6.1 **(b)**.) Now, let $V \in \mathbb{C}^{n\times m}$ be the matrix whose columns are $x_1, x_2, \ldots, x_m$. This matrix V is an isometry, since $x_1, x_2, \ldots, x_m$ are orthonormal. Moreover,

$$\begin{aligned}
V^*AV &= \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_m^* \end{pmatrix} A \begin{pmatrix} x_1 & x_2 & \cdots & x_m \end{pmatrix} \\
&= \begin{pmatrix} x_1^*Ax_1 & x_1^*Ax_2 & \cdots & x_1^*Ax_m \\ x_2^*Ax_1 & x_2^*Ax_2 & \cdots & x_2^*Ax_m \\ \vdots & \vdots & \ddots & \vdots \\ x_m^*Ax_1 & x_m^*Ax_2 & \cdots & x_m^*Ax_m \end{pmatrix},
\end{aligned}$$

so that

$$\begin{aligned}
\operatorname{Tr}(V^*AV) &= \sum_{j=1}^{m} x_j^* \underbrace{Ax_j}_{\substack{=\lambda_j(A)x_j \\ \text{(since } x_j \text{ is an eigenvector of } A \\ \text{at eigenvalue } \lambda_j(A)\text{)}}} = \sum_{j=1}^{m} \lambda_j(A) \underbrace{x_j^*x_j}_{\substack{=\left|\left|x_j\right|\right|^2=1 \\ \text{(since } (x_1,x_2,\ldots,x_n) \\ \text{is an orthonormal basis)}}} \\
&= \sum_{j=1}^{m} \lambda_j(A) = \lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A).
\end{aligned}$$

This is precisely what we needed. Thus, we conclude that

$$\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) = \min_{\text{isometries } V \in \mathbb{C}^{n\times m}} \operatorname{Tr}(V^*AV).$$

As we said above, this completes the proof. □

Corollary 4.6.5. Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian $n \times n$-matrix. Let $m \in \{0,1,\ldots,n\}$. Let $i_1, i_2, \ldots, i_m \in [n]$ be distinct. Then,

$$\begin{aligned}\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) &\leq A_{i_1,i_1} + A_{i_2,i_2} + \cdots + A_{i_m,i_m} \\ &\leq \lambda_{n-m+1}(A) + \lambda_{n-m+2}(A) + \cdots + \lambda_n(A).\end{aligned}$$

In words: For a Hermitian matrix A, each sum of m distinct diagonal entries of A is sandwiched between the sum of the m smallest eigenvalues of A and the sum of the m largest eigenvalues of A.

Proof. Let C be the matrix obtained from A by removing all but the i_1-st, i_2-nd, …, i_m-th rows and the corresponding columns of A. Then,

$$\operatorname{Tr} C = A_{i_1,i_1} + A_{i_2,i_2} + \cdots + A_{i_m,i_m}.$$

However, Cauchy's interlacing theorem for multiple deletions yields

$$\lambda_j(A) \leq \lambda_j(C) \qquad \text{for each } j \in [m].$$

Summing these up, we obtain

$$\begin{aligned}\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) &\leq \lambda_1(C) + \lambda_2(C) + \cdots + \lambda_m(C) \\ &= (\text{the sum of all eigenvalues of } C) \\ &= \operatorname{Tr} C = A_{i_1,i_1} + A_{i_2,i_2} + \cdots + A_{i_m,i_m}.\end{aligned}$$

So we have proved the first of the required two inequalities. The second follows by applying the first to $-A$ instead of A. □

The above corollary has a bunch of consequences that are obtained by restating it in terms of something called *majorization*. Let us define this concept and see what it entails.

4.7. Introduction to majorization theory ([HorJoh13, §4.3])

We will now give a brief introduction to majorization theory. Much more can be found in [MaOlAr11] (see also [Nathan21] for an elementary introduction).

4.7.1. Notations and definition

Convention 4.7.1. Let $x \in \mathbb{R}^n$ be a column vector with real entries. Then:

(a) For each $i \in [n]$, we let x_i denote the i-th coordinate of x. ("Coordinate" is just a synonym for "entry" in the context of a vector.) Thus,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)^T.$$

(b) We say that x is *weakly decreasing* if $x_1 \geq x_2 \geq \cdots \geq x_n$. We say that x is *weakly increasing* if $x_1 \leq x_2 \leq \cdots \leq x_n$.

(c) We let $x^{\downarrow}$ denote the weakly decreasing permutation of x (that is, the column vector obtained by sorting the entries of x in weakly decreasing order). In other words, $x^{\downarrow}$ is the unique weakly decreasing column vector that can be obtained by permuting the entries of x.

Thus, $x_i^{\downarrow}$ is the i-th largest entry of x for each $i \in [n]$; in particular, $x_1^{\downarrow} \geq x_2^{\downarrow} \geq \cdots \geq x_n^{\downarrow}$.

(d) We let $x^{\uparrow}$ denote the weakly increasing permutation of x (that is, the column vector obtained by sorting the entries of x in weakly increasing order). In other words, $x^{\uparrow}$ is the unique weakly increasing column vector that can be obtained by permuting the entries of x.

Thus, $x_i^{\uparrow}$ is the i-th smallest entry of x for each $i \in [n]$; in particular, $x_1^{\uparrow} \leq x_2^{\uparrow} \leq \cdots \leq x_n^{\uparrow}$.

For example, if $x = (3,5,2)^T$, then $x_1 = 3$ and $x_2 = 5$ and $x_3 = 2$ and

$$x^{\downarrow} = (5,3,2)^T \qquad \text{and } x_1^{\downarrow} = 5 \text{ and } x_2^{\downarrow} = 3 \text{ and } x_3^{\downarrow} = 2 \qquad \text{and}$$
$$x^{\uparrow} = (2,3,5)^T \qquad \text{and } x_1^{\uparrow} = 2 \text{ and } x_2^{\uparrow} = 3 \text{ and } x_3^{\uparrow} = 5.$$

Definition 4.7.2. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ be two column vectors with real entries. Then, we say that x *majorizes* y (and we write $x \succcurlyeq y$) if and only if we have

$$\sum_{i=1}^{m} x_i^{\downarrow} \geq \sum_{i=1}^{m} y_i^{\downarrow} \qquad \text{for each } m \in [n]$$

and

$$\sum_{i=1}^{n} x_i^{\downarrow} = \sum_{i=1}^{n} y_i^{\downarrow}.$$

In other words, x majorizes y if and only if we have

$$\begin{gathered}
x_1^{\downarrow} \geq y_1^{\downarrow}; \\
x_1^{\downarrow} + x_2^{\downarrow} \geq y_1^{\downarrow} + y_2^{\downarrow}; \\
x_1^{\downarrow} + x_2^{\downarrow} + x_3^{\downarrow} \geq y_1^{\downarrow} + y_2^{\downarrow} + y_3^{\downarrow}; \\
\cdots; \\
x_1^{\downarrow} + x_2^{\downarrow} + \cdots + x_{n-1}^{\downarrow} \geq y_1^{\downarrow} + y_2^{\downarrow} + \cdots + y_{n-1}^{\downarrow}; \\
x_1^{\downarrow} + x_2^{\downarrow} + \cdots + x_n^{\downarrow} = y_1^{\downarrow} + y_2^{\downarrow} + \cdots + y_n^{\downarrow}
\end{gathered}$$

(note that the last relation in this chain is an equality, not just an inequality).

Example 4.7.3. We have

$$\begin{pmatrix} 1 \\ 3 \\ 5 \\ 7 \end{pmatrix} \succcurlyeq \begin{pmatrix} 2 \\ 2 \\ 6 \\ 6 \end{pmatrix},$$

since

$$\begin{aligned} 7 &\geq 6; \\ 7+5 &\geq 6+6; \\ 7+5+3 &\geq 6+6+2; \\ 7+5+3+1 &= 6+6+2+2. \end{aligned}$$

Example 4.7.4. We don't have

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

since we don't have $7 \geq 8$.

The intuition behind majorization is the following: A column vector x majorizes a column vector y if and only if you can obtain y from x by "moving the entries closer together (while keeping the average equal)". We will soon see a rigorous way to state this.

Here are some more general examples of majorization:

Exercise 4.7.1. [2] Let $x \in \mathbb{R}^n$, and let $m = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$. Show that $x \succcurlyeq (m, m, \ldots, m)^T$.

Exercise 4.7.2. [2] Let $a, b, c \in \mathbb{R}$, and let $x = \dfrac{b+c}{2}$ and $y = \dfrac{c+a}{2}$ and $z = \dfrac{a+b}{2}$. Show that $(a, b, c)^T \succcurlyeq (x, y, z)^T$.

Exercise 4.7.3. [4] Let $a, b, c \in \mathbb{R}$, and let $x = \dfrac{b+c}{2}$ and $y = \dfrac{c+a}{2}$ and $z = \dfrac{a+b}{2}$ and $m = \dfrac{a+b+c}{3}$. Show that

$$(a, b, c, m, m, m)^T \succcurlyeq (x, x, y, y, z, z)^T.$$

Exercise 4.7.4. [3] Let $x, y \in \mathbb{R}^n$. Prove that $x \succcurlyeq y$ if and only if $-x \succcurlyeq -y$.

We note that the condition $\sum_{i=1}^{n} x_i^{\downarrow} = \sum_{i=1}^{n} y_i^{\downarrow}$ in the definition of majorization can be rewritten as $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, because the sum of all coordinates of a vector does not change when we permute the entries of the vector. However, the conditions $\sum_{i=1}^{m} x_i^{\downarrow} \geq \sum_{i=1}^{m} y_i^{\downarrow}$ cannot be rewritten as $\sum_{i=1}^{m} x_i \geq \sum_{i=1}^{m} y_i$.

Proposition 4.7.5. Majorization is a partial order: That is, the binary relation $\succcurlyeq$ on the set $\mathbb{R}^n$ is reflexive, antisymmetric and transitive.

Proof. This is straightforward to verify. For example, if $\sum_{i=1}^{m} x_i^{\downarrow} \geq \sum_{i=1}^{m} y_i^{\downarrow}$ and $\sum_{i=1}^{m} y_i^{\downarrow} \geq \sum_{i=1}^{m} z_i^{\downarrow}$, then $\sum_{i=1}^{m} x_i^{\downarrow} \geq \sum_{i=1}^{m} z_i^{\downarrow}$. □

However, majorization is not a total order: For example, the vectors

$$x = \begin{pmatrix} 2 \\ 2 \\ 4 \\ 6 \end{pmatrix} \qquad \text{and} \qquad y = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 5 \end{pmatrix}$$

satisfy neither $x \succcurlyeq y$ nor $y \succcurlyeq x$, since we have $6 > 5$ but $6 + 4 + 2 < 5 + 5 + 3$. For an even simpler example, if two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ have different sums of coordinates, then we have neither $x \succcurlyeq y$ nor $y \succcurlyeq x$.

4.7.2. Restating Schur's theorem as a majorization

Now we can restate Corollary 4.6.5 as follows:

Corollary 4.7.6 (Schur's theorem)**.** Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian $n \times n$-matrix. Then,

$$(\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A))^T \succcurlyeq (A_{1,1}, A_{2,2}, \ldots, A_{n,n})^T.$$

In words: The tuple of all eigenvalues of A majorizes the tuple of all diagonal entries of A.

Proof. We need to show that

- the sum of the m largest eigenvalues of A is $\geq$ to the sum of the m largest diagonal entries of A for each $m \in [n]$;
- the sum of all diagonal entries of A equals the sum of all eigenvalues of A.

But the second of these two statements follows from the well-known theorem that the trace of a matrix is the sum of its eigenvalues (Theorem 2.0.10 **(d)**). Thus, it remains to prove the first statement.

So let $m \in [n]$. We must prove that the sum of the m largest eigenvalues of A is $\geq$ to the sum of the m largest diagonal entries of A. However, Corollary 4.6.5 yields that

$$\begin{aligned}\lambda_1(A) + \lambda_2(A) + \cdots + \lambda_m(A) &\leq A_{i_1,i_1} + A_{i_2,i_2} + \cdots + A_{i_m,i_m} \\ &\leq \lambda_{n-m+1}(A) + \lambda_{n-m+2}(A) + \cdots + \lambda_{n-m}(A)\end{aligned}$$

for any distinct $i_1, i_2, \ldots, i_m \in [n]$. The second inequality here says that

$$\lambda_{n-m+1}(A) + \lambda_{n-m+2}(A) + \cdots + \lambda_{n-m}(A) \geq A_{i_1,i_1} + A_{i_2,i_2} + \cdots + A_{i_m,i_m}$$

for any distinct $i_1, i_2, \ldots, i_m \in [n]$. In other words, the sum of the m largest eigenvalues of A is $\geq$ to any sum of m distinct diagonal entries of A. Thus, in particular, the sum of the m largest eigenvalues of A is $\geq$ to the sum of the m largest diagonal entries of A. This completes the proof of Corollary 4.7.6. □

Exercise 4.7.5. [5] For each Hermitian matrix $A \in \mathbb{C}^{n \times n}$, let $\lambda(A) \in \mathbb{R}^n$ be the column vector $(\lambda_n(A), \lambda_{n-1}(A), \ldots, \lambda_1(A))^T$ consisting of all eigenvalues of A in decreasing order.

Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{C}^{n \times n}$ be two Hermitian matrices. Prove the following:

(a) (*Fan's theorem:*) We have

$$\lambda(A) + \lambda(B) \succcurlyeq \lambda(A + B).$$

(b) (*Lidskii's theorem:*) We have

$$\lambda(A + B) \succcurlyeq \lambda(A) + (\lambda(B))^{\uparrow}.$$

[**Hint:** For part **(a)**, use Corollary 4.6.4 **(b)**. For part **(b)**, observe first that three weakly decreasing vectors $x, y, z \in \mathbb{R}^n$ satisfying $x - z^{\uparrow} \succcurlyeq y$ must always satisfy $x \succcurlyeq y + z^{\uparrow}$.]

4.7.3. Robin Hood moves

Above, we briefly alluded to an intuition for majorization: We said that x majorizes y if and only if you can obtain y from x by "moving the entries closer together (while keeping the average equal)". Let us now turn this into an actual theorem. First, a simple lemma:

Lemma 4.7.7. Let $x, y \in \mathbb{R}^n$. Then, $x \succcurlyeq y$ if and only if $x^{\downarrow} \succcurlyeq y^{\downarrow}$.

Proof. The definition of $\succcurlyeq$ only involves $x^{\downarrow}$ and $y^{\downarrow}$. In other words, whether or not we have $x \succcurlyeq y$ does not depend on the order of the coordinates of x or of those of y. Thus, replacing x and y by $x^{\downarrow}$ and $y^{\downarrow}$ doesn't make any difference. □

Definition 4.7.8. Let $x \in \mathbb{R}^n$. Let i and j be two distinct elements of $[n]$ such that $x_i \leq x_j$. Let $t \in \left[x_i, x_j\right]$ (that is, $t \in \mathbb{R}$ and $x_i \leq t \leq x_j$). Let $y \in \mathbb{R}^n$ be the column vector obtained from x by

$$\text{replacing the } i\text{-th and } j\text{-th coordinates } x_i \text{ and } x_j \text{ by } u \text{ and } v$$

for some $u, v \in \left[x_i, x_j\right]$ satisfying $u + v = x_i + x_j$. In other words, we obtain y by picking two numbers $u, v \in \left[x_i, x_j\right]$ satisfying $u + v = x_i + x_j$ and setting

$$y_k = \begin{cases} u, & \text{if } k = i; \\ v, & \text{if } k = j; \\ x_k, & \text{otherwise} \end{cases} \qquad \text{for all } k \in [n].$$

Then, we say that y is obtained from x by a *Robin Hood move* (short: *RH move*), and we write

$$x \overset{\text{RH}}{\longrightarrow} y.$$

Moreover, if x and y are weakly decreasing, then this RH move is said to be an *order-preserving RH move* (short: *OPRH move*), and we write

$$x \overset{\text{OPRH}}{\longrightarrow} y.$$

A Robin Hood move is thus a "local transformation" that changes a column vector x by picking two of its entries (say, x_i and x_j) and replacing them by two new entries u and v that are "closer together" (that is, $u, v \in \left[x_i, x_j\right]$) and have the same sum (that is, $u + v = x_i + x_j$). If we regard the entries $x_1, x_2, \ldots, x_n$ of x as the wealths of n persons, then a Robin Hood move thus corresponds to redistributing some wealth from a richer person to a poorer person in such a way that the disparity between these two people does not increase. (Thus the name.) If the n people were ordered in the order of decreasing wealth at the beginning (that is, x was weakly decreasing) and this remains so after the RH move, then the RH move is an OPRH move.

Example 4.7.9. (a) Replacing two coordinates of a vector x by their average is an RH move. (Indeed, this corresponds to the case when $u = v = \dfrac{x_i + x_j}{2}$ in Definition 4.7.8.)

(b) Swapping two coordinates of a vector x is an RH move. (Indeed, this corresponds to the case when $u = x_j$ and $v = x_i$ in Definition 4.7.8.)

(c) If $x \in \mathbb{R}^n$ is weakly decreasing, then replacing two adjacent entries of x by their average is an OPRH move. (Indeed, if we replace x_i and x_{i+1} by their average $\dfrac{x_i + x_{i+1}}{2}$, then the vector remains weakly decreasing, since $x_1 \geq x_2 \geq \cdots \geq x_n$ entails $x_1 \geq x_2 \geq \cdots \geq x_{i-1} \geq \dfrac{x_i + x_{i+1}}{2} \geq \dfrac{x_i + x_{i+1}}{2} \geq x_{i+2} \geq x_{i+3} \geq \cdots \geq x_n$.)

(d) More generally: If $x \in \mathbb{R}^n$ is weakly decreasing, then replacing its coordinates x_i and x_{i+1} by u and $x_i + x_{i+1} - u$ is an OPRH move if and only if $u \in \left[\dfrac{x_i + x_{i+1}}{2}, x_i\right]$.

(e) Here is an example of an RH move that is not an OPRH move:

$$\begin{pmatrix} 6 \\ 5 \\ 2 \\ 1 \end{pmatrix} \xrightarrow{\text{RH}} \begin{pmatrix} 4 \\ 5 \\ 2 \\ 3 \end{pmatrix}.$$

This move replaces the two entries 1 and 6 by 3 and 4 (with $3, 4 \in [1, 6]$ and $3 + 4 = 1 + 6$), but it changes the relative order of the entries, so it is not order-preserving.

Proposition 4.7.10. If $x \xrightarrow{\text{RH}} y$, then the sum of the entries of x equals the sum of the entries of y.

Proof. Clear from the $u + v = x_i + x_j$ requirement in Definition 4.7.8. □

Lemma 4.7.11. Let $x, y \in \mathbb{R}^n$ be weakly decreasing column vectors such that y is obtained from x by a (finite) sequence of OPRH moves. Then, $x \succcurlyeq y$.

Proof. Let us first prove Lemma 4.7.11 in the case when y is obtained from x by a **single** OPRH move.

So let us assume that y is obtained from x by a **single** OPRH move. Let this move be replacing x_i and x_j by u and v, where $x_i \leq x_j$ and $u, v \in \left[x_i, x_j\right]$ with $u + v = x_i + x_j$. WLOG, we have $x_i < x_j$ (since otherwise, the OPRH move changes nothing, because we have $u = v = x_i = x_j$). Therefore, $i > j$ (since x is weakly decreasing (by the definition of an OPRH move)). Thus,

$$y = \left(x_1, x_2, \ldots, x_{j-1}, v, x_{j+1}, x_{j+2}, \ldots, x_{i-1}, u, x_{i+1}, x_{i+2}, \ldots, x_n\right)^T$$

(since y is obtained from x by replacing x_i and x_j by u and v).

Now, we must prove that $x \succcurlyeq y$. In other words, we must prove that

$$x_1 + x_2 + \cdots + x_m \geq y_1 + y_2 + \cdots + y_m \tag{94}$$

for each $m \in [n]$ (since x and y are weakly decreasing), and we must prove that

$$x_1 + x_2 + \cdots + x_n = y_1 + y_2 + \cdots + y_n. \tag{95}$$

The latter equality follows from $u + v = x_i + x_j$. So we only need to prove the former inequality. So let us fix an $m \in [n]$. We must prove the inequality (94). We are in one of the following cases:

1. We have $m < j$.
2. We have $j \leq m < i$.
3. We have $i \leq m$.

In Case 1, we have $x_1 + x_2 + \cdots + x_m = y_1 + y_2 + \cdots + y_m$ (because $x_p = y_p$ for all $p \leq m$ in this case). Thus, (94) is proved in Case 1.

In Case 2, we have

$$\begin{aligned} y_1 + y_2 + \cdots + y_m &= x_1 + x_2 + \cdots + x_{j-1} + v + x_{j+1} + x_{j+2} + \cdots + x_m \\ &= (x_1 + x_2 + \cdots + x_m) + \underbrace{v - x_j}_{\substack{\leq 0 \\ (\text{since } v \in [x_i, x_j])}} \\ &\leq x_1 + x_2 + \cdots + x_m. \end{aligned}$$

Thus, (94) is proved in Case 2.

In Case 3, we have

$$\begin{aligned} &y_1 + y_2 + \cdots + y_m \\ &= x_1 + x_2 + \cdots + x_{j-1} + v + x_{j+1} + x_{j+2} + \cdots + x_{i-1} + u + x_{i+1} + x_{i+2} + \cdots + x_m \\ &= (x_1 + x_2 + \cdots + x_m) + \underbrace{(u - x_i) + (v - x_j)}_{\substack{=0 \\ (\text{since } u+v=x_i+x_j)}} \\ &= x_1 + x_2 + \cdots + x_m. \end{aligned}$$

Thus, (94) is proved in Case 3.

So we have proved (94) in all cases. As we said, this concludes our proof of $x \succcurlyeq y$.

We are not completely done yet: We have only proved Lemma 4.7.11 in the case when y is obtained from x by a **single** OPRH move.

Now, let us prove the general case: Assume that y is obtained from x by a (finite) sequence of OPRH moves. That is, there is a finite sequence $x_{[0]}, x_{[1]}, \ldots, x_{[m]}$ of vectors in $\mathbb{R}^n$ such that $x_{[0]} = x$ and $x_{[m]} = y$ and

$$x_{[0]} \overset{\text{OPRH}}{\longrightarrow} x_{[1]} \overset{\text{OPRH}}{\longrightarrow} \cdots \overset{\text{OPRH}}{\longrightarrow} x_{[m]}.$$

Then, by the special case we have already proved, we conclude that

$$x_{[0]} \succcurlyeq x_{[1]} \succcurlyeq \cdots \succcurlyeq x_{[m]}.$$

Hence, $x_{[0]} \succcurlyeq x_{[m]}$ (since the relation $\succcurlyeq$ is reflexive and transitive). In other words, $x \succcurlyeq y$ (since $x_{[0]} = x$ and $x_{[m]} = y$). This completes the proof of Lemma 4.7.11. $\square$

We are now ready to state one version of the "moving the entries closer together" intuition for majorization:

Theorem 4.7.12 (RH criterion for majorization)**.** Let $x, y \in \mathbb{R}^n$ be two weakly decreasing column vectors. Then, $x \succcurlyeq y$ if and only if y can be obtained from x by a (finite) sequence of OPRH moves.

Example 4.7.13. (a) We have $(4,1,1)^T \succcurlyeq (2,2,2)^T$, and indeed $(2,2,2)^T$ can be obtained from $(4,1,1)^T$ by OPRH moves as follows:

$$(4,1,1)^T \overset{\text{OPRH}}{\longrightarrow} (3,2,1)^T \overset{\text{OPRH}}{\longrightarrow} (2,2,2)^T.$$

(b) We have $(7,5,2,0)^T \succcurlyeq (4,4,3,3)^T$, and indeed $(4,4,3,3)^T$ can be obtained from $(7,5,2,0)^T$ by OPRH moves as follows:

$$(7,5,2,0)^T \overset{\text{OPRH}}{\longrightarrow} (6,6,2,0)^T \overset{\text{OPRH}}{\longrightarrow} (6,5,3,0)^T \overset{\text{OPRH}}{\longrightarrow} (6,4,3,1)^T \overset{\text{OPRH}}{\longrightarrow} (4,4,3,3)^T.$$

Here is another way to do this:

$$(7,5,2,0)^T \overset{\text{OPRH}}{\longrightarrow} (7,4,3,0)^T \overset{\text{OPRH}}{\longrightarrow} (4,4,3,3)^T.$$

Proof of Theorem 4.7.12. $\Longleftarrow$: This follows from Lemma 4.7.11.

$\Longrightarrow$: Let $x \succcurlyeq y$. We must show that y can be obtained from x by a finite sequence of OPRH moves.

If $x = y$, then this is clear (just take the empty sequence). So we WLOG assume that $x \neq y$. We **claim** now that there is a further weakly decreasing vector $z \in \mathbb{R}^n$ such that

1. we have $x \overset{\text{OPRH}}{\longrightarrow} z$;
2. we have $z \succcurlyeq y$;
3. the vector z has more entries in common with y than x does; in other words, we have
$$|\{i \in [n] \mid z_i = y_i\}| > |\{i \in [n] \mid x_i = y_i\}|. \tag{96}$$

In other words, we claim that by making a strategic OPRH move starting at x, we can reach a vector z that still majorizes y but has at least one more entry in common with y than x does. If we can prove this claim, then we will automatically obtain a recursive procedure to transform x into y by a sequence of OPRH moves. (And in fact, this procedure will use at most n moves, because each move makes the vector agree with y in at least one more position.)

So let us prove our claim.

Since x is weakly decreasing, we have $x = x^{\downarrow}$. Similarly, $y = y^{\downarrow}$. Thus, from $x \succcurlyeq y$, we obtain

$$x_1 + x_2 + \cdots + x_m \geq y_1 + y_2 + \cdots + y_m \tag{97}$$

for all $m \in [n]$, as well as

$$x_1 + x_2 + \cdots + x_n = y_1 + y_2 + \cdots + y_n. \tag{98}$$

We define a *turn* to be a pair (a, b) of two elements of $[n]$ satisfying

$$x_a > y_a \qquad \text{and} \qquad x_b < y_b \qquad \text{and} \qquad a < b.$$

We claim that there exists at least one turn.

[*Proof:* We have $x \neq y$. Thus, there exists some $a \in [n]$ such that $x_a \neq y_a$. Consider the smallest such a. Thus, $x_i = y_i$ for each $i < a$. However, (97) (applied to $m = a$) yields $x_1 + x_2 + \cdots + x_a \geq y_1 + y_2 + \cdots + y_a$. Thus, $x_a \geq y_a$ (since $x_i = y_i$ for each $i < a$). Hence, $x_a > y_a$ (since $x_a \neq y_a$). Therefore, $x_1 + x_2 + \cdots + x_a > y_1 + y_2 + \cdots + y_a$.

Next, let us pick the smallest $b \in \{a+1, a+2, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_b = y_1 + y_2 + \cdots + y_b$. (This exists, because (98) shows that n is such a b.) Clearly, $a < b$ (since $b \in \{a+1, a+2, \ldots, n\}$).

We shall now show that $x_b < y_b$. Indeed, assume the contrary. Thus, $x_b \geq y_b$. Subtracting this inequality from the equality $x_1 + x_2 + \cdots + x_b = y_1 + y_2 + \cdots + y_b$, we obtain $x_1 + x_2 + \cdots + x_{b-1} \leq y_1 + y_2 + \cdots + y_{b-1}$. However, we have $x_1 + x_2 + \cdots + x_{b-1} \geq y_1 + y_2 + \cdots + y_{b-1}$ (by (97)). Combining these two inequalities, we obtain $x_1 + x_2 + \cdots + x_{b-1} = y_1 + y_2 + \cdots + y_{b-1}$. However, recall that b was defined to be the **smallest** element of $\{a+1, a+2, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_b = y_1 + y_2 + \cdots + y_b$. This contradicts the equality $x_1 + x_2 + \cdots + x_{b-1} = y_1 + y_2 + \cdots + y_{b-1}$ (after all, $b-1$ is smaller than b) unless $b-1$ is not an element of $\{a+1, a+2, \ldots, n\}$. Thus, $b-1$ must not be an element of $\{a+1, a+2, \ldots, n\}$. So we have $b \in \{a+1, a+2, \ldots, n\}$ but $b-1 \notin \{a+1, a+2, \ldots, n\}$. Therefore, $b = a+1$, so that $b-1 = a$. Thus, the equality $x_1 + x_2 + \cdots + x_{b-1} = y_1 + y_2 + \cdots + y_{b-1}$ rewrites as $x_1 + x_2 + \cdots + x_a = y_1 + y_2 + \cdots + y_a$. But this contradicts the inequality $x_1 + x_2 + \cdots + x_a > y_1 + y_2 + \cdots + y_a$ proved above. Thus, our proof of $x_b < y_b$ is complete.

We thus conclude that (a, b) is a turn (since $a < b$ and $x_a > y_a$ and $x_b < y_b$). This proves that there exists at least one turn.]

The *width* of a turn (a, b) shall denote the positive integer $b - a$. (This is a positive integer, since $a < b$ in a turn (a, b).)

Now, let us pick a turn (a, b) with the **smallest possible width**. Then, we have

$$x_a > y_a \qquad \text{and} \qquad x_b < y_b \qquad \text{and} \qquad a < b.$$

Moreover, for each $j \in \{a+1, a+2, \ldots, b-1\}$, we have $x_j = y_j$ (because if we had $x_j < y_j$, then (a, j) would be a turn of smaller width than (a, b), and if we had $x_j > y_j$, then (j, b) would be a turn of smaller width than (a, b)). Thus, we have

$$\begin{aligned} &x_a > y_a, \\ &x_j = y_j \qquad \text{for all } j \in \{a+1, a+2, \ldots, b-1\}, \\ &x_b < y_b. \end{aligned}$$

Since y is weakly decreasing (so that $y_a \geq y_{a+1} \geq \cdots \geq y_b$), we thus obtain the following chain of inequalities:

$$x_a > y_a \geq \left(\text{all of the } x_j \text{ and } y_j \text{ with } j \in \{a+1, a+2, \ldots, b-1\}\right) \geq y_b > x_b.$$

(If there are no $j \in \{a+1, a+2, \ldots, b-1\}$, then this is supposed to read $x_a > y_a \geq y_b > x_b$.) This shows, in particular, that y_a and y_b lie in the open interval (x_a, x_b).

Now, we make an RH move (on the vector x) that "squeezes x_a and x_b together" until either x_a reaches y_a or x_b reaches y_b (whatever happens first). In formal terms, this means that we do the following:

- If $x_a - y_a \leq y_b - x_b$, then we replace x_a and x_b by y_a and $x_a + x_b - y_a$.
- If $x_a - y_a \geq y_b - x_b$, then we replace x_a and x_b by $x_a + x_b - y_b$ and y_b.

(The two cases overlap, but this is not a problem, because if $x_a - y_a = y_b - x_b$, then both outcomes are identical.)

Let $z \in \mathbb{R}^n$ be the n-tuple resulting from this operation. We claim that z is weakly decreasing and satisfies the three requirements 1, 2, 3 above:

1. we have $x \overset{\text{OPRH}}{\longrightarrow} z$;
2. we have $z \succcurlyeq y$;
3. the vector z has more entries in common with y than x does; in other words, we have
$$\left|\{i \in [n] \ \mid \ z_i = y_i\}\right| > \left|\{i \in [n] \ \mid \ x_i = y_i\}\right|.$$

Indeed, let us prove that these three requirements hold. Requirement 3 is the easiest one to verify: The only entries of z that differ from the respective entries of x are the two entries z_a and z_b; among these two entries, at least one agrees with the corresponding entry of y (because we have either $z_a = y_a$ or $z_b = y_b$), whereas none of the corresponding entries of x agrees with the corresponding entry of y (since $x_a > y_a$ and $x_b < y_b$). Thus, going from x to z, we have increased the number of

entries that agree with the corresponding entries of y by at least 1. Requirement 3 is therefore satisfied.

Let us next check Requirement 1. (The reader should draw a picture of the numbers x_j and y_j for $j \in \{a, a+1, \ldots, b\}$ as points on the real axis. As we recall, z_a and z_b are obtained by moving x_a and x_b closer together (preserving their sum) until either x_a hits y_a or x_b hits y_b (whatever happens first). This picture should render some of the verifications below trivial.)

The definition of z shows that $z_a \geq y_a$ [42] and $z_b \leq y_b$ [43]. Moreover, the numbers z_a and z_b lie in the interval $[x_b, x_a]$ [44], and we have $z_a + z_b = x_a + x_b$ [45]. Thus, the operation that transformed x into z was an RH move (with b and a playing the roles of i and j from Definition 4.7.8). Therefore, $x \xrightarrow{\text{RH}} z$. It remains to prove that this RH move is order-preserving, i.e., that z is weakly decreasing. Since the only entries of x that changed in our RH move were x_a and x_b (and since we know that x is weakly decreasing), we only need to verify the inequalities

$$\begin{aligned}
x_{a-1} &\geq z_a && (\text{if } a > 1) && \text{and} \\
z_a &\geq x_{a+1} && (\text{if } a+1 \neq b) && \text{and} \\
z_a &\geq z_b && (\text{if } a+1 = b) && \text{and} \\
x_{b-1} &\geq z_b && (\text{if } a+1 \neq b) && \text{and} \\
z_b &\geq x_{b+1} && (\text{if } b \neq n)\,.
\end{aligned}$$

Fortunately, these inequalities all follow easily from the facts that we know (viz.,

[42] *Proof.* The definition of z_a shows that $z_a = y_a$ in the case when $x_a - y_a \leq y_b - x_b$, and that $z_a = x_a + x_b - y_b$ in the case when $x_a - y_a \geq y_b - x_b$. In the former case, the inequality $z_a \geq y_a$ is obvious (and is, in fact, an equality). Hence, it remains to prove $z_a \geq y_a$ in the latter case.

So let us assume that we are in the latter case – i.e., that we have $x_a - y_a \geq y_b - x_b$. Thus, $x_a \geq y_b - x_b + y_a$. Now,

$$z_a = \underbrace{x_a}_{\geq y_b - x_b + y_a} + x_b - y_b \geq y_b - x_b + y_a + x_b - y_b = y_a,$$

qed.

[43] *Proof.* The definition of z_b shows that $z_b = x_a + x_b - y_a$ in the case when $x_a - y_a \leq y_b - x_b$, and that $z_b = y_b$ in the case when $x_a - y_a \geq y_b - x_b$. In the latter case, the inequality $z_b \leq y_b$ is obvious (and is, in fact, an equality). Hence, it remains to prove $z_b \leq y_b$ in the former case.

So let us assume that we are in the former case – i.e., that we have $x_a - y_a \leq y_b - x_b$. Thus, $x_a \leq y_b - x_b + y_a$. Now,

$$z_b = \underbrace{x_a}_{\leq y_b - x_b + y_a} + x_b - y_a \leq y_b - x_b + y_a + x_b - y_a = y_b,$$

qed.

[44] *Proof.* The definition of z shows that we have either ($z_a = y_a$ and $z_b = x_a + x_b - y_a$) or ($z_a = x_a + x_b - y_b$ and $z_b = y_b$). Hence, we must show that the numbers y_a, $x_a + x_b - y_a$, $x_a + x_b - y_b$ and y_b all lie in the interval $[x_b, x_a]$. But this follows easily from $x_a > y_a \geq y_b > x_b$.

[45] *Proof.* The definition of z shows that we have either ($z_a = y_a$ and $z_b = x_a + x_b - y_a$) or ($z_a = x_a + x_b - y_b$ and $z_b = y_b$). Hence, we must show that $y_a + (x_a + x_b - y_a) = x_a + x_b$ and $(x_a + x_b - y_b) + y_b = x_a + x_b$. But both of these equalities are clearly true.

from the chain of inequalities

$$x_a > y_a \geq \left(\text{all of the } x_j \text{ and } y_j \text{ with } j \in \{a+1, a+2, \ldots, b-1\}\right) \geq y_b > x_b$$

and from the inequalities $z_a \geq y_a$ and $z_b \leq y_b$ and the fact that z_a and z_b lie in the interval $[x_b, x_a]$): The inequality $x_{a-1} \geq z_a$ (if $a > 1$) follows from $x_{a-1} \geq x_a \geq z_a$ (since $z_a \in [x_b, x_a]$). The inequality $z_a \geq x_{a+1}$ (if $a+1 \neq b$) follows from

$$z_a \geq y_a \geq \left(\text{all of the } x_j \text{ and } y_j \text{ with } j \in \{a+1, a+2, \ldots, b-1\}\right)$$

(since x_{a+1} is one of the latter x_j). Furthermore, $z_a \geq z_b$ follows from $z_a \geq y_a \geq y_b \geq z_b$ (since $z_b \leq y_b$). Next, $x_{b-1} \geq z_b$ (if $a+1 \neq b$) follows from

$$\left(\text{all of the } x_j \text{ and } y_j \text{ with } j \in \{a+1, a+2, \ldots, b-1\}\right) \geq y_b \geq z_b$$

(since x_{b-1} is one of those former x_j). Finally, $z_b \geq x_{b+1}$ follows by combining $z_b \geq x_b$ (this is because z_b lies in the interval $[x_b, x_a]$) and $x_b \geq x_{b+1}$. Thus, we have checked all the necessary inequalities to ensure that z is weakly decreasing. Thus, requirement 1 holds.

Finally, we need to verify requirement 2. In other words, we need to show that $z \succcurlyeq y$. Since z and y are weakly decreasing, this means that we need to verify the inequalities

$$z_1 + z_2 + \cdots + z_m \geq y_1 + y_2 + \cdots + y_m \tag{99}$$

for all $m \in [n]$, as well as the equality

$$z_1 + z_2 + \cdots + z_n = y_1 + y_2 + \cdots + y_n. \tag{100}$$

The equality (100) is easy to verify: Since $x \xrightarrow{\text{RH}} z$ (and since the sum of the entries of a vector does not change when we make an RH move), we have

$$z_1 + z_2 + \cdots + z_n = x_1 + x_2 + \cdots + x_n = y_1 + y_2 + \cdots + y_n$$

(since $x \succcurlyeq y$). Thus, it remains to prove the inequality (99). So let us fix $m \in [n]$. We must prove (99). We are in one of the following three cases:

Case 1: We have $m < a$.

Case 2: We have $a \leq m < b$.

Case 3: We have $m \geq b$.

Let us first consider Case 1. In this case, we have $m < a$. Hence, $z_i = x_i$ for each $i \leq m$. Thus,

$$z_1 + z_2 + \cdots + z_m = x_1 + x_2 + \cdots + x_m \geq y_1 + y_2 + \cdots + y_m$$

(since $x \succcurlyeq y$). This proves (99) in Case 1.

Let us next consider Case 2. In this case, we have $a \leq m < b$. Hence,

$$\begin{aligned} z_1 + z_2 + \cdots + z_m &= \underbrace{x_1 + x_2 + \cdots + x_{a-1}}_{\substack{\geq y_1+y_2+\cdots+y_{a-1} \\ (\text{since } x \succcurlyeq y)}} + \underbrace{z_a}_{\geq y_a} + \underbrace{x_{a+1} + x_{a+2} + \cdots + x_m}_{\substack{= y_{a+1}+y_{a+2}+\cdots+y_m \\ (\text{since } x_j = y_j \text{ for all } j \in \{a+1, a+2, \ldots, b-1\})}} \\ &\geq y_1 + y_2 + \cdots + y_{a-1} + y_a + y_{a+1} + y_{a+2} + \cdots + y_m \\ &= y_1 + y_2 + \cdots + y_m. \end{aligned}$$

This proves (99) in Case 2.

Finally, let us consider Case 3. In this case, we have $m \geq b$. Hence,

$$\begin{aligned}
&z_1 + z_2 + \cdots + z_m \\
&= x_1 + x_2 + \cdots + x_{a-1} + z_a + x_{a+1} + x_{a+2} + \cdots + x_{b-1} + z_b + x_{b+1} + x_{b+2} + \cdots + x_m \\
&= (x_1 + x_2 + \cdots + x_m) + \underbrace{(z_a - x_a) + (z_b - x_b)}_{\substack{=0 \\ (\text{since } z_a + z_b = x_a + x_b)}} \\
&= x_1 + x_2 + \cdots + x_m \geq y_1 + y_2 + \cdots + y_m \qquad (\text{since } x \succcurlyeq y).
\end{aligned}$$

This proves (99) in Case 3.

We have now proved (99) in all three Cases 1, 2 and 3. Hence, (99) always holds. Thus, we have verified Requirement 2. Now, all three requirements 1, 2 and 3 are satisfied. As we explained, this means that our vector z fits our bill, and this completes the proof of Theorem 4.7.12. □

Exercise 4.7.6. [2] Let $x, y \in \mathbb{R}^n$ be two weakly decreasing column vectors. Prove that $x \succcurlyeq y$ if and only if y can be obtained from x by a sequence of at most $n-1$ OPRH moves.

Theorem 4.7.12 formalizes our intuition about majorization as "moving entries closer together" for weakly decreasing vectors. There is a version for arbitrary vectors as well:

Theorem 4.7.14 (RH criterion for majorization, non-decreasing form)**.** Let $x, y \in \mathbb{R}^n$ be two column vectors. Then, $x \succcurlyeq y$ if and only if y can be obtained from x by a (finite) sequence of RH moves.

The proof of this theorem will be an exercise, but we delay this exercise until after Theorem 4.7.20, since the latter theorem provides a good tool for the proof.

Exercise 4.7.7. [7] Let $x, y \in \mathbb{R}^n$ be two column vectors such that $x \succcurlyeq y$. Prove that there exists a real symmetric matrix $A \in \mathbb{R}^{n \times n}$ with diagonal entries $y_1, y_2, \ldots, y_n$ and eigenvalues $x_1, x_2, \ldots, x_n$.

[**Remark:** This can be viewed as a converse to Corollary 4.7.6 (but is in fact even stronger, since A is not just Hermitian but real symmetric).]

4.7.4. Karamata's inequality

Now, what can we do with majorization? Probably the most important property of majorizing pairs of vectors is the so-called *Karamata inequality*. To state it, we recall the concept of a convex function:

Definition 4.7.15. Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a function. We say that f is *convex* if and only if it has the following property: For any $a, b \in I$ and any $\lambda \in [0,1]$, we have

$$\lambda f(a) + (1-\lambda) f(b) \geq f(\lambda a + (1-\lambda) b).$$

This property is best conceptualized using the plot of the function: A function $f : I \to \mathbb{R}$ is convex if and only if, for any two points $(a, f(a))$ and $(b, f(b))$ on the plot of f, the entire segment connecting these two points lies weakly above (i.e., on or above) the plot of f. (In fact, the segment connecting these two points can be parametrized as

$$\{(\lambda a + (1-\lambda) b,\ \lambda f(a) + (1-\lambda) f(b)) \mid \lambda \in [0,1]\},$$

so a typical point on this segment has the form

$$(\lambda a + (1-\lambda) b,\ \lambda f(a) + (1-\lambda) f(b))$$

for some $\lambda \in [0,1]$, whereas the corresponding point on the plot of f is

$$(\lambda a + (1-\lambda) b,\ f(\lambda a + (1-\lambda) b)).$$

Thus, the former point lies weakly above the latter point if and only if $\lambda f(a) + (1-\lambda) f(b) \geq f(\lambda a + (1-\lambda) b)$ holds.)

Here are some examples of convex functions:

- $f(t) = t^n$ defines a convex function $f : \mathbb{R} \to \mathbb{R}$ whenever $n \in \mathbb{N}$ is even.
- $f(t) = t^n$ defines a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ whenever $n \in \mathbb{R} \setminus (0,1)$. Otherwise, it defines a concave function[46].
- $f(t) = \sin t$ defines a concave function $f : [0, \pi] \to \mathbb{R}$ and a convex function $f : [\pi, 2\pi] \to \mathbb{R}$.

Before we state Karamata's inequality, let us recall three fundamental properties of convex functions:

Theorem 4.7.16 (second derivative test)**.** Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a twice differentiable function. Then, f is convex if and only if each $x \in I$ satisfies $f''(x) \geq 0$.

Note that there are convex functions that are not twice differentiable. For instance, the absolute value function $f : \mathbb{R} \to \mathbb{R}$ (given by $f(z) = |z|$ for each $z \in \mathbb{R}$) is convex. This cannot be proved by the second derivative test, but it is easy to check using the triangle inequality.

A classical property of convex functions is *Jensen's inequality*. Let us give it in two of its forms – a simple unweighted and a more general weighted one:

[46] A function $f : I \to \mathbb{R}$ is said to be *concave* if $-f$ is convex.

Theorem 4.7.17 (Jensen's inequality)**.** Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a convex function. Let $x_1, x_2, \ldots, x_n \in I$. Let $m = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$. Then,

$$f(x_1) + f(x_2) + \cdots + f(x_n) \geq nf(m).$$

Theorem 4.7.18 (weighted Jensen's inequality)**.** Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a convex function. Let $x_1, x_2, \ldots, x_n \in I$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be n nonnegative reals satisfying $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$. Then,

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n) \geq f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n).$$

The easiest way to prove these two theorems is to first prove Theorem 4.7.18 by induction on n (this is commonly done in textbooks on probability theory, where this inequality is used quite often[47]) and then obtain Theorem 4.7.17 from it by setting $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \dfrac{1}{n}$.

Note that both Jensen's inequalities are highly useful for proving various kinds of inequalities, even ones in which there is no convex function easily visible. See [Hung07, Chapter 4] for some such applications.

Jensen's inequality can be interpreted as saying that a sum of values of a convex function f at several points $x_1, x_2, \ldots, x_n$ becomes smaller (or at least not larger) if all these points are replaced by their average $m = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$. Karamata's inequality generalizes this by saying that we don't need to replace all the points by their average right away, but rather it suffices to "move them closer together" (not necessarily going all the way to the average). Here, "moving them closer together" is formalized using majorization:

Theorem 4.7.19 (Karamata's inequality)**.** Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a convex function. Let $x \in I^n$ and $y \in I^n$ be two vectors such that $x \succcurlyeq y$. Then,

$$f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n).$$

Karamata's inequality has many applications (see, e.g., [KDLM05, §2], which incidentally also gives a proof of Karamata's inequality different from the ones we shall give below). In particular, Jensen's inequality follows from Karamata's inequality, since Exercise 4.7.1 says that $(x_1, x_2, \ldots, x_n)^T \succcurlyeq (m, m, \ldots, m)^T$.

The weighted Jensen's inequality can, incidentally, be derived from a weighted Karamata's inequality (see Exercise 4.7.10 below).

Let us now prove Karamata's inequality:

[47] Or see `https://en.wikipedia.org/wiki/Jensen's_inequality` or `https://www.ucd.ie/mathstat/t4media/convex-sets-and-jensen-inequalities-mathstat.pdf` .

Proof of Theorem 4.7.19. It is enough to prove the claim in the case when x and y are weakly decreasing (because permuting the entries of any of x and y does not change anything).

Furthermore, it is enough to prove the claim in the case when $x \overset{\text{OPRH}}{\longrightarrow} y$ (this means that y is obtained from x by a single OPRH move). Indeed, if we have shown this, then it will mean that the sum $f(x_1) + f(x_2) + \cdots + f(x_n)$ decreases (weakly) every time we apply an OPRH move to the vector x. Therefore, if y is obtained from x by a (finite) sequence of OPRH moves, then $f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n)$. Hence, if if $x \succcurlyeq y$, then $f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n)$ (since Theorem 4.7.12 shows that y can be obtained from x by a (finite) sequence of OPRH moves).

So let us assume that $x \overset{\text{OPRH}}{\longrightarrow} y$. Thus, y is obtained from x by picking two entries x_i and x_j with $x_i \leq x_j$ and replacing them by u and v, where $u, v \in \left[x_i, x_j\right]$ with $u + v = x_i + x_j$. Consider these x_i, x_j, u, v. We must prove that

$$f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n).$$

It clearly suffices to show that

$$f(x_i) + f\left(x_j\right) \geq f(u) + f(v)$$

(since $y_k = x_k$ for all k other than i and j).

But showing this is easy: From $u \in \left[x_i, x_j\right]$, we obtain

$$u = \lambda x_i + (1 - \lambda) x_j \qquad \text{for some } \lambda \in [0, 1]$$

(namely, $\lambda = \dfrac{u - x_j}{x_i - x_j}$). Consider this λ. Then,

$$v = (1 - \lambda) x_i + \lambda x_j$$

(this follows easily from substituting $u = \lambda x_i + (1 - \lambda) x_j$ into $u + v = x_i + x_j$ and solving for v).

From $u = \lambda x_i + (1 - \lambda) x_j$, we obtain

$$f(u) = f\left(\lambda x_i + (1 - \lambda) x_j\right) \leq \lambda f(x_i) + (1 - \lambda) f\left(x_j\right) \qquad \text{(since } f \text{ is convex)}.$$

From $v = (1 - \lambda) x_i + \lambda x_j$, we obtain

$$f(v) = f\left((1 - \lambda) x_i + \lambda x_j\right) \leq (1 - \lambda) f(x_i) + \lambda f\left(x_j\right) \qquad \text{(since } f \text{ is convex)}.$$

Adding together these two inequalities, we obtain

$$\begin{aligned} f(u) + f(v) &\leq \left(\lambda f(x_i) + (1 - \lambda) f\left(x_j\right)\right) + \left((1 - \lambda) f(x_i) + \lambda f\left(x_j\right)\right) \\ &= f(x_i) + f\left(x_j\right), \qquad \text{qed.} \end{aligned}$$

Thus, Theorem 4.7.19 is proved. □

Karamata's inequality has a converse: If $x, y \in \mathbb{R}^n$ are two vectors such that **every** convex function $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n),$$

then $x \succcurlyeq y$. Even better, we don't even need to require this to hold for **every** convex function f; instead, it suffices to require for the special class of convex functions $f : \mathbb{R} \to \mathbb{R}$ that have the form $z \mapsto |z - t|$ for constants $t \in \mathbb{R}$. In other words, we have the following:

Theorem 4.7.20 (absolute-value criterion for majorization)**.** Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ be two vectors. Then, $x \succcurlyeq y$ if and only if all $t \in \mathbb{R}$ satisfy

$$|x_1 - t| + |x_2 - t| + \cdots + |x_n - t| \geq |y_1 - t| + |y_2 - t| + \cdots + |y_n - t|.$$

Proof. $\Longrightarrow$: Assume that $x \succcurlyeq y$. Let $t \in \mathbb{R}$. Consider the function

$$\begin{aligned} f_t : \mathbb{R} &\to \mathbb{R}, \\ z &\mapsto |z - t|. \end{aligned}$$

This function f_t is convex (this follows easily from the triangle inequality). Hence, Karamata's inequality (Theorem 4.7.19) yields

$$f_t(x_1) + f_t(x_2) + \cdots + f_t(x_n) \geq f_t(y_1) + f_t(y_2) + \cdots + f_t(y_n).$$

By the definition of f_t, this means

$$|x_1 - t| + |x_2 - t| + \cdots + |x_n - t| \geq |y_1 - t| + |y_2 - t| + \cdots + |y_n - t|.$$

So we have proved the "$\Longrightarrow$" direction of Theorem 4.7.20.

$\Longleftarrow$: We assume that the inequality

$$\begin{aligned} &|x_1 - t| + |x_2 - t| + \cdots + |x_n - t| \\ &\geq |y_1 - t| + |y_2 - t| + \cdots + |y_n - t| \end{aligned} \tag{101}$$

holds for all $t \in \mathbb{R}$. (Actually, it will suffice to assume that it holds for all $t \in \{x_1, x_2, \ldots, x_n\}$.)

We must prove that $x \succcurlyeq y$.

WLOG assume that x and y are weakly decreasing (since permuting the entries changes neither the inequality (101) nor the claim $x \succcurlyeq y$).

For each $t \in \mathbb{R}$, we have

$$\begin{aligned} \sum_{i=1}^{n} |x_i - t| &= |x_1 - t| + |x_2 - t| + \cdots + |x_n - t| \\ &\geq |y_1 - t| + |y_2 - t| + \cdots + |y_n - t| \qquad \text{(by (101))} \\ &= \sum_{i=1}^{n} |y_i - t|. \end{aligned} \tag{102}$$

Let $k \in \{0,1,\ldots,n\}$. Pick some $t \in \{x_1, x_2, \ldots, x_n\}$ satisfying $x_k \geq t \geq x_{k+1}$ [48]. Then, since x is weakly decreasing, we have

$$x_1 \geq x_2 \geq \cdots \geq x_k \geq t \geq x_{k+1} \geq x_{k+2} \geq \cdots \geq x_n.$$

Thus, each $i \in \{1,2,\ldots,k\}$ satisfies $x_i \geq t$ and therefore

$$|x_i - t| = x_i - t, \tag{103}$$

whereas each $i \in \{k+1, k+2, \ldots, n\}$ satisfies $t \geq x_i$ and therefore

$$|x_i - t| = t - x_i. \tag{104}$$

Now,

$$\begin{aligned}
\sum_{i=1}^{n} |x_i - t| &= \sum_{i=1}^{k} \underbrace{|x_i - t|}_{\substack{= x_i - t \\ \text{(by (103))}}} + \sum_{i=k+1}^{n} \underbrace{|x_i - t|}_{\substack{= t - x_i \\ \text{(by (104))}}} \\
&= \underbrace{\sum_{i=1}^{k} (x_i - t)}_{= \sum_{i=1}^{k} x_i - kt} + \underbrace{\sum_{i=k+1}^{n} (t - x_i)}_{= (n-k)t - \sum_{i=k+1}^{n} x_i} = \sum_{i=1}^{k} x_i - kt + (n-k)\,t - \underbrace{\sum_{i=k+1}^{n} x_i}_{= \sum_{i=1}^{n} x_i - \sum_{i=1}^{k} x_i} \\
&= \sum_{i=1}^{k} x_i - kt + (n-k)\,t - \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{k} x_i \right) \\
&= 2\sum_{i=1}^{k} x_i + (n-2k)\,t - \sum_{i=1}^{n} x_i
\end{aligned} \tag{105}$$

and

$$\begin{aligned}
\sum_{i=1}^{n} |y_i - t| &= \sum_{i=1}^{k} \underbrace{|y_i - t|}_{\substack{\geq y_i - t \\ \text{(since } |z| \geq z \text{ for each } z \in \mathbb{R}\text{)}}} + \sum_{i=k+1}^{n} \underbrace{|y_i - t|}_{\substack{\geq t - y_i \\ \text{(since } |z| \geq -z \text{ for each } z \in \mathbb{R}\text{)}}} \\
&\geq \underbrace{\sum_{i=1}^{k} (y_i - t)}_{= \sum_{i=1}^{k} y_i - kt} + \underbrace{\sum_{i=k+1}^{n} (t - y_i)}_{= (n-k)t - \sum_{i=k+1}^{n} y_i} = \sum_{i=1}^{k} y_i - kt + (n-k)\,t - \underbrace{\sum_{i=k+1}^{n} y_i}_{= \sum_{i=1}^{n} y_i - \sum_{i=1}^{k} y_i} \\
&= \sum_{i=1}^{k} y_i - kt + (n-k)\,t - \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{k} y_i \right) \\
&= 2\sum_{i=1}^{k} y_i + (n-2k)\,t - \sum_{i=1}^{n} y_i.
\end{aligned} \tag{106}$$

[48] Here is how to find such a t: If $k > 0$, then we pick $t = x_k$; otherwise, we pick $t = x_{k+1} = x_1$.

Now, (105) yields

$$\begin{aligned}
&2\sum_{i=1}^{k} x_i + (n-2k)\,t - \sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} |x_i - t| \geq \sum_{i=1}^{n} |y_i - t| \qquad \text{(by (102))} \\
&\geq 2\sum_{i=1}^{k} y_i + (n-2k)\,t - \sum_{i=1}^{n} y_i \qquad \text{(by (106))}\,.
\end{aligned}$$

Subtracting $(n-2k)\,t$ from both sides of this inequality, we obtain

$$2\sum_{i=1}^{k} x_i - \sum_{i=1}^{n} x_i \geq 2\sum_{i=1}^{k} y_i - \sum_{i=1}^{n} y_i. \tag{107}$$

Forget that we fixed k. We thus have proved the inequality (107) for all $k \in \{0,1,\ldots,n\}$.

Applying (107) to $k = 0$, we obtain

$$-\sum_{i=1}^{n} x_i \geq -\sum_{i=1}^{n} y_i \tag{108}$$

(since both $\sum_{i=1}^{k}$ sums are empty for $k = 0$). In other words,

$$\sum_{i=1}^{n} x_i \leq \sum_{i=1}^{n} y_i. \tag{109}$$

On the other hand, applying (107) to $k = n$, we obtain

$$\sum_{i=1}^{n} x_i \geq \sum_{i=1}^{n} y_i \tag{110}$$

(since the left hand side simplifies to $2\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i$, and likewise for the right hand side). Combining this inequality with (109), we obtain

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \tag{111}$$

Now, for each $k \in [n]$, we have

$$2\sum_{i=1}^{k} x_i \geq 2\sum_{i=1}^{k} y_i$$

(by adding the inequalities (107) and (110) together) and therefore

$$\sum_{i=1}^{k} x_i \geq \sum_{i=1}^{k} y_i$$

(by cancelling the factor 2 from both sides of the previous inequality). This fact, combined with (111), shows that $x \succcurlyeq y$ (since x and y are weakly decreasing). This proves the "$\Longleftarrow$" direction of Theorem 4.7.20. □

Exercise 4.7.8. [2] Let $x, y \in \mathbb{R}^n$ be two vectors such that $x \succcurlyeq y$. Let $u, v \in \mathbb{R}^m$ be two further vectors such that $u \succcurlyeq v$. Prove that $\begin{pmatrix} x \\ u \end{pmatrix} \succcurlyeq \begin{pmatrix} y \\ v \end{pmatrix}$. (We are using block-matrix notation here, so that $\begin{pmatrix} x \\ u \end{pmatrix}$ means the vector obtained by stacking x on top of u.)

Exercise 4.7.9. [3] Prove Theorem 4.7.14.

We note that the analogue of Exercise 4.7.6 for arbitrary (not necessarily weakly decreasing) column vectors is false: It is not hard to find two vectors $x, y \in \mathbb{R}^4$ such that $x \succcurlyeq y$ but it takes 4 (not 3) RH moves to transform x into y. (For example, $x = (5,3,2,0)^T$ and $y = (1,1,4,4)^T$ are two such vectors.) On the other hand, it is easy to see (piggybacking on Exercise 4.7.6) that for any two column vectors $x, y \in \mathbb{R}^n$ satisfying $x \succcurlyeq y$, it is possible to obtain y from x by at most $2n-2$ RH moves. Finding the minimum number of RH moves that always suffices to transform $x \in \mathbb{R}^n$ into $y \in \mathbb{R}^n$ when $x \succcurlyeq y$ appears to be an interesting question.

We can use Theorem 4.7.20 to define a "weighted" generalization of majorization. This leads to the following generalization of Theorem 4.7.19:

Theorem 4.7.21 (weighted Karamata's inequality)**.** Let $I \subseteq \mathbb{R}$ be an interval. Let $f : I \to \mathbb{R}$ be a convex function. Let $w_1, w_2, \ldots, w_n$ be n nonnegative reals. Let $x \in I^n$ and $y \in I^n$ be two vectors such that all $t \in \mathbb{R}$ satisfy

$$w_1 |x_1 - t| + w_2 |x_2 - t| + \cdots + w_n |x_n - t| \geq w_1 |y_1 - t| + w_2 |y_2 - t| + \cdots + w_n |y_n - t| .$$

(Note that this is a "weighted" version of the condition $x \succcurlyeq y$.) Then,

$$w_1 f(x_1) + w_2 f(x_2) + \cdots + w_n f(x_n) \geq w_1 f(y_1) + w_2 f(y_2) + \cdots + w_n f(y_n) .$$

Exercise 4.7.10. [7] Prove Theorem 4.7.21.

[Hint: Let S be a finite subset of I. For each $s \in S$, let $f_s : I \to \mathbb{R}$ be the piecewise-linear function that sends each $z \in I$ to $|s - z|$. Show that the convex function f can be interpolated on S by a linear combination $\sum_{s \in S} \alpha_s f_s$ of the

functions f_s with nonnegative coefficients α_s; that is, show that there exists a nonnegative real α_s for each $s \in S$ such that

$$f(z) = \sum_{s \in S} \alpha_s |s - z| \qquad \text{for each } z \in S.$$

Then, apply this to $S = \{x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n\}$.]

Exercise 4.7.11. [7] We define a new binary relation $\succcurlyeq'$ on the set $\mathbb{R}^n$ as follows: For two column vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we write $x \succcurlyeq' y$ (and say that x *weakly majorizes* y) if and only if we have

$$\sum_{i=1}^{m} x_i^{\downarrow} \geq \sum_{i=1}^{m} y_i^{\downarrow} \qquad \text{for each } m \in [n]$$

(but we do not require $\sum_{i=1}^{n} x_i^{\downarrow} = \sum_{i=1}^{n} y_i^{\downarrow}$).

Let x and y be two weakly decreasing column vectors in $\mathbb{R}^n$. Let I be an interval of $\mathbb{R}$ that contains all entries of x and of y. Prove that the following statements are equivalent:

- $\mathcal{A}$: We have $x \succcurlyeq' y$.
- $\mathcal{B}$: For any sufficiently low $\alpha \in \mathbb{R}$, we have $\begin{pmatrix} x \\ \alpha - \sum x \end{pmatrix} \succcurlyeq \begin{pmatrix} y \\ \alpha - \sum y \end{pmatrix}$, where $\sum x := \sum_{i=1}^{n} x_i$ and $\sum y := \sum_{i=1}^{n} y_i$ (and where we are using block-matrix notation, so that $\begin{pmatrix} x \\ \alpha - \sum x \end{pmatrix}$ denotes the result of appending a new entry $\alpha - \sum x$ to the bottom of the column vector x).
- $\mathcal{C}$: We can obtain y from x by a sequence of OPRH moves and OPD moves. Here, an "*OPD move*" (short for "order-preserving decrease move") means a move in which we decrease an entry of a weakly decreasing vector in such a way that the vector remains weakly decreasing (i.e., we replace an entry z_i of a decreasing vector $z \in \mathbb{R}^n$ by a smaller entry $z_i' \leq z_i$ such that we still have $z_1 \geq z_2 \geq \cdots \geq z_{i-1} \geq z_i' \geq z_{i+1} \geq z_{i+2} \geq \cdots \geq z_n$).
- $\mathcal{D}$: Every weakly increasing convex function $f : I \to \mathbb{R}$ satisfies
$$f(x_1) + f(x_2) + \cdots + f(x_n) \geq f(y_1) + f(y_2) + \cdots + f(y_n).$$
- $\mathcal{E}$: All $t \in \mathbb{R}$ satisfy
$$(x_1 - t)_+ + (x_2 - t)_+ + \cdots + (x_n - t)_+ \geq (y_1 - t)_+ + (y_2 - t)_+ + \cdots + (y_n - t)_+.$$
Here, the notation z_+ means the positive part of a real number z (that is, we have $z_+ = z$ when $z \geq 0$, and $z_+ = 0$ otherwise).

4.7.5. Doubly stochastic matrices

Majorizing pairs of vectors are closely related to *doubly stochastic matrices*:

Definition 4.7.22. A matrix $S \in \mathbb{R}^{n \times n}$ is said to be *doubly stochastic* if its entries $S_{i,j}$ satisfy the following three conditions:

1. We have $S_{i,j} \geq 0$ for all i, j.
2. We have $\sum_{j=1}^{n} S_{i,j} = 1$ for each $i \in [n]$.
3. We have $\sum_{i=1}^{n} S_{i,j} = 1$ for each $j \in [n]$.

In other words, a doubly stochastic matrix is an $n \times n$-matrix whose entries are nonnegative reals and whose rows and columns have sum 1 each.

Exercise 4.7.12. [2] Show that even if we allow S to be rectangular in Definition 4.7.22, the conditions 2 and 3 still force S to be a square matrix.

Example 4.7.23. (a) The matrix $\begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{5}{12} & \frac{7}{12} \end{pmatrix}$ is doubly stochastic.

(b) Each doubly stochastic 2×2-matrix has the form

$$\begin{pmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{pmatrix} \qquad \text{for some } \lambda \in [0,1].$$

(c) Any permutation matrix is doubly stochastic.

Proposition 4.7.24. Let $S \in \mathbb{R}^{n \times n}$ be a matrix whose entries are nonnegative reals. Let $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$. Then, S is doubly stochastic if and only if $Se = e$ and $e^T S = e^T$.

Corollary 4.7.25. Any product of doubly stochastic matrices is again doubly stochastic.

Exercise 4.7.13. [3] Prove Proposition 4.7.24 and Corollary 4.7.25.

Now, we can connect doubly stochastic matrices with majorization:

Theorem 4.7.26. Let $x, y \in \mathbb{R}^n$ be two vectors. Then, $x \succcurlyeq y$ if and only if there exists a doubly stochastic matrix $S \in \mathbb{R}^{n \times n}$ such that $y = Sx$.

Proof. $\Longrightarrow$: Assume that $x \succcurlyeq y$. We must prove that there exists a doubly stochastic matrix $S \in \mathbb{R}^{n \times n}$ such that $y = Sx$.

By Example 4.7.23 **(c)** and Corollary 4.7.25, it suffices to show this in the case when x and y are weakly decreasing (because any permutation of the entries of a vector can be effected by multiplying this vector with a permutation matrix).

Thus, we WLOG assume that x and y are weakly decreasing. We must prove that there exists a doubly stochastic matrix $S \in \mathbb{R}^{n \times n}$ such that $y = Sx$.

By Corollary 4.7.25, it suffices to show this in the case when $x \overset{\text{OPRH}}{\longrightarrow} y$ (because in the general case, y is obtained from x by a sequence of OPRH moves[49]).

So let us WLOG assume that $x \overset{\text{OPRH}}{\longrightarrow} y$. Thus, y is obtained from x by picking two entries x_i and x_j with $x_i \leq x_j$ and replacing them by u and v, where $u, v \in \left[x_i, x_j\right]$ with $u + v = x_i + x_j$. Consider these x_i, x_j, u, v.

From $u \in \left[x_i, x_j\right]$, we obtain

$$u = \lambda x_i + (1 - \lambda)\, x_j \qquad \text{for some } \lambda \in [0, 1]$$

(namely, $\lambda = \dfrac{u - x_j}{x_i - x_j}$). Consider this λ. Then,

$$v = (1 - \lambda)\, x_i + \lambda x_j \qquad \left(\text{since } u + v = x_i + x_j\right).$$

This entails that $y = Sx$, where $S \in \mathbb{R}^{n \times n}$ is the matrix defined by

$$\begin{aligned} &S_{i,i} = \lambda, \qquad S_{i,j} = 1 - \lambda, \qquad S_{j,i} = 1 - \lambda, \qquad S_{j,j} = \lambda, \\ &S_{k,k} = 1 \qquad \text{for each } k \notin \{i, j\}, \\ &S_{k,\ell} = 0 \qquad \text{for all remaining } k, \ell. \end{aligned}$$

For example, if $i = 2$ and $j = 4$, then

$$S = \begin{pmatrix} 1 & & & & \\ & \lambda & & 1 - \lambda & \\ & & 1 & & \\ & 1 - \lambda & & \lambda & \\ & & & & \end{pmatrix}$$

(where all empty cells are filled with zeroes). In this case,

$$Sx = S \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ \lambda x_2 + (1 - \lambda)\, x_4 \\ x_3 \\ (1 - \lambda)\, x_2 + \lambda x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ u \\ x_3 \\ v \end{pmatrix} = y.$$

[49] This is a consequence of Theorem 4.7.12.

So we have constructed a matrix $S \in \mathbb{R}^{n \times n}$ that satisfies $y = Sx$, and it is easy to see that this S is doubly stochastic. Thus, we have proved the "$\Longrightarrow$" direction of Theorem 4.7.26.

$\Longleftarrow$: Assume that $y = Sx$ for some doubly stochastic matrix $S \in \mathbb{R}^{n \times n}$. Then, for every $i \in [n]$, we have

$$y_i = (Sx)_i = \sum_{j=1}^{n} S_{i,j} x_j. \tag{112}$$

Hence, for every $i \in [n]$ and $t \in \mathbb{R}$, we have

$$\begin{aligned} y_i - t &= \sum_{j=1}^{n} S_{i,j} x_j - t = \sum_{j=1}^{n} S_{i,j} x_j - \sum_{j=1}^{n} S_{i,j} t \\ &\qquad \left(\begin{array}{c} \text{since condition 2 in Definition 4.7.22} \\ \text{yields } \sum\limits_{j=1}^{n} S_{i,j} = 1, \text{ so that } \sum\limits_{j=1}^{n} S_{i,j} t = \underbrace{\left(\sum\limits_{j=1}^{n} S_{i,j} \right)}_{=1} t = t \\ \text{and therefore } t = \sum\limits_{j=1}^{n} S_{i,j} t \end{array} \right) \\ &= \sum_{j=1}^{n} S_{i,j} \left(x_j - t \right) \end{aligned}$$

and therefore

$$\begin{aligned} |y_i - t| &= \left| \sum_{j=1}^{n} S_{i,j} \left(x_j - t \right) \right| \\ &\leq \sum_{j=1}^{n} \underbrace{\left| S_{i,j} \left(x_j - t \right) \right|}_{\substack{= S_{i,j} \cdot \left| x_j - t \right| \\ \text{(since } S_{i,j} \geq 0 \text{ (by condition 1 in Definition 4.7.22))}}} \\ &\qquad \left(\begin{array}{c} \text{by the triangle inequality, which says that } \left| \sum\limits_{j=1}^{n} \alpha_j \right| \leq \sum\limits_{j=1}^{n} \left| \alpha_j \right| \\ \text{for any } n \text{ reals } \alpha_1, \alpha_2, \ldots, \alpha_n \end{array} \right) \\ &= \sum_{j=1}^{n} S_{i,j} \cdot \left| x_j - t \right|. \end{aligned} \tag{113}$$

Thus, for every $t \in \mathbb{R}$, we have

$$\begin{aligned}
&|y_1 - t| + |y_2 - t| + \cdots + |y_n - t| \\
&= \sum_{i=1}^{n} |y_i - t| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} S_{i,j} \cdot |x_j - t| \qquad \text{(by (113))} \\
&= \sum_{j=1}^{n} \underbrace{\left(\sum_{i=1}^{n} S_{i,j} \right)}_{\substack{=1 \\ \text{(by condition 3} \\ \text{in Definition 4.7.22)}}} \cdot |x_j - t| = \sum_{j=1}^{n} |x_j - t| \\
&= |x_1 - t| + |x_2 - t| + \cdots + |x_n - t| .
\end{aligned}$$

In other words, for every $t \in \mathbb{R}$, we have

$$|x_1 - t| + |x_2 - t| + \cdots + |x_n - t| \geq |y_1 - t| + |y_2 - t| + \cdots + |y_n - t| .$$

Therefore, by Theorem 4.7.20, we have $x \succcurlyeq y$. This proves the "$\Longleftarrow$" direction of Theorem 4.7.26. □

5. Singular value decomposition ([HorJoh13, §2.6])

This will be just a brief introduction to singular value decomposition. For much more, see [TreBau97].

5.1. Some properties of A^*A

We first state some basic properties of matrices of the form A^*A:

Proposition 5.1.1 (the Ker (A^*A) lemma)**.** Let $A \in \mathbb{C}^{m \times n}$ be any $m \times n$-matrix with complex entries (not necessarily a square matrix). Then:

(a) The matrix A^*A is Hermitian and positive semidefinite.

(b) We have $\operatorname{Ker} A = \operatorname{Ker} (A^*A)$.

(c) We have $\operatorname{rank} A = \operatorname{rank} (A^*A)$.

Proof. **(a)** The matrix A^*A is Hermitian, since $(A^*A)^* = A^* \underbrace{(A^*)^*}_{=A} = A^*A$. Moreover, this matrix A^*A is positive semidefinite, since each vector $x \in \mathbb{C}^n$ satisfies

$$\langle A^*Ax,\ x \rangle = \underbrace{x^*A^*}_{=(Ax)^*} Ax = (Ax)^* Ax = ||Ax||^2 \geq 0.$$

Thus, Proposition 5.1.1 **(a)** is proven.

(b) Each $y \in \operatorname{Ker} A$ satisfies $y \in \operatorname{Ker}\left(A^*A\right)$ (because $y \in \operatorname{Ker} A$ entails $Ay = 0$, so that $A^* \underbrace{Ay}_{=0} = 0$ and thus $y \in \operatorname{Ker}\left(A^*A\right)$). In other words, $\operatorname{Ker} A \subseteq \operatorname{Ker}\left(A^*A\right)$.

Let us now show that $\operatorname{Ker}\left(A^*A\right) \subseteq \operatorname{Ker} A$. Indeed, let $x \in \operatorname{Ker}\left(A^*A\right)$. Thus, $x \in \mathbb{C}^n$ and $A^*Ax = 0$. Hence,

$$\begin{aligned} ||Ax||^2 &= \underbrace{(Ax)^*}_{=x^*A^*} Ax \qquad \left(\text{since } ||v||^2 = \langle v, v\rangle = v^*v \text{ for any } v \in \mathbb{C}^n\right) \\ &= x^* \underbrace{A^*Ax}_{=0} = 0. \end{aligned}$$

In other words, $||Ax|| = 0$. Hence, $Ax = 0$ (since a vector whose length is 0 must itself be 0). In other words, $x \in \operatorname{Ker} A$.

Forget that we fixed x. We thus have shown that $x \in \operatorname{Ker} A$ for each $x \in \operatorname{Ker}\left(A^*A\right)$. In other words, $\operatorname{Ker}\left(A^*A\right) \subseteq \operatorname{Ker} A$. Combining this with $\operatorname{Ker} A \subseteq \operatorname{Ker}\left(A^*A\right)$, we obtain $\operatorname{Ker} A = \operatorname{Ker}\left(A^*A\right)$. This proves Proposition 5.1.1 **(b)**.

(c) The rank-nullity theorem yields

$$\begin{aligned} \operatorname{rank} A &= n - \dim\left(\operatorname{Ker} A\right) \qquad \text{and} \\ \operatorname{rank}\left(A^*A\right) &= n - \dim\left(\operatorname{Ker}\left(A^*A\right)\right). \end{aligned}$$

The right hand sides of these two equalities are equal (since part **(b)** yields $\operatorname{Ker} A = \operatorname{Ker}\left(A^*A\right)$). Thus, the left hand sides are also equal. In other words, $\operatorname{rank} A = \operatorname{rank}\left(A^*A\right)$. This proves Proposition 5.1.1 **(c)**. □

Note that Proposition 5.1.1 **(b)** really requires a matrix with complex entries. It cannot be generalized to matrices over an arbitrary field.

5.2. The singular value decomposition

Definition 5.2.1. Let A and B be two $m \times n$-matrices with complex entries.

We say that A and B are *unitarily equivalent* if there exist unitary matrices $U \in \mathrm{U}_m\left(\mathbb{C}\right)$ and $V \in \mathrm{U}_n\left(\mathbb{C}\right)$ such that $A = UBV^*$.

Note that we could just as well require $A = UBV$ instead of $A = UBV^*$ here, since V is unitary if and only if V^* is unitary.

Note the difference between "unitarily equivalent" and "unitarily similar": The latter requires $A = UBU^*$, whereas the former only requires $A = UBV^*$.

Unitary equivalence is an equivalence relation.

Exercise 5.2.1. [2] Prove this!

A natural question is therefore: What is the "simplest" matrix in the equivalence class of a given matrix? The answer is pretty nice: Each matrix is unitarily equivalent to a "more or less diagonal" matrix. We are saying "more or less" because diagonal matrices are supposed to be square, but our matrices can have any dimensions; thus, we introduce a separate word for rectangular matrices that "would be diagonal if they were square":

Definition 5.2.2. Let $\mathbb{F}$ be a field. A rectangular matrix $A \in \mathbb{F}^{n\times m}$ is said to be *pseudodiagonal* if it satisfies

$$A_{i,j} = 0 \qquad \text{whenever } i \neq j.$$

This is just the straightforward generalization of diagonal matrices to non-square matrices. In particular, a square matrix is pseudodiagonal if and only if it is diagonal. A pseudodiagonal 2×3-matrix looks like this: $\begin{pmatrix} * & 0 & 0 \\ 0 & * & 0 \end{pmatrix}$. A pseudodiagonal 3×2-matrix looks like this: $\begin{pmatrix} * & 0 \\ 0 & * \\ 0 & 0 \end{pmatrix}$. (Of course, any of the $*$s can be a 0 too.)

Theorem 5.2.3 (SVD)**.** Let $A \in \mathbb{C}^{m\times n}$. Then:

(a) There exist unitary matrices $U \in \mathrm{U}_m(\mathbb{C})$ and $V \in \mathrm{U}_n(\mathbb{C})$ and a pseudodiagonal matrix $\Sigma \in \mathbb{C}^{m\times n}$ such that all diagonal entries of Σ are nonnegative reals and such that

$$A = U\Sigma V^*.$$

In other words, A is unitarily equivalent to a pseudodiagonal matrix whose diagonal entries are nonnegative reals.

(b) The matrix Σ is unique up to permutation of its diagonal entries. (The matrices U and V are usually not unique.)

(c) Let $k = \operatorname{rank} A$. Then, the matrix Σ has exactly k nonzero diagonal entries.

(d) Let $\sigma_1, \sigma_2, \ldots, \sigma_n$ be the square roots of the n eigenvalues of the Hermitian matrix A^*A, listed in decreasing order (so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$). Then, we have $\sigma_{k+1} = \sigma_{k+2} = \cdots = \sigma_n = 0$, and we can take

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{C}^{m\times n}$$

in part **(a)**.

Definition 5.2.4. The triple (U, V, Σ) in Theorem 5.2.3 **(a)** is called a *singular value decomposition* (short: *SVD*) of A. The numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ are called the *singular values* of A.

Before we prove the theorem, a few words are to be said about the use of an SVD. In practice, you often want to find a low-rank "approximation" for a given matrix A: that is, a matrix B that is "sufficiently close" to A and yet has low rank. One of the best ways to do this is by computing an SVD of A – that is, writing A in the form $A = U\Sigma V^*$ with U, Σ, V as in Theorem 5.2.3 **(a)** – and then setting all but the first few σ_i's to 0 in Σ. We will soon see why this "approximates" A.

Let us now prove the theorem.

Proof of Theorem 5.2.3. Let $k := \operatorname{rank} A$. Then, k is the rank of an $m \times n$-matrix (namely, A), and thus satisfies $k \leq m$ and $k \leq n$.

Proposition 5.1.1 **(a)** shows that the matrix A^*A is Hermitian and positive semidefinite. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of this matrix A^*A, listed in decreasing order. These eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are nonnegative reals (since A^*A is positive semidefinite). Thus, we can set

$$\sigma_i := \sqrt{\lambda_i} \qquad \text{for each } i \in [n].$$

Then, $\sigma_1, \sigma_2, \ldots, \sigma_n$ are nonnegative reals. Moreover, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ (since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are listed in decreasing order) and therefore $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ (since $\sigma_i = \sqrt{\lambda_i}$ for each i).

The matrix A^*A is Hermitian and thus normal. Hence, Corollary 2.6.2 (applied to A^*A instead of A) yields that there exists a spectral decomposition (V, D) of A^*A with $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Consider this spectral decomposition. Thus, $V \in \mathrm{U}_n(\mathbb{C})$ is a unitary matrix, and $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and we have

$$A^*A = VDV^*. \tag{114}$$

From (114), we obtain $A^*A \overset{\mathrm{us}}{\sim} D$ (since V is unitary). Now,

$$\begin{aligned} k &= \operatorname{rank} A = \operatorname{rank}(A^*A) && \text{(by Proposition 5.1.1 \textbf{(c)})} \\ &= \operatorname{rank}(\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)) && \left(\text{since } A^*A \overset{\mathrm{us}}{\sim} D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\right) \\ &= (\text{the number of } i \in [n] \text{ such that } \lambda_i \neq 0) \\ &= (\text{the number of } i \in [n] \text{ such that } \sigma_i \neq 0) \end{aligned}$$

(since $\sigma_i = \sqrt{\lambda_i}$ for each i). Hence, exactly k of the numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ are nonzero. Since $\sigma_1, \sigma_2, \ldots, \sigma_n$ are nonnegative reals and satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, this entails that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0 \qquad \text{and} \tag{115}$$

$$\sigma_{k+1} = \sigma_{k+2} = \cdots = \sigma_n = 0. \tag{116}$$

Let $v_1, v_2, \ldots, v_n$ be the columns of the unitary matrix V. Thus, $(v_1, v_2, \ldots, v_n)$ is an orthonormal basis of $\mathbb{C}^n$ (since V is unitary)[50]. Since V is the unitary matrix in a spectral decomposition of A^*A, we can easily see that the columns of V are eigenvectors of A^*A corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. In other words,

$$A^*Av_i = \lambda_i v_i \qquad \text{for each } i \in [n]. \tag{117}$$

[*Proof of (117):* Let $i \in [n]$. Then, we have[51] $V_{\bullet,i} = v_i$ (since $v_1, v_2, \ldots, v_n$ are the columns of V). However, from (114), we obtain $A^*AV = VD \qquad \underbrace{V^*V}_{\substack{=I_n \\ \text{(since } V \text{ is unitary)}}} = VD$.

Hence,

$$\begin{aligned}
(A^*AV)_{\bullet,i} &= (VD)_{\bullet,i} \\
&= (V \cdot \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n))_{\bullet,i} \qquad (\text{since } D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)) \\
&= \lambda_i \underbrace{V_{\bullet,i}}_{=v_i} \qquad \left(\begin{array}{c} \text{since multiplication by the diagonal} \\ \text{matrix } \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \text{ on the right} \\ \text{scales the } i\text{-th column of a matrix by } \lambda_i \end{array} \right) \\
&= \lambda_i v_i.
\end{aligned}$$

Comparing this with

$$\begin{aligned}
(A^*AV)_{\bullet,i} &= A^*A \underbrace{V_{\bullet,i}}_{=v_i} \qquad (\text{by the rules for matrix multiplication}) \\
&= A^*Av_i,
\end{aligned}$$

we obtain $A^*Av_i = \lambda_i v_i$. This proves (117).]

For each $j \in [k]$, we set

$$u_j := \frac{1}{\sigma_j} Av_j.$$

(The division by σ_j in this definition is legitimate, since (115) reveals that $\sigma_j \neq 0$.) We claim that the tuple $(u_1, u_2, \ldots, u_k)$ is orthonormal. Indeed:

[50] Here we are using the implication $\mathcal{A} \Longrightarrow \mathcal{E}$ of Theorem 1.5.3.

[51] Recall that the notation $M_{\bullet,i}$ denotes the i-th column of a matrix M.

- For any two distinct elements i and j of $[k]$, we have

$$\begin{aligned}\left\langle u_i, u_j \right\rangle &= \left\langle \frac{1}{\sigma_i} Av_i,\ \frac{1}{\sigma_j} Av_j \right\rangle \qquad \left(\text{by the definitions of } u_i \text{ and } u_j\right) \\ &= \frac{1}{\sigma_i \overline{\sigma_j}} \underbrace{\left\langle Av_i,\ Av_j \right\rangle}_{=\left(Av_j\right)^* Av_i} = \frac{1}{\sigma_i \overline{\sigma_j}} \underbrace{\left(Av_j\right)^*}_{=v_j^* A^*} Av_i \\ &= \frac{1}{\sigma_i \overline{\sigma_j}} v_j^* \underbrace{A^* A v_i}_{\substack{=\lambda_i v_i \\ \text{(by (117))}}} = \frac{1}{\sigma_i \overline{\sigma_j}} v_j^* \lambda_i v_i = \frac{\lambda_i}{\sigma_i \overline{\sigma_j}} \underbrace{v_j^* v_i}_{\substack{=\left\langle v_i, v_j \right\rangle \\ =0 \\ \text{(since } (v_1, v_2, \ldots, v_n) \\ \text{is orthonormal)}}} = 0\end{aligned}$$

  and thus $u_i \perp u_j$. Therefore, the tuple $(u_1, u_2, \ldots, u_k)$ is orthogonal.

- For any $i \in [k]$, we have

$$\begin{aligned}\left\langle u_i, u_i \right\rangle &= \left\langle \frac{1}{\sigma_i} Av_i,\ \frac{1}{\sigma_i} Av_i \right\rangle \qquad \left(\text{by the definition of } u_i\right) \\ &= \underbrace{\frac{1}{\sigma_i \overline{\sigma_i}}}_{\substack{=\frac{1}{\lambda_i} \\ \text{(since } \sigma_i \in \mathbb{R} \text{ and} \\ \text{thus } \sigma_i \overline{\sigma_i} = \sigma_i \sigma_i = \sigma_i^2 = \lambda_i \\ \text{(because } \sigma_i = \sqrt{\lambda_i}\text{))}}} \underbrace{\left\langle Av_i,\ Av_i \right\rangle}_{=(Av_i)^* Av_i} = \frac{1}{\lambda_i} \underbrace{\left(Av_i\right)^*}_{=v_i^* A^*} Av_i \\ &= \frac{1}{\lambda_i} v_i^* \underbrace{A^* A v_i}_{\substack{=\lambda_i v_i \\ \text{(by (117))}}} = \frac{1}{\lambda_i} v_i^* \lambda_i v_i = v_i^* v_i = \left\langle v_i, v_i \right\rangle = ||v_i||^2 = 1\end{aligned}$$

  (since $(v_1, v_2, \ldots, v_n)$ is orthonormal and thus $||v_i|| = 1$), and thus $||u_i|| = 1$. Hence, the orthogonal tuple $(u_1, u_2, \ldots, u_k)$ is orthonormal.

So $(u_1, u_2, \ldots, u_k)$ is an orthonormal tuple of vectors in $\mathbb{C}^m$. Hence, Corollary 1.2.9 shows that we can extend this tuple to an orthonormal basis $(u_1, u_2, \ldots, u_m)$ of $\mathbb{C}^m$ by appending $m - k$ new (appropriately chosen) vectors $u_{k+1}, u_{k+2}, \ldots, u_m$. Let us consider the orthonormal basis $(u_1, u_2, \ldots, u_m)$ of $\mathbb{C}^m$ obtained in this way. Let $U \in \mathbb{C}^{m \times m}$ be the $m \times m$-matrix with columns $u_1, u_2, \ldots, u_m$. Thus, the columns of this matrix U form an orthonormal basis of $\mathbb{C}^m$; hence, U is a unitary matrix (by the implication $\mathcal{E} \Longrightarrow \mathcal{A}$ of Theorem 1.5.3).

Let

$$\Sigma := \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{C}^{m \times n}.$$

(This is the $m \times n$-matrix whose (i,i)-th entries are σ_i for all $i \in [k]$, and whose all remaining entries are 0.) Clearly, this matrix Σ is pseudodiagonal. Moreover, this matrix Σ has exactly k nonzero diagonal entries (since (115) shows that $\sigma_1, \sigma_2, \ldots, \sigma_k$ are nonzero). All its diagonal entries are nonnegative reals.

Now, we claim that $A = U\Sigma V^*$. To prove this, we shall first show that $AV = U\Sigma$. It is sufficient to prove that[52]

$$(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j} \qquad \text{for each } j \in [n].$$

So let us fix $j \in [n]$ and try to prove that $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$. We note that $V_{\bullet,j} = v_j$ (since the columns of V are $v_1, v_2, \ldots, v_n$) and $U_{\bullet,j} = u_j$ (since the columns of U are $u_1, u_2, \ldots, u_m$). We distinguish between the cases $j \leq k$ and $j > k$:

- Assume that $j \leq k$. Then, by the rules for multiplying matrices, we have

$$\begin{aligned} (AV)_{\bullet,j} &= A \underbrace{V_{\bullet,j}}_{=v_j} = Av_j = \sigma_j \underbrace{u_j}_{=U_{\bullet,j}} \qquad \left(\text{since } u_j = \frac{1}{\sigma_j} Av_j \right) \\ &= \sigma_j U_{\bullet,j}. \end{aligned} \tag{118}$$

  On the other hand, $j \leq k$ shows that the j-th column of the matrix Σ has a σ_j in its j-th position and zeroes in all other positions. In other words, this column equals $\sigma_j e_j$ (where $(e_1, e_2, \ldots, e_m)$ denotes the standard basis of $\mathbb{C}^m$). In other words, $\Sigma_{\bullet,j} = \sigma_j e_j$ (since $\Sigma_{\bullet,j}$ is the j-th column of Σ). Now, by the rules for multiplying matrices, we have

$$(U\Sigma)_{\bullet,j} = U \underbrace{\Sigma_{\bullet,j}}_{=\sigma_j e_j} = U \cdot \sigma_j e_j = \sigma_j \underbrace{Ue_j}_{=U_{\bullet,j}} = \sigma_j U_{\bullet,j}.$$

  (since multiplying a matrix by e_j always produces the j-th column of the matrix)

  Comparing this with (118), we obtain $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$.

  Thus, $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$ is proved in the case when $j \leq k$.

[52] Recall that the notation $M_{\bullet,j}$ means the j-th column of a matrix M.

- Now, assume that $j > k$. Then, $\sigma_j = 0$ (by (116)), so that $\lambda_j = 0$ (since the definition of σ_j yields $\sigma_j = \sqrt{\lambda_j}$, so that $\lambda_j = \sigma_j^2$). However, applying (117) to $i = j$, we obtain

$$A^*Av_j = \underbrace{\lambda_j}_{\substack{=0 \\ (\text{since } j>k)}} v_j = 0,$$

so that $v_j \in \operatorname{Ker}(A^*A) = \operatorname{Ker} A$ (by Proposition 5.1.1 **(b)**). Now, by the rules for multiplying matrices, we have

$$(AV)_{\bullet,j} = A \underbrace{V_{\bullet,j}}_{=v_j} = Av_j = 0 \qquad \left(\text{since } v_j \in \operatorname{Ker} A\right).$$

Comparing this with

$$(U\Sigma)_{\bullet,j} = U \underbrace{\Sigma_{\bullet,j}}_{\substack{=0 \\ \text{(since } j>k\text{, so that all entries in} \\ \text{the } j\text{-th column of } \Sigma \text{ are 0)}}} = 0,$$

we obtain $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$.

Thus, $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$ is proved in the case when $j > k$.

Thus, we have proved $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$ in both cases.

Forget that we fixed j. We thus have shown that $(AV)_{\bullet,j} = (U\Sigma)_{\bullet,j}$ for each $j \in [n]$. In other words, each column of the matrix AV equals the corresponding column of the matrix $U\Sigma$. Hence, $AV = U\Sigma$. Therefore,

$$\underbrace{U\Sigma}_{=AV} V^* = A \underbrace{VV^*}_{\substack{=I_n \\ \text{(since } V \text{ is unitary)}}} = A,$$

so that $A = U\Sigma V^*$, as desired.

This proves parts **(a)**, **(c)** and **(d)** of Theorem 5.2.3.

(b) We must show that if P is a pseudodiagonal matrix such that all diagonal entries of P are nonnegative reals, and such that A is unitarily equivalent to P, then P and Σ have the same diagonal entries up to order.

Before we prove this, let us show two auxiliary results:

> *Claim 1:* Let $X \in \mathbb{C}^{m\times n}$ and $Y \in \mathbb{C}^{m\times n}$ be two unitarily equivalent matrices. Then, X and Y have the same singular values[53].

[53] Recall that the *singular values* of a matrix X are defined to be the square roots of the eigenvalues of X^*X.

[*Proof of Claim 1:* Since X and Y are unitarily equivalent, there exist two unitary matrices $U \in \mathrm{U}_m(\mathbb{C})$ and $V \in \mathrm{U}_n(\mathbb{C})$ such that $X = UYV^*$. (These U and V have nothing to do with the U and V from Theorem 5.2.3.) Consider these U and V. From $X = UYV^*$, we obtain

$$X^*X = \underbrace{(UYV^*)^*}_{=(V^*)^*Y^*U^*}(UYV^*) = \underbrace{(V^*)^*}_{=V}Y^*\underbrace{U^*U}_{\substack{=I_m \\ \text{(since } U \text{ is unitary)}}}YV^* = VY^*YV^*.$$

This shows that $X^*X \overset{\text{us}}{\sim} Y^*Y$ (since V is unitary). Therefore, $X^*X \sim Y^*Y$ (by Proposition 2.2.5). Thus, the matrices X^*X and Y^*Y have the same eigenvalues (by Proposition 2.1.5 **(e)**). Therefore, X and Y have the same singular values (because the singular values of X are defined as the square roots of the eigenvalues of X^*X, and likewise for Y). This proves Claim 1.]

> *Claim 2:* Let $D \in \mathbb{C}^{m\times n}$ be a pseudodiagonal matrix. Then, the nonzero singular values of D are the absolute values of the nonzero diagonal entries of D.

[*Proof of Claim 2:* We WLOG assume that $m \leq n$, since the case $m > n$ is similar but easier. Let $d_1, d_2, \ldots, d_m$ be the diagonal entries of D. Then,

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m & 0 & \cdots & 0 \end{pmatrix}.$$

Hence, it is easy to check that

$$\begin{aligned} D^*D &= \begin{pmatrix} \overline{d_1}d_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \overline{d_2}d_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{d_m}d_m & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \\ &= \operatorname{diag}\left(\overline{d_1}d_1, \overline{d_2}d_2, \ldots, \overline{d_m}d_m, \underbrace{0, 0, \ldots, 0}_{n-m \text{ entries}}\right) \\ &= \operatorname{diag}\left(|d_1|^2, |d_2|^2, \ldots, |d_m|^2, \underbrace{0, 0, \ldots, 0}_{n-m \text{ entries}}\right). \end{aligned}$$

Therefore, the eigenvalues of D^*D are $|d_1|^2, |d_2|^2, \ldots, |d_m|^2, \underbrace{0, 0, \ldots, 0}_{n-m \text{ entries}}$ (since the eigenvalues of a diagonal matrix are its diagonal entries). Hence, the singular values of D are $|d_1|, |d_2|, \ldots, |d_m|, \underbrace{0, 0, \ldots, 0}_{n-m \text{ entries}}$ (since the singular values of D are defined as the square roots of the eigenvalues of D^*D). Thus, the nonzero singular values of D are the nonzero numbers among $|d_1|, |d_2|, \ldots, |d_m|$. In other words, they are the absolute values of the nonzero diagonal entries of D. This proves Claim 2.]

Now, we can prove the claim we intended to prove. Let P be a pseudodiagonal matrix such that all diagonal entries of P are nonnegative reals, and such that A is unitarily equivalent to P. As we recall, our goal is to prove that P and Σ have the same diagonal entries up to order.

We know that the matrix A is unitarily equivalent to Σ and also to P. Thus, Σ is unitarily equivalent to P (because unitary equivalence is an equivalence relation). Therefore, Claim 1 (applied to $X = \Sigma$ and $Y = P$) shows that the matrices Σ and P have the same singular values. However, these two matrices are pseudodiagonal; thus, Claim 2 shows that their nonzero singular values are the absolute values of their nonzero diagonal entries. Since their diagonal entries are nonnegative reals, we can actually drop the "absolute values" part from this sentence, and conclude that their nonzero singular values are simply their nonzero diagonal entries. Thus, the matrices Σ and P have the same nonzero diagonal entries (because we have shown that they have the same singular values). Therefore, the matrices Σ and P have the same diagonal entries (since they have the same dimensions and therefore the same number of diagonal entries). Thus, we have shown that the matrices P and Σ have the same diagonal entries up to order. This completes the proof of Theorem 5.2.3 **(b)**. □

Example 5.2.5. Let $A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix} \in \mathbb{C}^{2\times 3}$. How do we find an SVD of A ?

This is not the way SVDs are computed in practice, but we can try following our above proof of Theorem 5.2.3. Thus, we compute a spectral decomposition of A^*A. (Since A^*A is Hermitian, this is equivalent to diagonalizing A^*A.) A simple computation yields that

$$A^*A = \begin{pmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{pmatrix}$$

and that A^*A has eigenvalues $25, 9, 0$ and a spectral decomposition (V, D) with

$$V = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/6 & -2/3 \\ \sqrt{2}/2 & -\sqrt{2}/6 & 2/3 \\ 0 & 2\sqrt{2}/3 & 1/3 \end{pmatrix} \qquad \text{and} \qquad D = \operatorname{diag}(25, 9, 0).$$

(Of course, we have handpicked A to make the eigenvalues integers; a random A would give rise to irrational eigenvalues.) Thus, $\lambda_1 = 25$ and $\lambda_2 = 9$ and $\lambda_3 = 0$ and $k = 2$. Hence, $\sigma_1 = \sqrt{25} = 5$ and $\sigma_2 = \sqrt{9} = 3$, so that $\Sigma = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}$. It remains to find U. To do so, we set $u_j := \dfrac{1}{\sigma_j} A v_j$ for all $j \in [k]$; thus,

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{5} \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} \qquad \text{and}$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{3} \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix} \begin{pmatrix} \sqrt{2}/6 \\ -\sqrt{2}/6 \\ 2\sqrt{2}/3 \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}.$$

The proof of Theorem 5.2.3 tells us to extend this orthonormal tuple $(u_1, u_2, \ldots, u_k)$ to an orthonormal basis $(u_1, u_2, \ldots, u_m)$ of $\mathbb{C}^m$, but this is unnecessary here, since it already is a basis (since $k = m$). Thus, we can now compute U as the matrix with columns $u_1, u_2, \ldots, u_m$; that is, $U = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{pmatrix}$. Hence, we obtain the SVD (U, V, Σ) of A with

$$U = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{pmatrix}, \qquad V = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/6 & -2/3 \\ \sqrt{2}/2 & -\sqrt{2}/6 & 2/3 \\ 0 & 2\sqrt{2}/3 & 1/3 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}.$$

Exercise 5.2.2. [3] Find an SVD of the matrix $A := \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

Exercise 5.2.3. [4] Find an SVD of the matrix $A := \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$.

Exercise 5.2.4. [3] Let (U, V, Σ) be an SVD of a matrix $A \in \mathbb{C}^{m \times n}$.

(a) Construct an SVD of the matrix $A^* \in \mathbb{C}^{n \times m}$.

(b) Now assume that A is invertible (so that $m = n$). Construct an SVD of the matrix A^{-1}.

Exercise 5.2.5. [3] Let A and B be two $m \times n$-matrices. Prove that A and B are unitarily equivalent if and only if the matrices A^*A and B^*B are unitarily similar.

A variant of the SVD is the so-called *compact SVD*, in which the unitary matrices U and V are replaced by isometries and the pseudodiagonal matrix Σ is replaced by a diagonal $k \times k$-matrix for $k = \operatorname{rank} A$:

Corollary 5.2.6. Let $A \in \mathbb{C}^{m \times n}$. Let $k = \operatorname{rank} A$. Then:

(a) There exist isometries $U \in \mathbb{C}^{m \times k}$ and $V \in \mathbb{C}^{n \times k}$ and a diagonal matrix $\Sigma \in \mathbb{C}^{k \times k}$ such that all diagonal entries of Σ are positive reals and such that

$$A = U \Sigma V^*.$$

(b) The matrix Σ is unique up to permutation of its diagonal entries. (The matrices U and V are usually not unique.)

(c) Let $\sigma_1, \sigma_2, \ldots, \sigma_n$ be the square roots of the n eigenvalues of the Hermitian matrix A^*A, listed in decreasing order (so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$). Then, we have $\sigma_{k+1} = \sigma_{k+2} = \cdots = \sigma_n = 0$, and we can take

$$\Sigma = \operatorname{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_k\right) \in \mathbb{C}^{k \times k}$$

in part **(a)**.

Exercise 5.2.6. [4] Prove Corollary 5.2.6.

Exercise 5.2.7. [4] Give a simple algorithm (without using eigenvalues or spectral decomposition) to compute a compact SVD of a given rank-1 matrix.

6. Positive and nonnegative matrices ([HorJoh13, Chapter 8])

6.1. Basics

Recall the triangle inequality:

Proposition 6.1.1 (triangle inequality)**.** Let $z_1, z_2, \ldots, z_n$ be n complex numbers. Then:

(a) We have the inequality

$$|z_1| + |z_2| + \cdots + |z_n| \geq |z_1 + z_2 + \cdots + z_n| .$$

(b) Equality holds in this inequality if and only if $z_1, z_2, \ldots, z_n$ have the same argument (i.e., there exists some $w \in \mathbb{C}$ such that $z_1, z_2, \ldots, z_n$ are nonnegative real multiples of w).

Definition 6.1.2. Let $A \in \mathbb{C}^{n \times m}$ be a matrix.

(a) We say that A is *positive* (and write $A > 0$) if all entries of A are positive reals.

(b) We say that A is *nonnegative* (and write $A \geq 0$) if all entries of A are nonnegative reals.

(c) We let $|A| \in \mathbb{R}^{n \times m}$ be the nonnegative matrix obtained by replacing each entry of A by its absolute value. In other words,

$$|A| := \begin{pmatrix} |A_{1,1}| & |A_{1,2}| & \cdots & |A_{1,m}| \\ |A_{2,1}| & |A_{2,2}| & \cdots & |A_{2,m}| \\ \vdots & \vdots & \ddots & \vdots \\ |A_{n,1}| & |A_{n,2}| & \cdots & |A_{n,m}| \end{pmatrix}.$$

Remark 6.1.3. Recall that row vectors and column vectors are matrices. Thus, the statements "$v > 0$" and "$v \geq 0$" and the notation $|v|$ are defined for them as well. If $v = (v_1, v_2, \ldots, v_k)^T$, then $|v| = (|v_1|, |v_2|, \ldots, |v_k|)^T$.

Warning 6.1.4. Do not mistake $|v|$ (a vector) for $||v||$ (a number). Also, when A is a matrix, do not mistake $|A|$ for (an old notation for) the determinant of A. (We always write $\det A$ for the determinant of A, so this confusion should not arise.)

Let us stress once again that positive matrices and nonnegative matrices are required to have real entries by definition.

Exercise 6.1.1. $\boxed{1}$ Let $v \in \mathbb{C}^m$ be a column vector. Prove that $|||v||| = ||v||$, where the left hand side means the length of $|v|$.

Exercise 6.1.2. $\boxed{1}$ Let $\lambda \in \mathbb{C}$ and $A \in \mathbb{C}^{n \times m}$. Prove that $|\lambda A| = |\lambda| \cdot |A|$.

Proposition 6.1.5. A matrix $A \in \mathbb{C}^{n \times m}$ is nonnegative if and only if $|A| = A$.

Proof. $\Longrightarrow$: If A is nonnegative, then each i and j satisfy $A_{i,j} \geq 0$ and thus $\left|A_{i,j}\right| = A_{i,j}$; therefore, $|A| = A$.

$\Longleftarrow$: If $|A| = A$, then A is nonnegative (since $|A|$ is always nonnegative). □

Definition 6.1.6. Let $A, B \in \mathbb{R}^{n \times m}$ be two matrices with real entries. Then:

(a) We say that $A \geq B$ if and only if $A - B \geq 0$ (or, equivalently, $A_{i,j} \geq B_{i,j}$ for all $i \in [n]$ and $j \in [m]$).

(b) We say that $A > B$ if and only if $A - B > 0$ (or, equivalently, $A_{i,j} > B_{i,j}$ for all $i \in [n]$ and $j \in [m]$).

(c) We say that $A \leq B$ if and only if $A - B \leq 0$ (or, equivalently, $A_{i,j} \leq B_{i,j}$ for all $i \in [n]$ and $j \in [m]$).

(d) We say that $A < B$ if and only if $A - B < 0$ (or, equivalently, $A_{i,j} < B_{i,j}$ for all $i \in [n]$ and $j \in [m]$).

Example 6.1.7. We have $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \geq \begin{pmatrix} 0 & 2 \\ 2 & 4 \end{pmatrix}$.

The relations $\geq$, $>$, $\leq$ and $<$ are known as *entrywise inequalities* (specifically, "entrywise greater or equal", "entrywise greater", etc.), since they are just saying that each entry of A is $\geq$, $>$, $\leq$ or $<$ to the corresponding entry of B.

Remark 6.1.8. Again, recall that row vectors and column vectors are matrices too; thus, Definition 6.1.6 applies to them as well.

Proposition 6.1.9. (a) The relations $\geq$, $>$, $\leq$ and $<$ on $\mathbb{R}^{n \times m}$ (introduced in Definition 6.1.6) are transitive.

(b) The relations $\geq$ and $\leq$ are reflexive and antisymmetric (so they are weak partial orders on $\mathbb{R}^{n \times m}$).

(c) Let A and B be two matrices in $\mathbb{R}^{n \times m}$. Then, the implications $(A > B) \implies (A \geq B)$ and $(A < B) \implies (A \leq B)$ as well as the equivalences $(A > B) \iff (B < A)$ and $(A \geq B) \iff (B \leq A)$ hold.

Proof. All of these are straightforward, since the relations $\geq$, $>$, $\leq$ and $<$ are just entrywise inequalities. □

Warning 6.1.10. The relations $\geq$ and $\leq$ are not total orders (unless $n \leq 1$). For instance, the row vector $(2, 1)$ is neither $\geq$ nor $\leq$ to $(3, 0)$.

Warning 6.1.11. Do not mistake the relation $\geq$ on column vectors for the relation $\succcurlyeq$ (majorization).

Warning 6.1.12. The trivial vector $v = () \in \mathbb{R}^0$ (with no entries at all) satisfies $v > v$ and $v < v$ and $v \geq v$ and $v \leq v$, because the "for all" statements in Definition 6.1.6 are vacuously true. However, this is the only case in which a vector v satisfies both $v > v$ and $v \leq v$.

Warning 6.1.13. Given two matrices A and B, the relation $A \geq B$ is **not** equivalent to "$A > B$ or $A = B$". For example, $(3, 1) \geq (2, 1)$ is true, but we have neither $(2, 1) > (3, 1)$ nor $(2, 1) = (3, 1)$.

Exercise 6.1.3. $\boxed{1}$ Let $A \in \mathbb{C}^{n \times n}$ be a doubly stochastic matrix (see Definition 4.7.22 for the meaning of this). Let J be the $n \times n$-matrix whose all entries equal 1. Prove that $J \geq A \geq 0$.

Two complex numbers z and w always satisfy $|z| \cdot |w| = |zw|$. For two matrices, however, this equality is not usually satisfied; however, it survives as an inequality:

Proposition 6.1.14. Let $A \in \mathbb{C}^{n \times m}$ and $B \in \mathbb{C}^{m \times p}$ be two matrices. Then,

$$|A| \cdot |B| \geq |AB| .$$

Proof. We must prove that $(|A| \cdot |B|)_{i,k} \geq |AB|_{i,k}$ for all $i \in [n]$ and $k \in [p]$.

So let $i \in [n]$ and $k \in [p]$. Then, the definition of the product of two matrices yields

$$(|A| \cdot |B|)_{i,k} = \sum_{j=1}^{m} \underbrace{|A|_{i,j}}_{=|A_{i,j}|} \cdot \underbrace{|B|_{j,k}}_{=|B_{j,k}|} = \sum_{j=1}^{m} \underbrace{|A_{i,j}| \cdot |B_{j,k}|}_{=|A_{i,j}B_{j,k}|} = \sum_{j=1}^{m} |A_{i,j}B_{j,k}| \geq \left| \sum_{j=1}^{m} A_{i,j}B_{j,k} \right|$$

(by the triangle inequality). In view of

$$|AB|_{i,k} = \left| (AB)_{i,k} \right| = \left| \sum_{j=1}^{m} A_{i,j}B_{j,k} \right| ,$$

we can rewrite this as $(|A| \cdot |B|)_{i,k} \geq |AB|_{i,k}$, qed. □

Corollary 6.1.15. Let $A \in \mathbb{C}^{n \times n}$ and $k \in \mathbb{N}$. Then, $|A|^k \geq |A^k|$.

Proof. Induction on k, using Proposition 6.1.14 (and the fact that $|I_n| = I_n$). □

It is not easy to characterize when the inequality in Proposition 6.1.14 becomes an equality. However, conclusions can be drawn in some cases. The following proposition considers the case when the matrix B is a column vector (which we call x to avoid unusual notations):

Proposition 6.1.16. Let $A \in \mathbb{C}^{n \times m}$ and $x \in \mathbb{C}^m$. Then:

(a) We have $|A| \cdot |x| \geq |Ax|$.

(b) If at least one row of A is positive and we have $A \geq 0$ and $|Ax| = A \cdot |x|$, then $|x| = \omega x$ for some $\omega \in \mathbb{C}$ satisfying $|\omega| = 1$.

(c) If $x > 0$ and $Ax = |A|\, x$, then $A = |A|$ (so that $A \geq 0$).

Proof. **(a)** follows from Proposition 6.1.14.

(b) Assume that at least one row of A is positive and we have $A \geq 0$ and $|Ax| = A \cdot |x|$.

We have assumed that at least one row of A is positive. Let the i-th row of A be positive. Thus, the numbers $A_{i,j}$ are positive reals for all $j \in [m]$.

Write $x = (x_1, x_2, \ldots, x_m)^T$. Thus, $|x| = (|x_1|, |x_2|, \ldots, |x_m|)^T$.

From $|Ax| = A \cdot |x|$, we obtain

$$(\text{the } i\text{-th entry of } |Ax|) = (\text{the } i\text{-th entry of } A \cdot |x|) = \sum_{j=1}^{m} \underbrace{A_{i,j} \cdot |x_j|}_{= |A_{i,j} x_j| \text{ (since } A \geq 0 \text{ and thus } A_{i,j} \geq 0)}$$

$$= \sum_{j=1}^{m} |A_{i,j} x_j| ,$$

so that

$$\sum_{j=1}^{m} |A_{i,j} x_j| = (\text{the } i\text{-th entry of } |Ax|) = |\text{the } i\text{-th entry of } Ax|$$

$$= \left| \sum_{j=1}^{m} A_{i,j} x_j \right|$$

(since the i-th entry of Ax is $\sum_{j=1}^{m} A_{i,j} x_j$). This is an equality case of the triangle inequality. Thus, the complex numbers $A_{i,j} x_j$ for all $j \in [m]$ have the same argument (by Proposition 6.1.1 **(b)**). In other words, the numbers x_j for all $j \in [m]$ have the same argument (since all the $A_{i,j}$ are positive reals and thus we have $\arg\left(A_{i,j} x_j\right) = \arg x_j$). Let φ be this argument, and let $\omega := e^{-i\varphi}$. Then, ω is a complex number satisfying $|\omega| = 1$, and the numbers $\omega x_1, \omega x_2, \ldots, \omega x_n$ are nonnegative reals. This shows that $\omega x \geq 0$, so that $|\omega x| = \omega x$. However, Exercise 6.1.2 yields $|\omega x| = \underbrace{|\omega|}_{=1} \cdot |x| = |x|$. Comparing these two equalities, we obtain $|x| = \omega x$. Theorem 6.1.16 **(b)** is thus proven.

(c) Suppose $x > 0$ and $Ax = |A|\, x$. We must show that $A = |A|$ (so that $A \geq 0$).

Write $x = (x_1, x_2, \ldots, x_m)^T$. Thus, $x_1, x_2, \ldots, x_m$ are positive reals (since $x > 0$).

Fix $i \in [n]$. Then,

$$(\text{the } i\text{-th entry of } Ax) = (\text{the } i\text{-th entry of } |A|\, x) .$$

In other words,

$$\sum_{j=1}^{m} A_{i,j} x_j = \sum_{j=1}^{m} \underbrace{|A_{i,j}|\, x_j}_{= |A_{i,j} x_j| \text{ (since } x_j \text{ is a positive real)}} = \sum_{j=1}^{m} |A_{i,j} x_j| .$$

This shows that $\sum_{j=1}^{m} A_{i,j}x_j$ is a nonnegative real. Furthermore, we obtain

$$\sum_{j=1}^{m} A_{i,j}x_j = \sum_{j=1}^{m} \left|A_{i,j}x_j\right| \geq \left|\sum_{j=1}^{m} A_{i,j}x_j\right| \qquad \text{(by the triangle inequality)}$$
$$\geq \sum_{j=1}^{m} A_{i,j}x_j.$$

This is a chain of inequalities in which the first and the last side are equal. Thus, all inequalities in it must be equalities. In particular, we thus have equality in the triangle inequality $\sum_{j=1}^{m} \left|A_{i,j}x_j\right| \geq \left|\sum_{j=1}^{m} A_{i,j}x_j\right|$. Hence, the complex numbers $A_{i,j}x_j$ for all $j \in [m]$ have the same argument (by Proposition 6.1.1 **(b)**). Their sum $\sum_{j=1}^{m} A_{i,j}x_j$ therefore has the same argument as them; but since we know that this sum $\sum_{j=1}^{m} A_{i,j}x_j$ is a nonnegative real, we thus conclude that this common argument is 0. In other words, the complex numbers $A_{i,j}x_j$ for all $j \in [m]$ are nonnegative reals. Since $x_1, x_2, \ldots, x_m$ are positive reals, this means that the $A_{i,j}$ for all $j \in [m]$ are nonnegative reals. Since we have proved this for all $i \in [n]$, we thus conclude that all entries of A are nonnegative reals. Hence, $A \geq 0$, so that $A = |A|$. This proves Theorem 6.1.16 **(c)**. □

Proposition 6.1.17. **(a)** If $A, B, C, D \in \mathbb{C}^{n\times m}$ satisfy $A \leq B$ and $C \leq D$, then $A + C \leq B + D$.

(b) If $A, B \in \mathbb{C}^{n\times m}$ and $C \in \mathbb{C}^{m\times p}$ satisfy $A \leq B$ and $0 \leq C$, then $AC \leq BC$.

(c) If $A, B \in \mathbb{C}^{n\times m}$ and $C \in \mathbb{C}^{p\times n}$ satisfy $A \leq B$ and $0 \leq C$, then $CA \leq CB$.

(d) If $A, B \in \mathbb{C}^{n\times m}$ and $C, D \in \mathbb{C}^{m\times p}$ satisfy $0 \leq A \leq B$ and $0 \leq C \leq D$, then $0 \leq AC \leq BD$.

(e) If $A, B \in \mathbb{C}^{n\times n}$ satisfy $0 \leq A \leq B$, and if $k \in \mathbb{N}$, then $0 \leq A^k \leq B^k$.

Proof. **(a)** For all i and j, we have $A_{i,j} \leq B_{i,j}$ and $C_{i,j} \leq D_{i,j}$ and therefore $A_{i,j} + C_{i,j} \leq B_{i,j} + D_{i,j}$. But this means $A + C \leq B + D$.

(b) Assume $A \leq B$ and $0 \leq C$. Let $i \in [n]$ and $k \in [p]$. The definition of the product of two matrices yields

$$(AC)_{i,k} = \sum_{j=1}^{m} A_{i,j}C_{j,k} \qquad \text{with} \qquad (BC)_{i,k} = \sum_{j=1}^{m} B_{i,j}C_{j,k}.$$

The right hand side of the first equality is $\leq$ to the right hand side of the second, because all $j \in [m]$ satisfy $A_{i,j} \leq B_{i,j}$ (since $A \leq B$) and $C_{j,k} \geq 0$ (since $0 \leq C$).

Thus, we obtain $(AC)_{i,k} \leq (BC)_{i,k}$. Since we have proved this for all i and k, we thus obtain $AC \leq BC$. This proves Proposition 6.1.17 **(b)**.

(c) Similar to **(b)**.

(d) Part **(b)** yields $AC \leq BC$. Part **(c)** (applied to C, D and B instead of A, B and C) yields $BC \leq BD$. Since the relation $\leq$ is transitive, we can conclude $AC \leq BD$ from these two inequalities.

(e) Follows from **(d)** by induction on k. □

Exercise 6.1.4. [2] Let n, m, p be three positive integers.

(a) Show that any two positive matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times p}$ satisfy $AB > 0$.

(b) Now, assume that $m > 1$. Find an example of two nonzero nonnegative matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times p}$ that nevertheless satisfy $AB = 0$.

6.2. The spectral radius

Definition 6.2.1. The *spectral radius* $\rho(A)$ of a matrix $A \in \mathbb{C}^{n \times n}$ (with $n > 0$) is defined to be the largest absolute value of an eigenvalue of A. That is,

$$\rho(A) := \max\{|\lambda| \mid \lambda \in \sigma(A)\}.$$

Note that $\rho(A)$ is always a nonnegative real.

Some examples:

- If $A = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then $\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$. More generally, this is true if A is a triangular matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$.
- By Exercise 3.4.2 (equivalence $\mathcal{A} \Longleftrightarrow \mathcal{C}$), a square matrix A satisfies $\rho(A) = 0$ if and only if A is nilpotent.

The following is obvious:

Lemma 6.2.2. Let $\lambda \in \mathbb{C}$ and $A \in \mathbb{C}^{n \times n}$ (where $n > 0$). Then, $\rho(\lambda A) = |\lambda| \cdot \rho(A)$.

It is furthermore easy to see that each $n \times n$-matrix $A \in \mathbb{C}^{n \times n}$ (with $n > 0$) satisfies $\rho(A^T) = \rho(A^*) = \rho(A)$.

Theorem 6.2.3. Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ be such that $B \geq |A|$ and $n > 0$. Then, $\rho(A) \leq \rho(B)$.

Proof of Theorem 6.2.3. If $\rho(A) = 0$, then this is obvious. So, WLOG assume that $\rho(A) > 0$.

We can thus scale both matrices A and B by the positive real $\dfrac{1}{\rho(A)}$. This does not break the inequality $B \geq |A|$, and also does not break the claim $\rho(A) \leq \rho(B)$ (by Lemma 6.2.2).

Thus, we WLOG assume that $\rho(A) = 1$. (This is achieved by the scaling we just mentioned.)

This yields that A has an eigenvalue λ with $|\lambda| = 1$. Let λ be such an eigenvalue, and let v be a nonzero λ-eigenvector. Thus, $Av = \lambda v$. Hence,

$$A^m v = \lambda^m v \qquad \text{for any } m \in \mathbb{N}. \tag{119}$$

(This follows easily by induction on m.)

Now, we must prove $\rho(A) \leq \rho(B)$. In other words, we must prove that $1 \leq \rho(B)$ (since $\rho(A) = 1$). Assume the contrary. Thus, $\rho(B) < 1$. Hence, all eigenvalues of B have absolute value < 1. Therefore, Corollary 3.5.2 (applied to B instead of A) shows that $\lim\limits_{m\to\infty} B^m = 0$. Therefore, $\lim\limits_{m\to\infty} B^m \cdot |v| = 0$.

However, let $m \in \mathbb{N}$. Then, $B \geq |A| \geq 0$ entails $B^m \geq |A|^m$ (by Proposition 6.1.17 **(e)**). Also, $|A|^m \geq |A^m|$ (by Corollary 6.1.15). Thus, $B^m \geq |A|^m \geq |A^m|$. Hence, using Proposition 6.1.17 **(b)**, we obtain

$$\begin{aligned}
B^m \cdot |v| &\geq |A^m| \cdot |v| && (\text{since } |v| \geq 0) \\
&\geq |A^m v| && (\text{by Proposition 6.1.14}) \\
&= |\lambda^m v| && (\text{by (119)}) \\
&= \underbrace{|\lambda^m|}_{\substack{=|\lambda|^m=1 \\ (\text{since } |\lambda|=1)}} \cdot |v| && (\text{by Exercise 6.1.2}) \\
&= |v| .
\end{aligned}$$

Taking limits as $m \to \infty$, we obtain $\lim\limits_{m\to\infty} B^m \cdot |v| \geq \lim\limits_{m\to\infty} |v| = |v| \neq 0$ (since v is nonzero). This contradicts $\lim\limits_{m\to\infty} B^m \cdot |v| = 0$. This contradiction shows that our assumption was false, and the proof of Theorem 6.2.3 is complete. □

Corollary 6.2.4. Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be such that $B \geq |A|$ and $n > 0$. Then, $\rho(A) \leq \rho(|A|) \leq \rho(B)$.

Proof. Applying Theorem 6.2.3 to $|A|$ instead of B, we get $\rho(A) \leq \rho(|A|)$.

Applying Theorem 6.2.3 to $|A|$ instead of A, we get $\rho(|A|) \leq \rho(B)$ (since $||A|| = |A|$).

Hence, Corollary 6.2.4 is proved. □

Corollary 6.2.5. Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ satisfy $B \geq A \geq 0$ and $n > 0$. Then, $\rho(A) \leq \rho(B)$.

Proof. We have $|A| = A$. Thus, we can apply Theorem 6.2.3. □

Corollary 6.2.6. Let $A \in \mathbb{R}^{n \times n}$ satisfy $A \geq 0$ and $n > 0$.

(a) If $\widetilde{A}$ is a principal submatrix of A (that is, a matrix obtained from A by removing a bunch of rows along with the corresponding columns), then $\rho\left(\widetilde{A}\right) \leq \rho(A)$.

(b) We have $\max\{A_{i,i} \mid i \in [n]\} \leq \rho(A)$.

(c) If $A_{i,i} > 0$ for some $i \in [n]$, then $\rho(A) > 0$.

Proof. **(a)** Let $\widetilde{A}$ be a principal submatrix of A. For simplicity, I assume that $\widetilde{A}$ is A with the n-th row and the n-th column removed[54]. Thus,

$$A = \begin{pmatrix} \widetilde{A} & y \\ x & \lambda \end{pmatrix} \qquad \text{(in block-matrix notation)} \tag{120}$$

for some nonnegative $x \in \mathbb{R}^{1 \times (n-1)}$, $y \in \mathbb{R}^{(n-1) \times 1}$ and $\lambda \in \mathbb{R}$. Let

$$B := \begin{pmatrix} \widetilde{A} & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{(in block-matrix notation)}, \tag{121}$$

where the three 0s have the same dimensions as the x, y and λ above. Comparing (120) with (121), we see that $A \geq B$ (since $x \geq 0$ and $y \geq 0$ and $\lambda \geq 0$). Also, $\widetilde{A} \geq 0$ (since $A \geq 0$) and thus $B \geq 0$. Thus, $A \geq B \geq 0$. Hence, Corollary 6.2.5 (applied to B and A instead of A and B) yields $\rho(B) \leq \rho(A)$.

However, it is easy to see from (121) that $\sigma(B) = \sigma\left(\widetilde{A}\right) \cup \{0\}$ (for example, because we can pick any Schur triangularization (U, T) of $\widetilde{A}$, and then obtain a Schur triangularization (U', T') of B by setting $U' = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix}$ and $T' = \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix}$). Hence, $\rho(B) = \rho\left(\widetilde{A}\right)$ (because inserting 0 into a set of nonnegative reals cannot change the maximum of this set). Hence, $\rho(B) \leq \rho(A)$ rewrites as $\rho\left(\widetilde{A}\right) \leq \rho(A)$. This proves Corollary 6.2.4 **(a)**.

(b) We must show that $A_{i,i} \leq \rho(A)$ for all $i \in [n]$.

So let $i \in [n]$. Then, the 1×1-matrix $\begin{pmatrix} A_{i,i} \end{pmatrix}$ is a principal submatrix of A (obtained by removing all rows of A other than the i-th one, and all columns of A other than the i-th one). Hence, part **(a)** yields $\rho\left(\begin{pmatrix} A_{i,i} \end{pmatrix}\right) \leq \rho(A)$. However, since the only eigenvalue of the 1×1-matrix $\begin{pmatrix} A_{i,i} \end{pmatrix}$ is $A_{i,i}$, we have $\rho\left(\begin{pmatrix} A_{i,i} \end{pmatrix}\right) =$

[54] The proof in the general case is similar; it just requires more notational work.

$|A_{i,i}| = A_{i,i}$ (since $A \geq 0$). Therefore, $A_{i,i} = \rho\left(\left(\ A_{i,i}\ \right)\right) \leq \rho\left(A\right)$. This proves Corollary 6.2.4 **(b)**.

(c) Follows from **(b)**. □

Exercise 6.2.1. [1] Let $A \in \mathbb{R}^{n\times n}$ satisfy $A > 0$ and $n > 0$. Prove that $\rho\left(A\right) > 0$.

Exercise 6.2.2. [2] Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be such that $B > |A|$ and $n > 0$. Prove that $\rho\left(A\right) < \rho\left(B\right)$.

[**Hint:** Show that there exists some real $\lambda > 1$ such that $B \geq \lambda \cdot |A|$.]

Let us next prove some more bounds for $\rho\left(A\right)$ when A is a nonnegative matrix. We will use the following notions:[55]

Definition 6.2.7. Let $\mathbb{F}$ be a field. Let $A \in \mathbb{F}^{n\times m}$.

(a) The *column sums* of A are the m sums

$$\sum_{i=1}^{n} A_{i,j} = \left(\text{the sum of all entries of the } j\text{-th column of } A\right)$$

for $j \in [m]$.

(b) The *row sums* of A are

$$\sum_{j=1}^{m} A_{i,j} = \left(\text{the sum of all entries of the } i\text{-th row of } A\right)$$

for $i \in [n]$.

(c) Now, assume that $\mathbb{F} = \mathbb{C}$ and $n > 0$ and $m > 0$. Then, we set

$$||A||_{\infty} := \left(\text{the largest row sum of } |A|\right) = \max_{i\in[n]} \sum_{j=1}^{m} \left|A_{i,j}\right|$$

and

$$||A||_{1} := \left(\text{the largest column sum of } |A|\right) = \max_{j\in[m]} \sum_{i=1}^{n} \left|A_{i,j}\right|.$$

These two numbers $||A||_{\infty}$ and $||A||_{1}$ are called the ∞-*norm* and the 1-*norm* of A (for reasons that will be explained in a later chapter).

[55] Recall that $\max\limits_{i\in I} a_i$ is a shorthand notation for $\max\left\{a_i \mid i \in I\right\}$ (when I is a set and a_i is a real number for each $i \in I$).

Example 6.2.8. The column sums of a 2×2-matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ are $a+c$ and $b+d$, whereas its row sums are $a+b$ and $c+d$.

Warning 6.2.9. Let A be a matrix. Then, the sum of all rows of A is a row vector whose entries are the column sums (not the row sums!) of A. Likewise, the sum of all columns of A is a column vector whose entries are the row sums (not the column sums!) of A.

The following is obvious:

Remark 6.2.10. Let $A \in \mathbb{F}^{n \times m}$ be a matrix over a field $\mathbb{F}$.

(a) The row sums of A are the column sums of A^T, and vice versa.

(b) If $\mathbb{F} = \mathbb{C}$ and $n > 0$ and $m > 0$, then $||A||_\infty = \left|\left|A^T\right|\right|_1$ and $||A||_1 = \left|\left|A^T\right|\right|_\infty$.

We can now bound the spectral radius of a matrix in terms of its 1-norm and its ∞-norm:

Lemma 6.2.11. Let $A \in \mathbb{C}^{n \times n}$ with $n > 0$. Then:

(a) We have $\rho(A) \leq ||A||_\infty$.

(b) If $A \geq 0$ and if all row sums of A are equal, then $\rho(A) = ||A||_\infty$.

(c) We have $\rho(A) \leq ||A||_1$.

(d) If $A \geq 0$ and if all column sums of A are equal, then $\rho(A) = ||A||_1$.

Proof. **(a)** We have $\rho(A) = |\lambda|$ for some eigenvalue λ of A (by the definition of $\rho(A)$). Consider this λ, and let $v = (v_1, v_2, \ldots, v_n)^T \in \mathbb{C}^n$ be a nonzero λ-eigenvector of A. Then, $Av = \lambda v$.

Choose an $i \in [n]$ such that $|v_i| = \max\{|v_1|, |v_2|, \ldots, |v_n|\}$. Then, $|v_i| > 0$ (since v is nonzero). Furthermore,

$$|v_j| \leq |v_i| \qquad \text{for each } j \in [n] \tag{122}$$

(since $|v_i| = \max\{|v_1|, |v_2|, \ldots, |v_n|\}$).

Now, the i-th entry of the column vector Av is $\sum_{j=1}^{n} A_{i,j} v_j$ (by the definition of the product Av); however, the same entry is λv_i (since $Av = \lambda v$). Comparing these two facts, we obtain

$$\lambda v_i = \sum_{j=1}^{n} A_{i,j} v_j.$$

Taking absolute values on both sides of this equality, we obtain

$$|\lambda v_i| = \left|\sum_{j=1}^{n} A_{i,j} v_j\right| \leq \sum_{j=1}^{n} \underbrace{\left|A_{i,j} v_j\right|}_{\substack{=|A_{i,j}| \cdot |v_j| \leq |A_{i,j}| \cdot |v_i| \\ \text{(by (122))}}} \qquad \text{(by the triangle inequality)}$$

$$\leq \sum_{j=1}^{n} |A_{i,j}| \cdot |v_i| .$$

Since $|\lambda v_i| = |\lambda| \cdot |v_i|$, we can rewrite this as

$$|\lambda| \cdot |v_i| \leq \sum_{j=1}^{n} |A_{i,j}| \cdot |v_i| .$$

Since $|v_i| > 0$, we can cancel $|v_i|$ from this inequality, and thus we obtain

$$|\lambda| \leq \sum_{j=1}^{n} |A_{i,j}| = (\text{the } i\text{-th row sum of } |A|)$$
$$\leq (\text{the largest row sum of } |A|) = ||A||_\infty .$$

Since $\rho(A) = |\lambda|$, this rewrites as $\rho(A) \leq ||A||_\infty$. This proves Lemma 6.2.11 **(a)**.

(b) Assume that $A \geq 0$ and that all row sums of A are equal. Let $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$, and let κ be the common value of the row sums of A. Then, all row sums of A equal κ; in other words, we have $Ae = \kappa e$ (since Ae is the column vector whose entries are the row sums of A, whereas κe is the column vector whose entries are $\underbrace{\kappa, \kappa, \ldots, \kappa}_{n \text{ times}}$). Hence, κ is an eigenvalue of A (since $e \neq 0$), so that $\rho(A) \geq |\kappa| = \kappa$ (since $A \geq 0$ entails $\kappa \geq 0$).

On the other hand, $|A| = A$ (since $A \geq 0$). Now, Lemma 6.2.11 **(a)** yields

$$\rho(A) \leq ||A||_\infty = \left(\text{the largest row sum of } \underbrace{|A|}_{=A}\right) = (\text{the largest row sum of } A) = \kappa$$

(since all row sums of A are κ). Combining this with $\rho(A) \geq \kappa$, we obtain $\rho(A) = \kappa = ||A||_\infty$. This proves Lemma 6.2.11 **(b)**.

(c) This follows by applying Lemma 6.2.11 **(a)** of the lemma to A^T instead of A, and recalling that $\left|\left|A^T\right|\right|_\infty = ||A||_1$ and $\rho\left(A^T\right) = \rho(A)$.

(d) This follows by applying Lemma 6.2.11 **(b)** of the lemma to A^T instead of A, and recalling that $\left|\left|A^T\right|\right|_\infty = ||A||_1$ and $\rho\left(A^T\right) = \rho(A)$ and the row sums of A^T are the column sums of A. □

Remark 6.2.12. We note that the converse of Lemma 6.2.11 **(b)** is false: For example, the 3×3-matrix $A := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ satisfies $A \geq 0$ and $\rho(A) = ||A||_\infty$, but the row sums of A are not all equal.

Next, let us bound the spectral radius $\rho(A)$ of a matrix A from both sides when $A \geq 0$:

Theorem 6.2.13. Let $A \in \mathbb{R}^{n \times n}$ satisfy $A \geq 0$ and $n > 0$. Then,

$$(\text{the smallest row sum of } A) \leq \rho(A) \leq (\text{the largest row sum of } A).$$

Proof. We have $|A| = A$ (since $A \geq 0$). Now, Lemma 6.2.11 **(a)** yields

$$\rho(A) \leq ||A||_\infty = \left(\text{the largest row sum of } \underbrace{|A|}_{=A} \right) = (\text{the largest row sum of } A).$$

Hence, it remains to prove that $(\text{the smallest row sum of } A) \leq \rho(A)$.

Let $r_1, r_2, \ldots, r_n$ be the row sums of A. Let r_i be the smallest among them. We must thus prove that $r_i \leq \rho(A)$. If $r_i = 0$, then this is obvious. So let WLOG assume that $r_i > 0$. Hence, all n numbers $r_1, r_2, \ldots, r_n$ are positive (since the smallest among them is $r_i > 0$) and thus nonzero.

Let B the $n \times n$-matrix whose (u, v)-th entry is $\dfrac{r_i}{r_u} A_{u,v}$ for all $u, v \in [n]$. [56] Thus, B is obtained from the matrix A by scaling each row by a certain positive real factor (namely, $\dfrac{r_i}{r_u}$ for the u-th row) chosen in such a way that the row sums all become r_i. Hence, the matrix B is ≥ 0 (since $A \geq 0$, and since all n numbers $r_1, r_2, \ldots, r_n$ are positive), and its row sums are all equal to r_i. Hence, Lemma 6.2.11 **(b)** (applied to B instead of A) yields

$$\begin{aligned} \rho(B) = ||B||_\infty &= \left(\text{the largest row sum of } \underbrace{|B|}_{\substack{=B \\ (\text{since } B \geq 0)}} \right) \\ &= (\text{the largest row sum of } B) = r_i \end{aligned} \tag{123}$$

(since all row sums of B are r_i). However, for each $u, v \in [n]$, we have $\dfrac{r_i}{r_u} A_{u,v} \leq A_{u,v}$ (since $r_i \leq r_u$ (because r_i is the smallest among the numbers $r_1, r_2, \ldots, r_n$)). In other words, $B \leq A$ (since the entries of B are the numbers $\dfrac{r_i}{r_u} A_{u,v}$, whereas the

[56] This is well-defined, since $r_u \neq 0$ (because all n numbers $r_1, r_2, \ldots, r_n$ are nonzero).

corresponding entries of A are $A_{u,v}$). Hence, Corollary 6.2.5 (applied to B and A instead of A and B) yields $\rho(B) \leq \rho(A)$. Thus, (123) becomes $r_i = \rho(B) \leq \rho(A)$. This proves Theorem 6.2.13. □

Corollary 6.2.14. Let $A \in \mathbb{R}^{n \times n}$ satisfy $A \geq 0$ and $n > 0$. Let $x_1, x_2, \ldots, x_n$ be any n positive reals. Then,

$$\min_{i \in [n]} \sum_{j=1}^{n} \frac{x_i}{x_j} A_{i,j} \leq \rho(A) \leq \max_{i \in [n]} \sum_{j=1}^{n} \frac{x_i}{x_j} A_{i,j}.$$

Proof. Let $D = \operatorname{diag}(x_1, x_2, \ldots, x_n)$. Then, DAD^{-1} is the $n \times n$-matrix whose (i,j)-th entry is $x_i A_{i,j} x_j^{-1} = \dfrac{x_i}{x_j} A_{i,j}$ for all $i, j \in [n]$. Thus, $DAD^{-1} \geq 0$ (since $A \geq 0$ and since $x_1, x_2, \ldots, x_n$ are positive). Hence, Theorem 6.2.13 (applied to DAD^{-1} instead of A) yields

$$\left(\text{the smallest row sum of } DAD^{-1}\right) \leq \rho\left(DAD^{-1}\right) \leq \left(\text{the largest row sum of } DAD^{-1}\right).$$

In view of $\rho\left(DAD^{-1}\right) = \rho(A)$ (which is a consequence of the fact that the matrices DAD^{-1} and A are similar and thus have the same spectrum), we can rewrite this as

$$\left(\text{the smallest row sum of } DAD^{-1}\right) \leq \rho(A) \leq \left(\text{the largest row sum of } DAD^{-1}\right).$$

Now, it remains only to notice that the row sums of DAD^{-1} are exactly the sums $\sum_{j=1}^{n} \dfrac{x_i}{x_j} A_{i,j}$ for $i \in [n]$. □

Remark 6.2.15. If the matrix A in Corollary 6.2.14 is positive, then there is a choice of $x_1, x_2, \ldots, x_n > 0$ such that both of the inequalities become equalities. (This follows from Theorem 6.3.2 **(c)** further below.)

Corollary 6.2.16. Let $A \in \mathbb{R}^{n \times n}$ satisfy $A \geq 0$ and $n > 0$.
Let $x \in \mathbb{R}^n$ satisfy $x > 0$.
Let α be a nonnegative real. Then:

(a) If $Ax \geq \alpha x$, then $\rho(A) \geq \alpha$.

(b) If $Ax > \alpha x$, then $\rho(A) > \alpha$.

(c) If $Ax \leq \alpha x$, then $\rho(A) \leq \alpha$.

(d) If $Ax < \alpha x$, then $\rho(A) < \alpha$.

Proof. Write x as $x = (x_1, x_2, \ldots, x_n)^T$. Then, the n numbers $x_1, x_2, \ldots, x_n$ are positive reals (since $x > 0$); hence, their reciprocals $1/x_1$, $1/x_2$, $\ldots$, $1/x_n$ are well-defined positive reals as well.

(a) Assume that $Ax \geq \alpha x$. Then, for each $i \in [n]$, we have

$$\begin{aligned}\sum_{j=1}^{n} A_{i,j} x_j &= (\text{the } i\text{-th entry of } Ax) \\ &\geq (\text{the } i\text{-th entry of } \alpha x) \qquad (\text{since } Ax \geq \alpha x) \\ &= \alpha x_i. \end{aligned} \tag{124}$$

However, Corollary 6.2.14 (applied to $1/x_k$ instead of x_k) yields

$$\min_{i\in[n]} \sum_{j=1}^{n} \frac{1/x_i}{1/x_j} A_{i,j} \leq \rho(A) \leq \max_{i\in[n]} \sum_{j=1}^{n} \frac{1/x_i}{1/x_j} A_{i,j}.$$

The first of these two inequalities yields

$$\begin{aligned}\rho(A) &\geq \min_{i\in[n]} \sum_{j=1}^{n} \underbrace{\frac{1/x_i}{1/x_j} A_{i,j}}_{= \frac{1}{x_i} A_{i,j} x_j} = \min_{i\in[n]} \sum_{j=1}^{n} \frac{1}{x_i} A_{i,j} x_j = \min_{i\in[n]} \frac{1}{x_i} \underbrace{\sum_{j=1}^{n} A_{i,j} x_j}_{\geq \alpha x_i \text{ (by (124))}} \\ &\geq \min_{i\in[n]} \underbrace{\frac{1}{x_i} \cdot \alpha x_i}_{=\alpha} = \min_{i\in[n]} \alpha = \alpha.\end{aligned}$$

This proves Corollary 6.2.16 **(a)**.

(b) The proof is analogous to the proof of Corollary 6.2.16 **(a)**, but uses $>$ signs instead of $\geq$ signs.

(c) The proof is similar to the proof of Corollary 6.2.16 **(a)**, but uses $>$ signs instead of $\geq$ signs and uses max instead of min.

(d) The proof is analogous to the proof of Corollary 6.2.16 **(c)**. □

Corollary 6.2.17. Let $A \in \mathbb{R}^{n\times n}$ satisfy $A > 0$ and $n > 0$ and $\rho(A) = 1$. Let $w \in \mathbb{R}^n$ satisfy $w \geq 0$ and $w \neq 0$. Then:

(a) We always have $Aw > 0$.

(b) If $Aw \geq w$, then $Aw = w > 0$.

Proof. Write the vector w as $w = (w_1, w_2, \ldots, w_n)^T$. Then, the numbers $w_1, w_2, \ldots, w_n$ are nonnegative reals (since $w \geq 0$). Moreover, at least one $k \in [n]$ satisfies $w_k \neq 0$ (since $w \neq 0$). Consider this k. Thus, $w_k > 0$ (since $w \geq 0$).

(a) For each $i \in [n]$, the i-th entry of Aw is $\sum_{j=1}^{n} A_{i,j} w_j$. This is a sum of nonnegative addends (since $A > 0$ and $w \geq 0$), and at least one of these addends is actually

positive (indeed, $A > 0$ entails $A_{i,k} > 0$ and thus $\underbrace{A_{i,k}}_{>0} \underbrace{w_k}_{>0} > 0$). Hence, this sum is positive. We have thus shown that for each $i \in [n]$, the i-th entry of Aw is positive. In other words, $Aw > 0$. This proves Corollary 6.2.17 **(a)**.

(b) Assume that $Aw \geq w$. Let $z := Aw - w$. Then, $z = Aw - w \geq 0$. However, Corollary 6.2.17 **(a)** also yields $Aw > 0$.

We claim that $z = 0$. Indeed, assume the contrary. Thus, $z \neq 0$. Hence, Corollary 6.2.17 **(a)** (applied to z instead of w) yields $Az > 0$. Therefore, $AAw > Aw$ (since $AAw - Aw = A\underbrace{(Aw - w)}_{=z} = Az > 0$). Also, $A \geq 0$ (since $A > 0$). Hence, Corollary 6.2.16 **(b)** (applied to $x = Aw$ and $\alpha = 1$) yields $\rho(A) > 1$, which contradicts $\rho(A) = 1$. This contradiction shows that our assumption was wrong. Hence, $z = 0$ is proved. Thus, $Aw = w$ (since $Aw - w = z = 0$). Hence, $w = Aw > 0$. This proves Corollary 6.2.17 **(b)**. □

6.3. Perron–Frobenius theorems

We now come to the most important results about nonnegative matrices: the Perron–Frobenius theorems.

6.3.1. Motivation

Let us first motivate the theorems using a less general (but more intuitive) setting.

Recall a standard situation in probability theory: Consider a system (e.g., a slot machine) that can be in one of n possible *states* $s_1, s_2, \ldots, s_n$. Every minute, the system randomly changes states according to the following rule: If the system is in state s_i, then it changes to state s_j with probability $P_{i,j}$, where P is a (fixed, predetermined) nonnegative $n \times n$-matrix whose row sums all equal 1 (such a matrix is called *row-stochastic*). This is commonly known as a *Markov chain*.

Given such a Markov chain, one often wonders about its "steady state": If you wait long enough, how likely is the system to be in a given state?

Example 6.3.1. Let $P = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$. This corresponds to a system that has two states s_1 and s_2, and the probability of going from state s_1 to state s_2 (any given minute) is 0.1, whereas the probability of going from state s_1 to state s_1 (that is, staying at state s_1) is 0.9, and the probability of going from state s_2 to either state is 0.5.

We encode the two states s_1 and s_2 as the basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$ of the vector space $\mathbb{R}^{1\times 2}$ (we work with row vectors here for convenience). Thus, a probability distribution on the set of states (i.e., a distribution of the form "state s_1 with probability a_1 and state s_2 with probability a_2") corresponds to a row vector $(a_1, a_2) \in \mathbb{R}^{1\times 2}$ satisfying $a_1 \geq 0$ and $a_2 \geq 0$ and $a_1 + a_2 = 1$.

If we start at state s_1 and let k minutes pass, then the probability distribution for the resulting state is $s_1 P^k$ (why?). More generally, if we start with a probability distribution $d \in \mathbb{R}^{1\times 2}$ and let k minutes pass, then the resulting state will be distributed according to dP^k (why?). So our question about the steady state can be rewritten as follows: What is $\lim\limits_{k\to\infty} dP^k$? Does this limit even exist?

We can notice one thing right away: If the limit $\lim\limits_{k\to\infty} dP^k$ exists, then this limit is a left 1-eigenvector of P, in the sense that it is a row vector y such that $yP = y$ (because if we set $y = \lim\limits_{k\to\infty} dP^k$, then we have $y = \lim\limits_{k\to\infty} dP^k = \lim\limits_{k\to\infty} d\underbrace{P^{k+1}}_{=P^k P} = \left(\lim\limits_{k\to\infty} dP^k\right) P = yP$). Since it is furthermore a probability distribution (because it is a limit of probability distributions), we can easily compute it (indeed, our matrix P has only one left 1-eigenvector up to scaling, and the scaling factor is uniquely determined by the requirement that it be a probability distribution). We obtain

$$\lim_{k\to\infty} dP^k = \left(\frac{5}{6}, \frac{1}{6}\right).$$

But does this limit actually exist? Yes: In our specific example, it does; but this isn't quite that obvious. Note that this limit (known as the *steady state* of the Markov chain) actually does not depend on the starting distribution d.

Does this generalize? Not always. Here are two examples where things go bad:

- If $P = I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, then $\lim\limits_{k\to\infty} dP^k = d$ for each d, so the limits do depend on d.
- If $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, then $\lim\limits_{k\to\infty} dP^k$ does not exist unless $d = (0.5,\ 0.5)$, since in all other cases the sequence $\left(dP^k\right)_{k\geq 0}$ oscillates between (a_1, a_2) and (a_2, a_1).

Perhaps surprisingly, such bad cases are an exception. For **most** row-stochastic matrices P (that is, nonnegative matrices whose row sums all equal 1), there is a **unique** steady state (i.e., left 1-eigenvector whose entries sum up to 1), and it can be obtained as $\lim\limits_{k\to\infty} dP^k$ for any starting distribution d. To be more precise, this holds whenever P is positive (i.e., all $P_{i,j} > 0$). Some weaker assumptions also suffice.

More general versions of these facts hold even if we don't assume P to be row-stochastic, but merely require $P > 0$ (or $P \geq 0$ with some extra conditions). These will be the Perron and Perron–Frobenius theorems.

6.3.2. The theorems

We can now state the Perron and Perron–Frobenius theorems; we will prove them later:

Theorem 6.3.2 (Perron theorem)**.** Let $A \in \mathbb{R}^{n \times n}$ satisfy $A > 0$ and $n > 0$. Then:

(a) We have $\rho(A) > 0$.

(b) The number $\rho(A)$ is an eigenvalue of A and has algebraic multiplicity 1 (and therefore geometric multiplicity 1 as well).

(c) There is a unique $\rho(A)$-eigenvector $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ of A with $x_1 + x_2 + \cdots + x_n = 1$. This eigenvector x is furthermore positive. (It is called the *Perron vector* of A.)

(d) There is a unique vector $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{C}^n$ such that $y^T A = \rho(A) y^T$ and $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$. This vector y is also positive.

(e) We have

$$\left(\frac{1}{\rho(A)} A \right)^m \to xy^T \qquad \text{as } m \to \infty.$$

(f) The only eigenvalue of A that has absolute value $\rho(A)$ is $\rho(A)$ itself.

We will prove this soon, but first we have two more theorems to state. The Perron theorem applies to positive matrices; but some parts of it can be adapted to the more general situation of a nonnegative matrix. If we require nothing other than nonnegativity, then only two statements hold:

Theorem 6.3.3 (Perron–Frobenius theorem 1)**.** Let $A \in \mathbb{R}^{n \times n}$ satisfy $A \geq 0$ and $n > 0$. Then:

(a) The number $\rho(A)$ is an eigenvalue of A.

(b) The matrix A has a nonzero nonnegative $\rho(A)$-eigenvector.

To get stronger statements without requiring $A > 0$, we need two further properties of A.

Definition 6.3.4. Let $A \in \mathbb{R}^{n \times n}$ be an $n \times n$-matrix with $n > 0$.

(a) We say that A is *reducible* if there exist two disjoint nonempty subsets I and J of $[n]$ such that $I \cup J = [n]$ and such that

$$A_{i,j} = 0 \qquad \text{for all } i \in I \text{ and } j \in J.$$

Equivalently, A is reducible if and only if there exists a permutation matrix $P \in \mathbb{R}^{n \times n}$ such that

$$P^{-1} A P = \begin{pmatrix} B & C \\ 0_{(n-r) \times r} & D \end{pmatrix} \qquad \text{for some } 0 < r < n$$

and some matrices B, C, D.

(Note that $P^{-1}AP$ is the matrix obtained from A by permuting the rows and then permuting the columns using the same permutation.)

(b) We say that A is *irreducible* if A is not reducible.

(c) We say that A is *primitive* if there exists some $m > 0$ such that $A^m > 0$.

Theorem 6.3.5 (Perron–Frobenius theorem 2)**.** Let $A \in \mathbb{R}^{n \times n}$ be nonnegative and irreducible and satisfy $n > 0$. Then:

(a) We have $\rho(A) > 0$.

(b) The number $\rho(A)$ is an eigenvalue of A and has algebraic multiplicity 1 (and therefore geometric multiplicity 1 as well).

(c) There is a unique $\rho(A)$-eigenvector $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ of A with $x_1 + x_2 + \cdots + x_n = 1$. This eigenvector x is furthermore positive. (It is called the *Perron vector* of A.)

(d) There is a unique vector $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{C}^n$ such that $y^T A = \rho(A) y^T$ and $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$. This vector y is also positive.

(e) Assume furthermore that A is primitive. We have

$$\left(\frac{1}{\rho(A)} A \right)^m \to x y^T \qquad \text{as } m \to \infty.$$

(f) Assume again that A is primitive. The only eigenvalue of A that has absolute value $\rho(A)$ is $\rho(A)$ itself.

Remark 6.3.6. If A is the row-stochastic matrix P corresponding to a Markov chain, then:

- A is irreducible if and only if there is no set of states from which you cannot escape (except for the empty set and for the set of all states);
- A is primitive if and only if there is an $m > 0$ such that we can get from any state to any state in exactly m minutes (this technical condition rules out the kind of "oscillation" that prevented us from finding a steady state for $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$).

6.3.3. Proof of Perron

We shall now approach the proof of the Perron theorem (Theorem 6.3.2). We begin with some notations:

Definition 6.3.7. For the rest of this subsection, we shall use the following notations:

Let n be a fixed positive integer. Let $e := (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Remark 6.3.8. **(a)** An $n \times n$-matrix A satisfies $Ae = e$ if and only if all row sums of A equal 1.

(b) An $n \times n$-matrix A satisfies $e^T A = e^T$ if and only if all column sums of A equal 1.

Proof of Remark 6.3.8. **(a)** This is because the entries of the column vector Ae are the row sums of A, while the entries of the column vector e are all 1.

(b) This is because the entries of the row vector $e^T A$ are the column sums of A, while the entries of the row vector e^T are all 1. □

Next, we state a few lemmas. The first one is an instance of the obvious idea that when some addends in a sum have different signs, they interfere destructively with each other:

Lemma 6.3.9. Let $y \in \mathbb{R}^n$ and $v = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ be two column vectors such that $y \geq 0$ and $y \neq 0$ and $y^T v = 0$. Let $a_1, a_2, \ldots, a_n$ be nonnegative reals. Then, there exists some **proper** subset K of $[n]$ such that

$$\left| \sum_{k=1}^{n} a_k v_k \right| \leq \sum_{k \in K} a_k \left| v_k \right|.$$

Proof. Let

$$P := \{k \in [n] \mid v_k > 0\} \qquad \text{and} \qquad N := \{k \in [n] \mid v_k < 0\}.$$

Then, P and N are two disjoint subsets of $[n]$, and every $k \in [n] \setminus (P \cup N)$ satisfies $v_k = 0$. Therefore, we can break up the sum $\sum_{k=1}^{n} a_k v_k$ as follows:

$$\begin{aligned} \sum_{k=1}^{n} a_k v_k &= \sum_{k \in P} a_k \underbrace{v_k}_{\substack{=|v_k| \\ \text{(since } k \in P \text{ and thus } v_k > 0\text{)}}} + \sum_{k \in N} a_k \underbrace{v_k}_{\substack{=-|v_k| \\ \text{(since } k \in N \text{ and thus } v_k < 0\text{)}}} \\ &= \sum_{k \in P} a_k \left| v_k \right| - \sum_{k \in N} a_k \left| v_k \right|. \end{aligned} \tag{125}$$

Note that both sums $\sum_{k \in P} a_k \left| v_k \right|$ and $\sum_{k \in N} a_k \left| v_k \right|$ are nonnegative (since $a_1, a_2, \ldots, a_n$ are nonnegative). However, it is easy to see that $|x - y| \leq \max\{x, y\}$ for any two nonnegative reals x and y. Applying this to $x = \sum_{k \in P} a_k \left| v_k \right|$ and $y = \sum_{k \in N} a_k \left| v_k \right|$, we thus obtain

$$\left| \sum_{k \in P} a_k \left| v_k \right| - \sum_{k \in N} a_k \left| v_k \right| \right| \leq \max\left\{ \sum_{k \in P} a_k \left| v_k \right|, \sum_{k \in N} a_k \left| v_k \right| \right\}.$$

Therefore,

$$\left| \sum_{k \in P} a_k \left| v_k \right| - \sum_{k \in N} a_k \left| v_k \right| \right| \leq \sum_{k \in K} a_k \left| v_k \right|, \tag{126}$$

where K is either P or N depending on whether $\sum_{k \in P} a_k \left| v_k \right| \geq \sum_{k \in N} a_k \left| v_k \right|$ or not. Consider this K. Clearly, K is a subset of $[n]$.

We shall now show that K is a **proper** subset of $[n]$. Indeed, assume the contrary. Thus, $K = [n]$. However, K is either P or N (by definition). We WLOG assume that $K = P$ (since the case $K = N$ is analogous). Thus, $P = K = [n]$. In other words, $v_k > 0$ for each $k \in [n]$ (by the definition of P). In other words, $v > 0$.

Write the vector y as $y = (y_1, y_2, \ldots, y_n)^T$. Then, at least one $k \in [n]$ satisfies $y_k \neq 0$ (since $y \neq 0$). Consider this k. Then, $y_k > 0$ (since $y \geq 0$).

However, $y^T v = \sum_{j=1}^{n} y_j v_j$. This is a sum of nonnegative reals (since $y \geq 0$ and $v > 0$), and at least one of its addends is actually positive (indeed, $\underbrace{y_k}_{>0} \underbrace{v_k}_{\substack{>0 \\ (\text{since } v>0)}} > 0$). Thus, the entire sum is positive. Therefore, $y^T v > 0$. But this contradicts $y^T v = 0$. This contradiction shows that our assumption was false. Hence, we have shown that K is a proper subset of $[n]$. Furthermore, from (125), we obtain

$$\left| \sum_{k=1}^{n} a_k v_k \right| = \left| \sum_{k \in P} a_k \left| v_k \right| - \sum_{k \in N} a_k \left| v_k \right| \right| \leq \sum_{k \in K} a_k \left| v_k \right|$$

(by (126)). This proves Lemma 6.3.9. □

The next three lemmas will help us derive some parts of Theorem 6.3.2 from others:

Lemma 6.3.10 (crucifix lemma, stochastic case)**.** Let $A \in \mathbb{R}^{n \times n}$ satisfy $A > 0$ and $Ae = e$. Let $y \in \mathbb{R}^n$ satisfy $y^T A = y^T$ and $y \geq 0$ and $y^T e = 1$ (that is, the sum of all entries of y is 1). Then,

$$A^m \to e y^T \qquad \text{as } m \to \infty.$$

Proof. Write the vector $y = (y_1, y_2, \ldots, y_n)^T$. Then, the entries $y_1, y_2, \ldots, y_n$ are nonnegative reals (since $y \geq 0$), and we have $y^T e = y_1 + y_2 + \cdots + y_n$ (since $e = (1, 1, \ldots, 1)^T$). Hence, $y_1 + y_2 + \cdots + y_n = y^T e = 1$.

Thus, the n numbers $y_1, y_2, \ldots, y_n$ are nonnegative reals whose sum is 1 (since $y_1 + y_2 + \cdots + y_n = 1$). Hence, all these n numbers $y_1, y_2, \ldots, y_n$ lie in the interval $[0, 1]$. In other words,

$$y_j \in [0, 1] \qquad \text{for each } j \in [n]. \tag{127}$$

From $Ae = e$, we conclude that all row sums of A equal 1 (by Remark 6.3.8 **(a)**). Since $A > 0$, this implies that all entries $A_{u,v}$ of A satisfy

$$0 < A_{u,v} \leq 1. \tag{128}$$

Let

$$\mu := 1 - \min\left\{A_{u,v} \mid u, v \in [n]\right\}.$$

Then, $0 \leq \mu < 1$ (by (128)). We claim the following:

Claim 1: For each $i \in [n]$ and each **proper** subset K of $[n]$, we have

$$\sum_{k \in K} A_{i,k} \leq \mu.$$

[*Proof of Claim 1:* Let $i \in [n]$. Let K be a proper subset of $[n]$. Then, there exists at least one element $j \in [n]$ satisfying $j \notin K$. Consider this j. Then, $\sum\limits_{\substack{k \in [n]; \\ k \notin K}} A_{i,k} \geq A_{i,j}$ (since $A \geq 0$ entails that all $A_{i,k}$ are nonnegative). However,

$$\sum_{k \in K} A_{i,k} = \underbrace{\sum_{k \in [n]} A_{i,k}}_{\substack{= (\text{the } i\text{-th row sum of } A) \\ = 1 \\ (\text{since all row sums of } A \text{ equal } 1)}} - \underbrace{\sum_{\substack{k \in [n]; \\ k \notin K}} A_{i,k}}_{\geq A_{i,j}} \leq 1 - \underbrace{A_{i,j}}_{\geq \min\{A_{u,v} \mid u,v \in [n]\}}$$

$$\leq 1 - v = \mu \qquad (\text{by the definition of } \mu).$$

This proves Claim 1.]

Next, we claim:

Claim 2: For any $i, j \in [n]$ and any $m \in \mathbb{N}$, we have

$$\left|\left(A^m - ey^T\right)_{i,j}\right| \leq \mu^m.$$

Once Claim 2 is proved, it will follow easily that $\left(A^m - ey^T\right)_{i,j} \to 0$ as $m \to \infty$ (because $0 \leq \mu < 1$), so that $A^m \to ey^T$, and the lemma will thus follow.

[*Proof of Claim 2:* We induct on m:

Base case: For any $i, j \in [n]$, we have

$$\left(A^0 - ey^T\right)_{i,j} = \underbrace{\left(A^0\right)_{i,j}}_{\substack{= (I_n)_{i,j} \\ = \delta_{i,j}}} - y_j = \underbrace{\delta_{i,j}}_{\in \{0,1\}} - \underbrace{y_j}_{\substack{\in [0,1] \\ (\text{by (127)})}} \in [-1, 1]$$

and therefore $\left|\left(A^0 - ey^T\right)_{i,j}\right| \leq 1 = \mu^0$. In other words, Claim 2 holds for $m = 0$.

Induction step: Let $p \in \mathbb{N}$. Assume (as the induction hypothesis) that Claim 2 holds for $m = p$. We must now show that Claim 2 also holds for $m = p+1$.

Let

$$B := A^p - ey^T \qquad \text{and} \qquad C := A^{p+1} - ey^T.$$

Our induction hypothesis says that Claim 2 holds for $m = p$. In other words, for all $i, j \in [n]$, we have $\left|\left(A^p - ey^T\right)_{i,j}\right| \leq \mu^p$. In other words, for all $i, j \in [n]$, we have

$$\left|B_{i,j}\right| \leq \mu^p \tag{129}$$

(since $B = A^p - ey^T$).

Our goal is to show that Claim 2 holds for $m = p+1$. In other words, our goal is to show that for all $i, j \in [n]$, we have $\left|\left(A^{p+1} - ey^T\right)_{i,j}\right| \leq \mu^{p+1}$. In other words, our goal is to show that for all $i, j \in [n]$, we have $\left|C_{i,j}\right| \leq \mu^{p+1}$ (since $C = A^{p+1} - ey^T$).

Fix $i, j \in [n]$. Thus we must prove that $\left|C_{i,j}\right| \leq \mu^{p+1}$.

From $B = A^p - ey^T$, we obtain

$$AB = A\left(A^p - ey^T\right) = \underbrace{AA^p}_{=A^{p+1}} - \underbrace{Ae}_{=e} y^T = A^{p+1} - ey^T = C.$$

Hence, $C = AB$, so that (by the definition of the product of two matrices we have)

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j}. \tag{130}$$

For each $m \in \mathbb{N}$, we have

$$y^T A^m = y^T$$

(indeed, this is easily proved by induction on m, using the fact that $y^T A = y^T$). Applying this to $m = p$, we obtain $y^T A^p = y^T$.

Recall that $B_{\bullet,j}$ denotes the j-th column of the matrix B. We have

$$y^T \underbrace{B}_{=A^p - ey^T} = y^T\left(A^p - ey^T\right) = \underbrace{y^T A^p}_{=y^T} - \underbrace{y^T e}_{=1} y^T = y^T - y^T = 0,$$

and thus $y^T B_{\bullet,j} = 0$ (since $y^T B_{\bullet,j}$ is the j-th entry of the row vector $y^T B$). Also, from $y^T e = 1 \neq 0$, we obtain $y \neq 0$. Hence, Lemma 6.3.9 (applied to $B_{\bullet,j}$ and $B_{k,j}$ and $A_{i,k}$ instead of v and v_k and a_i) yields that there exists some **proper** subset K of $[n]$ such that

$$\left|\sum_{k=1}^{n} A_{i,k} B_{k,j}\right| \leq \sum_{k \in K} A_{i,k} \left|B_{k,j}\right|. \tag{131}$$

Consider this K. Now, from (130), we obtain

$$\begin{aligned}
\left|C_{i,j}\right| &= \left|\sum_{k=1}^{n} A_{i,k} B_{k,j}\right| \leq \sum_{k \in K} A_{i,k} \underbrace{\left|B_{k,j}\right|}_{\substack{\leq \mu^p \\ \text{(by (129),} \\ \text{applied to } k \text{ instead of } i\text{)}}} && \text{(by (131))} \\
&\leq \underbrace{\sum_{k \in K} A_{i,k}}_{\substack{\leq \mu \\ \text{(by Claim 1)}}} \mu^p && \text{(since } A_{i,k} \geq 0 \text{ for all } k\text{)} \\
&\leq \mu \mu^p = \mu^{p+1}.
\end{aligned}$$

As we explained above, this completes the induction step. Thus, Claim 2 is proved.]

Finishing the proof of Lemma 6.3.10 is now easy: We have $0 \leq \mu < 1$ and therefore $\mu^m \to 0$ as $m \to \infty$. Thus, Claim 2 shows that for any $i, j \in [n]$, we have

$$\left(A^m - ey^T\right)_{i,j} \to 0 \qquad \text{as } m \to \infty.$$

In other words, $A^m - ey^T \to 0$ as $m \to \infty$. In other words, $A^m \to ey^T$ as $m \to \infty$. This proves Lemma 6.3.10. □

Next, we generalize Lemma 6.3.10 by replacing e by an arbitrary positive vector x:

Lemma 6.3.11 (crucifix lemma, general case)**.** Let $A \in \mathbb{R}^{n \times n}$ satisfy $A > 0$. Let $x \in \mathbb{R}^n$ satisfy $Ax = x$ and $x > 0$. Let $y \in \mathbb{R}^n$ satisfy $y^T A = y^T$ and $y \geq 0$ and $y^T x = 1$. Then,

$$A^m \to xy^T \qquad \text{as } m \to \infty.$$

Proof. We can easily reduce this to Lemma 6.3.10.

Indeed, write the vector x as $x = (x_1, x_2, \ldots, x_n)^T$. Thus, $x_1, x_2, \ldots, x_n$ are positive reals (since $x > 0$). Let

$$D := \operatorname{diag}(x_1, x_2, \ldots, x_n).$$

Thus, we easily obtain $De = x$. Moreover, the matrix D is diagonal, so that $D^T = D$. Furthermore, we have

$$D^{-1} A D > 0$$

(since the (u, v)-th entry of the matrix $D^{-1}AD$ is $\underbrace{x_u^{-1}}_{>0} \underbrace{A_{u,v}}_{>0} \underbrace{x_v}_{>0} > 0$ for each $u, v \in [n]$) and

$$D^{-1} A D e = e$$

(since $A \underbrace{De}_{=x} = Ax = x = De$) and

$$\underbrace{(Dy)^T}_{\substack{=y^T D^T = y^T D \\ (\text{since } D^T = D)}} D^{-1} A D = y^T \underbrace{DD^{-1}}_{=I_n} AD = \underbrace{y^T A}_{=y^T} D = y^T D$$

and

$$(Dy)^T = y^T \underbrace{D^T}_{=D} = y^T D \geq 0$$

(since $y \geq 0$ and $D \geq 0$) and

$$\underbrace{(Dy)^T}_{=y^T D} e = y^T \underbrace{De}_{=x} = y^T x = 1.$$

Hence, we can apply Lemma 6.3.10 to $D^{-1}AD$ and Dy instead of A and y. We thus obtain that

$$\left(D^{-1}AD\right)^m \to e\,(Dy)^T \qquad \text{as } m \to \infty.$$

Since we have $\left(D^{-1}AD\right)^m = D^{-1}A^m D$ for each $m \in \mathbb{N}$ (in fact, this is essentially the equality (30) we proved long ago), we can rewrite this as

$$D^{-1}A^m D \to e\,(Dy)^T \qquad \text{as } m \to \infty$$

Multiplying both sides by D from the left and by D^{-1} from the right, we can transform this into

$$A^m \to De\,(Dy)^T D^{-1} \qquad \text{as } m \to \infty.$$

In other words,

$$A^m \to xy^T \qquad \text{as } m \to \infty$$

(since $\underbrace{De}_{=x} \underbrace{(Dy)^T}_{=y^T D} D^{-1} = xy^T \underbrace{DD^{-1}}_{=I_n} = xy^T$). This proves Lemma 6.3.11. □

Lemma 6.3.12. Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times n}$ be two matrices such that $\rho(A) = 1$ and $\operatorname{rank} B \leq 1$. Assume that

$$A^m \to B \qquad \text{as } m \to \infty. \tag{132}$$

Then:

(a) The number 1 is an eigenvalue of A and has algebraic multiplicity 1 (and therefore geometric multiplicity 1 as well).

(b) There is at most one 1-eigenvector $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ of A with $x_1 + x_2 + \cdots + x_n = 1$.

(c) There is at most one vector $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{C}^n$ such that $y^T A = y^T$ and $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$.

(d) The only eigenvalue of A that has absolute value 1 is 1.

Proof. We know that A has a Schur triangularization. Let (U, T) be a Schur triangularization of A. Then, $U \in \mathbb{C}^{n\times n}$ is unitary and $T \in \mathbb{C}^{n\times n}$ is upper-triangular and $A = UTU^*$. Since U is unitary, we have $U^* = U^{-1}$, so that $A = UT\underbrace{U^*}_{=U^{-1}} = UTU^{-1}$. Moreover, Proposition 2.3.6 shows that the diagonal entries of T are the eigenvalues of A (with their algebraic multiplicities). Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be these diagonal entries, in the order in which they appear on the diagonal of T. Thus, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A.

For each $m \in \mathbb{N}$, we have

$$\begin{aligned} A^m &= \left(UTU^{-1}\right)^m \qquad \left(\text{since } A = UTU^{-1}\right) \\ &= UT^mU^{-1} \end{aligned}$$

(this is essentially the equality (30) we proved long ago). Hence, our assumption (132) can be rewritten as follows:

$$UT^mU^{-1} \to B \qquad \text{as } m \to \infty. \tag{133}$$

Thus, T^m converges to a limit (namely, to $U^{-1}BU$) as $m \to \infty$. Therefore, each entry of T^m converges to a limit as $m \to \infty$. In particular, $\lim\limits_{m\to\infty} (T^m)_{i,i}$ is well-defined for each $i \in [n]$. However, since T is an upper-triangular matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$, its m-th power T^m (for each $m \in \mathbb{N}$) is an upper-triangular matrix with diagonal entries $\lambda_1^m, \lambda_2^m, \ldots, \lambda_n^m$. In particular, we have $(T^m)_{i,i} = \lambda_i^m$ for each $i \in [n]$. Therefore, $\lim\limits_{m\to\infty} \lambda_i^m$ is well-defined for each $i \in [n]$ (since we have shown that $\lim\limits_{m\to\infty} (T^m)_{i,i}$ is well-defined for each $i \in [n]$). This shows that

$$|\lambda_i| < 1 \text{ or } \lambda_i = 1 \qquad \text{for each } i \in [n]\,. \tag{134}$$

If we had $|\lambda_i| < 1$ for each $i \in [n]$, then we would have $\rho(A) < 1$ (since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A); but this would contradict $\rho(A) = 1$. Hence, we cannot have $|\lambda_i| < 1$ for each $i \in [n]$. According to (134), this shows that at least one $i \in [n]$ satisfies $\lambda_i = 1$. In other words, 1 is an eigenvalue of A (since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A). Moreover, (134) shows that the only eigenvalue of A that has absolute value 1 is 1. This proves Lemma 6.3.12 **(d)**.

Let k be the number of all $i \in [n]$ that satisfy $\lambda_i = 1$. The matrix $\lim\limits_{m\to\infty} T^m$ is an upper-triangular matrix whose diagonal entries are $\lim\limits_{m\to\infty} \lambda_1^m$, $\lim\limits_{m\to\infty} \lambda_2^m$, $\ldots$, $\lim\limits_{m\to\infty} \lambda_n^m$ (since each T^m is an upper-triangular matrix whose diagonal entries are $\lambda_1^m, \lambda_2^m, \ldots, \lambda_n^m$). In view of (134), we conclude that exactly k of these diagonal entries are nonzero (since $|\lambda_i| < 1$ entails $\lim\limits_{m\to\infty} \lambda_i^m = 0$, whereas $\lambda_i = 1$ entails $\lim\limits_{m\to\infty} \lambda_i^m = \lim\limits_{m\to\infty} 1^m = 1$). Therefore, the matrix $\lim\limits_{m\to\infty} T^m$ has rank $\geq k$ (since the rank of an upper-triangular matrix is always $\geq$ to the number of its diagonal entries that are nonzero[57]). Therefore, the matrix $U\left(\lim\limits_{m\to\infty} T^m\right)U^{-1}$ has rank $\geq k$ as well (since U is invertible, so

[57] This can be proved in several ways; for example, nonzero diagonal entries can be used to create a nonzero principal minor.

that $\operatorname{rank}\left(U\left(\lim_{m\to\infty} T^m\right)U^{-1}\right) = \operatorname{rank}\left(\lim_{m\to\infty} T^m\right)$). In other words, the matrix B has rank $\geq k$ (since (133) yields $U\left(\lim_{m\to\infty} T^m\right)U^{-1} = B$). In view of the assumption $\operatorname{rank} B \leq 1$, this entails $k \leq 1$. In other words, at most one $i \in [n]$ satisfies $\lambda_i = 1$ (since k is the number of all $i \in [n]$ that satisfy $\lambda_i = 1$). In other words, the algebraic multiplicity of the eigenvalue 1 of A is at most 1 (since $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A). Since this multiplicity is at least 1 (because we know that 1 is an eigenvalue of A), we thus conclude that this multiplicity is 1. This proves Lemma 6.3.12 **(a)**.

The geometric multiplicity of the eigenvalue 1 of A is therefore also 1. Hence, the 1-eigenvectors $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ of A form a 1-dimensional vector subspace of $\mathbb{C}^n$. Thus, at most one of these 1-eigenvectors x satisfies $x_1 + x_2 + \cdots + x_n = 1$. This proves Lemma 6.3.12 **(b)**.

The matrices A and A^T have the same characteristic polynomial, and thus have the same eigenvalues with the same algebraic multiplicities. Hence, the algebraic multiplicity of the eigenvalue 1 of A^T equals the algebraic multiplicity of the eigenvalue 1 of A. Since the latter multiplicity is 1, we thus conclude that the former is 1 as well. Hence, the geometric multiplicity of the eigenvalue 1 of A^T must also equal 1.

The vectors $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{C}^n$ satisfying $y^T A = y^T$ are the 1-eigenvectors of A^T (since the equality $y^T A = y^T$ is equivalent to $A^T y = y$ (because $y^T A = \left(A^T y\right)^T$)). Hence, they form a 1-dimensional vector subspace of $\mathbb{C}^n$ (since the geometric multiplicity of the eigenvalue 1 of A^T is 1). Thus, at most one of these vectors y satisfies $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$. This proves Lemma 6.3.12 **(c)**. □

Proving Theorem 6.3.2 is now an easy matter of combining lemmas:

Proof of Theorem 6.3.2. **(a)** We have $A > 0$. Thus, all the more, we have $A_{i,i} > 0$ for some $i \in [n]$. Hence, Corollary 6.2.6 **(c)** yields $\rho(A) > 0$. This proves Theorem 6.3.2 **(a)**.

This also shows that the matrix $\dfrac{1}{\rho(A)}A$ is well-defined. Moreover, $\dfrac{1}{\rho(A)}A > 0$ (since $A > 0$). The matrix $\dfrac{1}{\rho(A)}A$ has the same eigenvectors as A (with the same algebraic multiplicities), while the corresponding eigenvalues are those of A multiplied by $\dfrac{1}{\rho(A)}$. Hence, if we replace A by $\dfrac{1}{\rho(A)}A$, then the claim of Theorem 6.3.2 does not substantially change (i.e., it gets replaced by an equivalent claim). Thus, let us replace A by $\dfrac{1}{\rho(A)}A$. This replacement causes $\rho(A)$ to become 1 (since $\rho\left(\dfrac{1}{\rho(A)}A\right) = \dfrac{1}{\rho(A)}\rho(A) = 1$). Thus, we have $\rho(A) = 1$ now.

Next, we shall show that A has a positive 1-eigenvector. Indeed, from $\rho(A) = 1$, we see that A has an eigenvalue $\lambda \in \mathbb{C}$ with $|\lambda| = 1$. Consider this λ. Pick any

nonzero λ-eigenvector $z = (z_1, z_2, \ldots, z_n)^T \in \mathbb{C}^n$ of A. Thus, $Az = \lambda z$. Moreover, $z \neq 0$ and thus $|z| \neq 0$. Also, clearly, $|z| \geq 0$.

From $A > 0$, we obtain $A = |A|$. Hence,

$$\begin{aligned} \underbrace{A}_{=|A|} |z| = |A| \cdot |z| &\geq \left| \underbrace{Az}_{=\lambda z} \right| \qquad \text{(by Proposition 6.1.16 \textbf{(a)})} \\ &= |\lambda z| = \underbrace{|\lambda|}_{=1} \cdot |z| = |z| . \end{aligned}$$

Thus, Corollary 6.2.17 **(b)** (applied to $w = |z|$) yields $A|z| = |z| > 0$. In other words, $|z|$ is a positive 1-eigenvector of A.

We have thus constructed a positive 1-eigenvector of A. The same argument (applied to A^T instead of A) lets us construct a positive 1-eigenvector of A^T (since $A > 0$ entails $A^T > 0$, and since $\rho\left(A^T\right) = \rho(A) = 1$). Let these two eigenvectors be x and y (with x being the 1-eigenvector of A and y being the one of A^T). Thus, $x \in \mathbb{R}^n$ satisfies $Ax = x$ and $x > 0$, whereas $y \in \mathbb{R}^n$ satisfies $A^T y = y$ and $y > 0$.

Write the vectors x and y as $x = (x_1, x_2, \ldots, x_n)^T$ and $y = (y_1, y_2, \ldots, y_n)^T$.

By scaling x by an appropriately chosen real scalar[58], we can achieve $x_1 + x_2 + \cdots + x_n = 1$. So we WLOG assume that $x_1 + x_2 + \cdots + x_n = 1$.

Moreover, by scaling y by an appropriately chosen positive real scalar [59], we can achieve $y^T x = 1$ (without disturbing the properties $A^T y = y$ and $y > 0$). So we WLOG assume that $y^T x = 1$. In other words, $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$ (since $y^T x = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$).

By taking transposes on both sides of the equality $A^T y = y$, we obtain $y^T A = y^T$ (since $\left(A^T y\right)^T = y^T A$). Thus, Lemma 6.3.11 yields

$$A^m \to x y^T \qquad \text{as } m \to \infty. \tag{135}$$

This proves Theorem 6.3.2 **(e)** (since $\dfrac{1}{\rho(A)} A = A$ (because $\rho(A) = 1$)).

The matrix xy^T has rank

$$\operatorname{rank}\left(xy^T\right) \leq \operatorname{rank} x \leq 1$$

(since x is a column vector). Hence, we can apply Lemma 6.3.12 to $B = xy^T$ (because of (135)).

Lemma 6.3.12 **(a)** tells us that the number 1 is an eigenvalue of A and has algebraic multiplicity 1 (and therefore geometric multiplicity 1 as well). In view of $\rho(A) = 1$, this proves Theorem 6.3.2 **(b)**.

[58] namely, by the scalar $\dfrac{1}{x_1 + x_2 + \cdots + x_n}$

[59] Specifically, we must scale y by $\dfrac{1}{y^T x}$ (which is a well-defined positive real scalar, since $x > 0$ and $y > 0$ entail $y^T x > 0$).

Lemma 6.3.12 **(d)** tells us that the only eigenvalue of A that has absolute value 1 is 1. In view of $\rho(A) = 1$, this proves Theorem 6.3.2 **(f)**.

We already know that $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{C}^n$ is an 1-eigenvector of A with $x_1 + x_2 + \cdots + x_n = 1$. Moreover, Lemma 6.3.12 **(b)** tells us that there is at most one such 1-eigenvector; therefore, x is the only such 1-eigenvector. This proves Theorem 6.3.2 **(c)** (since $\rho(A) = 1$, and since we also know that x is positive).

We already know that $y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{C}^n$ is a vector such that $y^T A = y^T$ and $x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = 1$. Moreover, Lemma 6.3.12 **(c)** tells us that there is at most one such vector; therefore, y is the only such vector. This proves Theorem 6.3.2 **(c)** (since $\rho(A) = 1$, and since we know that y is positive). □

We shall now prove the Perron–Frobenius theorems.

[...]

References

[AigZie14] Martin Aigner, Günter M. Ziegler, *Proofs from the Book*, 6th edition, Springer 2018.

[AndDos10] Titu Andreescu, Gabriel Dospinescu, *Problems from the Book*, 2nd edition, XYZ Press 2010.

[AndDos12] Titu Andreescu, Gabriel Dospinescu, *Straight from the Book*, XYZ Press 2012.

[Bartle14] Padraic Bartlett, *Math 108b: Advanced Linear Algebra, Winter 2014*, 2014. `http://web.math.ucsb.edu/~padraic/ucsb_2013_14/math108b_w2014/math108b_w2014.html`

[Bourba74] Nicolas Bourbaki, *Algebra I: Chapters 1–3*, Addison-Wesley 1974.

[Bourba03] Nicolas Bourbaki, *Algebra II: Chapters 4–7*, Springer 2003.

[BoyDip12] William E. Boyce, Richard C. DiPrima, *Elementary Differential Equations*, 10th edition, Wiley 2012.

[ChaSed97] Gengzhe Chang, Thomas W. Sederberg, *Over and Over Again*, Anneli Lax New Mathematical Library **39**, The Mathematical Association of America 1997.

[Conrad] Keith Conrad, *Expository notes ("blurbs")*. `https://kconrad.math.uconn.edu/blurbs/`

[Edward05] Harold M. Edwards, *Essays in Constructive Mathematics*, Springer 2005. See `https://www.math.nyu.edu/faculty/edwardsh/eserrata.pdf` for errata.

[Edward95] Harold M. Edwards, *Linear Algebra*, Springer 1995.

[Elman20] Richard Elman, *Lectures on Abstract Algebra*, 28 September 2020. `https://www.math.ucla.edu/~rse/algebra_book.pdf`

[GalQua20] Jean Gallier and Jocelyn Quaintance, *Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Engineering*, 11 November 2020. `https://www.cis.upenn.edu/~jean/gbooks/geomath.html`

[Geck20] Meinolf Geck, *On Jacob's construction of the rational canonical form of a matrix*, Electronic Journal of Linear Algebra **36** (2020), pp. 177–182.

[GelAnd17] Răzvan Gelca, Titu Andreescu, *Putnam and Beyond*, 2nd edition, Springer 2017.

[Goodma15] Frederick M. Goodman, *Algebra: Abstract and Concrete*, edition 2.6, 1 May 2015.
http://homepage.math.uiowa.edu/~goodman/algebrabook.dir/book.2.6.pdf .

[Grinbe15] Darij Grinberg, *Notes on the combinatorial fundamentals of algebra*, 10 January 2019.
http://www.cip.ifi.lmu.de/~grinberg/primes2015/sols.pdf
The numbering of theorems and formulas in this link might shift when the project gets updated; for a "frozen" version whose numbering is guaranteed to match that in the citations above, see https://github.com/darijgr/detnotes/releases/tag/2019-01-10 .

[Grinbe19] Darij Grinberg, *Notes on linear algebra*, 4th December 2019.
http://www.cip.ifi.lmu.de/~grinberg/t/16f/lina.pdf

[Grinbe19] Darij Grinberg, *The trace Cayley-Hamilton theorem*, 14 July 2019.
https://www.cip.ifi.lmu.de/~grinberg/algebra/trach.pdf

[Grinbe21] Darij Grinberg, *An Introduction to Algebraic Combinatorics [Math 701, Spring 2021 lecture notes]*, 10 September 2021.
https://www.cip.ifi.lmu.de/~grinberg/t/21s/lecs.pdf

[Heffer20] Jim Hefferon, *Linear Algebra*, 4th edition 2020.
http://joshua.smcvt.edu/linearalgebra

[Ho14] Law Ka Ho, *Variations and Generalisations to the Rearrangement Inequality*, Mathematical Excalibur **19**, Number 3, pp. 1–2, 4.

[HorJoh13] Roger A. Horn, Charles R. Johnson, *Matrix analysis*, Cambridge University Press, 2nd edition 2013.
See https://www.cambridge.org/us/files/7413/7180/9643/Errata_HJ_Matrix_Analysis_2nd_ed.pdf and https://www.kth.se/social/files/5707c3bef2765428eba786d3/errata.pdf for errata.

[Hung07] Pham Kim Hung, *Secrets in Inequalities, volume 1*, GIL 2007.

[Ivanov08] Nikolai V. Ivanov, *Linear Recurrences*, 17 January 2008.
https://nikolaivivanov.files.wordpress.com/2014/02/ivanov2008arecurrence.pdf

[KDLM05] Zoran Kadelburg, Dusan Dukic, Milivoje Lukic and Ivan Matic, *Inequalities of Karamata, Schur and Muirhead, and some applications*, The teaching of mathematics **VIII** (2005), issue 1, pp. 31–45.

[Knapp16] Anthony W. Knapp, *Basic Algebra*, digital second edition 2016.
http://www.math.stonybrook.edu/~aknapp/download.html

[Korner20] T. W. Körner, *Where Do Numbers Come From?*, Cambridge University Press 2020.
See `https://web.archive.org/web/20190813160507/https://www.dpmms.cam.ac.uk/~twk/Number.pdf` for a preprint.
See `https://www.dpmms.cam.ac.uk/~twk/` for errata and solutions.

[LaNaSc16] Isaiah Lankham, Bruno Nachtergaele, Anne Schilling, *Linear Algebra As an Introduction to Abstract Mathematics*, 2016.
`https://www.math.ucdavis.edu/~anne/linear_algebra/mat67_course_notes.pdf`

[Li99] Kin-Yin Li, *Rearrangement Inequality*, Mathematical Excalibur **4**, Number 3, pp. 1–2, 4.

[LibLav15] Leo Liberti, Carlile Lavor, *Six mathematical gems from the history of distance geometry*, International Transactions in Operational Research 2015.
`https://doi.org/10.1111/itor.12170`

[Loehr14] Nicholas Loehr, *Advanced Linear Algebra*, CRC Press 2014.

[MaOlAr11] Albert W. Marshall, Ingram Olkin, Barry C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, 2nd Edition, Springer 2011.

[Markus83] Aleksei Ivanovich Markushevich, *Recursion sequences*, Mir Publishers, Moscow, 2nd printing 1983.

[Mate16] Attila Máté, *The Cayley-Hamilton Theorem*, version 28 March 2016.
`http://www.sci.brooklyn.cuny.edu/~mate/misc/cayley_hamilton.pdf`

[Melian01] María Victoria Melián, *Linear recurrence relations with constant coefficients*, 9 April 2001.
`http://matematicas.uam.es/~mavi.melian/CURSO_15_16/web_Discreta/recurrence.pdf`

[Nathan21] Melvyn B. Nathanson, *The Muirhead-Rado inequality, 1 Vector majorization and the permutohedron*, arXiv:2109.01746v1.

[OmClVi11] Kevin C. O'Meara, John Clark, Charles I. Vinsonhaler, *Advanced Topics in Linear Algebra: Weaving Matrix Problems through the Weyr Form*, Oxford University Press 2011.

[PolSze78] George Pólya, Gabor Szegő, *Problems and Theorems in Analysis I*, Springer 1978 (reprinted 1998).

[Prasol94] Viktor V. Prasolov, *Problems and Theorems in Linear Algebra*, Translations of Mathematical Monographs, vol. #134, AMS 1994.

[Shapir15] Helene Shapiro, *Linear Algebra and Matrices: Topics for a Second Course*, Pure and applied undergraduate texts **24**, AMS 2015.

[Shurma15] Jerry Shurman, *The Cayley-Hamilton theorem via multilinear algebra*, `http://people.reed.edu/~jerry/332/28ch.pdf` . Part of the collection *Course Materials for Mathematics 332: Algebra*, available at `http://people.reed.edu/~jerry/332/mat.html`

[Silves00] John R. Silvester, *Determinants of Block Matrices*, The Mathematical Gazette, Vol. 84, No. 501 (Nov., 2000), pp. 460–467.

[Steele04] J. Michael Steele, *The Cauchy–Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Cambridge University Press 2004. See `http://www-stat.wharton.upenn.edu/~steele/Publications/Books/CSMC/CSMC_Errata.pdf` for errata.

[Steinb06] Mark Steinberger, *Algebra*, 31 August 2006. `https://web.archive.org/web/20180821125315/https://www.albany.edu/~mark/algebra.pdf`

[Straub83] Howard Straubing, *A combinatorial proof of the Cayley-Hamilton theorem*, Discrete Mathematics, Volume 43, Issues 2–3, 1983, pp. 273–279. `https://doi.org/10.1016/0012-365X(83)90164-4`

[Strick20] Neil Strickland, *Linear mathematics for applications*, 11 February 2020. `https://neilstrickland.github.io/linear_maths/notes/linear_maths.pdf`

[Swanso20] Irena Swanson, *Introduction with Analysis with Complex Numbers*, 2020. `https://web.archive.org/web/20201012174324/https://people.reed.edu/~iswanson/analysisconstructR.pdf`

[Tao07] Terence Tao, *The Jordan normal form and the Euclidean algorithm*, 12 October 2007. `https://terrytao.wordpress.com/2007/10/12/the-jordan-normal-form-and-the-euclidean-algorithm/`

[Taylor20] Michael Taylor, *Linear Algebra*, AMS 2020. See `https://mtaylor.web.unc.edu/wp-content/uploads/sites/16915/2018/04/linalg.pdf` for a preprint.

[TreBau97] Lloyd Nicholas Trefethen, David Bau III, *Numerical linear algebra*, SIAM 1997. See `https://people.maths.ox.ac.uk/trefethen/text.html` for the first five sections ("lectures").

[Treil15] Sergei Treil, *Linear Algebra Done Wrong*, 2017. `https://www.math.brown.edu/~treil/papers/LADW/LADW.html`

[Walker87] Elbert A. Walker, *Introduction to Abstract Algebra*, Random House-/Birkhauser, New York, 1987.

[Woerde16] Hugo J. Woerdeman, *Advanced Linear Algebra*, CRC Press 2016.

[Zeilbe85] Doron Zeilberger, *A combinatorial approach to matrix algebra*, Discrete Mathematics 56 (1985), pp. 61–72.

[Zill17] Dennis G. Zill, *A First Course in Differential Equations with Modeling Applications*, Cengage 2017.