LINEAR ALGEBRA AND MATRICES

Shmuel Friedland, Mohsen Aliabadi

University of Illinois at Chicago, USA

The first named author dedicates this book to his wife Margaret

The second named author dedicates this book to the memory of his parents

Contents

Preface

Linear algebra and matrix theory, abbreviated here as LAMT, is a foundation for many advanced topics in mathematics, and an essential tool for computer sciences, physics, engineering, bioinformatics, economics, and social sciences. A first course in linear algebra for engineers is like a cook book, where various results are given with very little rigorous justifications. For mathematics students, on the other hand, linear algebra comes at a time when they are being introduced to abstract concepts and formal proofs as the foundations of mathematics. In this case, very little of fundamental concepts of LAMT are covered. For a second course in LAMT there are a number of options. One option is to study the numerical aspects of LAMT, as for example in the book [8]. A totally different option, as in the popular book [11], which views LAMT as a part of a basic abstract course in algebra.

This book is aimed to be an introductory course in LAMT for beginning graduate students and an advanced (second) course in LAMT for undergraduate students. Reconciling such a dichotomy was made possible thanks to more than a decade of teaching the subject by the first author, in the Department of Mathematics, Statistics and Computer Science, the University of Illinois at Chicago, to both graduate students, and to advanced undergraduate students.

In this book, we used the abstract notions and arguments to give the complete proof of the Jordan canonical form, and more generally, the rational canonical forms of square matrices over fields. Also, we provide the notion of tensor products of vector spaces and linear transformations. Matrices are treated in depth: stability of matrix iterations, the eigenvalue properties of linear transformations in inner product space, singular value decomposition, and mini-max characterizations of Hermitian matrices and non-negative irreducible matrices.

We now briefly outline out the contents of this book. There are six chapters. The first chapter is a survey of basic notions. Some sections in this chapter are from other areas of mathematics, as elementary set theory, analysis, topology, and combinatorics. These sections can be assigned to students for self-study. Other sections deal with basic facts in LAMT, which may be skipped if the students are already familiar with them. The second chapter is a brief introduction to tensor products of finite dimensional vector spaces, tensor products of linear transformations, and their representations as Kronecker product. This chapter can either be skipped or can be taught later in the course The third chapter is a rigorous exposition of the Jordan canonical form over an algebraically closed field (which is usually the complex numbers in the engineering world), and a rational canonical form for linear operators and matrices. Again, the section dealing with cyclic subspaces and rational canonical forms can be skipped without losing consistency. Chapter 4 deals with applications of the Jordan canonical form of matrices with real and complex entries. First, we discuss the precise expression of $f(A)$, where A is a square matrix and f is a polynomial, in terms of the components of A. We then discuss the extension of this formula to functions f which are analytic at a neighborhood of the spectrum of A. The instructor may choose one particular application to teach from this chapter. Fifth chapter, the longest one, is devoted to properties of inner product spaces, and special linear operators such as normal, Hermitian and unitary. We bring the min-max and max-min characterizations of the eigenvalues of Hermitian matrices, the singular value decomposition and its minimal low rank approximation properties.

The last chapter deals with basic aspects of the Perron-Frobenius theory, which are not usually found in a typical linear algebra book.

One of the main objectives of this book is to show the variety of topics and tools that modern linear algebra and matrix theory encompass. To facilitate the reading of this book we have a good number of worked-out problems, helping the reader, especially those preparing for an exam such as a graduate preliminary exam, to better-understand the notions and results discussed in each section. We also provide a number of problems for instructors to assign, as well, to complement the material.

Perhaps, it will be hard to cover all these topics in a one-semester graduate course. However, as many sections of these book are independent, the instructor may choose appropriate sections as needed.

August 2016

Shmuel Friedland
Mohsen Aliabadi

Chapter 1

Preliminaries

1.1 Basic facts in abstract algebra

1.1.1 Groups

A *binary operation* on a set A is a map which sends elements of the Cartesian product $A \times A$ to A. A *group* denoted by G, is a set of elements with a binary operation $\oplus$, i.e. $a \oplus b \in G$, for each $a, b \in G$. This operation is

(i) associative: $(a \oplus b) \oplus c = a \oplus (b \oplus c)$;

(ii) there exists a *neutral* element $\bigcirc$ such that $a \oplus \bigcirc = a$, for each $a \in G$;

(iii) for each $a \in G$, there exists a unique $\ominus a$ such that $a \oplus (\ominus a) = \bigcirc$.

The group G is called *abelian (commutative)* if $a \oplus b = b \oplus a$, for each $a, b \in G$. Otherwise, it is called *nonabelian (noncommutative)*.

Examples of abelian groups

1. The following subsets of complex numbers where $\oplus$ is the standard addition $+$, $\ominus$ is the standard subtraction $-$ and the neutral element is 0.

   (a) The set of integers $\mathbb{Z}$.

   (b) The set of rational numbers $\mathbb{Q}$.

   (c) The set of real numbers $\mathbb{R}$.

   (d) The set of complex numbers $\mathbb{C}$

2. The following subsets of $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$, i.e. all non-zero complex numbers, where the operation $\oplus$ is the standard *product*, the neutral element is 1, and $\ominus a$ is a^{-1}.

   (a) $\mathbb{Q}^* := \mathbb{Q}\backslash\{0\}$.

   (b) $\mathbb{R}^* := \mathbb{R}\backslash\{0\}$.

   (c) $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$.

An example of nonabelian group

The *quaternion group* is a group with eight elements, which is denoted as $(Q_8, \oplus)$ and defined as follows:
$Q_8 = \{1, -1, i, -i, j, -j, k, -k\}$, 1 is the identity element and
$(-1) \oplus (-1) = 1$
$(-1) \oplus i = -i$
$(-1) \oplus j = -j$
$(-1) \oplus k = -k$
$i \oplus j = k$
$j \oplus i = -k$
$j \oplus k = i$
$k \oplus j = -i$
$k \oplus i = j$
$i \oplus k = -j$,
and the remaining relations can be deduced from these. Clearly, Q_8 is not abelian as for instance $i \oplus j \neq j \oplus i$.

1.1.2 Rings

From now on, we will denote the operation $\oplus$ by $+$. An abelian group $\mathbf{R}$ is called a *ring*, if there exists a second operation, called *product*, denoted by ab , for any $a, b \in \mathbf{R}$, which satisfies:

(i) associativity: $(ab)c = a(bc)$;

(ii) distributivity: $a(b + c) = ab + ac$, $(b + c)a = ba + ca$, $(a, b, c \in \mathbf{R})$;

(iii) existence of identity $1 \in \mathbf{R}$: $1a = a1 = a$, for all $a \in \mathbf{R}$.

Also, $\mathbf{R}$ is called a *commutative ring* if $ab = ba$, for all $a, b \in \mathbf{R}$. Otherwise, $\mathbf{R}$ is called a *noncommutative* ring.

Examples of rings

1. For a positive integer $n > 1$, the set of $n \times n$ complex valued matrices, denoted by $\mathbb{C}^{n\times n}$, with the addition $A + B$, product AB, and with the identity I_n, the $n \times n$ identity matrix . (We will introduce the concept of a matrix in section 1.5.) Furthermore, the following subsets of $\mathbb{C}^{n\times n}$:

   (a) $\mathbb{Z}^{n\times n}$, the ring of $n \times n$ matrices with integer entries. (Noncommutative ring).

   (b) $\mathbb{Q}^{n\times n}$, the ring of $n \times n$ matrices with rational entries. (Noncommutative ring).

   (c) $\mathbb{R}^{n\times n}$, the ring of $n \times n$ matrices with real entries. (Noncommutative ring).

   (d) $\mathbb{C}^{n\times n}$. (Noncommutative ring).

   (e) $\mathbf{D}(n, S)$, the set of $n \times n$ diagonal matrices with entries in $S = \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$. (Commutative ring).

   (Diagonal matrices are square matrices in which their entries outside the main diagonal are zero.)

2. $\mathbb{Z}_m = \mathbb{Z}/(m\mathbb{Z})$, all integers *modulo* a positive integer m, with the addition and multiplication modulo m. Clearly, $\#\mathbb{Z}_m$, the number of elements in $\mathbb{Z}_m$, is m and $\mathbb{Z}_m$ can be identified with $\{0, 1, \dots, m-1\}$.

A *subring* of a ring $\mathbf{R}$ is a subset S of $\mathbf{R}$ which contains 1 and is closed under subtraction and multiplication.
An *ideal* is a special kind of subring. A subring I of $\mathbf{R}$ is a *left ideal* if $a \in I$, $r \in \mathbf{R}$ imply $ra \in I$. A *right ideal* is defined similarly. A *two-sided ideal* (or just an ideal) is both left and right ideal. That is, $a, b \in I$, $r \in \mathbf{R}$ imply $a - b, ar, ra \in I$. It can be shown that if $\mathbf{R}$ is a commutative ring and $a \in \mathbf{R}$, then the set $I = \{ra; r \in \mathbf{R}\}$ is an ideal of $\mathbf{R}$. This ideal is called the *principal ideal generated by* a and is denoted by $\langle a \rangle$. If I is an ideal of $\mathbf{R}$ and $r \in \mathbf{R}$, $r + I$ is defined as $\{r + \mathbf{x}; \mathbf{x} \in I\}$. Consider the set $\mathbf{R}/I$ of all cosets $a + I$, where $a \in \mathbf{R}$. On this set, we define addition and multiplication as follows:

$$\begin{aligned}(a + I) + (b + I) &= (a + b) + I, \\ (a + I)(b + I) &= ab + I.\end{aligned}$$

With these two operations, $\mathbf{R}/I$ is a ring called the *quotient ring* by I. (Why $\mathbf{R}/I$ is a ring?)

1.1.3 Fields and division rings

A commutative ring $\mathbf{R}$ is called a *field* if each non-zero element $a \in \mathbf{R}$ has a unique inverse, denoted by a^{-1} such that $aa^{-1} = 1$. A field is usually denoted by $\mathbb{F}$.
A ring $\mathbf{R}$ is called a *division ring* if any non-zero element $a \in \mathbf{R}$ has a unique inverse.

Examples of fields

1. $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, with the standard addition and product.

2. $\mathbb{Z}_m$, where m is a prime integer.

An example of a division ring - the quaternion ring

Let $Q_{\mathbb{R}} = \{(x_1, x_2, x_3, x_4) \mid x_i \in \mathbb{R}, i = 1, 2, 3, 4\}$. Define $+$ and $\cdot$ on $Q_{\mathbb{R}}$ as follows:

$$\begin{aligned}(x_1, x_2, x_3, x_4) + (y_1, y_2, y_3, y_4) &= (x_1 + y_1, x_2 + y_2, x_3 + y_3, x_4 + y_4) \\ (x_1, x_2, x_3, x_4) \cdot (y_1, y_2, y_3, y_4) &= (x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4, \\ &\quad x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3, \\ &\quad x_1y_3 + x_3y_1 + x_4y_2 - x_2y_4, \\ &\quad x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1).\end{aligned}$$

One can view $Q_{\mathbb{R}}$ as four dimensional space vector space over $\mathbb{R}$, consisting of vectors $\mathbf{x} = x_1 + x_2\mathbf{i} + x_3\mathbf{j} + x_4\mathbf{k}$. Then the product $\mathbf{x} \cdot \mathbf{y}$ is determined by the product $(\oplus)$ in the nonabelian group Q_8 introduced in §1.1.1.

From the definition of $+$ and $\cdot$, it follows that $+$ and $\cdot$ are binary operations on $Q_{\mathbb{R}}$. Now $+$ is associative and commutative because addition is associative and commutative in $\mathbb{R}$. We also note that $(0, 0, 0, 0) \in Q_{\mathbb{R}}$ is the additive identity

and if $(x_1, x_2, x_3, x_4) \in Q_{\mathbb{R}}$, then $(-x_1, -x_2, -x_3, -x_4) \in Q_{\mathbb{R}}$ and $-(x_1, x_2, x_3, x_4) = (-x_1, -x_2, -x_3, -x_4)$. Hence, $(Q_{\mathbb{R}}, +)$ is a commutative group. Similarly, $\cdot$ is associative and $(1, 0, 0, 0) \in Q_{\mathbb{R}}$ is the multiplicative identity.
Let $(x_1, x_2, x_3, x_4) \in Q_{\mathbb{R}}$ be a nonzero element. Then, $N = x_1^2 + x_2^2 + x_3^2 + x_4^2 \neq 0$ and $N \in \mathbb{R}$. Thus, $(x_1/N, -x_2/N, -x_3/N, -x_4/N) \in Q_{\mathbb{R}}$. It is verified that $(x_1/N, -x_2/N, -x_3/N, -x_4/N)$ is the multiplicative inverse of (x_1, x_2, x_3, x_4). Thus, $Q_{\mathbb{R}}$ is a division ring and is called the ring of *real quaternions*. However, $Q_{\mathbb{R}}$ is not commutative because

$$(0, 1, 0, 0) \cdot (0, 0, 1, 0) = (0, 0, 0, 1) \neq (0, 0, 0, -1) = (0, 0, 1, 0) \cdot (0, 1, 0, 0).$$

Therefore, $Q_{\mathbb{R}}$ is not a field.

1.1.4 Vector spaces, modules and algebras

An abelian group $\mathbf{V}$ is called a *vector space* over a field $\mathbb{F}$, if for any $a \in \mathbb{F}$ and $\mathbf{v} \in \mathbf{V}$ the product $a\mathbf{v}$ is an element in $\mathbf{V}$, and this operation satisfies the following properties:

$$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}, (a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}, (ab)\mathbf{u} = a(b\mathbf{u}) \text{ and } 1\mathbf{v} = \mathbf{v},$$

for all $a, b \in \mathbb{F}, \mathbf{u}, \mathbf{v} \in \mathbf{V}$.

A *module* M is a vector space over a ring. The formal definition is exactly as above, but we use a ring $\mathbf{R}$ instead of a field. In this case, M is called to be an *R-module.*

A ring $\mathbf{R}$ is called an *algebra* over $\mathbb{F}$ if $\mathbf{R}$ is a vector space over $\mathbb{F}$ with respect to the addition operation $+$. Denote by $\cdot$ the product in $\mathbf{R}$. Then for any $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in $\mathbf{R}$ and a, b in $\mathbb{F}$ the following equalities hold:

(i) $(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \mathbf{x} \cdot \mathbf{z} + \mathbf{y} \cdot \mathbf{z}$,

(ii) $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$,

(iii) $(a\mathbf{x}) \cdot (b\mathbf{y}) = (ab)(\mathbf{x} \cdot \mathbf{y})$.

The algebra $\mathbf{R}$ is called a *division algebra* if for any $\mathbf{y} \in \mathbf{R} \setminus \{0\}$, there exists exactly one element $\mathbf{z}$ in $\mathbf{R}$ such that $\mathbf{z} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{z} = 1$. In what follows we denote for simplicity $\mathbf{x} \cdot \mathbf{y}$ by $\mathbf{xy}$ and no ambiguity will arise.

Examples of vector spaces, modules and algebras

1. If M is a vector space over the field $\mathbb{F}$, then M is an $\mathbb{F}$-module.

2. Let $M = \mathbf{R}^{m \times n}$ be the set of all $m \times n$ matrices with entries in the ring $\mathbf{R}$. Then M is an $\mathbf{R}$-module, where addition is ordinary matrix addition and multiplication of the scalar c by matrix A means the multiplication of each entry of A by c.

3. In the above example, if we change the ring $\mathbf{R}$ to a field $\mathbb{F}$, then $\mathbf{V} = \mathbb{F}^{m \times n}$ would be a vector space over $\mathbb{F}$.

4. Every abelian group A is a $\mathbb{Z}$-module. Addition and subtraction are defined according to the group structure of A; the point is that we can multiply $\mathbf{x} \in A$ by the integer n. If $n > 0$, then $n\mathbf{x} = \mathbf{x} + \mathbf{x} + \cdots + \mathbf{x}$ (n times); if $n < 0$, then $n\mathbf{x} = -\mathbf{x} - \mathbf{x} - \cdots - \mathbf{x}$ ($|n|$ times, where $|\,|$ denotes the absolute value function.)

5. Every commutative ring $\mathbf{R}$ is an algebra over itself.

6. An arbitrary ring $\mathbf{R}$ is always a $\mathbb{Z}$-algebra.

7. If $\mathbf{R}$ is a commutative ring, then $\mathbf{R}^{n\times n}$, the set of all $n\times n$ matrices with entries in $\mathbf{R}$ is an $\mathbf{R}$-algebra.

1.1.5 More about groups

A set G is called to be a *semigroup* if it has a binary operation satisfying the condition $(ab)c = a(bc)$, for any $a, b, c \in G$. (Here, the product operation is replaced by $\oplus$.)

A subset H of G is called a *subgroup* of G if H also forms a group under the operation of G. The set of the integers $\mathbb{Z}$ is a subgroup of the group of rational numbers $\mathbb{Q}$ under ordinary addition.

A subgroup N of a group G is called a *normal subgroup* if for each element n in N and each g in G, the element gng^{-1} is still in N. We use the notation $N \lhd G$ to denote that N is a normal subgroup of G. For example, all subgroups N of an abelian group G are normal (why?).

The *center of a group* G, denoted by $Z(G)$, is the set of elements that commute with every element of G, i.e.

$$Z(G) = \{z \in G;\ zg = gz,\ \forall g \in G\}.$$

Clearly, $Z(G) \lhd G$.

For a subgroup H of a group G and an element $\mathbf{x}$ of G, define $\mathbf{x}H$ to be the set $\{\mathbf{x}h; h \in H\}$. A subset of G of the form $\mathbf{x}H$ is said to be a *left coset* of H (a right coset of H is defined similarly.) For a normal subgroup N of the group G, the *quotient group* of N in G, written G/N and read "G modulo N", is the set of cosets of N in G. It is easy to see that G/N is a group with the following operation:

$$(Na)(Nb) = Nab, \quad \text{for all } a, b \in G.$$

A group G is called *finitely generated* if there exist $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in G such that every $\mathbf{x}$ in G can be written in the form $\mathbf{x} = \mathbf{x}_{i_1}^{\pm 1} \cdots \mathbf{x}_{i_m}^{\pm 1}$, where $i_1, \ldots, i_m \in \{1, \ldots, n\}$ and m ranges over all positive integers. In this case, we say that the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a *generating set* of G.
A *cyclic group* is a finitely generated group, which is generated by a single element.
A *group homomorphism* is a map $\varphi : G_1 \to G_2$ between two groups G_1 and G_2 such that the group operation is preserved, i.e. $\varphi(\mathbf{xy}) = \varphi(\mathbf{x})\varphi(\mathbf{y})$, for any $\mathbf{x}, \mathbf{y} \in G_1$. A group homomorphism is called an *isomorphism* if it is bijective. If φ is an isomorphism, we say that G_1 is isomorphic to G_2 and we use the notation $G_1 \cong G_2$. It is easy to show that the isomorphisms of groups form an equivalence relation on the class of all groups. (See Section 1.2 about equivalence relation.)
The *kernel* of a group homomorphism $\varphi : G_1 \to G_2$ is denoted by $\ker \varphi$ and it is the set of all elements of G_1 which are mapped to the identity element of G_2. The

kernel is a normal subgroup of G_1. Also, the *image* of φ is denoted by $\mathrm{Im}\varphi$ and is defined as follows:

$$\mathrm{Im}\varphi = \{\mathbf{y} \in G_2; \exists \mathbf{x} \in G_1 \text{ such that } \varphi(\mathbf{x}) = \mathbf{y}\}.$$

Clearly, $\mathrm{Im}\varphi$ is a subgroup of G_2. (and even its normal subgroup). Also, the *cokernel* of φ is denoted by $\mathrm{coker}\varphi$ and defined as $\mathrm{coker}\varphi = G_2/\mathrm{Im}\varphi$.

Now, we are ready to give the three isomorphism theorems in the context of groups. The interested reader is referred to [11] to see the proofs and more details about these theorems.

First isomorphism theorem

Let G_1 and G_2 be two groups and $\varphi : G_1 \to G_2$ be a group homomorphism. Then, $G_1/\ker\varphi \cong \mathrm{Im}\varphi$. In particular, if φ is surjective, then $G_1/\ker\varphi$ is isomorphic to G_2.

Second isomorphism theorem

Let G a group and S be a subgroup of G and N be a normal subgroup of G. Then

1. The product $SN = \{sn; s \in S \text{ and } n \in N\}$ is a subgroup of G.

2. $S \cap N \triangleleft S$.

3. $SN/N \simeq S/S \cap N$.

Third isomorphism theorem

1. If $N \triangleleft G$ and K is a subgroup of G such that $N \subseteq K \subseteq G$, then K/N is a subgroup of G/N.

2. Every subgroup of G/N is of the form K/N, for some subgroup K of G such that $N \subseteq K \subseteq G$.

3. If K is a normal subgroup of G such that $N \subseteq K \subseteq G$, then K/N is a normal subgroup of G/N.

4. Every normal subgroup of G/N is of the form K/N, for some normal subgroup K of G such that $N \subseteq K \subseteq G$.

5. If K is a normal subgroup of G such that $N \subseteq K \subseteq G$, then the quotient group $(G/N)\big/(K/N)$ is isomorphic to G/K.

Assume that G is a finite group and H its subgroup. Then G is a disjoint union of cosets of H. Hence the order of H, $(\#H)$, divides the order of G - Lagrange's theorem. Note that this is the case for vector spaces, i.e. if V is a finite vector space (this should not be assumed as a finite dimensional vector space.), and W is its subspace, then $\#W$ divides $\#V$. We will define the concept of "dimension" for vector spaces in subsection 1.6.2.

1.1.6 The group of bijections on a set $\mathcal{X}$

Let $\mathcal{X},\mathcal{Y}$ be two sets. Then, $\phi:\mathcal{X}\to\mathcal{Y}$ is called a mapping , i.e. for each $x\in\mathcal{X}$, $\phi(x)$ is an element in $\mathcal{Y}$. Moreover, $\phi:\mathcal{X}\to\mathcal{X}$ is called the *identity* map if $\phi(x)=x$, for each $x\in\mathcal{X}$. The identity map is denoted as id or $\mathrm{id}_{\mathcal{X}}$. Let $\psi:\mathcal{Y}\to\mathcal{Z}$. Then, one defines the composition map $\psi\circ\phi:\mathcal{X}\to\mathcal{Z}$ as follows $(\psi\circ\phi)(x)=\psi(\phi(x))$. A map $\phi:\mathcal{X}\to\mathcal{Y}$ is called *bijection* if there exists $\psi:\mathcal{Y}\to\mathcal{X}$ such that $\psi\circ\phi=\mathrm{id}_{\mathcal{X}},\phi\circ\psi=\mathrm{id}_{\mathcal{Y}}$ and ψ is denoted as ϕ^{-1}. Denote by $\mathcal{S}(\mathcal{X})$ the set of all bijections of $\mathcal{X}$ onto itself. It is easy to show that $\mathcal{S}(\mathcal{X})$ forms a group under the composition, with the identity element $\mathrm{id}_{\mathcal{X}}$. Assume that $\mathcal{X}$ is a finite set. Then, any bijection $\psi\in\mathcal{S}(\mathcal{X})$ is called a permutation and $\mathcal{S}(\mathcal{X})$ is called a *permutation group.*

Let $\mathcal{X}$ be a finite set. Then $\mathcal{S}(\mathcal{X})$ has $n!$ elements if n is the number of elements in $\mathcal{X}$. Assume that $\mathcal{X}=\{x_1,\dots,x_n\}$. We can construct a bijection ϕ on $\mathcal{X}$ as follows:

(1) Assign one of the n elements of $\mathcal{X}$ to $\phi(x_1)$. (There are n possibilities for $\phi(x_1)$ in $\mathcal{X}$.)

(2) Assign one of the $n-1$ elements of $\mathcal{X}-\{\phi(x_1)\}$ to $\phi(x_2)$. (There are $n-1$ possibilities for $\phi(x_2)$ in $\mathcal{X}-\{\phi(x_1)\}$.)

$\vdots$

(n) Assign the one remaining element to $\phi(x_n)$. (There is only one possibility for $\phi(x_n)$.)

This method can generate $n(n-1)\cdots1=n!$ different bijections of $\mathcal{X}$.

1.1.7 More about rings

If $\mathbf{R}$ and $\mathbf{S}$ are two rings, then a *ring homomorphism* from $\mathbf{R}$ to $\mathbf{S}$ is a function $\varphi:\mathbf{R}\to\mathbf{S}$ such that

(i) $\varphi(a+b)=\varphi(a)+\varphi(b)$,

(ii) $\varphi(ab)=\varphi(a)\varphi(b)$,

for each $a,b\in\mathbf{R}$.
A ring homomorphism is called an *isomorphism* if it is bijective.
Note that we have three isomorphism theorems for rings which are similar to isomorphism theorems in groups. See [11] for more details.

Remark 1.1.1 *If $\varphi:(\mathbf{R},+,\cdot)\to(\mathbf{S},+,\cdot)$ is a ring homomorphism, then $\varphi:(\mathbf{R},+)\to(\mathbf{S},+)$ is a group homomorphism.*

Remark 1.1.2 *If $\mathbf{R}$ is a ring and $\mathbf{S}\subset\mathbf{R}$ is a subring, then the inclusion $i:\mathbf{S}\hookrightarrow\mathbf{R}$ is a ring homomorphism. (Why?)*

1.1.8 More about fields

Let $\mathbb{F}$ be a field. Then by $\mathbb{F}[x]$ we denote the set of all polynomials $p(x) = a_n x^n + \cdots + a_0$, where $x^0 \equiv 1$. p is called monic if $a_n = 1$. The degree of p, denoted as $\deg p$, is n if $a_n \neq 0$. $\mathbb{F}[x]$ is a commutative ring under the standard addition and product of polynomials. $\mathbb{F}[x]$ is called the polynomial ring in x over $\mathbb{F}$. Here 0 and 1 are the constant polynomials having value 0 and 1, respectively. Moreover $\mathbb{F}[x]$ does not have zero divisors, i.e. if $p(x)g(x) = 0$, then either p or q is zero polynomial. A polynomial $p(x) \in \mathbb{F}[x]$ is called *irreducible (primitive)* if the decomposition $p(x) = q(x)r(x)$ in $\mathbb{F}[x]$ implies that either $q(x)$ or $r(x)$ is a constant polynomial.

A *subfield* $\mathbb{F}$ of a ring $\mathbf{R}$ is a subring of $\mathbf{R}$ that is a field. For example, $\mathbb{Q}$ is a subfield of $\mathbb{R}$ under the usual addition. Note that $\mathbb{Z}_2$ is not a subfield of $\mathbb{Q}$, even though $\mathbb{Z}_2 = \{0, 1\}$ can be viewed as a subset of $\mathbb{Q}$ and both are fields. (Why is $\mathbb{Z}_2$ not a subfield of $\mathbb{Q}$?) If $\mathbb{F}$ is a subfield of a field $\mathbb{E}$, one also says that $\mathbb{E}$ is a *field extension* of $\mathbb{F}$, and one writes $\mathbb{E}/\mathbb{F}$ is a field extension. Also, if C is a subset of $\mathbb{E}$, we define $\mathbb{F}(C)$ to be the intersection of all subfields of $\mathbb{E}$ which contains $\mathbb{F} \cup C$. It is verified that $\mathbb{F}(C)$ is a field and $\mathbb{F}(C)$ is called the *subfield of* $\mathbb{E}$ *generated by* C over $\mathbb{F}$. In the case $C = \{a\}$, we simply use the notation $\mathbb{F}(a)$ for $\mathbb{F}(C)$.
If $\mathbb{E}/\mathbb{F}$ is a field extension and $\alpha \in \mathbb{E}$, then α is called to be *algebraic* over $\mathbb{F}$, if α is a root of some polynomial with coefficients in $\mathbb{F}$; otherwise α is called to be *transcendental* over $\mathbb{F}$. If $m(x)$ is a monic irreducible polynomial with coefficients in $\mathbb{F}$ and $m(x) = 0$, then $m(x)$ is called a *minimal polynomial* of α over $\mathbb{F}$.

Theorem 1.1.3 *If $\mathbb{E}/\mathbb{F}$ is a field extension and $\alpha \in \mathbb{E}$ is algebraic over $\mathbb{F}$, then*

1. *The element α has a minimal polynomial over $\mathbb{F}$.*
2. *Its minimal polynomial is determined uniquely.*
3. *If $f(\alpha) = 0$, for some non-zero $f(\mathbf{x}) \in \mathbb{F}[\mathbf{x}]$, then $m(\mathbf{x})$ divides $f(\mathbf{x})$.*

The interested reader is referred to [11] to see the proof of Theorem 1.1.3.

1.1.9 The characteristic of a ring

Give a ring R and a positive integer n. For any $x \in R$, by $n \cdot x$, we mean

$$n \cdot x = \underbrace{x + \cdots + x}_{n \text{ terms}}.$$

It may happen that for a positive integer c we have

$$c \cdot 1 = \underbrace{1 + \cdots + 1}_{c \text{ terms}} = 0.$$

For example in $\mathbb{Z}_m = \mathbb{Z}/(m\mathbb{Z})$, we have $m \cdot 1 = m = 0$. On the other hand in $\mathbb{Z}$, $c \cdot 1 = 0$ implies $c = 0$, and then no such positive integer exists.
The smallest positive integer c for which $c \cdot 1 = 0$ is called the *characteristic* of R. If no such number c exists, we say that R has characteristic zero. The characteristic of R is denoted by $\mathbf{char} R$. It can be shown that any finite ring is of non-zero characteristic. Also, it is proven that the characteristic of a field is either zero or

prime. (See Worked-out Problem 1.5.1-2.) Notice that in a field $\mathbb{F}$ with $\mathbf{char}\mathbb{F} = 2$, we have $2\mathbf{x} = 0$, for any $\mathbf{x} \in \mathbb{F}$. This property makes fields of characteristic 2 exceptional. For instance, see Theorem 1.11.3 or Worked-out Problem 1.11.4-3.

1.2 Basic facts in set theory

A *relation from a set A to a set B* is a subset of $A \times B$ where

$$A \times B = \{(a, b);\ a \in A \text{ and } b \in B\}.$$

A *relation on a set A* is a relation from A to A, i.e. a subset of $A \times A$. Given a relation R on A, i.e. $R \subseteq A \times A$, we write $x \sim y$ if $(x, y) \in R$.
An *equivalence relation on a set A* is a relation on A that satisfies the following properties:

(i) Reflexivity: For all $x \in A$, $x \sim x$,

(ii) Symmetricity: For all $x, y \in A$, $x \sim y$ implies $y \sim x$,

(iii) Transitivity: For all $x, y, z \in A$, $x \sim y$ and $y \sim z$ imply $x \sim z$.

If $\sim$ is an equivalence relation on A and $x \in A$, the set $E_x = \{y \in A;\ x \sim y\}$ is called the *equivalence class of x*. Another notation for the equivalence class E_x of x is $[x]$.
A collection of non-empty subsets $A_1, A_2, \ldots$ of A is called a *partition of A* if it has the following properties:

(i) $A_i \cap A_j = \varnothing$ if $i \neq j$,

(ii) $\underset{i}{\cup} A_i = A$.

The following fundamental results are about the connection between equivalence relation and a partition of a set. See [19] for their proofs and more details.

Theorem 1.2.1 *Given an equivalence relation on a set A. The set of distinct equivalence classes forms a partition of A.*

Theorem 1.2.2 *Given a partition of A into sets $A_1, \ldots, A_n$. The relation defined by "$x \sim y$ if and only if x and y belong to the same set A_i from the partition" is an equivalence relation on A.*

1.3 Basic facts in analysis

A *metric space* is an ordered pair (X, d), where X is a set and d is a metric on X, i.e. a function $d : X \times X \to \mathbb{R}$, such that for any $x, y, z \in X$, the following conditions hold:

(i) non-negativity
$d(x, y) \geq 0$,

(ii) identity of indiscernibles
$d(x, y) = 0$ if and only if $x = y$,

(iii) symmetry
$d(x,y) = d(y,x)$,

(iv) triangle inequality
$d(x,z) \leq d(x,y) + d(y,z)$.

For any point x in X, we define the *open ball of radius* $r > 0$ (where r is a real number.) *about* x as the set $B(x,r) = \{y \in X : d(x,y) < r\}$.
A subset U of X is called *open* if for every x in U, there exists an $r > 0$ such that $B(x,r)$ is contained in U. The complement of an open set is called *closed.*
A metric space X is called *bounded* if there exists some real number r, such that $d(x,y) \leq r$, for all x, y in X.

Example 1.3.1 *$(\mathbb{R}^n, d)$ is a metric space for $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. This metric space is well-known as Euclidean metric space. The Euclidean metric on $\mathbb{C}^n$ is the Euclidean metric on $\mathbb{R}^{2n}$, where $\mathbb{C}^n$ viewed as $\mathbb{R}^{2n} \equiv \mathbb{R}^n \times \mathbb{R}^n$.*

If $\{s_n\} \subseteq X$ is a sequence and $s \in X$, it is called s_n *converges* to s if the following condition holds:
"for any positive real ε, there exists N such that $d(s_n, s) < \varepsilon$, for any $n > N$". This is denoted by $s_n \to s$ or $\lim_{n\to\infty} s_n = s$.
For $X = \mathbb{R}$ the *limit inferior* of $\{s_n\}$ is denoted by $\liminf_{n\to\infty} s_n$ and is defined by $\liminf_{n\to\infty} s_n := \lim_{n\to\infty}\left(\inf_{m\geq n} s_m\right)$. Similarly, the *limit superior* of $\{s_n\}$ is denoted by $\limsup_{n\to\infty} s_n$ and defined by $\limsup_{n\to\infty} s_n := \lim_{n\to\infty}\left(\sup_{m\geq n} s_m\right)$. Note that $\liminf_{n\to\infty} s_n$ and $\limsup_{n\to\infty} s_n$ can take the values $\pm\infty$.
The subset Y of X is called *compact* if every sequence in Y has a subsequence that converges to a point in Y.
It can be shown that if $F \subseteq \mathbb{R}^n$ (or $F \subseteq \mathbb{C}^n$), the following statements are equivalent:

(i) F is compact.

(ii) F is closed and bounded.

The interested reader is referred [18] to see the proof of the above theorem.

The sequence $\{s_n\} \subseteq X$ is called *Cauchy* if the following condition holds:
"For any positive real number ε, there is a positive integer N such that for all natural numbers $m, n > N$, $d(x_n, x_m) < \varepsilon$".
It can be shown that every convergent sequence is a Cauchy sequence. But the converse is not true necessarily. For example if $X = (0, \infty)$ with $d(x,y) = |x - y|$, then $s_n = \frac{1}{n}$ is a Cauchy sequence in X but not convergent. (As s_n converges to 0 and $0 \notin X$.)
A metric space X in which every Cauchy sequence converges in X is called *complete.* For example, the set of real numbers is complete under the metric induced by the usual absolute value but the set of rational numbers is not. (Why?)
We end up this section with *Big O notation.*
In mathematics, it is important to get a handle on the approximation error. For example it is written $e^x = 1 + x + \dfrac{x^2}{2} + O(x^3)$, to express the fact that the error is smaller in an absolute value than some constant times x^3 if x is close enough to

0. For formal definition, suppose $f(x)$ and $g(x)$ are two functions defined on some subsets of real numbers. We write

$$f(x) = O(g(x)) \ , \ x \to 0$$

if and only if there exist positive constants ε and C such that

$$|f(x)| < C|g(x)| \ , \text{ for all } |x| < \varepsilon.$$

1.4 Basic facts in topology

A *topological space* is a set X together with a collection of its subsets T, whose elements are called *open sets*, satisfy

(i) $\varnothing \in T$,

(ii) $X \in T$,

(iii) The intersection of a finite number of sets in T is also in T,

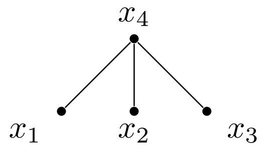
(iv) The union of an arbitrary number of sets in T is also in T.

(Here, $P(X)$ denotes the power set of X and $T \subset P(X)$.)
Note that a *closed* set is a set whose complement is an open set. Let (X, T) and (Y, T') be two topological spaces. A map $f : X \to Y$ is said to be *continuous* if $U \in T'$ implies $f^{-1}(U) \in T$, for any $U \in T'$.
The interested reader is encouraged to investigate the relation between continuity between metric spaces and topological spaces. In particular, the above definition inspires to check whether the inverse image of an open set under a continuous function is open or not, in metric space sense? We finish this section by defining path connectedness for topological spaces. A topological space X is said to be *path connected* if for any two points x_1 and $x_2 \in X$, there exists a continuous function $f : [0,1] \to X$ such that $f(0) = x_1$ and $f(1) = x_2$.
The reader is referred to [13] for more results on topological spaces.

1.5 Basic facts in graph theory

A *graph* G consists of a set V (or $V(G)$) of vertices, a set E (or $E(G)$) of edges, and a mapping associating to each edge $e \in E(G)$ an unordered pair x, y of vertices called the *ends* of e. The cardinality of $\mathbf{V}(G)$ is called the *order* of G. Also, the cardinality of $E(G)$ is called the *degree* of G. We say an edge is *incident* with its ends, and that is joints its ends. Two vertices are adjacent if they are jointed by a graph edge. The *adjacency matrix* , sometimes also called the *connection matrix* of a graph is a matrix with rows and columns labeled by graph vertices, with a 1 or 0 in position (v_i, v_j) according to whether v_i and v_j are adjacent or not. (Here v_i and v_j are two vertices of the graph.)

Example 1.5.1 *Consider the following graph:*



Two vertices x_1 and x_2 are not adjacent. Also, x_1 and x_4 are adjacent. Here, the adjacency matrix is

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

For the graph $G = (V, E)$, a *matching* M in G is a set of pairwise non-adjacent edges, that is, no two edges share a common vertex. A *perfect matching* is a matching which matches all vertices of the graph.
A *bipartite graph* (or bigraph) is a graph whose vertices can be divided into two disjoint sets V_1 and V_2 such that every edge connects a vertex in V_1 to one in V_2. Here, V_1 and V_2 are called *bipartite sets* of G. If $\#V_1 = \#V_2$, G is called *balanced bipartite.*
A *directed graph*, abbreviated as *digraph*, is denoted by $D = (V, E)$. V is the set of vertices and E is the set of directed edges, abbreviated as diedges, in G. So E is a subset of $V \times V = \{(v, w); v, w \in V\}$. Thus $(v, w) \in E$ is a directed edge from v to w. For example, the graph $D = ([4], \{(1, 2), (2, 1), (2, 3), (2, 4), (3, 3), (3, 4), (4, 1)\})$ has 4 vertices and 7 diedges.
The diedge $(v, v) \in E$ is called a *loop*, or selfloop.

$$\deg_{in} v := \#\{(w, v) \in E\}, \quad \deg_{out} v := \#\{(v, w) \in E\},$$

the number of diedges to v and out of v in D. Here, $\deg_{in}$, $\deg_{out}$ are called the *in* and *out* degrees, respectively. Clearly we have

$$\sum_{v \in V} \deg_{in} v = \sum_{v \in V} \deg_{out} v = \#E. \tag{1.5.1}$$

$v \in V$ is called *isolated* if $\deg_{in}(v) = \deg_{out}(v) = 0$.
A *multigraph* $G = (V, E)$ has undirected edges, which may be multiple, and may have multiple loops. A multidigraph $D = (V, E)$ may have multiple diedges.
Each multidigraph $D = (V, E)$ induces an undirected multigraph $G(D) = (V, E')$, where each *deidges* $(u, v) \in E$ is viewed as undirected edge $(u, v) \in E'$. (Each loop $(u, v) \in E$ will appear twice in E'.) Vice versa, a multigraph $G = (V, E')$ induces a multidigraph $D(G) = (V, E)$, where each undirected edge (u, v) induces diedges (u, v) and (v, u), when $u \neq v$. The loop (u, u) appears p times in $D(G)$ if it appears p times in G.
Most of the following notions are the same for graphs, digraphs, multigraphs or multidigraphs, unless stated otherwise. We state these notions for directed multidigraphs $D = (V, E)$ mostly.

Definition 1.5.2

1. *A walk in $D = (V, E)$ given by $v_0 v_1 \cdots v_p$, where $(v_{i-1}, v_i) \in E$ for $i = 1, \ldots, p$. One views it as a walk that starts at v_0 and ends at v_p. The length of the walk p, is the number of edges in the walk.*

2. *A path is a walk where $v_i \neq v_j$ for $i \neq j$.*

3. *A closed walk is walk where $v_p = v_0$.*

4. *A cycle is a closed walk where $v_i \neq v_j$, for $0 \le i < j < p$. A loop $(v, v) \in E$ is considered a cycle of length 1. Note that a closed walk vwv, where $v \neq w$, is considered as a cycle of length 2 in a digraph, but not a cycle in undirected multigraph!*

5. *A Hamiltonian cycle is a cycle through the graph that visits each node exactly once.*

6. *Two vertices $v, w \in V$, $v \neq w$ are called strongly connected if there exist two walks in D, the first starts at v and ends in w, and the second starts in w and ends in v. For multigraphs $G = (V, E)$ the corresponding notion is u, v are connected.*

7. *A multidigraph $D = ([n], E)$ is called strongly connected if either $n = 1$ and $(1, 1) \in E$, or $n > 1$ and any two vertices in D are strongly connected.*

8. *A multidiraph $g = (V, E)$ is called connected if either $n = 1$, or $n > 1$ and any two vertices in G are connected. (Note that a simple graph on one vertex $G = ([1], \varnothing)$ is considered connected directed graph $D(G) = G$ is not strongly connected.)*

9. *Two multidigraphs $D_1 = (V_1, E_1)$, $D_2 = (V_2, E_2)$ are called isomorphic if there exists a bijection $\phi : V_1 \to V_2$ which induces a bijection $\hat{\phi} : E_1 \to E_2$. That is if $(u_1, v_1) \in E_1$ is a diedge of multiplicity k in E_1, then $(\phi(u_1), \phi(v_1)) \in E_2$ is a diedge of multiplicity k and vice versa.*

The interested reader is referred to [3] to see more details about the above concepts.

1.5.1 Worked-out Problems

1. Show that the number e (the base of the natural logarithm) is transcendental over $\mathbb{Q}$.
Solution:
Assume that $I(t) = \int_0^t e^{t-u} f(u) du$, where t is a complex number and $f(\mathbf{x})$ is a polynomial with complex coefficients to be specified later. If $f(\mathbf{x}) = \sum_{j=0}^n a_j \mathbf{x}^j$, we set $\bar{f}(\mathbf{x}) = \sum_{j=0}^n |a_j| \mathbf{x}^j$, where $||$ denotes the norm of complex numbers. Integration by parts gives

$$I(t) = e^t \sum_{j=0}^{\infty} f^{(j)}(0) - \sum_{j=0}^{\infty} f^{(j)}(t) = e^t \sum_{j=0}^{n} f^{(j)}(0) - \sum_{j=0}^{n} f^{(j)}(t), \qquad (1.5.2)$$

where n is the degree of $f(\mathbf{x})$.
Assume that e is a root of $g(\mathbf{x}) = \sum_{i=0}^r b_i \mathbf{x}^i \in \mathbb{Z}[X]$, where $b_0 \neq 0$. Let p be a

prime greater than $\max\{r, |b_0|\}$, and define $f(\mathbf{x}) = \mathbf{x}^{p-1}(\mathbf{x}-1)^p(\mathbf{x}-2)^p \cdots (\mathbf{x}-r)^p$. Consider $J = b_0 I(0) + b_1 I(1) + \cdots + b_r I(r)$.
Since $g(e) = 0$, the contribution of the first summand on the right-hand side of (1.5.2) to J is 0. Thus,

$$J = -\sum_{k=0}^{r}\sum_{j=0}^{n} b_k f^{(j)}(k),$$

where $n = (r+1)p - 1$. The definition of $f(\mathbf{x})$ implies that many of the terms above are zero, and we can write

$$J = -\sum_{j=p-1}^{n} b_0 f^{(j)}(0) + \sum_{k=1}^{r}\sum_{j=p}^{n} b_k f^{(j)}(k).$$

Each of the terms on the right is divisible by $p!$ except for

$$f^{p-1}(0) = (p-1)!(-1)^{rp}(r!)^p,$$

where we have used that $p > r$. Thus, since $p > |b_0|$ as well, we see that J is an integer which is divisible by $(p-1)!$, but not by p. That is, J is an integer with

$$|J| \geq (p-1)!.$$

Since

$$\bar{f}(k) = k^{p-1}(k+1)^p(k+2)^p \cdots (k+r)^p \leq (2r)^n \qquad \text{for } 0 \leq k \leq r,$$

we deduce that

$$|J| \leq \sum_{j=0}^{r} |b_j||I(j)| \leq \sum_{j=0}^{r} |b_j| j e^j \bar{f}(j) \leq c\left((2r)^{(r+1)}\right)^p,$$

where c is a constant independent of p. This gives a contradiction.

2. Prove the following statements.

   (a) Any finite ring R is of non-zero characteristic.

   (b) The characteristic of a field $\mathbb{F}$ is either zero or prime.

   Solution:

   (a) Assume to the contrary $\mathbf{char}R = 0$. Without loss of generality, we may assume that $0 \neq 1$. For any $m, n \in \mathbb{Z}$, $m \cdot 1 = n \cdot 1$ implies $(m-n) \cdot 1 = 0$ and since we assumed $\mathbf{char}R = 0$, it follows $m = n$. Therefore, the set $\{n \cdot 1; n \in \mathbb{Z}\}$ is an infinite subset of R which is a contradiction.

   (b) Assume that $m = \mathbf{char}\mathbb{F}$ is non-zero and $m = nk$; then $0 = (nk) \cdot 1 = nk \cdot (1 \cdot 1) = (n \cdot 1)(k \cdot 1)$. Then, either $n = 0$ or $k = 0$ and this implies either $n = m$ or $k = m$. Hence, m is prime.

1.5.2 Problems

1. Let $\mathbb{Z}_6$ be the set of all integers modulo 6. Explain why this is not a field.

2. Assume that G and H are two groups. Let $\mathrm{Hom}(G,H)$ denote the set of all group homomorphisms from G to H. Show that

   (a) $\mathrm{Hom}(G,H)$ is a group under the composition of functions if H is a subgroup of G.

   (b) $\#\mathrm{Hom}(\mathbb{Z}_m,\mathbb{Z}_n) = \gcd(m,n)$, where $m,n \in \mathbb{N}$, where $\#$ denotes the cardinality of the set.

3. Prove that $\varphi : Q \to Q^{n\times n}$ by $\varphi(a) = \begin{bmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a \end{bmatrix}$ is a ring homomorphism.

1.6 Dimension and basis

As most algebraic structures, vector spaces contain substructures called subspace. Let $\mathbf{V}$ be a vector space over a field $\mathbb{F}$. (For simplicity of the exposition you may assume that $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.) A subset $\mathbf{U} \subset \mathbf{V}$ is called a *subspace* of $\mathbf{V}$, if $a_1\mathbf{u}_1 + a_2\mathbf{u}_2 \in \mathbf{U}$, for each $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{U}$ and $a_1, a_2 \in \mathbb{F}$. The subspace $\mathbf{U} := \{\mathbf{0}\}$ is called the *zero*, or the *trivial* subspace of $\mathbf{V}$. The elements of $\mathbf{V}$ are called *vectors*. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{V}$ be n vectors. Then, for any $a_1, \ldots, a_n \in \mathbb{F}$, called *scalars*, $a_1\mathbf{x}_1 + \ldots + a_n\mathbf{x}_n$ is called a *linear combinations* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Moreover, $\mathbf{0} = \sum_{i=1}^n 0\mathbf{x}_i$ is the trivial combination. Also, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *linearly independent*, if the equality $\mathbf{0} = \sum_{i=1}^n a_i\mathbf{x}_i$ implies that $a_1 = \ldots = a_n = 0$. Otherwise, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *linearly dependent*. The set of all linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is called the *span* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and denoted by $\mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. In particular, we use the notation $\mathrm{span}(\mathbf{x})$ to express $\mathrm{span}\{\mathbf{x}\}$, for any $\mathbf{x} \in V$. Clearly, $\mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a subspace of $\mathbf{V}$. (Why?) In general, the span of a set $\mathbf{S} \subseteq \mathbf{V}$ is the set of all linear combinations of $\mathbf{S}$, i.e. the set of all linear combinations of all finite subsets of $\mathbf{S}$.

1.6.1 More details on vector space

If V is a vector space, it is called *finitely generated* if $\mathbf{V} = \mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, for some vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a *spanning set* of $\mathbf{V}$. (In this book we consider only finitely generated vector spaces.) We will use the following notation for a positive integer n:

$$[n] := \{1, 2, \ldots, n-1, n\} \tag{1.6.1}$$

Observe the following "obvious" relation for any $n \in \mathbb{N}$. (Here, $\mathbb{N}$ is the set of all positive integers $1, 2, \ldots,$.)

$$\mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n\}\} \subseteq \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}, \text{ for each } i \in [n]. \tag{1.6.2}$$

It is enough to consider the case $i = n$. (We assume here that for $n = 1$, the span of empty set is the zero subspace $\mathbf{U} = \{\mathbf{0}\}$.) Then, for $n = 1$, (1.6.2) holds. Suppose that $n > 1$. Clearly, any linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$, which is $\mathbf{u} = \sum_{i=1}^{n-1} b_i\mathbf{u}_i$ is a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_n$: $\mathbf{u} = 0\mathbf{u}_n + \sum_{i=1}^{n-1} b_i\mathbf{u}_i$. Hence, (1.6.2) holds.

Theorem 1.6.1 *Suppose that $n > 1$. Then, there exists $i \in [n]$ such that*

$$\mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\} = \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n\}, \tag{1.6.3}$$

if and only if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly dependent vectors.

Proof. Suppose that (1.6.3) holds. Then, $\mathbf{u}_i = b_1\mathbf{u}_1 + \cdots + b_{i-1}\mathbf{u}_{i-1} + b_{i+1}\mathbf{u}_{i+1} + \cdots + b_n\mathbf{u}_n$. Hence, $\sum_{i=1}^n a_i\mathbf{u}_i = \mathbf{0}$, where $a_j = b_j$, for $j \neq i$ and $a_i = -1$. Then, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly dependent. Assume that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly dependent. Then, there exist scalars $a_1, \ldots, a_n \in \mathbb{F}$, not all of them are equal to zero, so that $a_1\mathbf{u}_1 + \ldots + a_n\mathbf{u}_n = 0$. Assume that $a_i \neq 0$. For simplicity of notation, (or by renaming indices in $[n]$), we may assume that $i = n$. This yields $\mathbf{u}_n = -\frac{1}{a_n}\sum_{i=1}^{n-1} a_i\mathbf{u}_i$. Let $\mathbf{u} \in \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Therefore,

$$\mathbf{u} = \sum_{i=1}^n b_i\mathbf{u}_i = b_n\mathbf{u}_n + \sum_{i=1}^{n-1} b_i\mathbf{u}_i = -b_n\sum_{i=1}^{n-1}\frac{a_i}{a_n}\mathbf{u}_i + \sum_{i=1}^{n-1} b_i\mathbf{u}_i = \sum_{i=1}^{n-1}\frac{a_nb_i - a_ib_n}{a_n}\mathbf{u}_i.$$

That is, $\mathbf{u} \in \{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}\}$. This proves the theorem. □

Corollary 1.6.2 *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbf{V}$. Assume that not all $\mathbf{u}_i$'s are zero vectors. Then, there exist $d \geq 1$ integers $1 \leq i_1 < i_2 < \ldots < i_d \leq n$ such that $\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_d}$ are linearly independent and $\mathrm{span}\{\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_d}\} = \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.*

Proof. Suppose that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly independent. Then, $d = n$ and $i_k = k$, for $k = 1, \ldots, n$ and we are done.

Assume that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly dependent. Applying Theorem 1.6.1, we consider now $n-1$ vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n$ as given by Theorem 1.6.1. Note that it is not possible that all vectors in $\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n\}$ are zero since this will imply that $\mathbf{u}_i = \mathbf{0}$. This will contradict the assumption that not all $\mathbf{u}_i$ are zero. Apply the previous arguments to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n$ and continue in this method until one gets d linearly independent vectors $\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_d}$. □

Corollary 1.6.3 *Let $\mathbf{V}$ be a finitely generated nontrivial vectors space, i.e. contains more than one element. Then, there exist $n \in \mathbb{N}$ and n linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ such that $\mathbf{V} = \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.*

The linear equation $\sum_{i=1}^n a_i\mathbf{x}_i = b$ with unknowns $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is called *homogeneous* if $b = 0$. Otherwise, it is called *non-homogeneous*. In the following lemma we use the fact that any m homogeneous linear equations with n variables have a nontrivial solution if $m < n$. (This will be proved later using REF of matrices.)

Lemma 1.6.4 *Let $n > m \geq 1$ be integers. Then, any $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ are linearly dependent.*

Proof. Observe that $\mathbf{w}_j = \sum_{i=1}^m a_{ij}\mathbf{u}_j$, for some $a_{ij} \in \mathbb{F}$, $1 \leq j \leq n$. Then

$$\sum_{j=1}^n x_j\mathbf{w}_j = \sum_{j=1}^n x_j\sum_{i=1}^m a_{ij}\mathbf{u}_i = \sum_{i=1}^m(\sum_{j=1}^n a_{ij}x_j)\mathbf{u}_i.$$

Consider the following m homogeneous equations in n unknowns: $\sum_{j=1}^{n} a_{ij}x_j = 0$, for $i = 1, \dots, m$. Since $n > m$, we have a nontrivial solution $(x_1, \dots, x_n)$. Thus, $\sum_{j=1}^{n} x_j\mathbf{w}_j = \mathbf{0}$. □

Theorem 1.6.5 *Let $\mathbf{V} = \operatorname{span}\{\mathbf{v}_1, \dots \mathbf{v}_n\}$ and assume that $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent. Then, the following holds.*

1. *Any vector $\mathbf{u}$ can be expressed as a unique linear combination of $\mathbf{v}_1, \dots, \mathbf{v}_n$.*

2. *For an integer $N > n$, any N vectors in $\mathbf{V}$ are linearly dependent.*

3. *Assume that $\mathbf{u}_1, \dots, \mathbf{u}_n \in \mathbf{V}$ are linearly independent. Then, $\mathbf{V} = \operatorname{span}\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$.*

Proof. Assume that $\sum_{i=1}^{n} x_i\mathbf{v}_i = \sum_{i=1}^{n} y_i\mathbf{v}_i$. Thus, $\sum_{i=1}^{n}(x_i - y_i)\mathbf{v}_i = \mathbf{0}$. As $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent, it follows that $x_i - y_i = 0$, for $i = 1, \dots, n$. Hence, 1 holds.

Lemma 1.6.4 implies 2.

Suppose that $\mathbf{u}_1, \dots, \mathbf{u}_n$ are linearly independent. Let $\mathbf{v} \in \mathbf{V}$ and consider $n+1$ vectors $\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}$. Next, 2 implies that $\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}$ are linearly dependent. Thus, there exist $n+1$ scalars $a_1, \dots, a_{n+1}$, not all of them zero, such that $a_{n+1}\mathbf{v} + \sum_{i=1}^{n} a_i\mathbf{u}_i = \mathbf{0}$. Assume first that $a_{n+1} = 0$. Then, $\sum_{i=1}^{n} a_i\mathbf{u}_i = \mathbf{0}$. Since not all a_i's are zero, it follows that $\mathbf{u}_1, \dots, \mathbf{u}_n$ are linearly dependent, which contradicts our assumption. Hence, $a_{n+1} \neq 0$ and $\mathbf{v} = \sum_{i=1}^{n} \frac{-a_i}{a_{n+1}}\mathbf{u}_i$. □

1.6.2 Dimension

A vector space $\mathbf{V}$ over the field $\mathbb{F}$ is called to be *finite dimensional* if there is a finite subset $\{x_1, \dots, x_n\}$ of $\mathbf{V}$ such that $\mathbf{V} = \operatorname{span}\{x_1, \dots, x_n\}$. Otherwise, $\mathbf{V}$ is called *infinite dimensional.* The *dimension* of a trivial vector space, consisting of the zero vector, $\mathbf{V} = \{\mathbf{0}\}$, is zero. The dimension of a finite dimensional non-zero vector space $\mathbf{V}$ is the number of vectors in any spanning linearly independent set, namely if $\mathbf{V}$ contains an independent set of n vectors but contains no independent set of $n+1$ vectors, we say that $\mathbf{V}$ has dimension n. Also, the dimension of an infinite dimensional vector space is infinity. The dimension of $\mathbf{V}$ is denoted by $\dim_{\mathbb{F}} \mathbf{V}$ or simply $\dim \mathbf{V}$ if there is no ambiguity in the background field. In this book $\mathbf{V}$ is assumed a vector space over the field $\mathbb{F}$ unless stated otherwise. See Worked-out Problem 1.6.3-2 on the existence of basis for vector spaces. Assume that $\dim \mathbf{V} = n$. Suppose that $\mathbf{V} = \operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Then, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a *basis* of $\mathbf{V}$, i.e. each vector $\mathbf{x}$ can be uniquely expressed as $\sum_{i=1}^{n} a_i\mathbf{x}_i$. Thus, for each $\mathbf{x} \in \mathbf{V}$, one corresponds a unique column vector $\mathbf{a} = (a_1, \dots, a_n)^\top \in \mathbb{F}^n$, where $\mathbb{F}^n$ is the vector space on column vectors with n coordinates in $\mathbb{F}$. It will be convenient to denote $\mathbf{x} = \sum_{i=1}^{n} a_i\mathbf{x}_i$ by the equality $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]\mathbf{a}$. (Note that we use the standard way to multiply row by column.)

Assume now that $\mathbf{y}_1, \dots, \mathbf{y}_n$ are n vectors in $\mathbf{V}$. Then

$$\mathbf{y}_i = \sum_{j=1}^{n} y_{ji}\mathbf{x}_j,$$

for some $y_{ji} \in \mathbb{F}$, $1 \le i, j \le n$.

Denote by $Y = [y_{ji}]_{i=j=1}^{n}$ the $n \times n$ matrix with the element y_{ji} in j-th row and i-th column. The above equalities are equivalent to the identity

$$[\mathbf{y}_1, \dots, \mathbf{y}_n] = [\mathbf{x}_1, \dots, \mathbf{x}_n]Y. \tag{1.6.4}$$

Then, $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ is a basis if and only if Y is invertible matrix. (See subsection 1.7.5 for details on invertible matrices.) In this case, Y is called the *matrix of the change of basis* from $[\mathbf{y}_1, \dots, \mathbf{y}_n]$ to $[\mathbf{x}_1, \dots, \mathbf{x}_n]$. The matrix of the change of basis from $[\mathbf{x}_1, \dots, \mathbf{x}_n]$ to $[\mathbf{y}_1, \dots, \mathbf{y}_n]$ is given by Y^{-1}. (Just multiply the identity (1.6.4) by Y^{-1} from the right.) Suppose also that $[\mathbf{z}_1, \dots, \mathbf{z}_n]$ is a basis in $\mathbf{V}$. Let $[\mathbf{z}_1, \dots, \mathbf{z}_n] = [\mathbf{y}_1, \dots, \mathbf{y}_n]Z$. Then, $[\mathbf{z}_1, \dots, \mathbf{z}_n] = [\mathbf{x}_1, \dots, \mathbf{x}_n](YZ)$. (Use (1.6.4)). Here, YZ is the matrix of the change of basis, from $[\mathbf{z}_1, \dots, \mathbf{z}_n]$ to $[\mathbf{x}_1, \dots, \mathbf{x}_n]$. See the next section about matrices and their properties.

We end up this section by investigating the linear analogue of a result in group theory. It is a well-known fact in elementary group theory that a subgroup of a finitely generated group is not finitely generated necessarily. See [9] for more details. Nevertheless, it is not the case for vector spaces, i.e. any subspace of a finite dimensional vector space is finite dimensional. Now, if $\mathbf{V}$ is a finite dimensional vector space over the field $\mathbb{F}$ with the basis $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and $\mathbf{W}$ is its subspace, this question can be asked that if there exists a subset $J \subseteq [n]$ for which $\mathbf{W} = \mathrm{span}\{\mathbf{x}_i : i \in J\}$? It is easy to show that this is not the case. For example if $\mathbf{V} = \mathbb{R}^3$ as a real vector space with the standard basis $\mathbf{X} = \{(1,0,0), (0,1,0), (1,1,0)\}$ and $\mathbf{W} = \mathrm{span}\left\{(1,1,1), (1,1,\frac{1}{2})\right\}$, then W is spanned by none of two elements of $\mathbf{X}$. (Why?)

1.6.3 Worked-out Problems

1. Let $C[-1,1]$ be the set of all real continuous functions on the interval $[-1,1]$

   (a) Show that $C[-1,1]$ is a vector space over $\mathbb{R}$.

   (b) Let $U(a,b) \subset C[-1,1]$ be all functions such that $f(a) = b$, for $a \in [-1,1]$ and $b \in \mathbb{R}$. For which values of b, $U(a,b)$ is a subspace?

   (c) Is $C[-1,1]$ finite dimensional?

   Solution:

   (a) We use the following elementary facts:
   "The sum of two continuous functions is continuous" and "Multiplying a continuous function by a real number is continuous".
   If $f(x) \in C[-1,1]$ and $a \in \mathbb{R}$, $(af)(x) = af(x)$. Obviously, af is also in $C[-1,1]$.
   Furthermore, if $f(x), g(x) \in C[-1,1]$, then $(f+g)(\mathbf{x}) = f(x)+g(x)$. Using the above facts, $f+g$ is also continuous.
   The zero function is in $C[-1,1]$ and works as neutral element for the operation "+" defined as above. Also, $g(x) = 1$ is in $C[-1,1]$ and $1f = f$, for any $f \in C[-1,1]$.
   Then, $C[-1,1]$ satisfies all properties of a vector space.

(b) If $U(a,b)$ is a subspace, it must contain 0-function. Then, $0(a) = 0$, i.e. $b = 0$. Now, if f and g are in $U(a,b)$ and c and d are real numbers, then $cf(a) + dg(a) = (c+d)b = 0$. Then, $cf + dg \in U(a,b)$. Since $U(a,b)$ is closed under addition and product by scaler, it is a subspace.

(c) No, $C[-1,1]$ is infinite dimensional. If n is an arbitrary positive integer, then $1, x, x^2, \ldots, x^n$ are linearly independent. Indeed, if $(a_0, \ldots, a_n)$ is non-zero, then the assumption that $p(x) = \sum_{i=0}^n a_i x^i$ is a zero function implies that $p(x)$ has infinitely many roots. This contradicts the well known statement that $p(x)$ has at most n roots.

2. Let $\mathbf{V}$ be vector space and $\dim \mathbf{V} = n$. show that

(a) A set $\mathbf{A}$ of n vectors in $\mathbf{V}$ spans $\mathbf{V}$ if and only if $\mathbf{A}$ is independent.

(b) $\mathbf{V}$ has a basis and every basis consists of n vectors.

Solution:

(a) Assume that $\mathbf{A} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. Since $\dim \mathbf{V} = n$, the set $\{\mathbf{x}, \mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is dependent, for every $\mathbf{x} \in V$. If $\mathbf{A}$ is independent, then $\mathbf{x}$ is in the span of $\mathbf{A}$. Therefore, $\mathbf{A}$ spans $\mathbf{V}$.
Conversely, if $\mathbf{A}$ is dependent, one of its elements can be removed without changing the span of $\mathbf{A}$. Thus, $\mathbf{A}$ cannot span $\mathbf{V}$, by Problem 1.6.4-9.

(b) Since $\dim \mathbf{V} = n$, $\mathbf{V}$ contains an independent set of n vectors, and (a) shows that every such set is a basis of $\mathbf{V}$; the statement follows from the definition of dimension and Problem 1.6.4-9.

1.6.4 Problems

1. * A complex number ζ is called m-th root of unity if $\zeta^m = 1$. ζ is called a *primitive unity root of order* $m > 1$ if $\zeta^m = 1$ and $\zeta^k \neq 1$, for $k \in [m-1]$. Let l be the number of integers in $[m-1]$ which are coprime with m. Let $\mathbb{Q}[\zeta] := \{a_0 + a_1\zeta + \cdots + a_{l-1}\zeta^{l-1},\ a_0, \ldots, a_{l-1} \in \mathbb{Q}\}$. Show that $\mathbb{Q}[\zeta]$ is a field. It is a finite extension of $\mathbb{Q}$. More precisely, $\mathbb{Q}[\zeta]$ can be viewed as a vector space over $\mathbb{Q}$ of dimension l.
(Hints:

(a) Let $\xi = e^{\frac{2\pi \mathbf{i}}{m}}$. Show that ζ is an m-th root of unity if and only if $\zeta = \xi^k$ and $k \in [m]$. Furthermore, ζ is primitive if and only if k and m are coprime.

(b) Let ζ be an m-th primitive root of unity. Show that $\zeta, \zeta^2, \ldots, \zeta^{m-1}, \zeta^m$ are all m-th root of unity.

(c) Show that $x^m - 1$ has simple roots.

(d) Let $p_1(x) \cdots p_j(x)$ be the decomposition of $x^m - 1$ to irreducible factors in $\mathbb{Q}[x]$. Assume that ζ, a primitive m-th root of unity, is a root of p_1. Let $l = \deg p_l$. Show that $\mathbb{Q}[\zeta] := \{a_0 + a_1\zeta + \cdots + a_{l-1}\zeta^{l-1},\ a_0, \ldots, a_{l-1} \in \mathbb{Q}\}$ is an extension field of $\mathbb{Q}$ of degree l which contains all m roots of unity.

(e) ** Show that l is the number of m-primitive roots of unity.

(f) ** Show that l is given by the Euler's function $\phi(m)$. Namely, assume that $m = p_1^{d_1} \cdots p_r^{d_r}$, where $1 < p_1 < \cdots < p_r$ are primes and $d_1, \ldots, d_r$ are positive integers. Then, $\phi(m) = (p_1 - 1)p_1^{d_1-1} \cdots (p_r - 1)p_r^{d_r-1}$.)

Special cases:

(a) $m = 4$: $x^4 - 1 = (x^2+1)(x+1)(x-1)$. The 4-th primitive roots of unity are $\pm i$. They are roots of x^2+1. $\mathbb{Q}[i]$ is called the *Gaussian field of rationals*. The set $\mathbb{Z}[i] = \{a + bi,\ a, b \in \mathbb{Z}\}$ is called the *Gaussian integers domain.*

(b) $m = 6$: $x^6 - 1 = (x^2 - x + 1)(x^2 + x + 1)(x + 1)(x - 1)$. The primitive roots of order 6 are $\frac{1}{2} \pm \frac{\sqrt{3}}{2} i$, which are the two roots of $x^2 - x + 1$.

2. *Quaternions* $\mathbf{Q}$ is a four dimensional noncommutative division algebra that generalizes complex numbers. Any quaternion is written as a vector $\mathbf{q} = a_0 + a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$. The product of two quaternions is determined by using the following equalities.

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1,\ \mathbf{ij} = -\mathbf{ji} = \mathbf{k},\ \mathbf{jk} = -\mathbf{kj} = \mathbf{i},\ \mathbf{ki} = -\mathbf{ik} = \mathbf{j}.$$

Let $\bar{\mathbf{q}} = a_0 - a_1\mathbf{i} - a_2\mathbf{j} - a_3\mathbf{k}$. Show

(a) The map $T(\mathbf{q}) = \bar{\mathbf{q}}$ is an *involution*. (T is a bijection preserving addition and multiplication and $T^2 = id$.)

(b) $\mathbf{Q}$ is a noncommutative division algebra, i.e. if $\mathbf{q} \neq \mathbf{0}$, then there exists a unique $\mathbf{r}$ such that $\mathbf{rq} = \mathbf{qr} = 1$. (You can try to use the identity $\mathbf{q}\bar{\mathbf{q}} = \bar{\mathbf{q}}\mathbf{q} = |\mathbf{q}|^2 = \sum_{i=0}^{3} a_i^2$).

3. Let $\mathcal{P}_m$ be the subset of all polynomials of degree m at most over the field $\mathbb{F}$. Show that the polynomials $1, x, \ldots, x^m$ form a basis in $\mathcal{P}_m$.

4. Let f_j be a non-zero polynomial in $\mathcal{P}_m$ of degree j, for $j = 0, \ldots, m$. Show that $f_0, \ldots, f_m$ form a basis in $\mathcal{P}_m$.

5. Find a basis in $\mathcal{P}_4$ such that each polynomial is of degree 4.

6. Does exist a basis in $\mathcal{P}_4$ such that each polynomial is of degree 3?

7. Write up the set of all vectors in $\mathbb{R}^4$ whose coordinates are two 0's and two 1's. (There are six such vectors.) Show that these six vectors span $\mathbb{R}^4$. Find all the collections of four vectors out of these six which form a basis in $\mathbb{R}^4$.

8. Prove the *Completion lemma*: let $\mathbf{V}$ be a vector space of dimension m. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbf{V}$ be n linearly independent vectors. (Here $m \geq n$) Then, there exist $m - n$ vectors $\mathbf{v}_{n+1}, \ldots, \mathbf{v}_m$ such that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a basis in $\mathbf{V}$.

9. Assume that a vector space $\mathbf{V}$ is spanned by a set of n vectors, ($n \in \mathbb{N}$). Show that $\dim \mathbf{V} \leq n$.

10. Let $\mathbf{V}$ be a vector space over a field $\mathbb{F}$ and $\mathbf{S} \subseteq \mathbf{V}$, show that

(a) $\text{span}(\text{span}(\mathbf{S})) = \text{span}(\mathbf{S})$

(b) $\text{span}(\mathbf{S})$ is the smallest subspace of $\mathbf{V}$ containing $\mathbf{S}$.

(c) $\text{span}(\mathbf{S}) = \mathbf{S}$ if and only if $\mathbf{S}$ is a subspace of $\mathbf{V}$.

1.7 Matrices

1.7.1 Basic properties of matrices

A matrix is a rectangular array of elements from the given set of elements S. The horizontal arrays of a matrix are called its *rows* and the vertical arrays called its *columns*. A matrix is said to have the *order* $m \times n$ if it has m rows and n columns. An $m \times n$ matrix A can be represented in the following form:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \\ \ddots & \vdots & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

where a_{ij} is the entry at the intersection of the i-th row and j-th column. In a more concise manner, we also write $A = [a_{ij}]_{m\times n}$ or $A = [a_{ij}]$.

The set of $m \times n$ matrices with entries in S is denoted by $\mathrm{S}^{m\times n}$. For a field $\mathbb{F}$, the set $\mathbb{F}^{m\times n}$ is a vector space over $\mathbb{F}$ of dimension mn. A standard basis of $\mathbb{F}^{m\times n}$ is given by $E_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$, where E_{ij} is a matrix with 1 in the entry (i, j) and 0 in all other entries. Also, $\mathbb{F}^{m\times 1} = \mathbb{F}^m$ the set of column vectors with m coordinates. A standard basis of $\mathbb{F}^m$ is $\mathbf{e}_i = (\delta_{1i}, \ldots, \delta_{mi})^\top$, where δ_{ji} denotes Kronecker's delta function defined as

$$\delta_{ji} = \begin{cases} 0 & if\ j \neq i \\ 1 & if\ j = i, \end{cases}$$

for $i = 1, \ldots, m$.

A matrix obtained by deleting some of the rows and/or columns of a matrix is said to be a *submatrix* of the given matrix.

For $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{p\times q}$, the product AB is defined if and only if the number of columns in A is equal to the number of rows in B, i.e. $n = p$. In that case, the resulting matrix $C = [c_{ij}]$ is $m \times q$. The entry c_{ij} is obtained by multiplying the row i of A by the column j of B. Recall that for $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{F}^n$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{F}^n$, $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i$. (This product can be regarded as a product of $1 \times n$ matrix $\mathbf{x}^\top$ with $n \times 1$ product matrix $\mathbf{y}$.) The product of corresponding sizes of matrices satisfies the properties:

(i) associativity: $(AB)C = A(BC)$;

(ii) distributivity: $(A_1 + A_2)B = A_1B + A_2B$ and $A(B_1 + B_2) = AB_1 + AB_2$;

(iii) $a(AB) = (aA)B = A(aB)$, for each $a \in \mathbb{F}$;

(iv) identities: $I_mA = AI_n$, where $A \in \mathbb{F}^{m\times n}$ and $I_m = [\delta_{ij}]_{i=j=1}^m \in \mathbb{F}^{m\times m}$ is the identity matrix of order m.

For $A = [a_{ij}] \in \mathbb{F}^{m\times n}$ denote by $A^\top \in \mathbb{F}^{n\times m}$ the transposed matrix of A, i.e. the (i, j) entry of $A^\top$ is the (j, i) entry of A. The following properties hold

$$(aA)^\top = aA^\top,\ (A + B)^\top = A^\top + B^\top,\ (AC)^\top = C^\top A^\top.$$

The matrix $A = [a_{ij}] \in \mathbb{F}^{m\times n}$ is called *diagonal* if $a_{ij} = 0$, for $i \neq j$. Denote by $\mathbf{D}(m,n) \subset \mathbb{F}^{m\times n}$ the vector subspace of diagonal matrices. A square diagonal matrix with the diagonal entries $d_1, \dots, d_n$ is denoted by $\operatorname{diag}(d_1, \dots, d_n) \in \mathbb{F}^{n\times n}$. A matrix A is called a *block matrix* if $A = [A_{ij}]_{i=j=1}^{p,q}$, where each entry A_{ij} is an $m_i \times m_j$ matrix. Then, $A \in \mathbb{F}^{m\times n}$ where $m = \sum_{i=1}^{p} m_i$ and $n = \sum_{j=1}^{q} n_j$. A block matrix A is called *block diagonal* if $A_{ij} = 0_{m_i \times m_j}$, for $i \neq j$. A block diagonal matrix with $p = q$ is denoted by $\operatorname{diag}(A_{11}, \dots, A_{pp})$. A different notation is $\oplus_{l=1}^{p} A_{ll} := \operatorname{diag}(A_{11}, \dots, A_{pp})$, and $\oplus_{l=1}^{p} A_{ll}$ is called the *direct sum* of the square matrices $A_{11}, \dots, A_{pp}$.

1.7.2 Elementary row operation

Elementary operations on the rows of $A \in \mathbb{F}^{m\times n}$ are defined as follows:

1. Multiply row i by a non-zero $\mathbf{a} \in \mathbb{F}$.

2. Interchange two distinct rows i and j in A.

3. Add to row j, row i multiplied by a, where $i \neq j$.

A matrix $A \in \mathbb{F}^{n\times n}$ is said to be in a *row echelon form* (REF) if it satisfies the following conditions:

1. All non-zero rows are above any rows of all zeros.

2. Each leading entry of a row is in a column to the right of the leading entry of the row above it.

3. All entries in a column below a leading entry are zeros.

Also, A can be brought to a unique canonical form, called the *reduced row echelon form*, abbreviated as RREF by the elementary row operation. The RREF of zero matrix $0_{m\times n}$ is $0_{m\times n}$. For $A \neq 0_{m\times n}$, RREF of A given by B is of the following form: In addition to the conditions 1,2 and 3 above, it must satisfy the followings:

1. The leading entry in each non-zero row is 1.

2. Each leading 1 is the only non-zero entry in its column.

A *pivot position* in the matrix A is a location in A that corresponds to a leading 1 in the RREF of A. First non-zero entry of each row of a REF of A is called a *pivot*, and the columns in which pivots appear are *pivot columns.* The process of finding pivots is called *pivoting.*

An elementary row-interchange matrix is an $n \times n$ matrix which can be obtained from the identity matrix I_n by performing on I_n a single elementary row operation.

1.7.3 Gaussian Elimination

Gaussian elimination is a method to bring a matrix either to is REF or its RREF or similar forms. It is often used for solving matrix equations of the form

$$A\mathbf{x} = b. \tag{1.7.1}$$

In the following, we may assume that A is full-rank, i.e. the rank of A equals the number of rows.

To perform Gaussian elimination starting with the system of equations

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & & & \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_k \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \tag{1.7.2}$$

compose the "augmented matrix equation"

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1k} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2k} & b_2 \\ \vdots & & & & \\ a_{k1} & a_{k2} & \cdots & a_{kk} & b_k \end{array}\right] \tag{1.7.3}$$

Here, the column vector is the variable $\mathbf{x}$ carried along for labeling the matrix rows. Now, perform elementary row operations to put the augmented matrix into the upper triangular form

$$\left[\begin{array}{cccc|c} a'_{11} & a'_{12} & \cdots & a'_{1k} & b'_1 \\ 0 & a'_{22} & \cdots & a'_{2k} & b'_2 \\ \vdots & & & & \\ 0 & 0 & \cdots & a'_{kk} & b'_k \end{array}\right]. \tag{1.7.4}$$

Solve the equation of the k-th row for $\mathbf{x}_k$, then substitute back into the equation of the $(k-1)$-st row to obtain a solution for $\mathbf{x}_{k-1}$, etc. We get the formula

$$\mathbf{x}_i = \frac{1}{a'_{ii}} \left(b'_i - \sum_{j=i+1}^{k} a'_{ij} \mathbf{x}_j \right).$$

Example 1.7.1 *We solve the following system using Gaussian elimination:*

$$\begin{cases} 2\mathbf{x} - 2\mathbf{y} & = -6 \\ \mathbf{x} - \mathbf{y} + \mathbf{z} & = 1 \\ 3\mathbf{y} - 2\mathbf{z} & = -5 \end{cases}$$

For this system, the augmented matrix is

$$\begin{bmatrix} 2 & -2 & 0 & -6 \\ 1 & -1 & 1 & 1 \\ 0 & 3 & -2 & -5 \end{bmatrix}$$

First, multiply row 1 by $\frac{1}{2}$:

$$\begin{bmatrix} 2 & -2 & 0 & -6 \\ 1 & -1 & 1 & 1 \\ 0 & 3 & -2 & -5 \end{bmatrix} \xrightarrow{\textit{Multiply } r_1 \textit{ by } \frac{1}{2}} \begin{bmatrix} 1 & -1 & 0 & -3 \\ 1 & -1 & 1 & 1 \\ 0 & 3 & -2 & -5 \end{bmatrix}$$

Now, adding -1 times the first row to the second row yields zeros below the first entry in the first column:

$$\begin{bmatrix} 1 & -1 & 0 & -3 \\ 0 & 0 & 1 & 1 \\ 0 & 3 & -2 & -5 \end{bmatrix} \xrightarrow{-r_1 \textit{ added to } r_2} \begin{bmatrix} 1 & -1 & 0 & -3 \\ 0 & 0 & 1 & 4 \\ 0 & 3 & -2 & -5 \end{bmatrix}$$

Interchanging the second and third rows then gives the desired upper-triangular coefficient matrix:

$$\begin{bmatrix} 1 & -1 & 0 & -3 \\ 0 & 0 & 1 & 4 \\ 0 & 3 & -2 & -5 \end{bmatrix} \xrightarrow{r_2 \leftrightarrow r_1} \begin{bmatrix} 1 & -1 & 0 & -3 \\ 0 & 3 & -2 & -5 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

The third row now says $\mathbf{z} = 4$*. Back-substituting this value into the second row gives* $\mathbf{y} = 1$*, and back-substitution of both these values into the first row yields* $\mathbf{x} = -2$*. The solution system is therefore* $(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (-2, 1, 4)$*.*

1.7.4 Solution of linear systems

Consider the following system of the linear equations:

$$\begin{aligned} & a_{11}\mathbf{x}_1 + a_{12}\mathbf{x}_2 + \cdots + a_{1n}\mathbf{x}_n = b_1 \\ & a_{21}\mathbf{x}_1 + a_{22}\mathbf{x}_2 + \cdots + a_{2n}\mathbf{x}_n = b_2 \\ & \vdots \\ & a_{n1}\mathbf{x}_1 + a_{n2}\mathbf{x}_2 + \cdots + a_{nn}\mathbf{x}_n = b_n \end{aligned} \tag{1.7.5}$$

Assume that $A = [a_{ij}]_{1 \le i,j \le n}$ and $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$.

Definition 1.7.2 *Let* $\hat{A} := [A|\mathbf{b}]$ *be the augmented* $m \times (n+1)$ *matrix representing the system* (1.7.5)*. Suppose that* $\hat{C} = [C|\mathbf{c}]$ *is a REF of* $\hat{A}$*. Assume that* C *has k-pivots in the columns* $1 \le \ell_1 < \cdots < \ell_k \le n$*. Then, the variable* $x_{\ell_1}, \ldots, x_{\ell_k}$ *corresponding to these pivots are called the lead variables. The other variables are called free variables.*

Definition 1.7.3 *A map* $f : \mathbb{R}^n \to \mathbb{R}$ *is called affine if* $f(\mathbf{x}) = a_1x_1 + \cdots + a_nx_n + b$*. Also,* f *is called a linear function if* $b = 0$*.*

The following theorem describes exactly the set of all solutions of (1.7.5).

Theorem 1.7.4 *Let* $\hat{A} := [A|\mathbf{b}]$ *be the augmented* $m \times (n+1)$ *matrix representing the system* (1.7.5)*. Suppose that* $\hat{C} = [C|\mathbf{c}]$ *be a REF of* $\hat{A}$*. Then, the system* (1.7.5) *is solvable if and only if* $\hat{C}$ *does not have a pivot in the last column* $n+1$*.*
Assume that (1.7.5) *is solvable. Then each lead variable is a unique affine function in free variables. These affine functions can be determined as follows:*

1. *Consider the linear system corresponding to* $\hat{C}$*. Move all the free variables to the right-hand side of the system. Then, one obtains a triangular system in lead variables, where the right-had side are affine functions in free variables.*

2. *Solve this triangular system by back substitution.*

In particular, for a solvable system, we have the following alternative.

1. *The system has a unique solution if and only if there are no free variables.*

2. *The system has infinitely many solutions if and only if there is at least one free variable.*

Proof. We consider the linear system equations corresponding to $\hat{C}$. As an ERO (elementary row operation) on $\hat{A}$ corresponds to an EO (elementary operation) on the system (1.7.5), it follows that the system represented by $\hat{C}$ is equivalent to (1.7.5). Suppose first that $\hat{C}$ has a pivot in the last column. Thus, the corresponding row of $\hat{C}$ which contains the pivot on the column $n+1$ is $(0,0,\ldots,0,1)^\top$. This equation is unsolvable, hence the whole system corresponding to $\hat{C}$ is unsolvable. Therefore, the system (1.7.5) is unsolvable.
Assume now that $\hat{C}$ does not have a pivot in the last column. Move all the free variables to the right-hand side of the system given by $\hat{C}$. It is a triangular system in the lead variables where the right-hand side of each equation is an affine function in the free variables. Now use back substitution to express each lead variable as an affine function of the free variables.
Each solution of the system is determined by the value of the free variables which can be chosen arbitrarily. Hence, the system has a unique solution if and only if it has no free variables. The system has many solutions if and only if it has at least one free variable. □

Consider the following example of $\hat{C}$:

$$\left[\begin{array}{cccc|c} 1 & -2 & 3 & -1 & 0 \\ 0 & 1 & 3 & 1 & 4 \\ 0 & 0 & 0 & 1 & 5 \end{array}\right], \tag{1.7.6}$$

x_1, x_2, x_4 are lead variables, x_3 is a free variable.

$$\begin{aligned} & x_4 = 5, \\ & x_2 + 3x_3 + x_4 = 4 \Rightarrow x_2 = -3x_3 - x_4 + 4 \Rightarrow \\ & x_2 = -3x_3 - 1, \\ & x_1 - 2x_2 + 3x_3 - x_4 = 0 \Rightarrow x_1 = 2x_2 - 3x_3 + x_4 = 2(-3x_3 - 1) - 3x_3 + 5 \Rightarrow \\ & x_1 = -9x_3 + 3. \end{aligned}$$

□

Notation 1.7.5 *Let S and T be two subsets of a set X. Then, the set $T \smallsetminus S$ is the set of elements of T which are not in S. ($T \smallsetminus S$ may be empty.)*

Theorem 1.7.6 *Let $\hat{A} := [A|\mathbf{b}]$ be the augmented $m \times (n+1)$ matrix representing the system* (1.7.5). *Suppose that $\hat{C} = [C|\mathbf{c}]$ is a RREF of $\hat{A}$. Then, the system* (1.7.5) *is solvable if and only if $\hat{C}$ does not have a pivot in the last column $n+1$.*
Assume that (1.7.5) *is solvable. Then, each lead variable is a unique affine function in free variables determined as follows. The leading variable x_{ℓ_i} appears only in the equation i, for $1 \le i \le r = \operatorname{rank} A$. Shift all other variables in the equation, (which are free variables), to the other side of the equation to obtain x_{ℓ_i} as an affine function in free variables.*

Proof. Since RREF is a row echelon form, Theorem 1.7.4 yields that (1.7.5) is solvable if and only if $\hat{C}$ does not have a pivot in the last column $n+1$. Assume that $\hat{C}$ does not have a pivot in the last column $n+1$. Then, all the pivots of $\hat{C}$ appear in C. Hence, rank $\hat{A}$ = rank $\hat{C}$ = rank $A(=r)$. The pivots of $C = [c_{ij}] \in \mathbb{R}^{m\times n}$ are located at row i and the column ℓ_i, denote as (i, ℓ_i), for $i = 1, \ldots, r$. Since C is also in RREF, in the column ℓ_i there is only one non-zero element which is equal 1 and is located in row i.

Consider the system of linear equations corresponding to $\hat{C}$, which is equivalent to (1.7.5). Hence, the lead variable ℓ_i appears only in the i-th equation. Left hand-side of this equation is of the form x_{ℓ_i} plus a linear function in free variables whose indices are greater than x_{ℓ_i}. The right-hand is c_i, where $\mathbf{c} = (c_1, \ldots, c_m)^\top$. Hence, by moving the free variables to the right-hand side, we obtain the exact form of x_{ℓ_i}.

$$x_{\ell_i} = c_i - \sum_{j \in \{\ell_i, \ell_i+1, \ldots, n\} \setminus \{j_{\ell_1}, j_{\ell_2}, \ldots, j_{\ell_r}\}} c_{ij} x_j, \quad \text{for } i = 1, \ldots, r. \tag{1.7.7}$$

(Therefore, $\{\ell_i, \ell_i+1, \ldots, n\}$ consists of $n - \ell_i + 1$ integers from j_{ℓ_i} to n, while $\{j_{\ell_1}, j_{\ell_2}, \ldots, j_{\ell_r}\}$ consists of the columns of the pivots in C.) □

Example 1.7.7 *Consider*

$$\left[\begin{array}{cccc|c} 1 & 0 & b & 0 & u \\ 0 & 1 & d & 0 & v \\ 0 & 0 & 0 & 1 & w \end{array}\right],$$

x_1, x_2, x_4 *lead variables* x_3 *free variable ;*

$$\begin{aligned} & x_1 + bx_3 = u \Rightarrow x_1 = -bx_3 + u, \\ & x_2 + dx_3 = v \Rightarrow x_2 = -dx_3 + v, \\ & x_4 = w. \end{aligned}$$

Definition 1.7.8 *The system* (1.7.5) *is called homogeneous if* $\mathbf{b} = \mathbf{0}$, *i.e.* $b_1 = \cdots = b_m = 0$. *A homogeneous system of linear equations has a solution* $\mathbf{x} = \mathbf{0}$, *which is called the trivial solution.*

Let A be a square $n \times n$, then A is called nonsingular if the corresponding homogeneous system of n equations in n unknowns has only solution $\mathbf{x} = \mathbf{0}$. *Otherwise, A is called singular.*

Theorem 1.7.9 *Let A be an $m \times n$ matrix. Then, its RREF is unique.*

Proof. Let U be a RREF of A. Consider the augmented matrix $\hat{A} := [A|\mathbf{0}]$ corresponding to the homogeneous system of equations. Clearly, $\hat{U} = [U|\mathbf{0}]$ is a RREF of $\hat{A}$. Put the free variables on the other side of the homogeneous system corresponding to $\hat{A}$, where each lead variable is a linear function of the free variables. Note that the exact formulas for the lead variables determine uniquely the column of the RREF which correspond to the free variables.

Assume that U_1 is another RREF of A. Then, U and U_1 have the same pivots. Hence, U and U_1 have the same pivots. Thus, U and U_1 have the same columns which correspond to pivots. By considering the homogeneous system of linear equations corresponding to $\hat{U}_1 = [U_1|\mathbf{0}]$, we find also the solution of the homogeneous system $\hat{A}$, by writing down the lead variables as linear functions in free variables. Since $\hat{U}$ and $\hat{U}_1$ give rise to the same lead and free variables, we deduce that the each linear function in free variables corresponding to a lead variable x_{ℓ_1} corresponding to $\hat{U}$ and $\hat{U}_1$ are equal. That is, the matrices U and U_1 have the same row i, for $i = 1, \ldots, \operatorname{rank} A$. All other rows of U and U_1 are zero rows. Hence, $U = U_1$. □

Corollary 1.7.10 *The matrix $A \in \mathbb{R}^{n\times n}$ is nonsingular if and only if* $\operatorname{rank} A = n$*, i.e. the RREF of A is the identity matrix*

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{1.7.8}$$

Proof. The matrix A is nonsingular if and only if no free variable. Thus, $\operatorname{rank} A = n$ and the RREF is I_n. □

1.7.5 Invertible matrices

A matrix $A \in \mathbb{F}^{m\times m}$ is called *invertible* if there exists $B \in \mathbb{F}^{m\times m}$ such that $AB = BA = I_m$. Note that such B is unique. If $AC = CA = I_m$, then $B = BI_m = B(AC) = (BA)C = I_mC = C$. We denote B, the inverse of A by A^{-1}. Denote by $\mathrm{GL}(m, \mathbb{F}) \subset \mathbb{F}^{m\times m}$ the set of all invertible matrices. Note that $\mathrm{GL}(m, \mathbb{F})$ is a group under the multiplication, with the unit I_n. (Observe that for A and $B \in \mathrm{GL}(m, \mathbb{F})$ $(AB)^{-1} = B^{-1}A^{-1}$.)

A matrix $E \in \mathbb{F}^{m\times m}$ is called *elementary* if it is obtained from I_m by applying one elementary row operation. Note that applying an elementary row operation on A is equivalent to EA, where E is the corresponding elementary row operation. By reversing the corresponding elementary operation we see that E is invertible, and E^{-1} is also an elementary matrix.

The following theorem is well-known as the fundamental theorem of invertible matrices; in the next subsections, we will see a more detailed version of the fundamental theorem of invertible matrices.

Theorem 1.7.11 *Let $A \in \mathbb{F}^{m\times m}$. The following statements are equivalent.*

1. *A is invertible.*

2. *$A\mathbf{x} = \mathbf{0}$ has only the trivial solution.*

3. *The RREF of A is I_m.*

4. *A is a product of elementary matrices.*

Proof. $1 \Rightarrow 2$. The equation $A\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = I_m\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{0} = \mathbf{0}$.
$2 \Rightarrow 3$. The system $A\mathbf{x} = \mathbf{0}$ does not have free variables, hence the number of pivots is the number of rows, i.e., the RREF of A is I_m.

3⇒ 4. There exists a sequence of elementary matrices so that $E_pE_{p-1}\cdots E_1A = I_m$. Hence, $A = E_1^{-1}E_2^{-1}\cdots E_p^{-1}$, and each E_j^{-1} is also elementary.
4⇒ 1. As each elementary matrix is invertible, so is their product, which is equal to A. □

Theorem 1.7.12 *Let $A \in \mathbb{F}^{n\times n}$. Define the matrix $B = [A\ I_n] \in F^{n\times(2n)}$. Let $C = [C_1\ C_2], C_1, C_2 \in F^{n\times n}$ be the RREF of B. Then, A is invertible if and only if $C_1 = I_n$. Furthermore, if $C_1 = I_n$ then $A^{-1} = C_2$.*

Proof. The fact that "A is invertible if and only if $C_1 = I_n$" follows straightforward from Theorem 1.7.11. Note that for any matrix $F = [F_1F_2], F_1, F_2 \in \mathbb{F}^{n\times p}, G \in \mathbb{F}^{l\times m}$, we have $GF = [(GF_1)(GF_2)]$. Let H be a product of elementary matrices such that $HB = [IC_2]$. Thus, $HA = I_n$ and $HI_n = C_1$. Hence, $H = A^{-1}$ and $C_1 = A^{-1}$. □

We can extract the following algorithm from Theorem 1.7.12 to calculate the inverse of a matrix if it was invertible:

Gauss-Jordan algorithm for A^{-1}

- Form the matrix $B = [A|I_n]$.
- Perform the ERO to obtain RREF of $B : C = [D|F]$.
- A is invertible ⇔ $D = I_n$.
- If $D = I_n$, then $A^{-1} = F$.

Example 1.7.13 *Let $A = \begin{bmatrix} 1 & 2 & -1 \\ -2 & -5 & 5 \\ 3 & 7 & -5 \end{bmatrix}$. Write $B = [A|I_3]$ observe that the (1,1) entry in B is a pivot:*

$$B = \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ -2 & -5 & 5 & 0 & 1 & 0 \\ 3 & 7 & -5 & 0 & 0 & 1 \end{array}\right].$$

Perform ERO: $R_2 + 2R_1 \to R_2$, $R_3 - 3R_1 \to R_3$:

$$B_1 = \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -1 & 3 & 2 & 1 & 0 \\ 0 & 1 & -2 & -3 & 0 & 1 \end{array}\right].$$

To make (2,2) entry pivot do: $-R_2 \to R_2$:

$$B_2 = \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & -1 & 0 \\ 0 & 1 & -2 & -3 & 0 & 1 \end{array}\right].$$

To eliminate (1,2), (1,3) entries do $R_1 - 2R_2 \to R_1$, $R_3 - R_2 \to R_3$

$$B_3 = \left[\begin{array}{ccc|ccc} 1 & 0 & 5 & 5 & 2 & 0 \\ 0 & 1 & -3 & -2 & -1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right].$$

Now, (3,3) entry is a pivot. To eliminate (1,3), (2,3) entries do: $R_1 - 5R_3 \to R_1$, $R_2 + 3R_3 \to R_2$

$$B_4 = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 10 & -3 & -5 \\ 0 & 1 & 1 & -5 & 2 & 3 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right].$$

Then, $B_4 = [I_3|F]$ *is RREF of* B*. Thus,* A *has the inverse:*

$$A^{-1} = \begin{bmatrix} 10 & -3 & -5 \\ -5 & 2 & 3 \\ -1 & 1 & 1 \end{bmatrix}.$$

1.7.6 Row and column spaces of A

Two matrices A and $B \in \mathbb{F}^{m\times n}$ are called *left equivalent* row equivalent) if A and B have the same reduced row echelon form or B can be obtained from A by applying elementary row operations. Equivalently, $B = UA$ for some $U \in \mathrm{GL}(m, F)$. This is denoted as $A \sim_l B$. It is straightforward to show that $\sim_l$ is an equivalence relation. Right equivalent (column equivalence) is defined similarly and denoted by $\sim_r$. Note that $A \sim_r B$ if and only if $A^\top \sim_l B^\top$. Also $A, B \in \mathbb{F}^{m\times n}$ are called *equivalent* if $B = UAV$, for some $U \in \mathrm{GL}(m,\mathbb{F})$, $V \in \mathrm{GL}(n,\mathbb{F})$. Equivalently, B is equivalent to A, if B can be obtained from A by applying elementary row and column operations. This is denoted as $A \sim B$. It is straightforward to show that $\sim$ is an equivalence relation.

Let $A \in \mathbb{F}^{m\times n}$. Denote by $\mathbf{c}_1, \dots, \mathbf{c}_n$ the n column of A. We write $A = [\mathbf{c}_1\ \mathbf{c}_2\ \dots\ \mathbf{c}_n]$. Then, $\mathrm{span}\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ is denoted by $R(A)$ and called the *column space* of A. Its dimension is rank A. Similarly, the dimension of the columns space of $A^\top$, which is equal to the dimension of the row space of A, is rank A, i.e. $\mathrm{rank}\, A^\top = \mathrm{rank}\, A$. Let $\mathbf{x} = (x_1, \dots, x_n)^\top$. Then, $A\mathbf{x} = \sum_{i=1}^n x_i \mathbf{c}_i$. Hence, the system $A\mathbf{x} = \mathbf{b}$ is solvable if and only if $\mathbf{b}$ is in the column space of A. The set of all $\mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \mathbf{0}$ is called the *null space* of A. It is a subspace of dimension $n - \mathrm{rank}\, A$ and is denoted as $\mathrm{N}(A) \subset \mathbb{F}^m$, $\dim \mathrm{N}(A)$ is called the *nullity* of A, and denoted by $\mathrm{null}\, A$.(See Worked-out Problem 1.9.2-5.) If $A \in \mathrm{GL}(n,\mathbb{F})$, then $A\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

Theorem 1.7.14 *Let* $A \in \mathbb{F}^{m\times n}$*. Assume that* $B \in \mathbb{F}^{m\times n}$ *is a row echelon form of* A*. Let* $k = \mathrm{rank}\, A$*. Then*

1. *The non-zero rows of* B *form a basis in the row space of* A*.*

2. *Assume that the pivots of* B *are the columns* $1 \le j_1 < \dots < j_k \le n$*, i.e.* $x_{j_1}, \dots, x_{j_k}$ *are the lead variables in the system* $A\mathbf{x} = \mathbf{0}$*. Then, the columns* $j_1, \dots, j_k$ *of* A *form a basis of the column space of* A*.*

3. *If* $n = \mathrm{rank}\, A$*, then the null space of* A *consists only of* $\mathbf{0}$*. Otherwise, the null space of* A *has the following basis. For each free variable* x_p*, let* $\mathbf{x}_p$ *be the unique solution of* $A\mathbf{x} = \mathbf{0}$*, where* $x_p = 1$ *and all other free variables are zero. Then, these* $n - \mathrm{rank}\, A$ *vectors form a basis in* $N(A)$*.*

Proof. 1. Note that if $E \in \mathbb{F}^{m\times m}$, then the row space of EA is contained in A, since any row in EA is a linear combination of the rows of A. Since $A = E^{-1}(EA)$, it follows that the row space of A is contained in the row space of EA. Hence, the row space of A and EA are the same. Therefore, the row space of A and B are the same. Since the last $m-k$ rows of B are zero rows, it follows that the row space of B spanned by the first k rows $\mathbf{b}_1^\top, \dots, \mathbf{b}_k^\top$. We claim that $\mathbf{b}_1^\top, \dots, \mathbf{b}_k^\top$ are linearly independent. Indeed, consider $x\mathbf{b}_1^\top + \dots + \mathbf{x}_k\mathbf{b}_k = \mathbf{0}^\top$. Since all the entries below the first pivot in B are zero, we deduce that $x_1 = 0$. Continue in the same manner to obtain that $x_2 = 0, \dots, x_k = 0$. Then, $\mathbf{b}_1^\top, \dots, \mathbf{b}_k^\top$ form a basis in the row space of B and A.

2. We first show that $\mathbf{c}_{j_1}, \dots, \mathbf{c}_{j_k}$ are linearly independent. Consider the equality $\sum_{j=1}^k x_{i_j}\mathbf{c}_{i_j} = \mathbf{0}$. This is equivalent to the system $A\mathbf{x} = \mathbf{0}$, where $\mathbf{x} = (x_1, \dots, x_n)^\top$ and $x_p = 0$ if x_p is a free variable. Since all free variables are zero, all lead variables are zero, i.e. $x_{j_i} = 0$, for $i = 1, \dots, k$. Thus, $\mathbf{c}_{j_1}, \dots, \mathbf{c}_{j_k}$ are linearly independent. It is left to show that $\mathbf{c}_p$, where x_p is a free variable, is a linear combination of $\mathbf{c}_{j_1}, \dots, \mathbf{c}_{j_k}$. Again, consider $A\mathbf{x} = \mathbf{0}$, where each $x_p = 1$ and all other free variables are zero. We have a unique solution to $A\mathbf{x} = \mathbf{0}$, which states that $\mathbf{0} = \mathbf{c}_p + \sum_{i=1}^k x_{j_i}\mathbf{c}_{j_i}$, i.e. $\mathbf{c}_p = \sum_{i=1}^k -x_{j_i}\mathbf{c}_{j_i}$.

3. This follows for the way we write down the general solution of the system $A\mathbf{x} = \mathbf{0}$ in terms of free variables. □

Remark 1.7.15 *Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{F}^m$. To find a basis in $\mathbf{U} := \operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ consisting of k vectors in $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, apply Theorem 1.7.14 to the column space of the matrix $A = [\mathbf{x}_1\ \mathbf{x}_2 \dots \mathbf{x}_n]$. To find an appropriate basis in $\mathbf{U}$, apply Theorem 1.7.14 to the row space of $A^\top$.*

Lemma 1.7.16 *Let $A \in \mathbb{F}^{m\times n}$. Then, A is equivalent to a block diagonal matrix $I_{\operatorname{rank} A} \oplus 0$.*

Proof. First, bring A to REF matrix B with all pivots equal to 1. Now perform the elementary column operations corresponding to the elementary row operations on $B^\top$. This will give a matrix C with r pivots equal to 1 and all other elements zero. Now, interchange the corresponding columns to obtain $I_r \oplus 0$. It is left to show that $r = \operatorname{rank} A$. First, recall that if $B = UA$, then the row space of A and B are the same. Then, $\operatorname{rank} A = \operatorname{rank} B$, which is the dimension of the row space. Next, if $C = BV$, then C and B have the same column space. Thus, $\operatorname{rank} B = \operatorname{rank} C$, which is the dimension of the column space. This means $r = \operatorname{rank} A$. □

The following theorem is a more detailed case of Theorem 1.7.11. Its proof is left as Problem 1.8.5-4.

Theorem 1.7.17 *Let $A \in \mathbb{F}^{n\times n}$. The following statements are equivalent:*

(1) A is invertible (non-singular),

(2) $A\mathbf{x} = b$ has a unique solution for every $b \in \mathbb{F}^n$,

(3) $A\mathbf{x} = 0$ has only the trivial (zero) solution,

(4) The reduced row echelon form of A is I_n,

(5) A is a product of elementary matrices,

(6) $\det A \neq 0$,

(7) $\operatorname{rank} A = n$,

(8) $N(A) = \{0\}$,

(9) $\operatorname{null} A = 0$,

(10) The column vectors of A are linearly independent,

(11) The column vectors of A span $\mathbb{F}^n$,

(12) The column vectors of A a basis for $\mathbb{F}^n$,

(13) The row vectors of A are linearly independent,

(14) The row vectors of A span $\mathbb{F}^n$,

(15) The row vectors of A a basis for $\mathbb{F}^n$.

1.7.7 Special types of matrices

The following are some special types of matrices in $\mathbb{F}^{n\times n}$.

1. A is *symmetric* if $A^\top = A$. Denote by $\mathrm{S}(n,\mathbb{F})$ the subspace of $n \times n$ symmetric matrices.

2. A is *skew-symmetric*, or *anti-symmetric*, if $A^\top = -A$. Denote by $\mathcal{A}(n,\mathbb{F})$ the subspace of the skew-symmetric matrices.

3. $A = [a_{ij}]$ is called *upper triangular* if $a_{ij} = 0$, for each $j < i$. Denote by $\mathrm{U}(n,\mathbb{F})$ the subspace of upper triangular matrices.

4. $A = [a_{ij}]$ is called *lower triangular* if $a_{ij} = 0$ for each $j > i$. Denote by $L(n,\mathbb{F})$ the subspace of lower triangular matrices.

   A *triangular matrix* is one that is either lower triangular or upper triangular.

   Let $\boldsymbol{\alpha} = \{\alpha_1 < \alpha_2 < \ldots < \alpha_p\} \subset [m], \boldsymbol{\beta} = \{\beta_1 < \beta_2 < \ldots < \beta_p\} \subset [n]$ Then $A[\boldsymbol{\alpha},\boldsymbol{\beta}]$ is an $p \times q$ submatrix of A, which is obtained by erasing in A all rows and columns which are not in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Assume that $p < m, q < n$. Denote by $A(\boldsymbol{\alpha},\boldsymbol{\beta})$ the $(m-p)\times(n-q)$ submatrix of A, which is obtained by erasing in A all rows and columns which are in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. For $i \in [m], j \in [n]$ denote by $A(i,j) := A(\{i\},\{j\})$. Assume that $A \in \mathbb{F}^{n\times n}$. Note that $A[\boldsymbol{\alpha},\boldsymbol{\beta}]$ is called a *principal* if and only if $\boldsymbol{\alpha} = \boldsymbol{\beta}$.

5. $A = [a_{ij}]$ is called *tridiagonal* if $a_{ij} = 0$ for all i, j with $|i-j| > 1$.

6. $A = [a_{ij}]$ is called *upper Hessenberg* if $a_{ij} = 0$ for $i > j+1$.

7. $A \in \mathbb{R}^{n\times n}$ is called *orthogonal* if $AA^\top = A^\top A = I_n$.

1.8 Sum of subspaces

1.8.1 Sum of two subspaces

Let $\mathbf{V}$ be a vector space and $\mathbf{U}$ and $\mathbf{W}$ be its two subspaces. To determine the smallest subspace of $\mathbf{V}$ containing $\mathbf{U}$ and $\mathbf{W}$, we make the following definition.

Definition 1.8.1 *For any two subspaces $\mathbf{U},\mathbf{W}\subseteq\mathbf{V}$ denote, $\mathbf{U}+\mathbf{W}:=\{\mathbf{v}:=\mathbf{u}+\mathbf{w},\mathbf{u}\in\mathbf{U},\mathbf{w}\in\mathbf{W}\}$, where we take all possible vectors $\mathbf{u}\in\mathbf{U}$, $\mathbf{w}\in\mathbf{W}$.*

Theorem 1.8.2 *Let $\mathbf{V}$ be a vector space and $\mathbf{U},\mathbf{W}$ be subspaces in $\mathbf{V}$. Then*

(a) $\mathbf{U}+\mathbf{W}$ and $\mathbf{U}\cap\mathbf{W}$ are subspaces of $\mathbf{V}$.

(b) Assume that $\mathbf{V}$ is finite dimensional. Then

1. *$\mathbf{U},\mathbf{W},\mathbf{U}\cap\mathbf{W}$ are finite dimensional. Let $l=\dim\mathbf{U}\cap\mathbf{W}\geq 0$, $p=\dim\mathbf{U}\geq 1$, $q=\dim\mathbf{W}\geq 1$*
2. *There exists a basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}$ in $\mathbf{U}+\mathbf{W}$ such that $\{\mathbf{v}_1,\ldots,\mathbf{v}_l\}$ is a basis in $\mathbf{U}\cap\mathbf{W}$, $\{\mathbf{v}_1,\ldots,\mathbf{v}_p\}$ is a basis in $\mathbf{U}$ and $\{\mathbf{v}_1,\ldots,\mathbf{v}_l,\mathbf{v}_{p+1},\ldots,\mathbf{v}_{p+q+1}\}$ is a basis in $\mathbf{W}$.*
3. $\dim(\mathbf{U}+\mathbf{W})=\dim\mathbf{U}+\dim\mathbf{W}-\dim\mathbf{U}\cap\mathbf{W}$
   Identity $\#(A\cup B)=\#A+\#B-\#(A\cap B)$ for finite sets A, B is analogous to 3.
4. *$\mathbf{U}+\mathbf{W}\subseteq\operatorname{span}(\mathbf{U}\cup\mathbf{W})$, i.e. $\mathbf{U}+\mathbf{W}$ is the smallest subspaces of $\mathbf{V}$ containing $\mathbf{U}$ and $\mathbf{W}$.*

Proof.
(a) 1. Let $\mathbf{u},\mathbf{w}\in\mathbf{U}\cap\mathbf{W}$. Since $\mathbf{u},\mathbf{w}\in\mathbf{U}$, it follows $a\mathbf{u}+b\mathbf{w}\in\mathbf{U}$. Similarly $a\mathbf{u}+b\mathbf{w}\in\mathbf{W}$. Hence, $a\mathbf{u}+b\mathbf{w}\in\mathbf{U}\cap\mathbf{W}$ and $\mathbf{U}\cap\mathbf{W}$ is a subspace.
(a) 2. Assume that $\mathbf{u}_1,\mathbf{u}_2\in\mathbf{U}$, $\mathbf{w}_1,\mathbf{w}_2\in\mathbf{W}$. Then, $a(\mathbf{u}_1+\mathbf{w}_1)+b(\mathbf{u}_2+\mathbf{w}_2)=(a\mathbf{u}_1+b\mathbf{u}_2)+(a\mathbf{w}_1+b\mathbf{w}_2)\in\mathbf{U}+\mathbf{W}$. Hence, $\mathbf{U}+\mathbf{W}$ is a subspace.
(b) 1. Any subspace of an m-dimensional space has dimension m at most.
(b) 2. Let $\{\mathbf{v}_1,\ldots,\mathbf{v}_l\}$ be a basis in $\mathbf{U}\cap\mathbf{W}$. Complete this linearly independent set in $\mathbf{U}$ and $\mathbf{W}$ to a basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_p\}$ in $\mathbf{U}$ and a basis $\{\mathbf{v}_1,\ldots,\mathbf{v}_l,\mathbf{v}_{p+1},\ldots,\mathbf{v}_{p+q-l}\}$ in $\mathbf{W}$. Hence, for any $\mathbf{u}\in\mathbf{U}$, $\mathbf{w}\in\mathbf{W}$, $\mathbf{u}+\mathbf{w}\in\operatorname{span}(\mathbf{v}_1,\ldots,\mathbf{v}_{p+q-l})$. Hence $\mathbf{U}+\mathbf{W}=\operatorname{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{p+q-l}\}$.
We now show that $\mathbf{v}_1,\ldots,\mathbf{v}_{p+q-l}$ are linearly independent. Suppose that $a_1\mathbf{v}_1+\cdots+a_{p+q-l}\mathbf{v}_{p+q-l}=0$. So $\mathbf{u}:=a_1\mathbf{v}_1+\cdots+a_p\mathbf{v}_p=-a_{p+1}\mathbf{v}_{p+1}+\cdots-a_{p+q-l}\mathbf{v}_{p+q-l}:=\mathbf{w}$. Note that $\mathbf{u}\in\mathbf{U}$, $\mathbf{w}\in\mathbf{W}$. So $\mathbf{w}\in\mathbf{U}\cap\mathbf{W}$. Hence, $\mathbf{w}=b_1\mathbf{v}_1+\cdots+b_l\mathbf{v}_l$. Since $\mathbf{v}_1,\ldots,\mathbf{v}_l,\mathbf{v}_{p+1},\ldots,\mathbf{v}_{p+q-l}$ are linearly independent, then $a_{p+1}=\cdots=a_{p+q-l}=b_1=\cdots=b_l=0$. So $\mathbf{w}=0=\mathbf{u}$. Since $\mathbf{v}_1,\ldots,\mathbf{v}_p$ are linearly independent, then $a_1=\cdots=a_p=0$. Hence, $\mathbf{v}_1,\ldots,\mathbf{v}_{p+q-l}$ are linearly independent.
(b) 3. Note that from (b) 2, $\dim(\mathbf{U}+\mathbf{W})=p+q-l$.
Observe that $\mathbf{U}+\mathbf{W}=\mathbf{W}+\mathbf{U}$.
(b) 4. Clearly, every element $\mathbf{u}+\mathbf{w}$ of $\mathbf{U}+\mathbf{W}$ is in $\operatorname{span}(\mathbf{U}+\mathbf{W})$. Thus, $\mathbf{U}+\mathbf{W}\subseteq\operatorname{span}(\mathbf{U}+\mathbf{W})$. □

Definition 1.8.3 *The subspace $\mathbf{X} := \mathbf{U}+\mathbf{W}$ of $\mathbf{V}$ is called a direct sum of $\mathbf{U}$ and $\mathbf{W}$, if any vector $\mathbf{v} \in \mathbf{U}+\mathbf{W}$ has a unique representation of the form $\mathbf{v} = \mathbf{u}+\mathbf{w}$, where $\mathbf{u} \in \mathbf{U}$, $\mathbf{w} \in \mathbf{W}$. Equivalently, if $\mathbf{u}_1 + \mathbf{w}_1 = \mathbf{u}_2 + \mathbf{w}_2$, where $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{U}$, $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{W}$, then $\mathbf{u}_1 = \mathbf{u}_2$, $\mathbf{v}_1 = \mathbf{v}_2$.*
A direct sum of $\mathbf{U}$ and $\mathbf{W}$ is denoted by $\mathbf{U} \oplus \mathbf{W}$. Here, $\mathbf{U}$ is said to be a complement of $\mathbf{W}$ in $\mathbf{U} \oplus \mathbf{W}$. Also, $\mathbf{U}$ (or $\mathbf{W}$) is called a summand of $\mathbf{X}$.

Proposition 1.8.4 *For two finite dimensional vector subspaces $\mathbf{U}, \mathbf{W} \subseteq \mathbf{V}$, the following are equivalent:*

(a) $\mathbf{U} + \mathbf{W} = \mathbf{U} \oplus \mathbf{W}$

(b) $\mathbf{U} \cap \mathbf{W} = \{0\}$

(c) $\dim \mathbf{U} \cap \mathbf{W} = 0$

(d) $\dim(\mathbf{U} + \mathbf{W}) = \dim \mathbf{U} + \dim \mathbf{W}$

(e) For any bases $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$, $\{\mathbf{w}_1, \dots, \mathbf{w}_q\}$ in $\mathbf{U}$, $\mathbf{W}$, respectively $\{\mathbf{u}_1, \dots, \mathbf{u}_p, \mathbf{w}_1, \dots, \mathbf{w}_q\}$ is a basis in $\mathbf{U} + \mathbf{W}$.

Proof. Straightforward. □

Example 1.8.5 *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{l\times n}$. Then, $N(A) \cap N(B) = N(\begin{pmatrix} A \\ B \end{pmatrix})$. Note that $\mathbf{x} \in N(A) \cap N(B)$ if and only if $A\mathbf{x} = B\mathbf{x} = 0$.*

Example 1.8.6 *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times l}$. Then, $R(A) + R(B) = R((A\, B))$. Note that $\mathbf{x} \in R(A) + R(B)$ is the span of the columns of A and B.*

1.8.2 Sums of many subspaces

Definition 1.8.7 *Let $\mathbf{U}_1, \dots, \mathbf{U}_k$ be k subspaces of $\mathbf{V}$. Then, $\mathbf{X} := \mathbf{U}_1 + \dots + \mathbf{U}_k$ is the subspace consisting all vectors of the form $\mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_k$, where $\mathbf{u}_i \in \mathbf{U}_i$, $i = 1, \dots, k$. $\mathbf{U}_1 + \dots + \mathbf{U}_k$ is called a direct sum of $\mathbf{U}_1, \dots, \mathbf{U}_k$, and denoted by $\oplus_{i=1}^k \mathbf{U}_1 := \mathbf{U}_1 \oplus \dots \oplus \mathbf{U}_k$ if any vector in $\mathbf{X}$ can be represented in a unique way as $\mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_k$, where $\mathbf{u}_i \in \mathbf{U}_i$, $i = 1, \dots, k$.*

Proposition 1.8.8 *For finite dimensional vector subspaces $\mathbf{U}_i \subseteq \mathbf{V}$, $i = 1, \dots, k$, the following statements are equivalent:*

(a) $\mathbf{U}_1 + \dots + \mathbf{U}_k = \oplus_{i=1}^k \mathbf{U}_i$,

(b) $\dim(\mathbf{U}_1 + \dots + \mathbf{U}_k) = \sum_{i=1}^k \dim \mathbf{U}_i$

(c) For any bases $\{u_{1,i}, \dots, \mathbf{u}_{p_i,i}\}$ in $\mathbf{U}_i$, the vectors $\mathbf{u}_{j,i}$ form a basis in $\mathbf{U}_1 + \dots + \mathbf{U}_k$, where $1 \le i \le k$ and $1 \le j \le p_i$.

Proof. $(a) \Rightarrow (c)$. Choose a basis $\{\mathbf{u}_{1,i}, \dots, \mathbf{u}_{p_i,i}\}$ in $\mathbf{U}_i$, $i = 1, \dots, k$. Sine every vector in $\oplus_{i=1}^k \mathbf{U}_i$ has a unique representation as $\mathbf{u}_1 + \cdots + \mathbf{u}_k$, where $\mathbf{u}_i \in \mathbf{U}_i$ for $i = 1, \dots, k$, it follows that $\mathbf{0}$ can be written in the unique form as a trivial linear combination of all $\mathbf{u}_{1,i}, \dots, \mathbf{u}_{p_i,i}$, for $i \in [k]$. So all these vectors are linearly independent and span $\oplus_{i=1}^k \mathbf{U}_i$.
Similarly, $(c) \Rightarrow (a)$.
Clearly $(c) \Rightarrow (b)$.
$(b) \Rightarrow (c)$. It is left as an exercise. □

Note that Proposition 1.8.8 states that the dimension of a direct sum is the sum of the dimensions of its summands.

1.8.3 Dual spaces and annihilators

Let $\mathbf{V}$ be a vector space over the field $\mathbb{F}$. Then $\mathbf{V}' = L(\mathbf{V}, \mathbb{F})$, the space of the linear functionals on $\mathbf{V}$, is called the *dual space*. For $S \subseteq \mathbf{V}$ we define the annihilator of S as

$$S^\perp = \{\mathbf{f} \in \mathbf{V}' : \mathbf{f}(\mathbf{v}) = 0, \text{ for all } \mathbf{v} \in S\}.$$

Theorem 1.8.9 *Let $\mathbf{V}$ be a vector space and $S \subseteq \mathbf{V}$.*

(i) $S^\perp$ is a subspace of $\mathbf{V}'$ (although S does not have to be subspace of $\mathbf{V}$).

(ii) $S^\perp = (\mathrm{span} S)^\perp$.

(iii) Assume that $\mathbf{V}$ is finite dimensional with a basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. Let $i \in \{1, \dots, n\}$ and $V_i = \{\mathbf{v}_1, \dots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \dots, \mathbf{v}_n\}$. Then $\dim V_i^\perp = 1$. Let $\mathbf{f}_i$ be a basis in $V_i^\perp$. Then $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ is a basis in $\mathbf{V}'$. (It is called the dual basis of $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$.) Furthermore, $\mathbf{f}_{i+1}, \dots, \mathbf{f}_n$ is a basis of $\mathrm{span}(\mathbf{v}_1, \dots, \mathbf{v}_i)^\perp$.

Proof. First we show that $S^\perp$ is a subspace of $\mathbf{V}'$. Note that the zero functional obviously sends every vector in S to zero, so $\mathbf{0} \in S^\perp$. If $c, d \in \mathbb{F}$ and $\mathbf{f}_1, \mathbf{f}_2 \in S^\perp$, then for each $\mathbf{v} \in S$

$$(c\mathbf{f}_1 + d\mathbf{f}_2)(\mathbf{v}) = c\mathbf{f}_1(\mathbf{v}) + d\mathbf{f}_2(\mathbf{v}) = 0.$$

Then, $c\mathbf{f}_1 + d\mathbf{f}_2 \in S^\perp$ and $S^\perp$ is a subspace of $\mathbf{V}'$.
Next we show that $S^\perp = (\mathrm{span}\, S)^\perp$. Take $\mathbf{f} \in S^\perp$. Then if $\mathbf{v} \in \mathrm{span} S$ we can write

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m,$$

for scalars $c_i \in \mathbb{F}$ and $\mathbf{v}_i \in S$. Thus

$$\mathbf{f}(\mathbf{v}) = c_1 f(\mathbf{v}_1) + \cdots + c_m\mathbf{f}(\mathbf{v}_m) = 0,$$

so $\mathbf{f} \in (\mathrm{span} S)^\perp$. On the other hand if $\mathbf{f} \in (\mathrm{span} S)^\perp$ then clearly $\mathbf{f}(\mathbf{v}) = 0$, for all $\mathbf{v} \in S$. This completes the proof of (ii).
(iii) Let $\mathbf{f}_i \in \mathbf{V}'$ be given by the equalities $\mathbf{f}_i(\mathbf{v}_j) = \delta_{ij}$ for $j = 1, \dots, n$. Clearly, $\mathbf{f}_i \in V_i^\perp$. Assume that $\mathbf{f} \in V_i^\perp$. Let $\mathbf{f}(\mathbf{v}_i) = a$. Consider $\mathbf{g} = \mathbf{f} - a\mathbf{f}_i$. So $\mathbf{g} \in V_i^\perp$ and $\mathbf{g}(\mathbf{v}_i) = \mathbf{f}(\mathbf{v}_i) - a\mathbf{f}_i(\mathbf{v}_i) = 0$. So $\mathbf{g} \in \{\mathbf{v}_1, \dots, \mathbf{v}_n\}^\perp$. Hence $\mathbf{g} = \mathbf{0}$ and $\mathbf{f} = a\mathbf{f}_i$. Thus $\{\mathbf{f}_i\}$ is a basis in $V_i^\perp$. Observe next that $\mathbf{f}_1, \dots, \mathbf{f}_n$ are linearly independent. Assume that $\sum_{j=1}^n a_j\mathbf{f}_j = \mathbf{0}$. Then $0 = (\sum_{j=1}^n a_j\mathbf{f}_j)(\mathbf{v}_i) = a_i\mathbf{f}_i(\mathbf{v}_i) = a_i$ for $i = 1, \dots, n$. We now show

that $\mathbf{f}_1, \dots, \mathbf{f}_n$ is a basis in $\mathbf{V}'$. Let $\mathbf{f} \in \mathbf{V}'$ and assume that $\mathbf{f}(\mathbf{v}_i) = a_i$ for $i = 1, \dots, n$. Then $g = \mathbf{f} - \sum_{i=1}^n a_i \mathbf{f}_i \in \{\mathbf{v}_1, \dots, \mathbf{v}_n\}^\perp$. Hence $\mathbf{g} = \mathbf{0}$ and $\mathbf{f} = \sum_{i=1}^n \mathbf{f}(\mathbf{v}_i)\mathbf{f}_i$. Assume that $\mathbf{f} \in V_i^{\text{span}}$. Let $\mathbf{f} = \sum_{j=1}^n a_j \mathbf{f}_j$. Assume that $k \in \{1, \dots, i\}$. Hence $0 = \mathbf{f}(\mathbf{v}_k) = a_k$. Hence $\mathbf{f} = \sum_{j=i+1}^n a_j \mathbf{f}_j$. Vice versa, suppose that $\mathbf{f} = \sum_{j=i+1}^n a_j \mathbf{f}_j$. Then $\mathbf{f} \in V_i^{\text{span}}$. Hence $\{\mathbf{f}_{i+1}, \dots, \mathbf{f}_n\}$ is a basis of $V_i^\perp$. □

1.8.4 Worked-out Problems

1. Consider the set of all polynomials of degree two at most with real coefficients and denote it by P_2. P_2 is a real vector space with the usual addition and scalar multiplication. Show that the polynomials $1, x-1, (x-1)^2$ form a basis in this vector space.

   Solution:

   Using Taylor series at $x = 1$, we have $p(x) = p(1) + p'(1)(x-1) + \frac{p''(1)}{2!}(x-1)^2$, where all other terms in Taylor series are zero, i.e. $p^{(k)}(x) = 0$, for $k \geq 3$. Then, $\{1, x-1, (x-1)^2\}$ spans P_2. Suppose that a linear combination of $1, x-1, (x-1)^2$ is identically zero; $p(x) = a + b(x-1) + c(x-1)^2 = 0$. Then, $0 = p(1) = a$, $0 = p'(1) = b$ and $0 = p''(1) = 2c$. Hence, $1, x-1$ and $(x-1)^2$ are linearly independent. Thus, it is a basis for P_2.

2. Let $\mathbf{V}$ be the set of all complex numbers of the form $a + ib\sqrt{5}$, where a, b are rational numbers and $i = \sqrt{-1}$. It is easy to see that $\mathbf{V}$ is a vector space over $\mathbb{Q}$.

   (a) Find the dimension of $\mathbf{V}$ over $\mathbb{Q}$.

   (b) Define a product on $\mathbf{V}$ such that $\mathbf{V}$ is a field.

   Solution:

   (a) Let $\mathbf{u} = 1 + i0\sqrt{5}$ and $\mathbf{v} = 0 + i\sqrt{5}$. Then, $a + ib\sqrt{5} = a\mathbf{u} + b\mathbf{v}$ and so $\{1, i\sqrt{5}\}$ spans $\mathbf{V}$. In addition, $\mathbf{u}$ and $\mathbf{v}$ are linearly independent. As $\mathbf{u}$ and $\mathbf{v}$ are non-zero, suppose that $\mathbf{v} = t\mathbf{u}$, for some $t \in \mathbb{Q}$. Thus, $t = \frac{\mathbf{v}}{\mathbf{u}} = i\sqrt{5}$, which is not a rational number. Hence, $\dim_{\mathbb{Q}} \mathbf{V} = 2$.

   (b) Define
   $$(a + ib\sqrt{5})(c + id\sqrt{5}) = (ac - 5bd) + i(ad + bc)\sqrt{5}.$$
   If $(a, b) \neq 0$, then $(a + ib\sqrt{5})^{-1} = \frac{a - ib\sqrt{5}}{a^2 + 5b^2}$.

3. Let $A \in \mathbb{F}^{n \times n}$ be an invertible matrix. Show that

   (a) $(A^{-1})^\top = (A^\top)^{-1}$

   (b) A is symmetric if and only if A^{-1} is symmetric.

   Solution:

   (a) $(A^{-1})^\top A^\top = (AA^{-1})^\top = (I_n)^\top = I_n$. Then, $(A^{-1})^\top = (A^\top)^{-1}$.

   (b) Assume that A is symmetric. Using part (a), we have $(A^{-1})^\top = (A^\top)^{-1}$. On the other hand $A^\top = A$. Then, $(A^{-1})^\top = A^{-1}$. The Converse is obtained by considering $(A^{-1})^{-1} = A$ and using the same argument.

4. If $A \in \mathbb{F}^{m\times n}$, show that the dimension of its row space equals the dimension of its column space.
   Solution:
   Use the row operations to find an equivalent echelon form matrix C. Using Theorem 1.7.14, we know that the dimension of the row space of A equals the number of non-zero rows of C. Note that each non-zero row of C has exactly one pivot and the different rows have pivots in different columns. Therefore, the number of pivot columns equals the number of non-zero rows. But by Theorem 1.7.14, the number of pivot columns of C equals the number vector in a basis for the column space of A. Thus, the dimension of the column space equals the number of non-zero rows of B. This means the dimensions of the row space and column space are the same.

5. Let A and B be $n \times n$ matrices over the field $\mathbb{F}$. Prove that if $I - AB$ is invertible, then $I - BA$ is in invertible and $(I - BA)^{-1} = I + B(I - AB)^{-1}A$.
   Solution:

$$(I - BA)(I + B(I - AB)^{-1}A) = I - BA + B(I - AB)^{-1}A - BAB(I - AB^{-1})A.$$

   After the identity matrix, we can factor B on the left and A on the right, and we get:

$$\begin{aligned}(I - BA)(I + B(I - AB)^{-1}A) &= I - B[-I + (I - AB)^{-1} - AB(I - AB)^{-1}]A \\ &= I - B[-I + (I - AB)(I - AB)^{-1}]A \\ &= I + 0 = I.\end{aligned}$$

1.8.5 Problems

1. Let $A \in \mathbb{F}^{n\times n}$ be an invertible matrix. Show that
   (a) A is skew-symmetric if and only if A^{-1} is skew-symmetric.
   (b) A is lower triangular if and only if A^{-1} is lower triangular.
   (c) A is upper triangular if and only if A^{-1} is upper triangular.

2. Let $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ with the following property. Its row echelon form (REF) is an upper triangular matrix U with 1's on the main diagonals and there is no pivoting. That is, first $a_{11} \neq 0$. Apply the Gauss elimination to obtain a matrix $A_1 = [a_{ij}^{(1)}]$, where the first column of A_1 is $\mathbf{e}_1$ the first column of the identity. Let $B_1 = [a_{ij}^{(1)}]_{i=j=2}^{n}$ be the matrix obtained by deleting the first row and column of A_1. Then, B_1 satisfies the same assumptions as A, i.e. $a_{22}^{(1)} \neq 0$. Now continue as above to obtain U. Show that $A = LU$, where L is a corresponding nonsingular lower triangular matrix.

3. Let $A \in \mathbb{F}^{l\times m}, B \in \mathbb{F}^{m\times n}$. Show rank $AB \leq \min\{\text{rank } A, \text{rank } B\}$. Furthermore, if A or B are invertible, then equality holds. Give an example where inequality holds strictly.
   (Hint: Use the fact that each column of AB is a combination of the columns of A and then use Worked-out Problem 1.8.4-4.)

4. Prove Theorem 1.7.17.

5. Let $\mathbf{V}$ be a finite dimensional vector space and $\mathbf{U}$ be a subspace of $\mathbf{V}$. Prove that $\mathbf{U}$ has a complement in $\mathbf{V}$.

6. Let $\mathbf{V}$ be a finite dimensional vector space and $\mathbf{U}$ is a subspace of $\mathbf{V}$. Suppose $\mathbf{W}_1$ and $\mathbf{W}_2$ are complements to $\mathbf{U}$ in $\mathbf{V}$. Prove that $\mathbf{W}_1$ and $\mathbf{W}_2$ are isomorphic.

7. Prove the following statements:

   (a) Row equivalence is an equivalence relation.

   (b) Row equivalent matrices have the same row space.

   (c) Let $\mathbb{F}$ be a field. Then the set of matrices of rank r, whose rows, (viewed as column vectors), span a given subspace $\mathbf{U} \subseteq \mathbb{F}^n$ of dimension r, correspond exactly one row equivalence class in $\mathbb{F}^{m\times n}$.

   The set of subspaces of given dimension in a fixed vector space is called *Grassmannian*. In part (c) of the above problem, we have constructed a bijection between the Grassmannian of r-dimensional subspaces of $\mathbb{F}^n$ (denoted by $\mathrm{Gr}(r, \mathbb{F}^m)$) and the set of reduced row *e*chelon matrices with n columns and r non-zero rows. We will study this notion in Section 5.10

1.9 Permutations

1.9.1 The permutation group

Denote by $\mathcal{S}_n$ the group of the permutations of the set $[n] := \{1, \dots, n\}$ onto itself. Indeed $\mathcal{S}_n = \mathcal{S}([n])$. The smallest integer $N > 0$ such that for $\omega \in \mathcal{S}_n$, $\omega^N = id$ is called the *order* of ω. If $a_1, \dots, a_r \in [n]$ are distinct, then the symbol $(a_1, \dots, a_r)$ denotes the permutation ω of $[n]$ which sends a_1 to a_2, sends a_2 to a_3, ..., sends a_{r-1} to a_r, sends a_r to a_1, and fixes all the other numbers in $[n]$. In other words, $\omega(a_1) = a_2$, $\omega(a_2) = a_3$, ..., $\omega(a_r) = a_1$ and $\omega(i) = i$, if $i \in \{a_1, \dots, a_r\}$. Such a permutation is called *cyclic*. The number r is called the *length* of the cycle. For example the permutation of $\mathcal{S}_4$ that sends 1 to 3, 3 to 2, 2 to 4 and 4 to 1 is cyclic, while the permutation that sends 1 to 3, 3 to 1, 2 to 4 and 4 to 2 is not.

Remark 1.9.1 *The cyclic permutation ω by $(a_1, \dots, a_r)$ has order r. Note that $\omega(a_1) = a_2$, $\omega^2(a_1) = a_3$, ..., $\omega^{r-1}(a_1) = a_r$, $\omega^r(a_1) = a_1$, by definition of ω. Likewise, for any $i = 1, \dots, r$, we have $\omega^r(a_i) = a_i$.*

Now, introduce the following polynomials

$$P_\omega(\mathbf{x}) := \prod_{1\le i<j\le n} (x_{\omega(i)} - x_{\omega(j)}), \mathbf{x} = (x_1, \dots, x_n), \text{ for each } \omega \in \mathcal{S}_n. \tag{1.9.1}$$

Define

$$\operatorname{sign}(\omega) := \frac{P_\omega(\mathbf{x})}{P_{\mathrm{id}}(\mathbf{x})}. \tag{1.9.2}$$

Theorem 1.9.2 *For each* $\omega \in \mathcal{S}_n$*,* $\mathrm{sign}(\omega) \in \{1,-1\}$*. The map* $\mathrm{sign} : \mathcal{S}_n \to \{1,-1\}$ *is a group homomorphism, i.e.*

$$\mathrm{sign}(\omega \circ \sigma) = \mathrm{sign}(\omega)\mathrm{sign}(\sigma), \textit{ for each } \omega, \sigma \in \mathcal{S}_n. \tag{1.9.3}$$

Proof. Clearly, if $\omega(i) < \omega(j)$, then the factor $(x_{\omega(i)} - x_{\omega(j)})$ appears in $P_{\mathrm{id}}(\mathbf{x})$. If $\omega(i) > \omega(j)$, then the factor $-(x_{\omega(i)} - x_{\omega(j)})$ appears in $P_{\mathrm{id}}(\mathbf{x})$. Hence $\mathrm{sign}(\omega) \in \{1,-1\}$. Observe next that

$$\frac{P_{\omega\circ\sigma}(\mathbf{x})}{P_{\mathrm{id}}(\mathbf{x})} = \frac{P_{\omega\circ\sigma}(\mathbf{x})}{P_{\sigma}(\mathbf{x})}\frac{P_{\sigma}(\mathbf{x})}{P_{\mathrm{id}}(\mathbf{x})} = \mathrm{sign}(\omega)\mathrm{sign}(\sigma).$$

(To show the equality $\frac{P_{\omega\circ\sigma}(\mathbf{x})}{P_{\sigma}(\mathbf{x})} = \mathrm{sign}(\omega)$, introduce new variables $y_i = x_{\sigma(i)}$, for $i \in [n]$.)

□

An element $\tau \in \mathcal{S}_n$ is called a *transposition* if there exists a pair of integers $1 \le i < j \le n$ such that $\tau(i) = j, \tau(j) = i$. Furthermore, $\tau(k) = k$ for $k \ne i, j$. (Transposition is a cycle of length 2.) Note that $\tau \circ \tau$ =id, i.e. $\tau^{-1} = \tau$.

Theorem 1.9.3 *For an integer* $n \ge 2$*, each* $\omega \in \mathcal{S}_n$ *is a product of transpositions*

$$\omega = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_m. \tag{1.9.4}$$

The parity of m *is unique. More precisely,* $\mathrm{sign}(\omega) = (-1)^m$*.*

Proof. We agree that in (1.9.4), $m = 0$ if and only if $\omega = \mathrm{id}$. (This is true for any $n \in \mathbb{N}$.) We prove the theorem by induction on n. For $n = 2$, $\mathcal{S}_2$ consists of two elements id and a unique permutation τ, which satisfies $\tau(1) = 2, \tau(2) = 1$.) Thus, $\tau \circ \tau$ =id. In this case, the lemma follows straightforward.

Suppose that theorem holds for $n = N \ge 2$ and assume that $n = N+1$. Let $\omega \in \mathcal{S}_n$. Suppose first that $\omega(n) = n$. Then, ω can be identified with the bijection $\omega' \in \mathcal{S}_{n-1}$, where $\omega'(i) = \omega(i)$, for $i = 1, \dots, n-1$. Use the induction hypothesis to express ω' as a product of m transposition $\tau_1' \circ \tau_2' \circ \cdots \circ \tau_m'$ in $\mathcal{S}_{n-1}$. Clearly, each τ_i' extends to a transposition τ_i in $\mathcal{S}_n$. Hence, (1.9.4) holds.

Suppose now that $\omega(n) = i < n$. Let τ be the transposition that interchange i and n. Let $\omega' = \tau \circ \omega$. Then, $\omega'(n) = \tau(\omega(n)) = n$. The previous arguments show that $\omega' = \tau_1 \circ \ldots \circ \tau_l$. Therefore, $\omega = \tau \circ \omega' = \tau \circ \tau_1 \circ \ldots \circ \tau_l$.

It is left to show that the parity of m is unique. First observe that $\mathrm{sign}(\tau) = -1$. Using Problem 1.9.3-4 and Theorem 1.9.2, we obtain $\mathrm{sign}(\omega) = (-1)^m$. Hence, the parity of m is fixed. □

Remark 1.9.4 *Note that the product decomposition is not unique if* $n > 1$*. For example for* $(1,2,3,4,5)$ *we have the following product decompositions:*

$$\begin{aligned}\underbrace{(5,4)(5,6)(2,1)(2,5)(2,3)(1,3)}_{6 \textit{ transpositions}} &= \underbrace{(1,2)(2,3)(3,4)(4,5)}_{4 \textit{ transpositions}} \\ &= \underbrace{(1,5)(1,4)(1,3)(1,2)}_{4 \textit{ transpositions}} = \cdots\end{aligned}$$

We finish this section with the definition of a permutation matrix.

Definition 1.9.5 *A permutation matrix is a square matrix obtained from the same size identity matrix by a permutation of rows. A permutation matrix is called elementary if it is obtained by permutation of exactly two distinct rows.*

Clearly, every permutation matrix has exactly one 1 in each row and column. It is easy to show that every elementary permutation matrix is symmetric. Note that an elementary permutation matrix corresponds to a transposition in $\mathcal{S}_n$. Theorem 1.9.3 implies that every permutation matrix is a product of elementary row-interchange matrices. Note that a general permutation matrix is not symmetric.

We denote by $\mathcal{P}_n$ the set of all permutation matrices. It is easy to see that $\#\mathcal{P}_n = n!$. Indeed, there is a one-to-one correspondence between $\mathcal{P}_n$ and $\mathcal{S}_n$, the set of all permutations of $[n]$ and every $P = [p_{ij}] \in \mathcal{P}_n$ is associated to a permutation $\sigma \in \mathcal{S}_n$ for which $\sigma(i) = j$ if $p_{ij} = 1$.
In Section 1.13, we will be familiarized with a new family of matrices called doubly stochastic matrices and will see their relation with permutation matrices.

1.9.2 Worked-out Problems

1. In $\mathcal{S}_n$ consider all elements ω such that $\text{sign}(\omega) = 1$, denote it by A_n.

   (a) Show that it is a subgroup of $\mathcal{S}_n$.

   (b) How many elements it has?

   (c) Show that $\mathcal{S}_3$ is not a commutative group.

   Solution:

   (a) Since $id \in A_n$, then A_n is a non-empty subset of $\mathcal{S}_n$. Now, if σ and ω are two elements of A_n, then

   $$\text{sign}(\sigma\omega^{-1}) \overset{(1)}{=} \text{sign}(\sigma)\text{sign}(\omega^{-1}) \overset{(2)}{=} \text{sign}(\sigma)\text{sign}(\omega) = 1.$$

   (1): Use Theorem 1.9.3.
   (2): Use Problem 1.9.3-2.
   Then, $\sigma\omega^{-1} \in A_n$ and this means A_n is a subgroup of $\mathcal{S}_n$. Note that A_n is called to be the *alternating group* on n elements.

   (b) Let $\sigma \in \mathcal{S}_n$ be a transposition. Then, for any $\omega \in A_n$, $\text{sign}(\sigma\omega) = \text{sign}(\sigma)\text{sign}(\omega) = (-1)(1) = -1$. It follows $\sigma A_n \cup A_n = \mathcal{S}_n$. Since $\#\sigma A_n = \#A_n$ and $A_n \cap \sigma A_n = \varnothing$, then $\#A_n = \frac{\#\mathcal{S}_n}{2} = \frac{n!}{2}$.

   (c) Take $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$, $\omega(1) = 2$, $\omega(2) = 1$ and $\omega(3) = 3$. We have

   $$\begin{aligned} (\sigma\omega)(1) &= \sigma(\omega(1)) = \sigma(2) = 3, \\ (\omega\sigma)(1) &= \omega(\sigma(1)) = \omega(2) = 1. \end{aligned}$$

   Thus, $\sigma\omega \neq \omega\sigma$ and $\mathcal{S}_3$ is not commutative.

1.9.3 Problems

Assume that ω and$\tau \in \mathrm{S}_n$. Show that

1. $\mathrm{sign}(\mathrm{id}) = 1$.

2. $\mathrm{sign}(\omega^{-1}) = \mathrm{sign}(\omega)$.
(Hint: Use the previous problem and the fact that $\mathrm{sign}(\omega\omega^{-1}) = \mathrm{sign}(\mathrm{id})$.)

3. Assume that $1 \le i < j \le n$. We say that ω changes the order of the pair (i, j) if $\omega(i) > \omega(j)$. Let $N(\omega)$ be number of pairs (i, j) such that ω changes their order. Then, $\mathrm{sign}(\omega) = (-1)^{N(\omega)}$.

4. Assume that $1 \le i < j \le n$. Let τ be the transposition $\tau(i) = j, \tau(j) = i$. Then, $N(\tau) = 2(j-i) - 1$. Hence, $\mathrm{sign}(\tau) = -1$ for any transposition τ.

5. Give an example of permutation matrix which is not symmetric.

6. Let $\mathcal{X}$ be a set. Show that $\mathcal{S}(\mathcal{X})$ is a group.

7. Show that if $\mathcal{X}$ has one or two elements, then $\mathcal{S}(\mathcal{X})$ is a commutative group.

8. Show that if $\mathcal{X}$ has at least three elements, then $\mathcal{S}(\mathcal{X})$ is not a commutative group.

9. Show that $Z(S_n) = \{1\}$, for $n \ge 3$.

10. Let $A_n = \{\sigma \in S_n; \mathrm{sign}(\omega) = 1\}$. Show that

    (a) $A_n \triangleleft S_n$,

    (b) $S_n/A_n \approx \mathbb{Z}_2$,

    (c) $\#A_n = \frac{n!}{2}$.

    Note that A_n is called an *alternating group* on $[n]$. See Worked-out Problem 1.9.2-1.

1.10 Linear, multilinear maps and functions

Roughly speaking, a linear operator is a function from one vector space to another that preserves the vector space operations. In what follows, we give its precise definition and related concepts.

Let $\mathbf{U}$ and $\mathbf{V}$ be two vector spaces over a field $\mathbb{F}$. A map $T : \mathbf{U} \to \mathbf{V}$ is called *linear* or a *linear operator* (linear transformation) if $T(a\mathbf{u} + b\mathbf{v}) = aT(\mathbf{u}) + bT(\mathbf{v})$, for all $a, b \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in \mathbf{V}$. The set of linear maps from $\mathbf{U}$ to $\mathbf{V}$ is denoted by $L(\mathbf{U}, \mathbf{V})$. For the case $\mathbf{U} = \mathbf{V}$, we simply use the notation $L(\mathbf{V})$. A linear operator $T : \mathbf{U} \to \mathbf{V}$ is called *linear isomorphism* if T is bijective. In this case, $\mathbf{U}$ is called to be *isomorphic* to $\mathbf{V}$ and it is denoted by $\mathbf{U} \cong \mathbf{V}$. Also, the *kernel* and *image* of T are denoted by $\ker T$ and $\mathrm{Im}\, T$, respectively and defined as follows:

$$\begin{aligned} \ker T &= \{\mathbf{x} \in \mathbf{U};\ T(\mathbf{x}) = \mathbf{0}\}, \\ \mathrm{Im}\, T &= \{T(\mathbf{x});\ \mathbf{x} \in \mathbf{U}\}. \end{aligned}$$

It is easy to show that $\ker T$ and $\mathrm{Im}\ T$ are subspaces of $\mathbf{U}$ and $\mathbf{V}$, respectively. Furthermore, the *nullity* and the *rank* of T are denoted by $\mathrm{null}\, T$ and $\mathrm{rank}\, T$ and defined as follows:

$$\begin{aligned} \mathrm{null}\, T &= \dim(\ker T), \\ \mathrm{rank}\, T &= \dim(\mathrm{Im}\, T). \end{aligned}$$

Let $\mathbf{U}_1, \ldots, \mathbf{U}_k$ be vector spaces ($k \ge 2$). Denote by $\overset{k}{\underset{i=1}{\times}} \mathbf{U}_i := \mathbf{U}_1 \times \ldots \times \mathbf{U}_k$ the set of all tuples $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$, where $\mathbf{u}_i \in \mathbf{U}_i$, for $i \in [k]$. A map $T : \overset{k}{\underset{i=1}{\times}} \mathbf{U}_i \to \mathbf{V}$ is called *multilinear* or *a multilinear operator* if it is linear with respect to each variable $\mathbf{u}_i$, while all other variable are fixed, i.e.

$$T(\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, a\mathbf{v}_i + b\mathbf{w}_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_k) =$$
$$aT(\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{v}_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_k) + bT(\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{w}_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_k),$$

for all $a, b \in \mathbb{F}$, $\mathbf{u}_j \in \mathbf{U}_j, j \in [k] \smallsetminus \{i\}, \mathbf{v}_i, \mathbf{w}_i \in \mathbf{U}_i$ and $i \in [k]$.

Note that if $k = 2$, then T is called *bilinear*.

Suppose that $\mathbf{U}_1 = \ldots = \mathbf{U}_k = \mathbf{U}$. Denote $\underset{k}{\times}\mathbf{U} := \underbrace{\mathbf{U} \times \ldots \times \mathbf{U}}_{k}$. A map $T : \underset{k}{\times}\mathbf{U} \to \mathbf{V}$ is called a *symmetric* map if

$$T(\mathbf{u}_{\sigma(1)}, \ldots, \mathbf{u}_{\sigma(k)}) = T(\mathbf{u}_1, \ldots, \mathbf{u}_k),$$

for each permutation $\sigma \in \mathcal{S}_k$ and each $\mathbf{u}_1, \ldots, \mathbf{u}_k$. A map $T : \underset{k}{\times}\mathbf{U} \to \mathbf{V}$ is called a *skew-symmetric* map if

$$T(\mathbf{u}_{\sigma(1)}, \ldots, \mathbf{u}_{\sigma(k)}) = \mathrm{sign}(\sigma) T(\mathbf{u}_1, \ldots, \mathbf{u}_k), \text{ for each permutation } \sigma \in \mathcal{S}_k.$$

Since each permutation is a product of transpositions, and $\mathrm{sign}(\tau) = -1$, for any transposition, a map T is skew-symmetric if and only if

$$T(\mathbf{u}_{\tau(1)}, \ldots, \mathbf{u}_{\tau(k)}) = -T(\mathbf{u}_1, \ldots, \mathbf{u}_k), \text{ for each transposition } \tau \in \mathcal{S}_k. \tag{1.10.1}$$

In this book most of the maps are linear or multilinear. If $\mathbf{V} = \mathbb{F}$, then the map T is called linear, multilinear, symmetric and skew-symmetric functions, respectively.

Example 1.10.1 *Consider the multilinear map $T : \mathbb{F}^m \times \mathbb{F}^n \to \mathbb{F}^{m\times n}$ given by $T(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^\top$. Then, $T(\mathbb{F}^m \times \mathbb{F}^n)$ is the set of all matrices of rank one and zero.*

Example 1.10.2 *Consider T and $Q : \mathbb{F}^m \times \mathbb{F}^m \to \mathbb{F}^{m\times m}$ given by $T(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^\top + \mathbf{y}\mathbf{x}^\top$, $Q(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^\top - \mathbf{y}\mathbf{x}^\top$. Then, T and Q are multilinear symmetric and skew-symmetric map, respectively. Also, $T(\mathbb{F}^m \times \mathbb{F}^m)$ is the set of all symmetric matrices of rank two, one and zero.*

Examples of linear operators

1. If $A \in \mathbb{F}^{m\times n}$, then the map $\begin{aligned} T_A : \mathbb{F}^n &\to \mathbb{F}^m \\ \mathbf{x} &\mapsto A\mathbf{x} \end{aligned}$ is linear.

2. The derivative map $D : \mathbb{R}[z] \to \mathbb{R}[z]$ given by $D(f) = \frac{df}{dz}$ is linear (Here, $\mathbb{R}[z]$ denotes all polynomials with variable z and real coefficients and $f \in \mathbb{R}[z]$.)

3. The integral map $T : \mathbb{R}[z] \to \mathbb{R}[z]$ given by $T(f) = \int_0^1 f(z)dz$ is linear.

Remark 1.10.3 *Consider the linear operator given in example 1. The function T_A is called the linear operator associated to the matrix A.*

Notation 1.10.4 *The vectors $\mathbf{e}_1 = (1,0,\dots,0)^\top$, $\mathbf{e}_2 = (0,1,\dots,0)^\top$, …, $\mathbf{e}_n = (0,\dots,0,1)^\top$ in $\mathbb{F}^n$ will be called the standard basis vectors. Generally, all the components of $\mathbf{e}_i$'s are zero except for the i-th component which is one. Any vector $\mathbf{x} = (x_1,\dots,x_n)^\top \in \mathbb{F}^n$ may be written $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$.*

Theorem 1.10.5 *Let $T : \mathbb{F}^n \to \mathbb{F}^m$ be a linear operator. Then, $T = T_A$ for some $A \in \mathbb{F}^{m\times n}$. The j-th column of A equals $T(\mathbf{e}_j)$, for $j = 1,\dots,n$.*

Proof. For any $\mathbf{x} = (x_1,\dots,x_n)^\top \in \mathbb{F}^n$, we have $T(\mathbf{x}) = T\left(\sum_{i=1}^n x_i\mathbf{e}_i\right) = \sum_{i=1}^n x_iT(\mathbf{e}_i)$. Consequently, T is entirely determined by its value on the standard basis vectors, which we may write

$$T(\mathbf{e}_1) = (a_{11}, a_{21}, \dots, a_{m1})^\top, \dots, T(\mathbf{e}_n) = (a_{1n}, a_{2n}, \dots, a_{mn})^\top.$$

Combining this with the previous formula, we obtain

$$T(\mathbf{x}) = x_1\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + x_n\begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n a_{ij}x_j \\ \vdots \\ \sum_{j=1}^n a_{mj}x_j \end{bmatrix}$$

Then, $T = T_A$ for the matrix $A = [a_{ij}]_{m\times n}$. □

Corollary 1.10.6 *There is a one-to-one correspondence between linear operators $T : \mathbb{F}^n \to \mathbb{F}^m$ and $m \times n$ matrices, i.e. $\mathbb{F}^{m\times n}$.*

Definition 1.10.7 *An inner product space is a vector space $\mathbf{V}$ over the field $\mathbb{F}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$), together with an inner product, i.e. with a map $\langle\cdot,\cdot\rangle : \mathbf{V}\times\mathbf{V} \to \mathbb{F}$ that satisfies the following three axioms for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$ and all scalars $a \in \mathbb{F}$:*

(i) Conjugate symmetry:

$$\langle\mathbf{x},\mathbf{y}\rangle = \overline{\langle\mathbf{y},\mathbf{x}\rangle}$$

(See Section 1.9 for the above notation).

(ii) Linearity in the first component:

$$\langle a\mathbf{x}, \mathbf{y}\rangle = a\langle \mathbf{x}, \mathbf{y}\rangle$$
$$\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle = \langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle$$

(iii) Positive-definiteness:

$$\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$$
$$\langle \mathbf{x}, \mathbf{x}\rangle = 0 \Rightarrow \mathbf{x} = 0$$

We will discuss about this map with more details in chapter 5.

Definition 1.10.8 *For vectors* $\mathbf{u} = (u_1, u_2, u_3)^\top$ *and* $\mathbf{v} = (v_1, v_2, v_3)^\top$ *in* $\mathbb{R}^3$*, the cross product is defined by*

$$\mathbf{u} \times \mathbf{v} = (u_2v_3 - u_3v_2, -u_1v_3 + u_3v_1, u_1v_2 - u_2v_1)^\top.$$

Example 1.10.9 *Assume that* $\mathbf{U} = \mathbf{V} = \mathbb{R}^3$*. Let* $T(\mathbf{x}, \mathbf{y}) = \mathbf{x} \times \mathbf{y}$ *be the cross product in* $\mathbb{R}^3$*. Then,* $T : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ *is a bilinear skew-symmetric map.*

Example 1.10.10 *If a vector space* $\mathbf{V}$ *over the real numbers carries an inner product, then the inner product is a bilinear map* $\mathbf{V} \times \mathbf{V} \to \mathbb{R}$*.*

Example 1.10.11 *Matrix multiplication is a bilinear map,* $\mathbb{F}^{m\times n} \times \mathbb{F}^{n\times p} \to \mathbb{F}^{m\times p}$*.*

Remark 1.10.12 *Note that in the definition of a linear operator, the background fields of the vector spaces of the domain and codomain must be the same. For example, consider the identity linear operator* $\begin{array}{rl}\mathrm{id} : \mathbf{V} & \to \mathbf{V} \\ \mathbf{x} & \mapsto \mathbf{x}\end{array}$*. If the background fields of the vector space* $\mathbf{V}$ *as domain and codomain are not the same, then the identity function is not a linear operator necessarily. If we look at the identity function* $\mathrm{id} : \mathbb{C} \to \mathbb{C}$*, where* $\mathbb{C}$ *in the domain is a vector space over* $\mathbb{R}$ *and it is a vector space over* $\mathbb{Q}$ *for codomain. If* id *is a linear operator, then it is a linear isomorphism. This contradicts the case* $\dim_{\mathbb{R}} \mathbb{C} = 2 < \dim_{\mathbb{Q}} \mathbb{C}$*. (See the first Worked-out Problem.)*

1.10.1 Worked-out Problems

1. Let $\mathbf{V}$ and $\mathbf{W}$ be finite dimensional vector spaces over the field $\mathbb{F}$. Show that $\mathbf{V} \cong \mathbf{W}$ if and only if $\dim \mathbf{V} = \dim \mathbf{W}$.
Solution:
Assume that $T : \mathbf{V} \to \mathbf{W}$ is a linear isomorphism and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for $\mathbf{V}$. We claim that $\{T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)\}$ is a basis for $\mathbf{W}$. First, we verify that $T(\mathbf{v}_i)$'s are linearly independent. Suppose that $\sum_{i=1}^n c_iT(\mathbf{v}_i) = \mathbf{0}$, for some $c_i \in \mathbb{F}$. Since T is linear, then $T\left(\sum_{i=1}^n c_i\mathbf{v}_i\right) = \mathbf{0} = T(\mathbf{0})$ and as T is injective, so $\sum_{i=1}^n c_i\mathbf{v}_i = 0$. Since $\mathbf{v}_i$'s are linearly independent, thus $c_i = 0$, $1 \leq i \leq n$. This verifies that $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$ are linearly independent. Next, we show that $T(\mathbf{v}_i)$'s span $\mathbf{W}$. Choose $\mathbf{y} \in \mathbf{W}$, since T is surjective, one can find $\mathbf{x} \in \mathbf{V}$, for which $T(\mathbf{x}) = \mathbf{y}$. Now, if $\mathbf{x} = \sum_{i=1}^n c_i\mathbf{v}_i$, for some $c_i \in \mathbb{F}$, then $\mathbf{y} = \sum_{i=1}^n c_iT(\mathbf{v}_i)$. Thus, $\dim \mathbf{W} = n$. Conversely, assume that $\dim \mathbf{V} = \dim \mathbf{W}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ are bases for $\mathbf{V}$ and $\mathbf{W}$, respectively. Then, the map $\mathbf{v}_i \mapsto \mathbf{w}_i$ is a linear isomorphism from $\mathbf{V}$ to $\mathbf{W}$.

2. Prove the rank-nullity theorem:
   For any $A \in \mathbb{F}^{m\times n}$ we have

$$\operatorname{rank} A + \operatorname{null} A = n.$$

   Solution:
   If $\operatorname{rank} A = 0$, then by Theorem 1.7.17, the only solution to $A\mathbf{x} = 0$ is the trivial solution $\mathbf{x} = 0$. Hence, in this case, $\operatorname{null} A = 0$ and the statement holds. Now suppose that $\operatorname{rank} A = r < n$. In this case, there are $n - r > 0$ free variables in the solution to $A\mathbf{x} = 0$. Let $t_1, \ldots, t_{n-r}$ denote these free variables (chosen as those variable not attached to a leading one in any row-echelon form of A), and let $\mathbf{x}_1, \ldots, \mathbf{x}_{n-r}$ denote the solution obtained by sequentially setting each free variable to 1 and the remaining free variables to zero. Note that $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\}$ is linearly independent. Moreover, every solution to $A\mathbf{x} = 0$ is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_{n-r}$ which shows that $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-r}\}$ spans $N(A)$. Then, $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-r}\}$ is a basis for $N(A)$ and $\operatorname{null} A = n - r$.

1.10.2 Problems

1. If $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are vector spaces over the field $\mathbb{F}$, show that

   (a) If $T_1 \in L(\mathbf{U}, \mathbf{V})$ and $T_2 \in L(\mathbf{V}, \mathbf{W})$, then $T_2 \circ T_1 \in L(\mathbf{U}, \mathbf{W})$.

   (b) $\operatorname{null} T_2 \circ T_1 \leq \operatorname{null} T_1 + \operatorname{null} T_2$.

   (Hint: Use Worked-out Problem 1.10.1-2.)

2. Assume that $\mathbf{V}$ is a vector space over the field $\mathbb{F}$. Let $\mathbf{W}$ be a subspace of $\mathbf{V}$. Show that there exists a $T_1 \in L(\mathbf{V}, \mathbb{F})$ such that $\ker(T_1) = \mathbf{W}$. Show also that there is a $T_2 \in L(\mathbf{V}, \mathbb{F})$ such that $\operatorname{Im}(T_2) = \mathbf{W}$.

3. If $\mathbf{U}$ and $\mathbf{V}$ are vector spaces over the field $\mathbb{F}$, show that $L(\mathbf{U}, \mathbf{V})$ is a vector space over $\mathbb{F}$, too.

4. Consider the matrices $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \in \mathbb{R}^{2\times 2}$ and $B = \begin{bmatrix} e^{i\theta} \end{bmatrix} \in \mathbb{C}^{1\times 1} = \mathbb{C}$. Show that T_A and T_B both are counterclockwise rotation through angle θ. (Then, A and B represent the same motion.)

1.11 Definition and properties of the determinant

Given a square matrix A. It is important to be able to determine whether or not A is invertible, and if it is invertible, how to compute A^{-1}. This arises, for example, when trying to solve the non-homogeneous equation $A\mathbf{x} = \mathbf{y}$, or trying to find eigenvalues by determining whether or not $A - \lambda I$ is invertible. (See Section 3.1 to find the definition and details on eigenvalues.)

1.11.1 Geometric interpretation of determinant (First encounter)

Let A be a $n \times n$ real matrix. As we mentioned already, we can view A as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ given by $\mathbf{x} \to A\mathbf{x}$.

If A is a 2×2 matrix with column vectors $\mathbf{a}$ and $\mathbf{b}$, then the linearity means that A transforms the unit square in $\mathbb{R}^2$ into the parallelogram in $\mathbb{R}^2$ determined by $\mathbf{a}$ and $\mathbf{b}$. (See Figure 1.1)

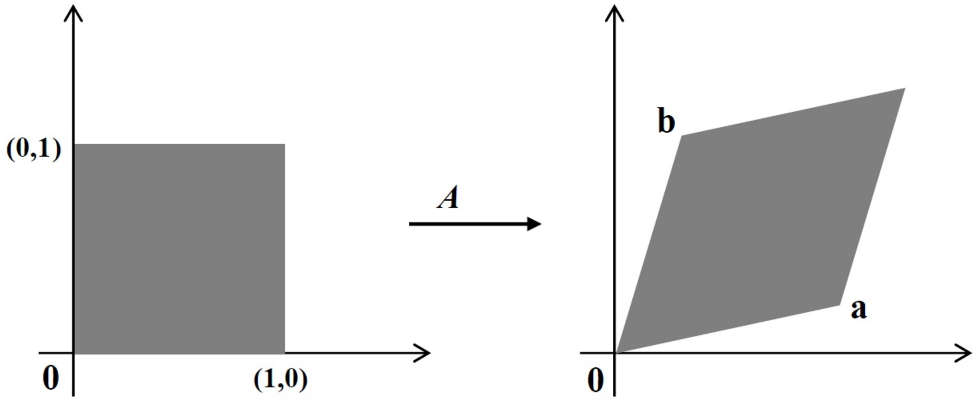


Figure 1.1: Effect of the matrix $A = [\mathbf{a}|\mathbf{b}]$ on the unit square.

Similarly, in the 3×3 case, A maps the unit in $\mathbb{R}^3$ into the parallelepiped (or solid parallelogram) in $\mathbb{R}^3$ determined by the column vectors of A. In general, an $n \times n$ matrix A maps the unit n-cube in $\mathbb{R}^n$ into the n-dimensional parallelogram determined by the column vectors of A.
Other squares (or cubes, or hypercubes, etc.) are transformed in much the same way and scaling the sides of the squares merely scales the sides of the parallelograms (or parallelepipeds, or higher dimensional parallelograms) by the same amount. In particular, the magnification factor

$$\frac{\text{area (or volume) of image region}}{\text{area (or volume) of original region}}$$

is always the same, no matter which squares (or cubes, or hypercubes) we start with. Indeed, since we can calculate the areas (or volumes) of reasonably nice regions by covering them with little squares (or cubes) and taking limits, the above ratio will still be the same for these regions, too.

Definition 1.11.1 *The absolute value of the determinant of the matrix A is the above magnification factor.*

For example, since the unit square has area 1, the determinant of a 2×2 matrix A is the area of the parallelogram determined by the columns of A. Similarly, the determinant of a 3×3 matrix A is the volume of the parallelepiped determined by columns of A. See Section 5.6 for more results on geometric approach to determinant. Moreover, for more results see [10]. See also Section 5.7. In what follows, we give an explicit formula to calculate determinant of a matrix on an arbitrary field.

1.11.2 Explicit formula of the determinant and its properties

We define the determinant over an arbitrary field in the standard way.

Definition 1.11.2 *For $A \in \mathbb{F}^{n \times n}$, we define the determinant of A by the formula:*

$$\det A = \sum_{\omega \in \mathcal{S}_n} \operatorname{sign}(\omega) a_{1\omega(1)} a_{2\omega(2)} \dots a_{n\omega(n)}, \quad A = [a_{ij}] \in \mathbb{F}^{n \times n}. \tag{1.11.1}$$

In what follows, we verify that the formulas given in Definition 1.11.1 and Definition 1.11.2 are the same for 2×2 real matrices. This can be generalized for real matrices of any size. We assume that $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. We are going to calculate the area of the parallelogram determined by $\mathbf{a}$ and $\mathbf{b}$. We have

$$\begin{aligned} \mathbf{a} &= A(1,0)^\top = (a_{11}, a_{21})^\top \\ \mathbf{b} &= A(0,1)^\top = (a_{12}, a_{22})^\top \end{aligned}$$

The equation of the line passing through the origin and $\mathbf{a}$ is $f(x,y) = \frac{a_{21}}{a_{11}} x - y$. Next, we have

$$\begin{aligned} d(\mathbf{b}, f) &= \frac{\left| \frac{a_{21} a_{12}}{a_{11}} - a_{22} \right|}{\sqrt{\left(\frac{a_{21}}{a_{11}} \right)^2 + 1}} \\ d(\mathbf{a}, 0) &= \sqrt{a_{11}^2 + a_{21}^2} \end{aligned}$$

Here, d denotes the distance of two points or a point and a line. Then, the area of image region in Figure 1.1 is equal to

$$\sqrt{a_{11}^2 + a_{21}^2} \cdot \frac{\left| \frac{a_{21} a_{12}}{a_{11}} - a_{22} \right|}{\sqrt{\left(\frac{a_{21}}{a_{11}} \right)^2 + 1}} = |a_{11} a_{22} - a_{21} a_{12}|.$$

As the area of the original region (unit square) in Figure 1.1 is 1, then the magnification factor ($|\det A|$) is $|a_{11} a_{22} - a_{21} a_{21}|$. Clearly, we obtain the same value for $|\det A|$ by formula 1.11.1.
Note that here we used the distance formula between a line $L = ax + by + c$ and a point $p = (x_0, y_0)$ which is denoted by $d(p, L)$ and given as $d(p, L) = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}$.
In the following theorem, we study the main properties of the determinant function.

Theorem 1.11.3 *Assume that the characteristic of $\mathbb{F}$ is different from 2, i.e. $2 \neq 0$ in $\mathbb{F}$. Then, the determinant function $\det : \mathbb{F}^{n \times n} \to \mathbb{F}$ is the unique skew-symmetric multilinear functions in the rows of A satisfying the normalization condition:*

$$\det I_n = 1.$$

Furthermore, it satisfies the following properties:

1. $\det A = \det A^\top$.
2. $\det A$ *is a multilinear function in rows or columns of A.*
3. *The determinant of a lower triangular or an upper triangular matrix is a product of the diagonal entries of A.*

4. *Let B obtained from A by permuting two rows or columns of A. Then* $\det B = -\det A$.

5. *If A has two equal rows or columns, then* $\det A = 0$.

6. *A is invertible if and only if* $\det A \neq 0$.

7. *Let $A, B \in \mathbb{F}^{n\times n}$. Then,* $\det AB = \det A \det B$.

8. *(*Laplace row and column expansion for determinants*) For $i, j \in [n]$ denote by $A(i,j) \in \mathbb{F}^{(n-1)\times(n-1)}$ the matrix obtained from A by deleting its i-th row and j-th column. Then*

$$\det A = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det A(i,j) = \sum_{i=1}^{n} a_{ji}(-1)^{i+j} \det A(i,j), \tag{1.11.2}$$

for $i = 1, \dots, n$.

(If $A = [a_{ij}] \in \mathbb{F}^{n\times n}$, we sometimes denote $\det A$ *by* $\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$*.)*

Proof. First, observe that if the determinant function exists, then it must be defined by (1.11.1). Indeed,

$$\mathbf{r}_i = \sum_{j=1}^{n} a_{ij}\mathbf{e}_j^\top, \quad \mathbf{e}_j = (\delta_{j1}, \dots, \delta_{jn})^\top.$$

Let $T : \bigoplus_n \mathbb{F}^n \to \mathbb{F}$ be a multilinear skew-symmetric function. Use multilinearity of T to deduce

$$T(\mathbf{r}_1, \dots, \mathbf{r}_n) = \sum_{i_1,\dots,i_n \in [n]} a_{1i_1} \dots a_{ni_n} T(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_n}).$$

Since T is skew-symmetric, $T(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_n}) = 0$ if $i_p = i_q$, for some $1 \le p < q \le n$. Indeed, if we interchange $\mathbf{e}_{i_p} = \mathbf{e}_{i_q}$, then we do not change the value of the T. On the other hand, since T is skew-symmetric, then the value of T is equal to $-T$, when we interchange $\mathbf{e}_{i_p} = \mathbf{e}_{i_q}$. As the characteristic of $\mathbb{F}$ is not 2, we deduce that $T(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_n}) = 0$, if $i_p = i_q$. Hence, in the above expansion of $T(\mathbf{r}_1, \dots, \mathbf{r}_n)$, the only non-zero terms are where $\{i_1, \dots, i_n\} = [n]$. Each such set of n indices $\{i_1, \dots, i_n\}$ corresponds uniquely to a permutation $\omega \in \mathcal{S}_n$, where $\omega(j) = i_j$, for $j \in [n]$. Now $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ can be brought to $\{\mathbf{e}_{\sigma(1)}, \dots, \mathbf{e}_{\sigma(n)}\}$ by using transpositions. The composition of this transpositions yields σ. Since T is skew-symmetric, it follows that $T(\mathbf{e}_{\sigma(1)}, \dots, \mathbf{e}_{\sigma(n)}) = \mathrm{sign}(\omega) T(\mathbf{e}_1, \dots, \mathbf{e}_n)$. As $T(I_n) = 1$, we deduce that $T(\mathbf{e}_{\sigma(1)}, \dots, \mathbf{e}_{\sigma(n)}) = \mathrm{sign}(\omega)$.

1. In (1.11.1) note that if $j = \omega(i)$, then $i = \omega^{-1}(j)$. Since ω is a bijection,we deduce that when $i = 1, \dots, n$, then j takes each value in $[n]$. Since $\mathrm{sign}(\omega) = \mathrm{sign}(\omega^{-1})$, we obtain

$$\det A = \sum_{\omega \in \mathcal{S}_n} \mathrm{sign}(\omega) a_{\omega^{-1}(1)1} a_{\omega^{-1}(2)2} \dots a_{\omega^{-1}(n)n} =$$
$$\sum_{\omega \in \mathcal{S}_n} \mathrm{sign}(\omega^{-1}) a_{\omega^{-1}(1)1} a_{\omega^{-1}(2)2} \dots a_{\omega^{-1}(n)n} = \det A^\top.$$

(Note that since $\mathcal{S}_n$ is a group, when ω varies over $\mathcal{S}_n$, so is ω^{-1}.)
2. Fix all rows of A except the row i. From (1.11.1) it follows that $\det A$ is a linear function in the row i of A, i.e. $\det A$ is a multilinear function in the rows of A. In view of the identity $\det A = \det A^\top$, we deduce that $\det A$ is a multilinear function of the columns of A.
3. Assume that A is upper triangular. Then, $a_{n\omega(n)} = 0$ if $\omega(n) \neq n$. Therefore, all non-zero terms in (1.11.1) are zero unless $\omega(n) = n$. Thus, assume that $\omega(n) = n$. Then, $a_{(n-1)\omega(n-1)} = 0$ unless $\omega(n-1) = n-1$. Hence, all non-zero terms in (1.11.1) must come from all ω satisfying $\omega(n) = n, \omega(n-1) = n-1$. Continuing in the same manner we deduce that the only non-zero term in (1.11.1) comes from $\omega =$id. As sign(id) = 1, it follows that $\det A = \prod_{i=1}^n a_{ii}$. Since $\det A = \det A^\top$, we deduce the claim for lower triangular matrices.
4. Note that a permutation of two elements $1 \le i < j \le n$ in $[n]$ is achieved by a transposition $\tau \in \mathcal{S}_n$. Therefore, $B = [b_{pq}]$ and $b_{pq} = a_{\tau(p)q}$. As in the proof of 1 we let $\tau(i) = j$, then $j = \tau^{-1}(i) = \tau(i)$. Hence

$$\det B = \sum_{\omega\in\mathcal{S}_n} \operatorname{sign}(\omega) a_{\tau(1)\omega(1)} a_{\tau(2)\omega(2)} \dots a_{\tau(n)\omega(n)} =$$
$$\sum_{\omega\in\mathcal{S}_n} \operatorname{sign}(\omega) a_{1\omega(\tau(1))} a_{2\omega(\tau(2))} \dots a_{n\omega(\tau(n))} =$$
$$\sum_{\omega\in\mathcal{S}_n} -\operatorname{sign}(\omega\circ\tau) a_{1(\omega\circ\tau)(1)} a_{2(\omega\circ\tau)(2)} \dots a_{n(\omega\circ\tau)(n)} = -\det A.$$

5. Suppose A has two identical rows. Interchange these two rows to obtain $B = A$. Then, $\det A = \det B = -\det A$, where the last equality is established in 4. Thus, $2\det A = 0$ and this means $\det A = 0$ as $\mathbf{char}\mathbb{F} \neq 2$.
6. Use 2, 3 and 5 to deduce that $\det EA = \det E \det A$ if E is an elementary matrix. (Note that $\det E \neq 0$.) Hence, if $E_1, \dots, E_k$ are elementary matrices, we deduce that $\det(E_k \dots E_1) = \prod_{i=1}^k \det E_i$. Let B be the reduced row echelon form of A. Therefore, $B = E_k E_{k-1} \dots E_1 A$. Hence, $\det B = \left(\prod_{i=1}^n \det E_i\right) \det A$. If $I_n \neq B$, then the last row of B is zero, so $\det B = 0$ which implies that $\det A = 0$. If $B = I_n$, then $\det A = \prod_{i=1}^n (\det E_i)^{-1}$.
7. Assume that either $\det A = 0$ or $\det B = 0$. We claim that $(AB)\mathbf{x} = \mathbf{0}$ has a nontrivial solution. Suppose $\det B = 0$. Using 6 and Theorem 1.7.11, we conclude that the equation $B\mathbf{x} = 0$ has a nontrivial solution which satisfies $AB\mathbf{x} = \mathbf{0}$. Suppose that B is invertible and $A\mathbf{y} = \mathbf{0}$, for some $\mathbf{y} \neq \mathbf{0}$. Then, $AB(B^{-1}\mathbf{y}) = \mathbf{0}$ which implies that $\det AB = 0$. Hence in these cases $\det AB = 0 = \det A \det B$. Suppose that A and B are invertible. Then, each of them is a product of elementary matrices. Use the arguments in the proof of 6 to deduce that $\det AB = \det A \det B$.
8. First we prove the first part of formula 1.11.2 for $i = n$. Clearly, (1.11.1) yields the equality

$$\det A = \sum_{j=1}^n a_{nj} \sum_{\omega\in\mathcal{S}_n, \omega(n)=j} \operatorname{sign}(\omega) a_{1\omega(1)} \dots a_{(n-1)\omega(n-1)}. \tag{1.11.3}$$

In the above sum, let $j = n$. Hence, $\omega(n) = n$ and then ω can be viewed as $\omega' \in \mathcal{S}_{n-1}$. Also $\operatorname{sign}(\omega) = \operatorname{sign}(\omega')$. Hence, $\sum_{\omega'\in\mathcal{S}_{n-1}} \operatorname{sign}(\omega) a_{1\omega'(1)} \dots a_{(n-1)\omega'(n-1)} = \det A(n,n)$. Note that $(-1)^{n+n} = 1$. This justifies the form of the last term of expansion 1.11.2 for $i = n$. To justify the sign of any term in 1.11.2 for $i = n$, we take the column j and interchange it first with column $j+1$, then with column $j+2$, and at last with the column n. The sign of $\det A(n,j)$ is $(-1)^{n+j}$. This proves the case $i=n$.

By interchanging any row $i<n$ with row $i+1$, row $i+2$, and finally with row n, we deduce the first part of formula 1.11.2, for any i. By considering $\det A^\top$, we deduce the second part of formula 1.11.2.

□

1.11.3 Matrix inverse

Observe that $\det I_n = 1$. Hence

$$1 = \det I_n = \det AA^{-1} = \det A \det A^{-1} \Rightarrow \det A^{-1} = \frac{1}{\det A}.$$

For $A=[a_{ij}]\in\mathbb{F}^{n\times n}$ denote by A_{ij} the determinant of the matrix obtained from A by deleting i-th row and j-th column and multiplied by $(-1)^{i+j}$. (This is called the (i,j) *cofactor* of A.)

$$A_{ij} := (-1)^{i+j}\det A(i,j). \tag{1.11.4}$$

Then, the expansion of $\det A$ by the row i and the column j, respectively is given by the equalities

$$\det A = \sum_{j=1}^{n} a_{ij}A_{ij} = \sum_{i=1}^{n} a_{ij}A_{ij}. \tag{1.11.5}$$

Then, the *adjoint matrix* of A is defined as follows:

$$\text{adj } A := \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}. \tag{1.11.6}$$

Proposition 1.11.4 *Let $A\in\mathbb{F}^{n\times n}$. Then*

$$A(\text{adj } A) = (\text{adj } A)A = (\det A)I_n. \tag{1.11.7}$$

Hence, A is invertible if and only if $\det A\neq 0$. Furthermore, $A^{-1}=(\det A)^{-1}\text{adj } A$.

Proof. Consider an (i,k) entry of $A(\text{adj } A)$. It is given as $\sum_{j=1}^n a_{ij}A_{kj}$. For $i=k$, (1.11.5) yields that $\sum_{j=1}^n a_{ij}A_{ij}=\det A$. Suppose that $i\neq k$. Let B_k be the matrix obtained from A by replacing the row k of A by the row i. Then, B_k has two identical rows, hence $\det B_k=0$. On the other hand, expand B_k by the row k to obtain that $0=\det B_k=\sum_{j=1}^n a_{ij}A_{kj}$. This shows that $A(\text{adj } A)=(\det A)I_n$. Similarly, one shows that $(\text{adj } A)A=(\det A)I_n$.

Using Theorem 1.11.3 part 6, we see that A is invertible if and only if $\det A\neq 0$. Hence, for invertible A, $A^{-1}=\frac{1}{\det A}\text{adj } A$. □

Proposition 1.11.5 *(Cramer's rule) Let $A\in\mathrm{GL}(n,\mathbb{F})$ and consider the system $A\mathbf{x}=\mathbf{b}$, where $\mathbf{x}=(x_1,\dots,x_n)^\top$. Denote by B_k the matrix obtained from A by replacing the column k in A by $\mathbf{b}$. Then, $x_k=\frac{\det B_k}{\det A}$, for $k=1,\dots,n$.*

Proof. Clearly, $\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{\det A}(\mathrm{adj}\ A)\mathbf{b}$. Hence, $x_k = (\det A)^{-1}\sum_{j=1}^{n} A_{jk}b_j$, where $\mathbf{b} = (b_1, \dots, b_k)^\top$. Expand B_k by the column k to deduce that $\det B_k = \sum_{j=1}^{n} b_j A_{jk}$. □

Example 1.11.6 *We solve the following system of equations by Cramer's rule:*

$$\begin{cases} 2x_1 + x_2 + x_3 = 3 \\ x_1 - x_2 - x_3 = 0 \\ x_1 + 2x_2 + x_3 = 0 \end{cases}$$

We rewrite this system in the coefficient matrix form;

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

Assume that D denotes the determinant of the coefficient matrix:

$$D = \begin{vmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 1 \end{vmatrix}$$

Denote by D_1 the determinant of the matrix obtained from the coefficient matrix by replacing the first column with the right-hand side column of the given system:

$$D_1 = \begin{vmatrix} 3 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 2 & 1 \end{vmatrix}$$

Similarly, D_2 and D_3 would then be:

$$D_2 = \begin{vmatrix} 2 & 3 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{vmatrix} \quad \textit{and} \quad D_3 = \begin{vmatrix} 2 & 1 & 3 \\ 1 & -1 & 0 \\ 1 & 2 & 0 \end{vmatrix}.$$

Evaluating each determinant, we get:

$$D = 3,\ D_1 = 3,\ D_2 = -6 \textit{ and } D_3 = 9.$$

Cramer's Rule says that $x_1 = \frac{D_1}{D} = \frac{3}{3} = 1$, $x_2 = \frac{D_2}{D} = -\frac{6}{3} = -2$ and $x_3 = \frac{D_3}{D} = \frac{9}{3} = 3$. Note that the point of Cramer's Rule is that you don't have to solve the whole system to get the one value we need.

For large matrices, Cramer's rule does not provide a practical method for computing the inverse, because it involves n^2 determinants, and the computation via Gauss-Jordan algorithm given in 1.7.5 is significantly faster. Nevertheless, Cramer's rule has important theoretical importance.

1.11.4 Worked-out Problems

1. If $A \in \mathbb{F}^{n\times n}$, show that the rank of adj A is n, 1 or 0.
   Solution:
   Since A is an $n \times n$ matrix, then

   $$\begin{aligned} &A(\text{adj } A) = (\det A)I_n \Rightarrow \\ &(\det A)(\det(\text{adj } A)) = (\det A)(\det I_n) = \det A. \end{aligned}$$

   Assume that A is a full-rank matrix, then $\det A \neq 0$ and so: $(\det A)(\det(\text{adj } A)) = \det A$ gives $\det(\text{adj } A) = 1$. Therefore, the matrix adj A is also invertible and it is of rank n. If the rank of A is $n-1$, then at least one minor of order $n-1$ of A is non-zero and this means adj A is non-zero and so the rank of adj A is greater than zero. Since rank $A = n-1$, then $\det A = 0$ and so $A(\text{adj } A) = 0$. This tells us rank $A(\text{adj } A) = 0$. Now, by Problem 1.10.2-2, we conclude that: $0 \geq \text{rank } A + \text{rank adj } A - n$ or $n - 1 + \text{rank adj } A \leq n$ or rank adj $A \leq 1$.
   But we showed that rank adj $A > 0$. Then, rank adj $A = 1$. Finally, if rank $A \leq n-1$, then all minors of orders $n-1$ of A are zero. Thus, adj $A = 0$ and then rank adj $A = 0$.

2. Assume that the matrix $A \in \mathbb{F}^{(n+1)\times(n+1)}$ is given as follows:

   $$A = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}$$

   Show that

   $$\det A = \prod_{i>j}(x_i - x_j). \tag{1.11.8}$$

   The determinant of A is called the *Vandermonde determinant of order* $n+1$.
   Solution:
   We prove by induction. First, we can verify that the result holds in the 2×2 case. Indeed

   $$\begin{vmatrix} 1 & x_0 \\ 1 & x_1 \end{vmatrix} = x_1 - x_0.$$

   We now assume the result for $n-1$ and consider n. We note that the index n corresponds to a matrix of order $(n+1)\times(n+1)$, hence our induction hypothesis is that the claim (1.12.1) holds for any Vandermonde determinant of order $n \times n$. We subtract the first row from all other rows, and expand the determinant along the first column

   $$\begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix} = \begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 0 & x_1 - x_0 & \cdots & x_1^n - x_0^n \\ \vdots & \vdots & & \vdots \\ 0 & x_n - x_0 & \cdots & x_n^n - x_0^n \end{vmatrix} = \begin{vmatrix} x_1 - x_0 & \cdots & x_1^n - x_0^n \\ \vdots & & \vdots \\ x_n - x_0 & \cdots & x_n^n - x_0^n \end{vmatrix}.$$

For every row k in the last determinant, we factor out a term $x_k - x_0$ to get

$$\begin{vmatrix} x_1 - x_0 & \cdots & x_1^n - x_0^n \\ \vdots & & \vdots \\ x_n - x_0 & \cdots & x_n^n - x_0^n \end{vmatrix} = \prod_{k=1}^{n} (x_k - x_0) \begin{vmatrix} 1 & x_1 + x_0 & \cdots & \sum\limits_{i=0}^{n} x_1^{n-1-i} x_0^i \\ 1 & x_2 + x_0 & \cdots & \sum\limits_{i=0}^{n} x_2^{n-1-i} x_0^i \\ \vdots & \vdots & & \vdots \\ 1 & x_1 + x_0 & \cdots & \sum\limits_{i=0}^{n} x_n^{n-1-i} x_0^i \end{vmatrix}.$$

Here we use the expansion

$$x_k^n - x_0^n = (x_k - x_0)(x_k^{n-1} + x_k^{n-2} x_0 + x_k^{n-3} x_0^2 + \cdots + x_0^{n-1}).$$

For example, the first row of the last determinant is

$$1 \quad x_1 + x_0 \quad x_1^2 + x_1 x_0 + x_0^2 \quad x_1^3 + x_1^2 x_0 + x_1 x_0^2 + x_0^3 \quad \cdots.$$

Now for every column l, starting from the second one, subtracting the sum of x_0^i times column i, we end up with

$$\prod_{k=1}^{n} (x_k - x_0) \begin{vmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{vmatrix}. \tag{1.11.9}$$

Next step, using the first row as the example, means

$$(x_1^2 + x_1^2 x_0 + x_1 x_0^2 + x_0^3) - (x_1^2 + x_1 x_0 + x_0^2) x_0 = x_1^3, \ (x_1^2 + x_1 x_0 + x_0^2) - (x_1 + x_0) x_0 = x_1^2,$$

$$(x_1 + x_0) - 1 \times x_0 = x_1.$$

Now we have on the right-hand side of (1.11.9) a Vandermonde determinant of dimension $n \times n$, we can use the induction to conclude with the desired result. Note that the matrix given in this problem is well-known as a *Vandermonde matrix*, named after Alexander-Theophile Vandermonde.

3. Prove that Theorem 1.10.2-3 is the case for any field without any condition on characteristic.
Solution:
Assume that **char**$\mathbb{F} = 2$, i.e. $2 = 0$. In this case, $\det A$ equals to the *permanent* of A;

$$\operatorname{perm} A = \sum_{\omega \in \mathcal{S}_n} a_{1\omega(1)} a_{2\omega(2)} \cdots a_{n\omega(n)}.$$

(See the next section for more details on the permanent of a matrix.)
Since A has two identical rows, each term appears twice. For example, if row one is equal to row two, then $a_{1i} a_{2j} = a_{1j} a_{2i}$. Thus, we have only $\frac{n!}{2}$ terms and each term is multiplied by $2 = 0$. Hence, $\det A = 0$. Use $\det A = \det A^\top$ to deduce that $\det A = 0$ if A has two identical columns.

1.11.5 Problems

1. Let $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ be a symmetric tridiagonal matrix, . If B is a matrix formed from A by deleting the first two rows and columns, show that
$$\det A = a_{11}\det A(1,1) - a_{12}^2 \det B$$

2. If $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times p}$, show that
$$\text{rank } A + \text{rank } B - n \le \text{rank } AB \tag{1.11.10}$$
(Hint: Use Problem 1.10.2-1.b and Worked-out Problem 1.10.1-2.)
The combination of the inequality (1.11.10) with the inequality mentioned in Problem 1.8.5-3, is well-known as Sylvester's Inequality.

3. Let $A, B \in \mathbb{F}^{n\times n}$. Prove that
$$\text{rank } (A + B) \le \text{rank } A + \text{rank } B.$$

4. If $A, B \in \mathbb{F}^{n\times n}$, prove or disprove the following statement:
$$\text{"rank } (A + B) \le \min\{\text{rank } A, \text{rank } B\}\text{."}$$
Compare this statement with Problem 1.8.5-3.

5. If $A, B \in \mathbb{F}^{n\times n}$ and $A - B = AB$, prove that $AB = BA$.

6. Let $A \in \mathbb{F}^{m\times n}$ and denote by A_k the upper left $k\times k$ submatrix of A. Assume that $\det A_i \ne 0$ for $i = 1,\dots,k$. Show that the i-th pivot of A is $\frac{\det A_i}{\det A_{i-1}}$ for $i = 1,\dots,k$. (Assume that $\det A_0 = 1$.)

7. Let $A \in \mathbb{R}^{n\times n}$. Prove that rank $AA^\top$ = rank $A^\top A$.

8. Let $A \in \mathbb{R}^{2\times 2}$ be an orthogonal matrix. Show the followings:
   (a) if $\det A = 1$, then $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, for some θ, $0 \le \theta < 2\pi$. That is, A counterclockwise rotates every point in $\mathbb{R}^2$ by an angle θ.
   (b) if $\det A = -1$, then $A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$, for some θ, $0 \le \theta < 2\pi$. That is, A reflects every point in $\mathbb{R}^2$ about a line passing through the origin. Determine this line. Or equivalently, there exists an invertible matrix P such that $P^{-1}AP = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

   (See problem 1.10.2-4.)

1.12 Permanents

1.12.1 A combinatorial approach of Philip Hall Theorem

The following theorem is well-known as Philip Hall Theorem and gives us the necessary and sufficient conditions to have a perfect matching in a bipartite graph. See [23] for more results on this theorem and its proof.

Theorem 1.12.1 *If $G = (V, E)$ is a bipartite graph with the bipartite sets X and Y, then G has a perfect matching if and only if $\#X = \#Y$ and for any subset $\mathcal{S} \subset X$, $\#\mathcal{S} \le \#N(\mathcal{S})$, where $N(\mathcal{S})$ denotes the neighborhood of $\mathcal{S}$ in Y, i.e. the set of all vertices in Y adjacent to some elements of $\mathcal{S}$.*

If $\mathcal{A} = \{A_i\}_{i=1}^n$ is a family of non-empty subsets of a finite set $\mathcal{S}$, a *system of distinct representative* (SDR) of $\mathcal{A}$ is a tuple $(a_1, \dots, a_n)$ with $a_i \in A_i$ and $a_i \ne a_j$, $1 \le i, j \le n$, $i \ne j$. This is also called a *transversal* of $\mathcal{A}$.

Theorem 1.12.2 *Let $\mathcal{A} = \{A_i\}_{i=1}^n$ is a family of non-empty subsets of a finite set A. Then, $\mathcal{A}$ has an SDR if and only if*

$$\#\bigcup_{i \in J} A_i \ge \#J,$$

for any $J \subseteq [n]$.

It is easy to check that the above theorem is equivalent to Philip Hall Theorem. Indeed, construct a bipartite graph with $X = \{a_1, \dots, a_n\}$ and $Y = \{A_1, \dots, A_n\}$. Furthermore, there is an edge (a_i, A_j) if and only if $a_i \in A_j$.

The interested reader can verify that the number of perfect matchings in a bipartite graph G with bipartite sets X and Y and $\#X = \#Y = n$ is equal to perm A, where A is the adjacency matrix of G.

We end up this section with the linear version of Philip Hall Theorem.
Let $\mathbf{V}$ be a vector space over the field $\mathbb{F}$ and let $\mathcal{V} = \{\mathbf{V}_1, \dots, \mathbf{V}_n\}$ be a family of vector subspaces of $\mathbf{V}$. A free transversal for $\mathcal{V}$ is a family of linearly independent vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ such that $\mathbf{x}_i \in \mathbf{V}_i$, $i = 1, \dots, n$. The following result of Rado [16] gives a linear version of Philip Hall Theorem for the existence of a free transversal for $\mathcal{V}$. The interested reader is referred to [1] to see the proof and more details.

Theorem 1.12.3 *Let $\mathbf{V}$ be a vector space over the field $\mathbb{F}$ and let $\mathcal{V} = \{\mathbf{V}_1, \dots, \mathbf{V}_n\}$ be a family of vector subspaces of $\mathbf{V}$. Then, $\mathcal{V}$ admits a free transversal if and only if*

$$\dim \operatorname{span}\left(\bigcup_{i \in J} \mathbf{V}_i\right) \ge \#J,$$

for all $J \subset [n]$.

1.12.2 Permanents

Recall that if $A \in \mathbb{F}^{n \times n}$, the *permanent* of $A = [a_{ij}]$ is denoted by perm A and defined as

$$\operatorname{perm} A = \sum_{w \in \mathcal{S}_n} \prod_{i=1}^n a_{iw(i)}.$$

From the definition of permanent, it follows that perm$A$ = perm$A^{\top}$.

The matrix $A \in \mathbb{F}^{m\times n}$ whose entries are either 1 or 0 is called a $(0,1)$-*matrix.* A special class of $(0,1)$-matrices, namely the $(0,1)$-matrices in $\mathbb{F}^{n\times n}$ that have exactly k ones in each row and column. We denote this class by $\mathcal{C}(n,k)$. We define

$$\begin{aligned} S(n,k) &:= \max\{\mathrm{perm}A;\ A \in \mathcal{C}(n,k)\},\\ s(n,k) &:= \min\{\mathrm{perm}A;\ A \in \mathcal{C}(n,k)\}. \end{aligned}$$

The following result, known as Fekete's lemma. It will be used to find an upper bound for the permanent function. See [23] for more details about Fekete's lemma.

Lemma 1.12.4 *Let $f : \mathbb{N} \to N$ be a function for which $f(m+n) \geq f(m)f(n)$, for all $m,n \in \mathbb{N}$. Then $\lim_{n\to\infty} f(n)^{\frac{1}{n}}$ exists (possibly ∞).*

It is easy to check the following inequalities:

$$\begin{aligned} &S(n_1+n_2,k) \geq S(n_1,k)S(n_2,k),\\ &s(n_1+n_2,k) \leq s(n_1,k)s(n_2,k). \end{aligned}$$

By applying these inequalities to Fekete's lemma, we can define:

$$\begin{aligned} S(k) &:= \lim_{n\to\infty}\{S(n,k)\}^{\frac{1}{n}},\\ s(k) &:= \lim_{n\to\infty}\{s(n,k)\}^{\frac{1}{n}}. \end{aligned}$$

Note that a function $f : \mathbb{R}^n \to \mathbb{R}$ is called to be *convex* if $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$, for any $x,y \in \mathbb{R}^n$ and $t \in [0,1]$. Note that it can be shown that the necessary and sufficient condition for a continuous function f on $[0,1]$, such that f' and f'' exist on$(0,1)$, is that the second derivative is always non-negative. (Try to justify!)

Lemma 1.12.5 *If $t_1,\ldots,t_n$ are non-negative real numbers, then*

$$\left(\frac{\sum_{i=1}^n t_i}{n}\right)^{\sum_{i=1}^n t_i} \leq \prod_{i=1}^n t_i^{t_i}.$$

Proof. Since $x\ln x$ is a convex function, we have

$$\frac{\sum_{i=1}^n t_i}{n}\ln\left(\frac{\sum_{i=1}^n t_i}{n}\right) \leq \frac{\sum_{i=1}^n t_i\ln t_i}{n},$$

which proves the assertion. □

It was conjectured by H. Minc that if $A \in \mathbb{R}^{n\times n}$ is a (0,1)-matrix with row-sums $r_1,\ldots,r_n$, then

$$\mathrm{perm}\, A \leq \prod_{j=1}^n (r_j!)^{\frac{1}{r_j}}.$$

This conjecture was proved by L.M. Bregman [4].
The following short proof of this theorem was given by A. Schrijver [20].

Proof by induction on n: For $n = 1$ the statement is obvious. Suppose that it satisfies for $(n-1)\times(n-1)$ matrices. We will show that

$$(\mathrm{perm}A)^{n\mathrm{perm}A} \leqslant \left(\prod_{i=1}^{n} r_i!^{\frac{1}{r_i}}\right)^{n\mathrm{perm}A},$$

which implies the above inequality. In the following inequalities, the variables i, j and k range from 1 to n. Let $\mathcal{S}$ be the set of all permutations $\omega \in \mathcal{S}_n$ for which $a_{i\omega(i)} = 1$, for $i = 1,\dots,n$. Clearly, $\#\mathcal{S} = \mathrm{perm}\ A$. Using Lemma 1.12.5,

$$\prod_i (\mathrm{perm}\ A)^{\mathrm{perm}\ A} \leqslant \prod_i \left(r_i^{\mathrm{perm}\ A} \prod_{\substack{k\\ a_{ik}=1}} \mathrm{perm}\ A_{ik}^{\mathrm{perm}\ A_{ik}} \right),$$

where A_{ik} denotes the minor obtained from A by deleting row i and column k. Now,

$$\prod_i \left(r_i^{\mathrm{perm}A} \prod_{\substack{k\\ a_{ik}=1}} \mathrm{perm}A_{ik}^{\mathrm{perm}A_{ik}} \right) = \prod_{\omega\in\mathcal{S}} \left(\left(\prod_i r_i\right)\left(\prod_i \mathrm{perm}A_{i\omega(i)}\right)\right),$$

because the number of factors r_i equals perm$A$ on both sides, while the number of factors permA_{ik} equals the number of $\omega \in \mathcal{S}$ for which $\omega(i) = k$. Applying the induction hypothesis to each $A_{i\omega(i)}$, we conclude that the recent expression is less than or equal to

$$\prod_{\omega\in\mathcal{S}} \left(\left(\prod_i r_i\right) \left(\prod_i \left(\prod_{\substack{j\\ j\neq i\\ a_{i\omega(i)=0}}} r_j!^{\frac{1}{r_j}} \right) \left(\prod_{\substack{j\\ j\neq i\\ a_{i\omega(i)=1}}} (r_j-1)!^{\frac{1}{(r_j-1)}} \right) \right) \right). \tag{1.12.1}$$

Changing the order of multiplication and considering that the number of i such that $i \neq j$ and $a_{j\omega(j)} = 0$ is $n - r_i$, whereas the number of i such that $i \neq j$ and $a_{j\omega(i)} = 1$ is $r_j - 1$ we get

$$(1.12.1) = \prod_{\omega\in\mathcal{S}} \left(\left(\prod_i r_i\right) \left(\prod_j r_j!^{\frac{(n-r_j)}{r_j}} (r_j-1)!^{\frac{(r_j-1)}{(r_j-1)}} \right) \right). \tag{1.12.2}$$

That (1.12.2) equals $\prod_{\omega\in\mathcal{S}} \left(\prod_i r_i!^{\frac{n}{r_i}} \right) = \left(\prod_i r_i!^{\frac{1}{r_i}} \right)^{n\mathrm{perm}\ A}$ is trivial. Since $(\mathrm{perm}\ A)^{n\mathrm{perm}A} = \prod_i (\mathrm{perm}A)^{\mathrm{perm}A}$, the proof is complete.

Now, if G is a balanced bipartite graph with adjacency matrix $A \in \{0,1\}^{n\times n}$ and row-sums $r_1,\dots,r_n$, then by Bregman 's theorem, $\prod_{j=1}^{n} (r_j!)^{\frac{1}{r_j}}$ is an upper bound for the number of perfect matchings in G.

1.12.3 Worked-out Problems

1. Assume that $T_n \subseteq (\mathbb{N}\cup\{0\})^{n\times n}$ is the set of matrices with row-sums and column-sums 3; $t_n = \min\{\mathrm{perm}\ A;\ A \in T_n\}$. Denote by X_n the set of all

matrices obtained from elements of T_n by decreasing one positive entry by 1; $x_n = \min\{\text{perm } A;\ A \in X_n\}$. Show that $t_n \geqslant \lceil \frac{3}{2} x_n \rceil$, where $\lceil\ \rceil$ denotes the ceiling function.
Solution:
Choose $A \in T_n$ with first row $y = (y_1, y_2, y_3, 0, \ldots, 0)$, where $y_i \geqslant 0$, for $i = 1, 2, 3$. Then

$$\begin{aligned} 2y &= y_1(y_1 - 1, y_2, y_3, 0, \ldots, 0) + \\ &\quad y_2(y_1, y_2 - 1, y_3, 0, \ldots, 0) + \\ &\quad y_3(y_1, y_2, y_3 - 1, 0, \ldots, 0). \end{aligned}$$

Since $S(n_1 + n_2, k) \geq S(n_1, k)S(n_2, k)$, then $2t_n \geqslant (y_1 + y_2 + y_3)x_n = 3x_n$.

1.12.4 Problems

1. Show that

   (a) $S(n, k) \geq k!$

   (b) $S(k) \leq (k!)^{\frac{1}{k}}$

   (c) Show by example that $S(k) \geq (k!)^{\frac{1}{k}}$. This shows that $S(k) = (k!)^{\frac{1}{k}}$.

   (**Hint:** Use Bregman's theorem.)

1.13 An application of Philip Hall Theorem

Recall that a permutation matrix is a square matrix that has exactly one entry of 1 in each row and each column and zero elsewhere. Now, we define a more general family of matrices called doubly stochastic as mentioned in Section 1.9.

Definition 1.13.1 *A doubly stochastic matrix is a square matrix $A = [a_{ij}]$ of non-negative real entries, each of whose rows and columns sum 1, i.e.*

$$\sum_i a_{ij} = \sum_j a_{ij} = 1.$$

The set of all $n \times n$ doubly stochastic matrices is denoted by Ω_n. If we denote all $n \times n$ permutation matrices by $\mathcal{P}_n$, then clearly $\mathcal{P}_n \subset \Omega_n$.

Definition 1.13.2 *A subset A of a real vector space $\mathbf{V}$ is said to be convex if $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in A$, for all vectors $\mathbf{x}, \mathbf{y} \in A$ and all scalars $\lambda \in [0, 1]$. Via induction, this can be seen to be equivalent to the requirement that $\sum_{i=1}^n \lambda_i \mathbf{x}_i \in A$, for all vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in A$ and all scalars $\lambda_1, \ldots, \lambda_n \geqslant 0$ with $\sum_{i=1}^n \lambda_i = 1$. A point $\mathbf{x} \in A$ is called an extreme point of A if $\mathbf{y}, \mathbf{z} \in A$, $0 < t < 1$, and $\mathbf{x} = t\mathbf{y} + (1 - t)\mathbf{z}$ imply $\mathbf{x} = \mathbf{y} = \mathbf{z}$. With this restrictions on λ_i's, an expression of the form $\sum_{i=1}^n \lambda_i \mathbf{x}_i$ is said to be a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The convex hull of a set $B \subset \mathbf{V}$ is defined as $\{\sum \lambda_i \mathbf{x}_i : \mathbf{x}_i \in B,\ \lambda_i \geq 0$ and $\sum \lambda_i = 1\}$. The convex hull of B can also be defined as the smallest convex set containing B. (Why?)*

The importance of extreme points can be seen from the following theorem whose proof can be found in [2].

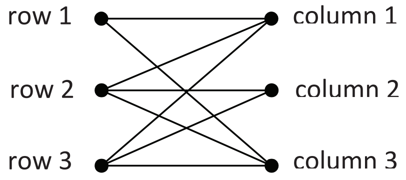
Theorem 1.13.3 (Krein-Milman) *Let $A \subset \mathbb{R}^n$ be a nonempty compact convex set. Then*

1. *The set of all extreme points of A is non-empty.*
2. *The convex hull of the set of all extreme points of A is A itself.*

The following theorem is a direct application of matching theory to express the relation between two sets of matrices $\mathcal{P}_n$ and Ω_n.

Theorem 1.13.4 (Birkhoff) *Every doubly stochastic matrix can be written as a convex combination of permutation matrices.*

Proof. We use Philip Hall Theorem to prove this theorem. We associate to our doubly stochastic matrix $A = [a_{ij}]$ a bipartite graph as follows. We represent each row and each column with a vertex and we connect the vertex representing row i with the vertex representing row j if the entry a_{ij} is non-zero.
For example if $A = \begin{bmatrix} \frac{7}{12} & 0 & \frac{5}{12} \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$, the graph associated to A is given in the picture below.



We claim that the associated graph of any doubly stochastic matrix has a perfect matching. Assume to the contrary, A has no perfect matching. Then, by Philip Hall Theorem, there is a subset E of the vertices in one part such that the set $R(E)$ of all vertices connected to some vertex in E has strictly less than $\#E$ elements. Without loss of generality, we may assume that A is a set of vertices representing rows, the set $R(A)$ consists then of vertices representing columns. Consider now the sum $\sum_{i \in E, j \in R(E)} a_{ij} = \#E$, the sum of all entries located in columns belonging to $R(E)$. (by the definition of the associated graph). Thus

$$\sum_{i \in E, j \in R(E)} a_{ij} = \#E.$$

Since the graph is doubly stochastic and the sum of elements located in any of given $\#E$ rows is $\#E$. On the other hand, the sum of all elements located in all columns belonging to $R(E)$ is at least $\sum_{i \in E, j \in R(E)} a_{ij}$, since the entries not belonging to a row in E are non-negative. Since the matrix is doubly stochastic, the sum of all elements located in all columns belonging to $R(E)$ is also exactly $\#R(E)$. Thus, we obtain

$$\sum_{i \in E, j \in R(E)} a_{ij} \leq \#R(E) < \#E = \sum_{i \in E, j \in R(E)} a_{ij},$$

a contradiction. Then, A has a perfect matching.
Now, we are ready to prove the theorem. We proceed by induction on the number

of non-zero entries in the matrix. As we proved, associated graph of A has a perfect matching. Underline the entries associated to the edges in the matching. For example in the associated graph above, $\{(1,3),(2,1),(3,2)\}$ is a perfect matching so we underline a_{13}, a_{23} and a_{32}. Thus, we underline exactly one element in each row and each column. Let α_0 be the minimum of the underlined entries. Let P_0 be the permutation matrix that has a 1 exactly at the position of the underlined elements. If $\alpha_0 = 1$, then all underlined entries are 1, and $A = P_0$ is a permutation matrix. If $\alpha_0 < 1$, then the matrix $A - \alpha_0 P_0$ has non-negative entries, and the sum of the entries in any row or any column is $1 - \alpha_0$. Dividing each entry by $(1 - \alpha_0)$ in $A - \alpha_0 P_0$ gives a doubly stochastic matrix A_1. Thus, we may write $A = \alpha_0 P_0 + (1 - \alpha_0) A_1$, where A_1 is not only doubly stochastic but has less non-zero entries than A. By our induction hypothesis, A_1 may be written as $A_1 = \alpha_1 P_1 + \cdots + \alpha_n P_n$, where $P_1, \ldots, P_n$ are permutation matrices, and $\alpha_1 P_1 + \cdots + \alpha_n P_n$ is a convex combination. But then we have

$$A = \alpha_0 P_0 + (1 - \alpha_0)\alpha_1 P_1 + \cdots + (1 - \alpha_0)\alpha_n P_n,$$

where $P_0, P_1, \ldots, P_n$ are permutation matrices and we have a convex combination. Since $\alpha_0 \geq 0$, each $(1 - \alpha_0)\alpha_i$ is non-negative and we have

$$\alpha_0 + (1 - \alpha_0)\alpha_1 + \cdots + (1 - \alpha_0)\alpha_n = \alpha_0 + (1 - \alpha_0)(\alpha_1 + \ldots + \alpha_n) = \alpha_0 + (1 - \alpha_0) = 1.$$

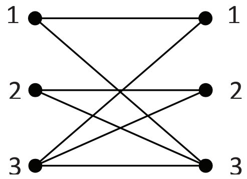
In our example

$$P_0 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

and $\alpha_0 = \frac{1}{6}$. Thus, we get

$$A_1 = \frac{1}{1 - \frac{1}{6}}\left(A - \frac{1}{6}P_0\right) = \frac{6}{5}\begin{bmatrix} \frac{7}{12} & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{7}{10} & 0 & \frac{3}{10} \\ 0 & \frac{3}{5} & \frac{2}{5} \\ \frac{3}{10} & \frac{2}{5} & \frac{3}{10} \end{bmatrix}.$$

The graph associated to A_1 is the following:



A perfect matching is $\{(1,1),(2,2),(3,3)\}$, the associated permutation matrix is

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Remark 1.13.5 *Let $\mathbf{e} \in \mathbb{R}^n$ be the column vector with each coordinate equal to 1. Then, for a matrix $A \in \mathbb{R}^{n \times n}$, the condition that each sum of entries in every row and column can be described by $\mathbf{e}^\top A = \mathbf{e}^\top$ and $A\mathbf{e} = \mathbf{e}$. It follows that the product of finitely many doubly stochastic matrices is a doubly stochastic matrix.*

1.13.1 Worked-out Problems

1. If we denote by $\Omega_{n,s}$, the subset of symmetric doubly stochastic matrices, show that each $A \in \Omega_{n,s}$ can be written as a convex combination of $\frac{1}{2}(P+P^\top)$, where $P \in \mathcal{P}_n$.
   Solution:
   As $A \in \Omega_{n,s} \subset \Omega_n$, by Theorem 1.13.4, one can find $P_1, \dots, P_N \in \mathcal{P}_n$ and $t_1, \dots, t_n \in (0,1)$ such that $A = \sum_{j=1}^N t_j P_j$ and $\sum_{j=1}^N t_j = 1$. Since A is symmetric, then $\sum_{j=1}^n t_j P_j = \left(\sum_{j=1}^n t_j P_j\right)^\top$. This implies $\sum_{j=1}^N t_j(P_j - P_j^\top) = 0$. Since t_j's are positive and P_j's are non-negative matrices, we conclude that $(P_j - P_j^\top)$'s are zero matrices and so $P_j = P_j^\top$, $1 \le j \le N$. Therefore, $P_j = \frac{1}{2}(P_j + P_j^\top)$ and then, $A = \sum_{j=1}^N t_j \frac{1}{2}(P_j + P_j^\top)$.

1.13.2 Problems

1. Show that in the decomposition of Worked-out Problem 1, $N \leqslant \frac{n^2-n+2}{2}$, for $n > 2$.
   (**Hint:** Use Carathéodary's theorem [22].)

2. Let A be a doubly stochastic matrix. Show that perm $A > 0$.

3. Show that Ω_n is a compact set.

4. Show that $\mathcal{P}_n$ is a group with respect to the multiplication of matrices, with I_n the identity and $P^{-1} = P^\top$.

5. Let $A \in \Omega_n$ and $B \in \Omega_m$. Show that $A \oplus B \in \Omega_{n+m}$.

6. Assume that A is an invertible $n \times n$ doubly stochastic matrix and that A^{-1} is doubly stochastic. Prove A is a permutation matrix.

7. Let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. The $\mathbf{x}$ is said to be a *stochastic vector* if its entries are non-negative that add up to one. Denote by Π_n the set of all probability vectors in $\mathbb{R}^n$. Prove that Π_n is a compact set.

1.14 Polynomial rings

1.14.1 Polynomials

Let $\mathbb{F}$ be a field, (usually $\mathbb{F} = \mathbb{R}, \mathbb{C}$). By the *ring polynomials* in the indeterminate, z, written as $\mathbb{F}[z]$, we mean the set of all polynomials $p(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n$, where n can be any non-negative integer and coefficients $a_0, \dots, a_n$ are all in $\mathbb{F}$. The *degree* of $p(z)$, denoted by $\deg p$, is the maximal degree $n-j$ of a monomial $a_j x^{n-j}$ which is not identically zero, i.e. $a_j \ne 0$. Then, $\deg p = n$ if and only if $a_0 \ne 0$, the degree of a non-zero constant polynomial $p(z) = a_0$ is zero, and the degree of the zero polynomial is agreed to be equal to $-\infty$. For two polynomials $p, q \in \mathbb{F}[z]$ and two scalars $a, b \in \mathbb{F}$, $ap(z) + bq(z)$ is a well-defined polynomial. Hence, $\mathbb{F}[z]$ is a vector space over $\mathbb{F}$, whose dimension is infinite. The set of polynomials of degree at most n, is $n+1$ dimensional subspace of $\mathbb{F}[z]$. Given two polynomials $p = \sum_{i=0}^{n} a_i z^{n-i}, q = \sum_{j=0}^{m} b_j z^{m-j} \in \mathbb{F}[z]$, one can form the product

$$p(z)q(z) = \sum_{k=0}^{n+m} (\sum_{i=0}^{k} a_i b_{k-i}) z^{n+m-k}, \text{ where } a_i = b_j = 0, \text{ for } i > n \text{ and } j > m.$$

Note that $pq = qp$ and $\deg pq = \deg p + \deg q$. The addition and the product in $\mathbb{F}[z]$ satisfy all the nice distribution identities as the addition and multiplication in $\mathbb{F}$. This implies that $\mathbb{F}[z]$ is a commutative ring with the addition and product defined above. Here, the constant polynomial $p \equiv 1$ is the identity element, and the zero polynomial as the zero element. (That is the reason for the name *ring* of polynomials in one indeterminate (variable) over $\mathbb{F}$.)

Given two polynomials $p, q \in \mathbb{F}[z]$, one can divide p by $q \not\equiv 0$ with the residue r, i.e. $p = tq + r$, for some unique $t, r \in \mathbb{F}[z]$, where $\deg r < \deg q$. For $p, q \in \mathbb{F}[z]$, let (p, q) denote *greatest common divisor* of p, q. If p and q are identically zero, then (p, q) is the zero polynomial. Otherwise, (p, q) is a polynomial s of the highest degree that divides p and q. Note that s is determined up to a multiple of a non-zero scalar and it can be chosen as a unique *monic* polynomial:

$$s(z) = z^l + s_1 z^{l-1} + \ldots + s_l \in \mathbb{F}[z]. \tag{1.14.1}$$

For $p, q \not\equiv 0$, s can be found using the *Euclid* algorithm:

$$p_i(z) = t_i(z) p_{i+1}(z) + p_{i+2}(z), \ \deg p_{i+2} < \deg p_{i+1} \quad i = 1, \ldots \tag{1.14.2}$$

Start this algorithm with $p_1 = p, p_2 = q$. Continue it until $p_k = 0$ the first time. (Note that $k \ge 3$). Then, $p_{k-1} = (p, q)$. It is easy to show, for example by induction, that each p_i is of the form $u_i p + v_i q$, for some polynomials u_i, v_i. Hence, the Euclidean algorithm yields

$$(p(z), q(z)) = u(z)p(z) + v(z)q(z), \text{ for some } u(z), v(z) \in \mathbb{F}[z]. \tag{1.14.3}$$

(This formula holds for any $p, q \in \mathbb{F}[z]$.) Note that $p, q \in \mathbb{F}[z]$ are called *coprime* if $(p, q) \in \mathbb{F}$.

Note that if we divide $p(z)$ by $z - a$, we get the residue $p(a)$, i.e. $p(z) = (z - a)q(z) + p(a)$. Hence, $z - a$ divides $p(z)$ if and only if $p(a) = 0$, i.e. a is the root of p. A monic $p(z)$ *splits* to a product of linear factors if

$$p(z) = (z - z_1)(z - z_2)\ldots(z - z_n) = \prod_{i=1}^{n} (z - z_i). \tag{1.14.4}$$

Note that $z_1, \ldots, z_n$ are the *roots* of p.

Let $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{F}^n$. Denote

$$\sigma_k(\mathbf{z}) := \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} z_{i_1} z_{i_2} \ldots z_{i_k}, \quad k = 1, \ldots, n. \tag{1.14.5}$$

Then, $\sigma_k(\mathbf{z})$ is called the $k - th$ *elementary symmetric polynomial* in $z_1, \ldots, z_n$. Observe that

$$\sigma_1(\mathbf{z}) = z_1 + z_2 + \ldots + z_n, \ n \text{ summands},$$

$$\sigma_2(\mathbf{z}) = z_1 z_2 + \ldots + z_1 z_n + z_2 z_3 + \ldots + z_n z_n + \cdots + z_{n-1} z_n, \ \frac{n(n+1)}{2} \text{ summands},$$

$$\sigma_n(\mathbf{z}) = z_1 z_2 \ldots z_n, \ n \text{ terms in the product}.$$

A straightforward calculation shows

$$\prod_{i=1}^{n}(z-z_i)=z^n+\sum_{i=1}^{n}(-1)^i\sigma_i(\mathbf{z})z^{n-i}. \tag{1.14.6}$$

Remark 1.14.1 *For a field $\mathbb{F}$, the ring of polynomials $p(z_1,\ldots,z_n)$ in variables $z_1,\ldots,z_n$ and with coefficients in $\mathbb{F}$ is denoted by $\mathbb{F}[z_1,\ldots,z_n]$. The field of fractions of the polynomial ring $\mathbb{F}[z_1,\ldots,z_n]$ over $\mathbb{F}$ is the field of fractions $\left\{\frac{f(z_1,\ldots,z_n)}{g(z_1,\ldots,z_n)};\ f,g\in\mathbb{F}[z_1,\ldots,z_n], g\neq 0\right\}$ and this is denoted by $\mathbb{F}(z_1,\ldots,z_n)$.*

1.14.2 Finite fields and finite extension of fields

A finite field of order q exists if and only if the order q is a prime power p^k. (Why?) The field of order q is denoted by $\mathbb{F}_q$. For the prime number p, $\mathbb{F}_p$ is isomorphic to $\mathbb{Z}_p$, while $\mathbb{F}_{p^k}$ can be viewed as a vector space of dimension k over $\mathbb{F}_p$. Moreover, it can be proved that for any prime p and positive integer n, there exists an irreducible polynomial $\pi(x)\in\mathbb{F}_p[x]$ of degree n such that $\mathbb{F}_p[x]/\langle\pi(x)\rangle\cong\mathbb{F}_{p^n}$.

Suppose that E/$\mathbb{F}$ is a field extension. Then, E may be considered as a vector space over $\mathbb{F}$. The dimension of this vector space is called the *degree* of the field extension, and it is denoted by $[\mathrm{E}:\mathbb{F}]$. The degree may be finite or infinite, the field extension being called a *finite extension* or *infinite extension* accordingly. An extension E/$\mathbb{F}$ is also sometimes said to be simply finite if it is a finite extension; this should not be confused with the fields themselves being finite fields (fields with finitely many elements). For example, $[\mathbb{C}:\mathbb{R}]=2$ as $\dim_{\mathbb{R}}\mathbb{C}=2$.

A polynomial $p(z)\in\mathbb{F}[z]$ is called *irreducible*, if all polynomials q that divide p are either constant non-zero polynomials or polynomials of the form $ap(z)$, where $a\in\mathbb{F}\smallsetminus\{0\}$. The field $\mathbb{F}$ is called an *algebraically closed field* if any monic polynomial $p(z)\in\mathbb{F}[z]$ splits to linear factors in $\mathbb{F}[z]$. It is easy to see that $\mathbb{F}$ is algebraically closed if and only if the only irreducible monic polynomials are $z-a$, for all $a\in\mathbb{F}$. Then, $\mathbb{F}$ is not algebraically closed if and only if there exists an irreducible monic polynomial in $\mathbb{F}[z]$ of degree greater than 1. For example, $\mathbb{R}$ is not algebraically closed as $x^2+1\in\mathbb{R}[x]$ is irreducible over $\mathbb{R}$.

If $f(z)=a_0+a_1z+\cdots+a_nz^n\in\mathbb{F}[z]$, an extension field E of $\mathbb{F}$ is called a *splitting field* for $f(z)$ over $\mathbb{F}$ if there exist elements $c_1,\ldots,c_n\in E$ such that

(i) $f(z)=a_n(z-c_1)\cdots(z-c_n)$,

(ii) $E=\mathbb{F}(c_1,\ldots,c_n)$.

For example $Q(\sqrt{2})$ is a splitting field of $x^2-2\in Q[x]$ over Q.

Theorem 1.14.2 *Let $\mathbb{F}$ be a field. Assume that $p(z)=z^d+\sum_{i=1}^{d}a_iz^{d-i}$ be an irreducible polynomial, where $d>1$. Denote by $\mathbb{F}[z]/\langle p(z)\rangle$ the set of all polynomials modulo $p(z)$. That is, for $f(z),g(z)\in\mathbb{F}[z]$, $f(z)\equiv g(z)$ if the polynomial $f(z)-g(z)$ is divided by $p(z)$. Then, this set is a field, denoted by $\mathbb{F}_{p(z)}$, under the addition and product modulo $p(z)$. Moreover, $\mathbb{F}_{p(z)}$ is a vector space over $\mathbb{F}$ of dimension d. The set of all constant polynomials in $\mathbb{F}_{p(z)}$ is isomorphic to $\mathbb{F}$. ($\mathbb{F}_{p(z)}$ is called a finite extension of $\mathbb{F}$, and more precisely an extension of degree d).*

1.14.3 Worked-out Problems

1. An element $\lambda \in \mathbb{F}$ is called an *eigenvalue* of $A \in \mathbb{F}^{n\times n}$ if there exists $0 \neq \mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$. This $\mathbf{x} \neq 0$ is called an *eigenvector* of A corresponding to λ. It can be shown that λ is an eigenvalue of A if and only if it is a root of the polynomial $\det(zI_n - A)$. (See Problem 1.14.4-5.) Also, $\det(zI_n - A)$ is called the characteristic polynomial of A and it is denoted by $P_A(z)$ or simply $P(z)$.

   (a) Give an example of $A \in \mathbb{F}_3^{2\times 2}$ that does not have an eigenvalue in $\mathbb{F}_3$.

   (b) Does this matrix have a multiple eigenvalue in some field extension of $\mathbb{F}_3$, where the characteristic polynomial of A splits?

   Solution:

   (a) Consider $A = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} \in \mathbb{F}_3^{2\times 2}$. If A has an eigenvalue in $\mathbb{F}_3$, then by Problem 1.14.4-5, $\det(zI - A) = 0$ has solution in $\mathbb{F}_3$. Then, $z^2 - 2 = 0$ has root in $\mathbb{F}_3$. But $0^2 = 0$, $1^2 = 1$ and $2^2 = 1$ in $\mathbb{F}_3$. Thus, A has no eigenvalue in $\mathbb{F}_3$.

   (b) Assume that λ is a multiple eigenvalue of A, then $z^2 - 2 = (z-a)^2 = z^2 - 2az + a^2$. This means $2a = 0$ and since $\mathrm{char}\mathbb{F}_3) \neq 2$, then $a = 0$, which is impossible. Therefore, A does not have any multiple eigenvalue in any extension of $\mathbb{F}_3$.

2. Let $A \in \mathbb{R}^{n\times n}$ be a skew-symmetric matrix. Show that if n is odd, then A is singular.
   Solution:
   We know that $\det A = \det A^\top$. On the other hand, $-A = A^\top$ as A is skew-symmetric. This means $\det A = \det A^\top = \det(-A) = (-1)^n \det A$. Since n is odd, then $(-1)^n = -1$. Thus, $\det A = -\det A$ and so $\det A = 0$. Therefore, A is singular.

3. If A is an invertible matrix and B is a matrix such that AB exists, prove that $\operatorname{rank} AB = \operatorname{rank} B$.
   Solution:
   Assume that $C = AB$. Since A is invertible, then $B = A^{-1}C$. Using Problem 1.8.5-3, we have
   $$\operatorname{rank} C \leq \operatorname{rank} B,$$
   and
   $$\operatorname{rank} B = \operatorname{rank}\,(A^{-1}C) \leq \operatorname{rank} C.$$
   Then, $\operatorname{rank} B = \operatorname{rank} C = \operatorname{rank} AB$.

4. If $A \in \mathbb{F}^{n\times n}$ satisfies $A^2 = A$, show that $\operatorname{rank} A + \operatorname{rank}\,(I_n - A) = n$.
   Solution:
   Since $A - A^2 = 0$, then $A(I_n - A) = 0$. As the sum of the matrices A and $I_n - A$ is the matrix I_n, using Sylvester Inequality we have:
   $$n = \operatorname{rank}\,(A + I_n - A) \leq \operatorname{rank} A + \operatorname{rank}\,(I_n - A). \tag{1.14.7}$$

Again since $0 = \text{rank}\,(A - A^2) = \text{rank}\,(A(I_n - A))$, by Sylvester Inequality $0 \geq \text{rank}\,A + \text{rank}\,(I_n - A) - n$, i.e.

$$n \geq \text{rank}\,A + \text{rank}\,(I_n - A). \tag{1.14.8}$$

Hence from (1.14.7) and (1.14.8), we get $\text{rank}\,A + \text{rank}\,(I_n - A) = n$.

5. Show that a finite dimensional vector space over an infinite field cannot be written as a finite union of its proper subspaces.
   Solution:
   Assume that $\mathbf{V}$ is a vector space over the infinite field $\mathbb{F}$. Assume to the contrary, $\mathbf{V} = \bigcup_{i=1}^{n} \mathbf{V}_i$, where $\mathbf{V}_i$'s are proper subspaces of $\mathbf{V}$. Pick a non-zero vector $\mathbf{x} \in \mathbf{V}_1$. Choose $\mathbf{y} \in \mathbf{V} \setminus \mathbf{V}_1$, and note that there are infinitely many vectors of the form $\mathbf{x} + c\mathbf{y}$, with $c \in \mathbb{F}^*$. Now, $\mathbf{x} + c\mathbf{y}$ in never in $\mathbf{V}_1$ and so there is $\mathbf{V}_j$, $j \neq 1$, with infinitely many of these vectors, so it contains $\mathbf{y}$, and thus contains $\mathbf{x}$. Since $\mathbf{x}$ was arbitrary, we see $\mathbf{V}_1$ is contained in $\bigcup_{i=2}^{n} \mathbf{V}_i$; clearly this process can be repeated to find a contradiction.

6. Show that a finite dimensional vector space over an arbitrary filed (not an infinite field necessarily) cannot be written as the union of two proper subspaces.
   First Solution:
   Bearing in mind the previous problem, the statement is clear for vector spaces over infinite fields. Assume that the background field $\mathbb{F}$ is finite and $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2$, where $\mathbf{V}_1$ and $\mathbf{V}_2$ are proper subspaces. Using Problem 1.14.4-3, it follows $\#\mathbb{F} = q$, where q is a prime power. Assume that $\dim_{\mathbb{F}} \mathbf{V} = n$. Thus, $\#\mathbf{V} = q^n$. Clearly, the cardinality of $\mathbf{V}_i$'s can be at most q^{n-1} (because of Lagrange's theorem). Since $\mathbf{V}_i$'s have at least the zero element in common, $\#\mathbf{V}_1 \cup \mathbf{V}_2 \leqslant 2q^{n-1} - 1$ which is strictly less than q^n as $q \geq 2$. This shows that the statement is valid for vector spaces over finite fields as well.
   Second Solution:
   Assume to the contrary, there exist two proper subspaces $\mathbf{V}_1$ and $\mathbf{V}_2$ of $\mathbf{V}$ for which $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2$. Choose the elements $\mathbf{x}$ and $\mathbf{y}$ of $\mathbf{V}_1 \setminus \mathbf{V}_2$ and $\mathbf{V}_2 \setminus \mathbf{V}_1$, respectively. Then, $\mathbf{x} + \mathbf{y} \in \mathbf{V}_1 \cup \mathbf{V}_2$ since $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2$. This contradicts the cases $\mathbf{x} \notin \mathbf{V}_2$ and $y \notin \mathbf{V}_1$.

1.14.4 Problems

1. Prove Theorem 1.14.2.

2. Show that there is only one monic irreducible polynomial of degree two over $\mathbb{F}_2$. Describe the extension of $\mathbb{F}_2$ of degree 2.

3. Show that every finite field is of order p^m, for some prime p and positive integer m.
   (Hint: Use Lagrange's theorem for groups to show that there is no prime other than p which divides $\#\mathbb{F}$, where $\mathbb{F}$ is considered as a finite field.)

4. Show that any finite extension of $\mathbb{F}_p$, where $p \geq 2$ is prime, has p^d elements.

5. The characteristic polynomial of $A \in \mathbb{F}^{n \times n}$ is defined as $\det(zI_n - A)$.

(a) Show that the characteristic polynomial is a monic polynomial of degree n.

(b) Prove that the coefficient of z^{n-k} of the characteristic polynomial of A is the sum of all minors $\det A[\alpha, \alpha]$, where α runs over all subsets of $[n]$ of cardinality k.

(c) Show that λ is an eigenvalue of A if and only if λ is a zero of the characteristic polynomial of A.

6. Find an $A \in \mathbb{F}_2^{2\times 2}$, which does not have eigenvalues in $\mathbb{Z}_2$.

7. Show that $\mathbb{C}$ is a 2-extension of $\mathbb{R}$, i.e. the degree of $\mathbb{C}$ over $\mathbb{R}$ as a field extension is 2. What is the corresponding irreducible polynomial in $\mathbb{R}[z]$?

8. Let $p \geq 3$ be a prime and consider the polynomial $f_p = \sum_{i=0}^{p-1} z^i \in \mathbb{Q}[z]$. Show that this polynomial is irreducible over $\mathbb{Q}[z]$.

9. If $\mathbb{F}$ is a finite field, show that $\mathbb{F}^* = \mathbb{F} \setminus \{0\}$ is a *cyclic group* under multiplication.

10. Show that any finite field has prime characteristic.
(Hint: Use Worked-out Problem 1.5.1-2)

11. Find a non-zero symmetric matrix $A \in \mathbb{F}_p^{n\times n}$ (p a prime) such that $A^2 = 0$.

12. Find a non-zero matrix $A \in \mathbb{C}^{2\times 2}$ such that $AA^\top = 0$.

13. Prove that if the union of two subspaces is a subspace, then one of the two subspaces contains the other.

1.15 The general linear group

If $\mathbb{F}$ is a field, then $\mathrm{GL}(n, \mathbb{F})$, the subset of $n \times n$ invertible matrices of $\mathbb{F}^{n\times n}$, is a group under matrix multiplication. (Here, n is a positive integer.) This group is called the *general linear group of degree n*.

It is not immediately clear whether $\mathrm{GL}(n, \mathbb{F})$ is an infinite group when $\mathbb{F}$ is. However, such is the case. If $a \in \mathbb{F}$ is non-zero, then aI_n is an invertible $n \times n$ matrix with inverse $a^{-1}I_n$. Indeed, the set of all such matrices forms a subgroup of $\mathrm{GL}(n, \mathbb{F})$ that is isomorphic to $\mathbb{F}^* = \mathbb{F} \setminus \{0\}$.

Obviously, if $\mathbb{F}$ is a finite field, $\mathrm{GL}(n, \mathbb{F})$ is. An interesting question: how many elements this group has. Before answering this question completely, let's look at particular case $n = 1$; clearly $\mathrm{GL}(1, \mathbb{F}_q) \cong \mathbb{F}_q^*$, which has $q-1$ elements (note that here q is a prime power and $\mathbb{F}_q$ denotes a field with q elements.)

Theorem 1.15.1 *The number of elements of* $\mathrm{GL}(n, \mathbb{F}_q)$ *is* $\prod_{k=0}^{n-1}(q^n - q^k)$.

Proof. We will count the $n \times n$ matrices whose rows are linearly independent. The first row can be anything other than the zero row, so there are $q^n - 1$ possibilities. The second row must be linearly independent from the first, which is to say that it must not be a multiple of the first. Since there are q multiples of the first row, there are $q^n - q$ possibilities for the second row. In general, the ith row must be linearly independent from the first $i - 1$ rows, which means that it cannot be a linear combination of the first $i - 1$ rows. There are q^{i-1} linear combinations of the first $i - 1$ rows, so there are $q^n - q^{i-1}$ possibilities for the ith row. Once we build the entire matrix this way, we know that the rows are all linearly independent by choice. Also, we can build any $n \times n$ matrix whose rows are linearly independent in this fashion. Thus, there are $(q^n - 1)(q^n - q)\cdots(q^n - q^{n-1}) = \prod_{k=0}^{n-1}(q^n - q^k)$ such matrices. □

Now, we consider an interesting subgroup of $\mathrm{GL}(n, \mathbb{F})$. The determinant function, $\det : \mathrm{GL}(n, \mathbb{F}) \to \mathbb{F}^*$ is a group homomorphism; it maps the identity matrix to 1, and it is multiplicative, as desired. We define the *special linear group* $\mathrm{SL}(n, \mathbb{F})$, to be the kernel of this homomorphism. Put another way, $\mathrm{SL}(n, \mathbb{F}) = \{M \in \mathrm{GL}(n, \mathbb{F});\ \det M = 1\}$.

1.15.1 Matrix Groups

This subsection is devoted to subgroups of general linear groups. We will show that any subgroup of $\mathrm{GL}(n, \mathbb{C})$ is isomorphic to a subgroup of $\mathrm{GL}(m, \mathbb{R})$, for some m.

Definition 1.15.2 *A subgroup G of $\mathrm{GL}(n, \mathbb{F})$ with the operation of matrix multiplication is called a matrix group over $\mathbb{F}$.*

Example 1.15.3 $\mathrm{SL}(n, \mathbb{F})$ *is a matrix group over* $\mathbb{F}$.

Definition 1.15.4 *A subgroup H of a matrix group G is called a matrix subgroup of G.*

Example 1.15.5 *We can consider $\mathrm{GL}(n, \mathbb{F})$ as a subgroup of $\mathrm{GL}(n + 1, \mathbb{F})$ by identifying the matrix $A = [a_{ij}]$ with*

$$\begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix}.$$

Hence, $\mathrm{GL}(n, \mathbb{F})$ is a matrix subgroup of $\mathrm{GL}(n + 1, \mathbb{F})$. Similarly, it is verified that $\mathrm{SL}(n, \mathbb{F})$ is a matrix subgroup of $\mathrm{SL}(n + 1, \mathbb{F})$.

Example 1.15.6 *Consider the upper triangular matrix $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ with $a_{ii} = 1$, $1 \le i \le n$. Then, A is called unipotent. The upper trinangular subgroup of $\mathrm{GL}(n, \mathbb{F})$ is $UT(n, \mathbb{F}) = \{A \in \mathrm{GL}(n, \mathbb{F});\ A$ is upper trinangular$\}$, while the unipotent subgroup of $\mathrm{GL}(n, \mathbb{F})$ is*

$$SUT(n, \mathbb{F}) = \{A \in \mathrm{GL}(n, \mathbb{F});\ A \ \ \textit{is unipotent}\}.$$

Note that $SUT(n, \mathbb{F})$ is a matrix subgroup of $UT(n, \mathbb{F})$.
For the case $SUT(2, \mathbb{F}) = \left\{\begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix} \in \mathrm{GL}(2, \mathbb{F});\ c \in \mathbb{F}\right\}$, the function $\sigma : \mathbb{F} \to SUT(2, \mathbb{F})$

defined as $\sigma(c) = \begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix}$*, is a group isomorphism. Then, we can view* $\mathbb{F}$ *as a matrix group.*

Note that the complex numbers can be viewed as a 2-dimensional real vector space. Similarly, every $A \in \mathbb{C}^{n\times n}$ *can be viewed as a* $2n \times 2n$ *real matrix as follows:*

Consider the function $f : \mathbb{C} \to \mathbb{R}^{2\times 2}$ *defined as* $\mathbf{x} + \mathbf{y}i \mapsto \begin{bmatrix} \mathbf{x} & -\mathbf{y} \\ \mathbf{y} & \mathbf{x} \end{bmatrix}$*. It is easily verified that* f *is an injective ring homomorphism. Thus, we can view* $\mathbb{C}$ *as a subring of* $\mathbb{R}^{2\times 2}$*. Note that* $f(\bar{z}) = f(z)^\top$*. (See Section 1.15 for the conjugate of a complex number.)*

Given $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ *with* $a_{rs} = \mathbf{x}_{rs} + \mathbf{y}_{rs}i$*, we write* $A = [\mathbf{x}_{ij}] + i[\mathbf{y}_{ij}]$*, where the matrices* $\mathbf{X} = [\mathbf{x}_{ij}]$ *and* $\mathbf{Y} = [\mathbf{y}_{ij}]$ *are real symmetric. Define a function* $f_n : \mathbb{C}^{n\times n} \to \mathbb{R}^{2n\times 2n}$ *by* $A \mapsto \begin{bmatrix} \mathbf{X} & -\mathbf{Y} \\ \mathbf{Y} & \mathbf{X} \end{bmatrix}$*, which is an injective ring homomorphism.*

Let $J_{2n} = \begin{bmatrix} O_n & -I_n \\ I_n & O_n \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$*. Then* $J_{2n}^2 = -I_{2n}$ *and* $J_{2n}^\top = -J_{2n}$*. We have*

$$\begin{aligned} f_n(A) &= \begin{bmatrix} \mathbf{X} & O_n \\ O_n & \mathbf{X} \end{bmatrix} + \begin{bmatrix} \mathbf{Y} & O_n \\ O_n & \mathbf{Y} \end{bmatrix} J_{2n}; \\ f_n(\bar{A}) &= f_n(A)^\top. \end{aligned}$$

Note that $f_n(\mathrm{GL}(n,\mathbb{C}))$ *is a subgroup of* $\mathrm{GL}(2n,\mathbb{R})$*, so any subgroup* G *of* $\mathrm{GL}(n,\mathbb{C})$ *can be viewed as a matrix subgroup of* $\mathrm{GL}(2n,\mathbb{R})$ *by identifying it with* $f_n(G)$*. (Here, we use the fact that* f_n *is continuous.)*

1.15.2 Worked-out Problems

1. Determine the number of elements of $\mathrm{SL}(n,\mathbb{F}_q)$.
   Solution:
   Consider the group homomorphism $\det : \mathrm{GL}(n,\mathbb{F}_q) \to \mathbb{F}_q^*$. This map is surjective and sine $\mathrm{SL}(n,\mathbb{F}_q)$ is the kernel of the homomorphism, it follows from the First Isomorphism Theorem that $\mathrm{GL}(n,\mathbb{F}_q)/\mathrm{SL}(n,\mathbb{F}_q) \cong \mathbb{F}_q^*$. Therefore,

   $$\#\,\mathrm{SL}(n,\mathbb{F}_q) = \frac{\#\,\mathrm{GL}(n,\mathbb{F}_q)}{\#\mathbb{F}_q^*} = \frac{\prod_{k=0}^{n-1}(q^n - q^k)}{q-1}.$$

   Note that since $\ker(\det) = \mathrm{SL}(n,\mathbb{F}_q)$, then $\mathrm{SL}(n,\mathbb{F}_q) \lhd \mathrm{GL}(n,\mathbb{F}_q)$. Also, this is the case for infinite fields.

1.15.3 Problems

1. Prove that $Z(\mathrm{GL}(n,\mathbb{F})) = \{a \cdot I_n;\ a \in \mathbb{F}^*\}$.

2. Prove that $Z(\mathrm{SL}(n,\mathbb{F})) = \{a \cdot I_n;\ a \in \mathbb{F}^* \text{ and } a^n = 1\}$.

1.16 Complex numbers

Denote by $\mathbb{C}$, the field of complex numbers. A *complex number* z can be written in the form $z = x + \mathbf{i}y$, where $x, y \in \mathbb{R}$. Here, $\mathbf{i}^2 = -1$. Sometimes in this book we denote $\mathbf{i}$ by $\sqrt{-1}$. Then, $\mathbb{C}$ can be viewed as $\mathbb{R}^2$, where the vector $(x,y)^\top$ represents z. Note that $x = \Re z$, the *real part* of z, and $y = \Im z$, the *imaginary part* of z. In addition, $\bar{z} = x - \mathbf{i}y$ is the *conjugate* of z. Note that $|z| = \sqrt{x^2 + y^2}$ is the absolute value of z or the modulus of z. For $z \neq 0$, the *argument* of z is defined as $\arctan \frac{y}{x}$. The polar representation of z is $z = re^{\mathbf{i}\theta} = r(\cos\theta + \mathbf{i}\sin\theta)$. Here, r and θ are the *modulus* and the *argument* of $z(\neq 0)$, respectively. Let $w = u + \mathbf{i}v = R(\cos\psi + \mathbf{i}\sin\psi)$, where $u, v \in \mathbb{R}$. Then

$$z + w = (x+u) + \mathbf{i}(y+v), \quad zw = (xu - yv) + \mathbf{i}(xv + yu) = rRe^{i(\theta+\psi)},$$
$$\frac{w}{z} = \frac{1}{z\bar{z}} w\bar{z} = \frac{R}{r} e^{\mathbf{i}(\psi-\theta)} \text{ if } z \neq 0.$$

For a complex number $w = Re^{\mathbf{i}\psi}$ and a positive integer $n \geq 2$, the equation $z^n - w = 0$ has n-complex distinct roots for $w \neq 0$, which are $R^{\frac{1}{n}} e^{\mathbf{i}\frac{\psi + 2k\pi}{n}}$ for $k = 0, 1, \ldots, n-1$.

The fundamental theorem of algebra states that any monic polynomial $p(z) \in \mathbb{C}[z]$ of degree $n \geq 2$ splits to linear factors, i.e. $p(z) = \prod_{i=1}^n (z - z_i)$. See [6] for more details about fundamental theorem of algebra. The interested reader is referred to [21] to see an improvement of the fundamental theorem of algebra.

1.16.1 Worked-out Problems

1. Find the real part of $(\cos 0.7 + i \sin 0.7)^{53}$.
   Solution:
   This is the same as $\left(e^{0.7i}\right)^{53} = e^{37.1i} = \cos(37.1) + i\sin(37.1)$.
   Then, the real part is simply $\cos(37.1)$.

2. Write $(1-i)^{100}$ as $a + ib$, where a and b are real.
   Solution:
   The complex number $1 - i$ has modulus $\sqrt{2}$ and argument $-\frac{\pi}{4}$. That is

$$\begin{aligned}
1 - i &= \sqrt{2}\left(\cos\left(-\frac{\pi}{4}\right) + i\sin\left(-\frac{\pi}{4}\right)\right) \Rightarrow \\
(1-i)^{100} &= (\sqrt{2})^{100}\left(\cos\left(-\frac{100\pi}{4}\right) + i\sin\left(-\frac{100\pi}{4}\right)\right) \\
&= 2^{50}\left(\cos\left(-25\pi\right) + i\sin\left(-25\pi\right)\right) \\
&= 2^{50}\left(-1 + 0i\right) = -2^{50}.
\end{aligned}$$

1.16.2 Problems

1. Show that $p(z) = z^2 + bz + c \in \mathbb{R}[z]$ is irreducible over $\mathbb{R}$ if and only if $b^2 < 4c$.

2. Show that any monic polynomial $p(z) \in \mathbb{R}[z]$ of degree at least 2 splits to a product of irreducible linear and quadratic monic polynomials over $\mathbb{R}[z]$.

3. Deduce from the previous problem that any $p(z) \in \mathbb{R}[z]$ of odd degree must have a real root.

4. Show that for $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$, $(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta$.

1.17 Linear operators (Second encounter)

Let $\mathbf{V}$ and $\mathbf{U}$ be two finite dimensional subspaces over $\mathbb{F}$, where $\dim \mathbf{V} = n, \dim \mathbf{U} = m$. Recall that a map $T : \mathbf{V} \to \mathbf{U}$ is called a *linear transformation* (or *linear operator*) if $T(a\mathbf{u} + b\mathbf{v}) = aT(\mathbf{u}) + bT(\mathbf{v})$, for all $a, b \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in \mathbf{V}$. Assume that $\alpha = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\beta = \{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ are two bases in $\mathbf{V}$ and $\mathbf{U}$, respectively. Then, T is completely determined by $T(\mathbf{v}_j) = \sum_{i=1}^m a_{ij}\mathbf{u}_i, j = 1, \dots, n$. Let $A = [a_{ij}]_{j=i=1}^{m,n} \in \mathbb{F}^{m\times n}$. Then, the above equality is equivalent to

$$T[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] = [T(\mathbf{v}_1), T(\mathbf{v}_2), \dots, T(\mathbf{v}_n)] = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]A. \tag{1.17.1}$$

The matrix A is called the *representation matrix* of T in the ordered bases α and β; we write $[T]_\alpha^\beta$. Note that if $\mathbf{V} = \mathbf{U}$ and $\alpha = \beta$, we write $[T]_\beta$ instead of $[T]_\beta^\beta$. Assume that

$$[\mathbf{y}_1, \dots, \mathbf{y}_n] = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]Y,\ [\mathbf{x}_1, \dots, \mathbf{x}_m] = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]X, \tag{1.17.2}$$

where $Y \in \mathrm{GL}(n, \mathbb{F})$ and $X \in \mathrm{GL}(m, \mathbb{F})$ are other bases in $\mathbf{V}$ and $\mathbf{U}$, respectively. Then

$$T[\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]X^{-1}AY. \tag{1.17.3}$$

Denote by $\mathbf{V}' := L(\mathbf{V}, \mathbb{F})$, the set of all linear functions on $\mathbf{V}$. Then, $\mathbf{V}'$ is called the *dual space* of $\mathbf{V}$. If $\mathbf{v}, \mathbf{w} \in \mathbf{U}'$, then $a\mathbf{v} + b\mathbf{w}$ is the linear transformation, (also called linear functional), defined as follows: $(a\mathbf{v} + b\mathbf{w})(\mathbf{u}) = a\mathbf{v}(\mathbf{u}) + b\mathbf{w}(\mathbf{u})$. It is straightforward to show that $\mathbf{U}'$ is a vector space over the background field of $\mathbf{V}$ as a vector space.

Proposition 1.17.1 *Let $\mathbf{U}$ be a finite dimensional vector space. Then, $\mathbf{U}'$ is isomorphic to $\mathbf{U}$.*

Proof. Choose a basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for $\mathbf{U}$. Let $\mathbf{v}_i \in \mathbf{U}'$ be the following linear transformation $\mathbf{v}_i(\mathbf{u}_j) = \delta_{ij}$, for $i, j \in [n]$. (This basis is called a dual basis in $\mathbf{U}$.) It is straightforward to show that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis in $\mathbf{U}'$. □

If $\mathbf{U}$ is finite dimensional, then all three spaces $\mathbf{U}, \mathbf{U}', \mathbf{U}'' := (\mathbf{U}')'$ are isomorphic. There is a *natural* isomorphism $\phi : \mathbf{U} \to \mathbf{U}''$. Namely

$$\phi(\mathbf{u})(\mathbf{v}) := \mathbf{v}(\mathbf{u}), \quad \mathbf{u} \in \mathbf{U},\ \mathbf{v} \in \mathbf{U}'.$$

Let $T \in L(\mathbf{V}, \mathbf{U})$. Then, there exists a unique $T' \in L(\mathbf{U}', \mathbf{V}')$ defined as follows.

$$(T'\mathbf{w})(\mathbf{v}) = \mathbf{w}(T\mathbf{v}), \quad \mathbf{w} \in \mathbf{U}'. \tag{1.17.4}$$

Assume that $\mathbf{U}$ and $\mathbf{V}$ are finite dimensional. Choose the bases $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ in $\mathbf{U}$ and $\mathbf{V}$, respectively. Assume that T is represented in these bases by $A \in \mathbb{F}^{m\times n}$. Choose dual bases in $\mathbf{U}'$ and $\mathbf{V}'$, respectively. Then, T' is represented in these dual bases by $A^\top$.

Denote by $L(\mathbf{V}) := L(\mathbf{V}, \mathbf{V})$. Any $T \in L(\mathbf{V})$ is represented by a matrix $A \in \mathbb{F}^{n\times n}$ in a basis $[\mathbf{v}_1, \dots, \mathbf{v}_n]$ of $\mathbf{V}$ as follows:

$$T[\mathbf{v}_1, \dots, \mathbf{v}_n] = [\mathbf{v}_1, \dots, \mathbf{v}_n], \text{ i.e. } T(\mathbf{v}_j) = \sum_{i=1}^n a_{ij}\mathbf{v}_i, \text{ for } j \in [n]. \tag{1.17.5}$$

1.17.1 Worked-out Problems

1. If A and $B \in \mathbb{F}^{n\times n}$, then A is called to be *similar* to B if there exists an invertible matrix P such that $B = P^{-1}AP$. It is straightforward to show that the relation "A similar to B" is an equivalence relation in $\mathbb{F}^{n\times n}$. Equivalently, we say that A and B are similar.
   If A and $B \in \mathbb{F}^{n\times n}$ are similar, show that A and B have the same characteristic polynomials. Give an example where the opposite claim does not hold.
   Solution:
   Since A and B are similar, then $B = PAP^{-1}$, for some $P \in \mathrm{GL}(n,\mathbb{F})$. Thus , the characteristic polynomial of $B = \det(zI - B) = \det(zI - PAP^{-1}) = \det(P(zI - A)P^{-1}) = \det P.\det(zI-A).(\det P)^{-1} = \det(zI-A) =$ characteristic polynomial of A.
   For the second part, consider the matrices $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ in $\mathbb{R}^{2\times 2}$. Then, $\det(zI-A) = \det(zI-B) = z^2$. If A and B are similar, then $XAX^{-1} = 0$, for any $X \in \mathrm{GL}(2,\mathbb{R})$. But $B \neq 0$ and then B is not similar to A.

2. Assume that $\mathbf{V}$ is a finite dimensional vector space. Let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation. Show that the following statements are equivalent:

   (a) T is not one-to-one.

   (b) T is not onto.

   (c) $\det T = 0$.

   (d) T has 0 as an eigenvalue.

   Solution:
   Assume that T is represented by $A \in \mathbb{F}^{n\times n}$, where $n = \dim \mathbf{V}$. Put $r = \operatorname{rank} A$ and $m = \operatorname{null} A$.
   (a) $\Rightarrow$ (b) Since $m > 0$, then $r < n$. Then, $r = \dim(\mathrm{Rang}(T)) < n$ and so T is not onto.
   (b) $\Rightarrow$ (c) Since $r < n$, then $\det A = 0$ and so $\det T = 0$.
   (c) $\Rightarrow$ (d) Since $\det T = 0$, then $\det A = 0$ and so $\det(0I - A) = 0$. Therefore $\lambda = 0$ is an eigenvalue of T.
   (d) $\Rightarrow$ (a) Since T has 0 as an eigenvalue, then $A\mathbf{x} = 0\mathbf{x} = 0$, for some non-zero $\mathbf{x}$. Thus, $\operatorname{null} A > 0$ and so T is not one-to-one.

1.17.2 Problems

1. Show that any n-dimensional vector space $\mathbf{V}$ over the field $\mathbb{F}$ is isomorphic to $\mathbb{F}^n$.

2. Show that any finite dimensional vector space $\mathbf{V}$ over $\mathbb{F}$ is isomorphic to $(\mathbf{V}')'$. Define an explicit isomorphism from $\mathbf{V}$ to $(\mathbf{V}')'$.

3. Show that any $T \in L(\mathbf{V},\mathbf{U})$ is an isomorphism if and only if T is represented by an invertible matrix in some bases of $\mathbf{V}$ and $\mathbf{U}$.

4. Recall that the matrices A and $B \in \mathbb{F}^{n\times n}$ are called similar if $B = P^{-1}AP$, for some $P \in \mathrm{GL}(n,\mathbb{F})$. Show that similarity is an equivalence relation on $\mathbb{F}^{n\times n}$.

5. Show that the matrices A and $B \in \mathbb{F}^{n\times n}$ are similar if and only if they represent the same linear transformation $T \in L(\mathbf{V})$ in different bases, for a given n-dimensional vector space $\mathbf{V}$ over $\mathbb{F}$.

6. Suppose that A and $B \in \mathbb{F}^n$ have the same characteristic polynomial, which has n distinct roots in $\mathbb{F}$. Show that A and B are similar.

7. Show that every square matrix is similar (over the splitting field of its characteristic polynomial) to an upper triangular matrix.

8. Show that the eigenvalues of a triangular matrix are the entries on its main diagonal.

9. Assume that $A \in \mathbb{R}^{n\times n}$ and $\lambda \in \mathbb{C}$ is an eigenvalue of A. Show that $\bar{\lambda}$ is also an eigenvalue of A.

10. Let $\mathbf{V}$ and $\mathbf{W}$ two vector spaces over a field $\mathbb{F}$ and $T \in L(\mathbf{V}, \mathbf{W})$. Show that

    (a) If $\dim \mathbf{V} > \dim \mathbf{W}$, then T is not one-to-one.

    (b) If $\dim \mathbf{V} = \dim \mathbf{W}$, then T is one-to-one if and only if T is onto.

11. Let $A, B \in \mathbb{F}^{n\times n}$ are similar. Prove that $\det A = \det B$.

1.17.3 Trace

As we saw before, the determinant is a function that assigns a scalar value to every square matrix. Another important scalar-valued function is the trace.

The *trace* of $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ is denoted as $\operatorname{tr} A$ and defined to be the sum of the elements on the main diagonal of A, i.e.

$$\operatorname{tr} A = \sum_{i=0}^{n} a_{ii}$$

Clearly, $\det(zI - A) = z^n - \operatorname{tr} A z^{n-1} + \cdots$. Assume that A is similar to $B \in \mathbb{F}^{n\times n}$. As A and B have the same characteristic polynomial, it follows that $\operatorname{tr} A = \operatorname{tr} B$. Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$. Assume that $T \in L(\mathbf{V})$ is represented by A and B, respectively, in two different bases in $\mathbf{V}$. Therefore A and B are similar. Hence $\operatorname{tr} A = \operatorname{tr} B$. Then, the trace of T is denoted by $\operatorname{tr} T$ and defined to be $\operatorname{tr} A$.

Note that the trace is only defined for a square matrix. The following properties are obvious about the trace:
If $A, B \in \mathbb{F}^{n\times n}$ and $c \in \mathbb{F}$, then

1. $\operatorname{tr} A + B = \operatorname{tr} A + \operatorname{tr} B$
2. $\operatorname{tr} cA = c \operatorname{tr} A$
3. $\operatorname{tr} A = \operatorname{tr} A^\top$
4. $\operatorname{tr} AB = \operatorname{tr} BA$

Furthermore, if $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times m}$, then $\operatorname{tr} AB = \operatorname{tr} BA$. Also, the trace is invariant under cyclic permutations, i.e.

$$\operatorname{tr} ABCD = \operatorname{tr} BCDA = \operatorname{tr} CDAB = \operatorname{tr} DABC,$$

where $A, B, C, D \in \mathbb{F}^{n\times n}$. This is known as the cyclic property. Note that arbitrary permutations are not allowed; in general $\operatorname{tr} ABC \neq \operatorname{tr} ACB$.

1.17.4 Worked-out Problems

1. If $A, B \in \mathbb{F}^{n\times n}$, char $\mathbb{F} \neq 2$, and A is symmetric and B is skew-symmetric, prove that $\operatorname{tr} AB = 0$.
   Solution:

$$\operatorname{tr} AB \overset{(3)}{=} \operatorname{tr}(AB)^\top = \operatorname{tr} B^\top A^\top = \operatorname{tr} -BA \overset{(2)}{=} -\operatorname{tr} BA = -\operatorname{tr} AB.$$

   Then, $\operatorname{tr} AB = 0$.
   (Here, (2) and (3) mean the second and third properties mentioned for trace.)

1.17.5 Problems

1. Deduce the cyclic property by proving $\operatorname{tr} ABC = \operatorname{tr} BCA = \operatorname{tr} CAB$, where $A, B, C \in \mathbb{F}^{n\times n}$.

2. Let $A, B \in \mathbb{F}^n$. Prove or disprove the following statements:

   (a) $\operatorname{tr} AB = \operatorname{tr} A \operatorname{tr} B$,

   (b) $\operatorname{tr}(A^{-1}) = \frac{1}{\operatorname{tr} A}$, if A is invertible,

   (c) $\det(\mathbf{x}I - A) = \mathbf{x}^2 - \operatorname{tr} A + \det A$, if $A \in \mathbb{F}^{2\times 2}$.

3. Determine if the map $\varphi : \mathbb{F}^{n\times n} \to \mathbb{F}$ by $\varphi(A) = \operatorname{tr} A$ is a ring homomorphism.

Chapter 2

Tensor products

2.1 Universal property of tensor products

Let $\mathbf{U}_1, \ldots, \mathbf{U}_d$ and $\mathbf{V}$ be given finite dimensional subspaces over $\mathbb{F}$, where $d > 1$ is an integer. Denote by $\mathbf{U}_1 \times \cdots \mathbf{U}_d$ the product set, all d-tuples of the form $(\mathbf{u}_1, \ldots, \mathbf{u}_d)$. Let $L : \mathbf{U}_1 \times \cdots \times \mathbf{U}_d \to \mathbf{V}$ be a multilinear map. That is, for any fixed $i \in [d]$ and vectors $\mathbf{u}_j \in \mathbf{U}_j$, for $j \in [d] \setminus \{i\}$, the restriction of L to $(\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{U}_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_d)$ is linear in $\mathbf{u} \in \mathbf{U}_i$. Note that for any $1 \le p < q \le d$, we have equality

$$\begin{aligned} L(\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}, a\mathbf{u}_p, \mathbf{u}_{p+1}, \ldots, \mathbf{u}_d) &= aL(\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}, \mathbf{u}_p, \mathbf{u}_{p+1}, \ldots, \mathbf{u}_d) \\ &= L(\mathbf{u}_1, \ldots, \mathbf{u}_{q-1}, a\mathbf{u}_q, \mathbf{u}_{q+1}, \ldots, \mathbf{u}_d)(2.1.1) \end{aligned}$$

Theorem 2.1.1 *Let $\mathbf{U}_1, \ldots, \mathbf{U}_d$ be given finite dimensional subspaces over $\mathbb{F}$ ($d > 1$). Then, there exists a unique vector space $\mathbf{U}$ over $\mathbb{F}$, up to an isomorphism, of dimension $\prod_{i \in [d]} \dim \mathbf{U}_i$ with the following properties.*

1. *There exists a multilinear map $\iota : \mathbf{U}_1 \times \cdots \times \mathbf{U}_d \to \mathbf{U}$ with the following properties.*

   (a) $\iota((\mathbf{u}_1, \ldots, \mathbf{u}_d)) = 0$ if and only if $\mathbf{u}_i = 0$, for some $i \in [d]$.

   (b) Assume that $\mathbf{u}_i, \mathbf{v}_i \neq 0$, for $i \in [d]$. Then, $\iota((\mathbf{u}_1, \ldots, \mathbf{u}_d)) = \iota((\mathbf{v}_1, \ldots, \mathbf{v}_d))$ if and only if there exist scalars $t_i \in \mathbb{F}$, $i \in [d]$ satisfying $\prod_{i \in [d]} t_i = 1$ such that $\mathbf{v}_i = t_i \mathbf{u}_i$, for $i \in [d]$.

2. *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Assume that $L : \mathbf{U}_1 \times \cdots \times \mathbf{U}_d \to \mathbf{V}$ is a multilinear map. Then, there exists a unique linear map $\hat{L} : \mathbf{U} \to \mathbf{V}$ such that $L = \hat{L} \circ \iota$.*

Usually one denotes:

$$\mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_d = \otimes_{i \in [d]} \mathbf{U}_i := \mathbf{U}, \quad \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_d = \otimes_{i \in [d]} \mathbf{u}_i := \iota((\mathbf{u}_1, \ldots, \mathbf{u}_d)). \tag{2.1.2}$$

Property 2 is called the lifting property. It is usually started in terms of a commutative diagram.

$$\begin{tikzcd} \mathbf{U}_1 \times \cdots \times \mathbf{U}_d \arrow[rr, "\iota"] \arrow[dr, "L"'] & & \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_d \arrow[dl, "\hat{L}"] \\ & \mathbf{V} & \end{tikzcd}$$

Proof.

1. Assume that $\{\mathbf{e}_{1,i}, \dots, \mathbf{e}_{m_i,i}\}$ is a basis in $\mathbf{U}_i$, for $i \in [d]$. Let $\mathbf{U}$ be a finite vector space of dimension $m := m_1 \cdots m_d$. Choose a basis of cardinality m and denote the elements of this basis by $\mathbf{e}_{j_1,1} \otimes \cdots \otimes \mathbf{e}_{j_d,d}$, where $j_i \in [m_i]$ for $i \in [d]$. Assume that
$$\mathbf{u}_i = \sum_{j_i \in [m_i]} u_{j_i,i} \mathbf{e}_{j_i,i}, \qquad i \in [m_i]. \tag{2.1.3}$$
Define ι by the formula
$$\iota((\mathbf{u}_1, \dots, \mathbf{u}_d)) := \sum_{j_1 \in [m_1], \dots, j_d \in [m_d]} u_{j_1,1} \cdots u_{j_d,d} \mathbf{e}_{j_1,1} \otimes \cdots \otimes \mathbf{e}_{j_d,d}. \tag{2.1.4}$$
It is straightforward to see that ι is multilinear. Suppose that some $\mathbf{u}_i = 0$. Clearly, $\iota((\mathbf{u}_1, \dots, \mathbf{u}_d)) = 0$. Assume that each $\mathbf{u}_i \neq 0$. Then, $u_{j_i,i} \neq 0$, for some $j_i \in [m_i]$ for each $i \in [d]$. Thus, the coefficient $u_{j_1,1} \cdots u_{j_d,d}$ of the basis element $\mathbf{e}_{j_1,1} \otimes \cdots \otimes \mathbf{e}_{j_d,d}$ is non-zero, hence $\iota((\mathbf{u}_1, \dots, \mathbf{u}_d)) \neq 0$. To show part 1b, consider first the case $d = 2$. Let $\mathbf{u}_1 = \mathbf{x} = (x_1, \dots, x_p)^\top$, $\mathbf{u}_2 = \mathbf{y} = (y_1, \dots, y_q)^\top$, where $p = m_1$, $q = m_2$. The coefficients of $\mathbf{e}_{1,i}\mathbf{e}_{2,j}$ correspond to the rank one matrix $\mathbf{x}\mathbf{y}^\top$. Similarly, let $\mathbf{v}_1 = \mathbf{z}$, $\mathbf{v}_2 = \mathbf{w}$. Then, the equality $\iota((\mathbf{u}_1, \mathbf{u}_2)) = \iota((\mathbf{v}_1, \mathbf{v}_2))$ is equivalent to $\mathbf{x}\mathbf{y}^\top = \mathbf{z}\mathbf{w}^\top$. Hence, $\mathbf{z} = t_1\mathbf{x}$, $\mathbf{w} = t_2\mathbf{y}$ and $t_1 t_2 = 1$. The case $d \geq 3$ follows by induction, where we use the obvious identity:
$$\mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_d = \mathbf{U}_1 \otimes (\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_d). \tag{2.1.5}$$

2. Let $\hat{L}$ be define by $\hat{L}(\mathbf{e}_{j_1,1} \otimes \cdots \otimes \mathbf{e}_{j_d,d}) := L((\mathbf{e}_{j_1,1}, \dots, \mathbf{e}_{j_d,d})$, for $j_i \in [m_i]$ and $i \in [d]$. Then, $\hat{L}$ extends uniquely to a linear map from $\mathbf{U}$ to $\mathbf{V}$. It is straightforward to show that $L = \hat{L} \circ \iota$.
The isomorphism to two representation of a tensor product follows from 2. □

In quantum mechanics the tensor product $\mathbf{U} = \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_d$ over $\mathbb{C}$ is associated with the d-partite system. That is, a vector in $\mathbf{U}_i$ of length one represents the state of the particle i, while a tensor in $\mathbf{U}$ of length one represents the d-partite system, which is usually quantum entanglement, [14] .

2.2 Matrices and tensors

We start with the case $d = 2$. Assume that $\mathbf{U}_1 = \mathbf{E}$ and $\mathbf{U}_2 = \mathbf{F}$ are two vector spaces, and assume that $\dim \mathbf{E} = m$ and $\dim \mathbf{F} = n$. Let $\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ and $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ be bases in $\mathbf{E}$ and $\mathbf{F}$, respectively. Then, it follows from 2.1.1 that $\mathbf{E} \otimes \mathbf{F}$ has a basis $\mathbf{e}_i \otimes \mathbf{f}_j$, $i \in [m]$, $j \in [n]$. Thus, any vector in $\mathbf{E} \otimes \mathbf{F}$ is of the from $\sum_{i=j=1}^{m,n} a_{ij} \mathbf{e}_i \otimes \mathbf{f}_j$. Let $A = [a_{ij}]_{i=j=1}^{m,n} \in \mathbb{F}^{m \times n}$. Hence, $\mathbb{F}^m \otimes \mathbb{F}^n$ is isomorphic to the space of $m \times n$ matrices over the field $\mathbb{F}$, denoted as $\mathbb{F}^{m \times n}$. In particular
$$\mathbf{e} \otimes \mathbf{f} \leftrightarrow \mathbf{e}\mathbf{f}^\top, \quad \mathbf{e} \in \mathbb{F}^m, \ \mathbf{f} \in \mathbb{F}^n. \tag{2.2.1}$$
Hence, all tensors of the form $\mathbf{e} \otimes \mathbf{f}$ are called *rank one tensors*.
A more functorial isomorphism is the following. Let $\mathbf{F}'$ denote the dual space to $\mathbf{F}$,

i.e. all linear transformations from $\mathbf{F}$ to $\mathbf{F}$, denoted as $L(\mathbf{F},\mathbf{F})$. Denote $\mathbf{G} := \mathbf{F}'$. Then, $\mathbf{g} : \mathbf{F} \to \mathbf{F}$ is a linear functional. Recall next that $\mathbf{G}'$ is isomorphic to $\mathbf{F}$, where we define $\mathbf{f}(\mathbf{g}) := \mathbf{g}(\mathbf{f})$. Then, $\mathbf{e}\otimes\mathbf{f}$ can be viewed as an element of the space of linear transformations $L(\mathbf{F}',\mathbf{E})$, namely

$$(\mathbf{e}\otimes\mathbf{f})(\mathbf{g}) = \mathbf{f}(\mathbf{g})\mathbf{e} = \mathbf{g}(\mathbf{f})\mathbf{e}.$$

Hence, $\mathbf{E}\otimes\mathbf{F}$ is isomorphic to $L(\mathbf{F}',\mathbf{E})$.
Let $3 \le d \in \mathbb{N}$ and $m_1,\dots,m_d \in \mathbb{N}$. Denote by $\mathbb{F}^{m_1\times\cdots\times m_d}$ the linear space of all d-mode tensors $\mathcal{A} := [a_{j_1\cdots j_d}]$, where $j_i \in [m_i]$ for $i \in [d]$. Let $\mathbf{U}_i = \mathbb{F}^{m_i}$, for $i \in [d]$. By assuming that $\{\mathbf{e}_{1,i},\dots,\mathbf{e}_{m_i,i}\}$ is a standard basis in $\mathbb{F}^{m_i}$, for $i \in [d]$, the proof of Theorem 2.1.1 yields that $\otimes_{i\in[d]}\mathbb{F}^{m_i}$ is isomorphic to $\mathbb{F}^{m_1\times\cdots\times m_d}$.
Note that $\mathbf{u} \in \otimes_{i\in[d]}\mathbf{U}_i$ is called a *rank one tensor*, or *decomposable*, if $\mathbf{u} = \otimes_{i\in[d]}\mathbf{u}_i$, where each $\mathbf{u}_i$ is a non-zero vector. (If $\mathbf{u}_i$ is allowed to be zero, then $\mathbf{u}$ is at most rank one matrix.) Zero vector has rank zero. Each $\mathbf{0} \ne \mathbf{u} \in \otimes_{i\in[d]}\mathbf{U}_i$ is a sum of rank one tensors. Moreover, rank $\mathbf{u} = k$ if $\mathbf{u}$ is a sum of rank one tensors, and if $\mathbf{u}$ is a sum of k' rank one tensors than $k' \ge k$.

2.3 Tensor product of linear operators and Kronecker product

Let $T_i \in L(\mathbf{U}_i,\mathbf{V}_i)$ be linear operators for $i \in [d]$. Then, $\otimes_{i\in[d]}T_i$ acts on the rank one tensors in $\otimes_{i\in[d]}\mathbf{U}_i$ as follows:

$$(\otimes_{i\in[d]}T_i)(\otimes_{i\in[d]}\mathbf{u}_i) := \otimes_{i\in[d]}T_i(\mathbf{u}_i). \tag{2.3.1}$$

It is straightforward to check that that the above action of $\otimes_{i\in[d]}T_i$ on rank one tensors extends to a linear operator in $L(\otimes_{i\in[d]}\mathbf{U}_i,\otimes_{i\in[d]}\mathbf{V}_i)$, denoted as $\otimes_{i\in[d]}T_i$. Assume that $Q_i \in L(\mathbf{V}_i,\mathbf{W}_i)$, for $i \in [d]$. Then

$$(\otimes_{i\in[d]}Q_i)((\otimes_{i\in[d]}T_i)(\otimes_{i\in[d]}\mathbf{u}_i)) = (\otimes_{i\in[d]}Q_i)(\otimes_{i\in[d]}T_i(\mathbf{u}_i)) = \otimes_{i\in[d]}(Q_iT_i)(\mathbf{u}_i).$$

Hence

$$(\otimes_{i\in[d]}Q_i)(\otimes_{i\in[d]}T_i) = \otimes_{i\in[d]}Q_iT_i. \tag{2.3.2}$$

In particular, if $\mathbf{U}_i = \mathbf{V}_i$ and T_i is invertible for $i \in [d]$, then

$$(\otimes_{i\in[d]}T_i)^{-1} = \otimes_{i\in[d]}T_i^{-1}.$$

We now discuss a particular case of the tensor product $T_1\otimes T_2$. Assume that

$$\mathbf{U}_1 = \mathbb{F}^n,\ \ \mathbf{V}_1 = \mathbb{F}^m,\ \ \mathbf{U}_2 = \mathbb{F}^q,\ \ \mathbf{V}_2 = \mathbb{F}^p.$$

By choosing the standard bases in all the above four vector spaces, we have that T_1, T_2 are represented by the matrices $A = [a_{ij}] \in \mathbb{F}^{m\times n}$, $B = [b_{kl}] \in \mathbb{F}^{p\times q}$, respectively. Thus, the action of T_1 and T_2 can be described respectively as

$$\mathbf{x} \mapsto A\mathbf{x},\ \mathbf{x} = (x_1,\dots,x_n)^\top, \quad \mathbf{y} \mapsto B\mathbf{y},\ \mathbf{y} = (y_1,\dots,y_q)^\top.$$

Recall that in 2.2 we associated with $\mathbf{u}_1\otimes\mathbf{u}_2$ the matrix $\mathbf{x}\mathbf{y}^\top$. Hence, in view of (2.3.1) the action of $T_1\otimes T_2$ on $\mathbf{u}_1\otimes\mathbf{u}_2$ is given by $\mathbf{x}\mathbf{y}^\top \mapsto (A\mathbf{x})(B\mathbf{y})^\top = A(\mathbf{x}\mathbf{y}^\top)B^\top$.

Thus, if we identify $\mathbb{F}^n \otimes \mathbb{F}^q$ with $\mathbb{F}^{n\times q}$, the space of $n \times q$ matrices, then the action of $T_1 \otimes T_2$ is equivalent to

$$X \mapsto AXB^\top, \qquad X \in \mathbb{F}^{n\times q}. \tag{2.3.3}$$

In order to represent $T_1 \otimes T_2$ as a matrix, we need to convert the matrix $X = [x_{jl}] \in \mathbb{F}^{n\times q}$ to a column vector $\mathbb{F}^{nq}$. This task is achieved by arranging the pairs (i,j), $j \in [n]$, $l \in [q]$ in the lexicographical order:

$$(1,1), \dots, (1,q), (2,1), \dots, (2,q), \dots, (n,1), \dots, (n,q).$$

Let

$$\mathbf{c}(X) \coloneqq (x_{11}, \dots, x_{1q}, x_{21}, \dots, x_{2q}, \dots, x_{n1}, \dots, x_{nq})^\top. \tag{2.3.4}$$

Note that

$$\mathbf{c}(\mathbf{x}\mathbf{y}^\top) = \begin{bmatrix} x_1\mathbf{y} \\ x_2\mathbf{y} \\ \vdots \\ x_n\mathbf{y} \end{bmatrix}. \tag{2.3.5}$$

As the transformation (2.3.3) is a linear transformation from $\mathbb{F}^{n\times q}$ to $\mathbb{F}^{m\times p}$, it follows that we must have an equality of the form $\mathbf{c}(AXB^\top) = C\mathbf{c}(X)$, for some matrix $C \in \mathbb{F}^{(mp)\times(nq)}$. Use the choice $X = \mathbf{x}\mathbf{y}^\top$ and (2.3.5) to deduce that C has the following block matrix from:

$$\begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}. \tag{2.3.6}$$

The above matrix is called the *Kronecker (tensor) product* of A and B and denoted by $A\hat{\otimes}B$. In most references, $A\hat{\otimes}B$ is just identified with $A \otimes B$. Observe that $\mathbf{c}(\mathbf{x}\mathbf{y}^\top) = \mathbf{x}\hat{\otimes}\mathbf{y}$. Similarly, we can define the Kronecker tensor product of more than two matrices by the formula:

$$\otimes_{i\in[d]} A_i = (\otimes_{i\in[d-1]} A_i) \otimes A_d.$$

Equivalently, we can define $\otimes_{i\in[d]} A_i$ as follows; assume that $A_i = [a_{j_i k_i, i}] \in \mathbb{F}^{m_i \times n_i}$, for $i \in [d]$. Let

$$A = [a_{(j_1, j_2, \dots, d_d)(k_1, k_2, \dots, k_d)}] \in \mathbb{R}^{(m_1 m_2 \cdots m_d)\times(n_1 n_2 \cdots n_d)}, \tag{2.3.7}$$

$$a_{(j_1, j_2, \dots, d_d)(k_1, k_2, \dots, k_d)} = a_{j_1 k_1, 1} a_{j_2 k_2, 2}, \cdots a_{j_d k_d, d}.$$

Here, we arrange the indices $(j_1, j_2, \dots, j_d)$ and $(k_1, k_2, \dots, k_d)$ in the lexicographical order.

Another possibility is to view each A_i as a linear operator in $L(\mathbb{F}^{n_i}, \mathbb{F}^{m_i})$. Then, $\otimes_{i\in[d]} A_i$ is viewed as a tensor product of operators.

Let $A_i \in \mathbb{F}^{m_i\times n_i}$, $B_i \in \mathbb{F}^{n_i \times p_i}$ for $i \in [d]$. Then, we have identity

$$(\otimes_{i\in[d]} A_i)(\otimes_{i\in[d]} B_i) = \otimes_{i\in[d]} A_i B_i. \tag{2.3.8}$$

This identity holds if the above product are Kronecker products.

2.3.1 Worked-out Problems

1. A square matrix A is called *diagonalizable* if it is similar to a diagonal matrix, i.e. if there exists an invertible matrix P such that $P^{-1}AP$ is a diagonal matrix. Otherwise, A is called *nondiagonalizable.*
   Let $A \in \mathbb{F}^{m\times m}$ and $B \in \mathbb{F}^{n\times n}$. Assume that the characteristic polynomials of A and B split in $\mathbb{F}$:
   $$\det(\lambda I_m - A) = \prod_{i\in[m]}(\lambda - \alpha_i), \quad \det(\lambda I_n - B) = \prod_{j\in[n]}(\lambda - \beta_j).$$
   Show that
   $$\det(\lambda I_{mn} - A\otimes B) = \prod_{i\in[m],j\in[n]}(\lambda - \alpha_i\beta_j). \tag{2.3.9}$$
   Furthermore, if A and B are diagonalizable, prove that $A\otimes B$ is diagonalizable.
   Solution:
   Since the characteristic polynomials of A and B split, it is well-known that A and B are similar to upper triangular matrices: $A = UA_1U^{-1}$, $B = VB_1V^1$ (Problem 1.17.2-7). As the characteristic polynomials of A_1, B_1 are equal to A, B, respectively, it follows that the diagonal entries of A_1 and B_1 are $\{\alpha_1,\dots,\alpha_m\}$ and $\{\beta_1,\dots,\beta_n\}$, respectively. Observe that
   $$A\otimes B = (U\otimes V)(A_1\otimes A_2)(U^{-1}\otimes V^{-1}) = (U\otimes V)(A_1\otimes A_2)(U\otimes V)^{-1}.$$
   Thus, the characteristic polynomials of $A\otimes B$ and $A_1\otimes B_1$ are equal. Observe that $A_1\hat{\otimes}B_1$ is also upper triangular with the diagonal entries
   $$\alpha_1\beta_1,\dots,\alpha_1\beta_n,\alpha_2\beta_1,\dots,\alpha_2\beta_n,\dots,\alpha_m\beta_1,\dots,\alpha_m\beta_n.$$
   Hence, the characteristic polynomial of $A_1\otimes B_1$ is $\prod_{i\in[m],j\in[n]}(\lambda-\alpha_i\beta_j)$. This establishes (2.3.9).
   Assume that A and B are diagonalizable, i.e. one can assume that A_1 and A_2 are diagonal matrices. Clearly, $A_1\hat{\otimes}B_1$ is a diagonal matrix. Hence, $A\otimes B$ is a diagonal matrix.

   Note that *diagonalization* is the process of finding a corresponding diagonal matrix for a diagonalizable matrix or linear transformation.

2.3.2 Problems

1. Let $A \in \mathbb{F}^{m\times m}$ and $B \in \mathbb{F}^{n\times n}$. Assume that $\mathbf{x}_1,\dots,\mathbf{x}_k$ and $\mathbf{y}_1,\dots,\mathbf{y}_l$ are linearly independent eigenvectors of A and B with the corresponding eigenvalues $\alpha_1,\dots,\alpha_k$ and $\beta_1,\dots,\beta_l$, respectively. Show that $\mathbf{x}_1\otimes\mathbf{y}_1,\dots,\mathbf{x}_1\otimes\mathbf{y}_l,\dots,\mathbf{x}_k\otimes\mathbf{y}_1,\dots,\mathbf{x}_k\otimes\mathbf{y}_l$ are kl linearly independent vectors of $A\otimes B$ with the corresponding $\alpha_1\beta_1,\alpha_1\beta_l,\dots,\alpha_k\beta_l$, respectively.

2. Let $A_i \in \mathbb{F}^{m_i\times m_i}$ for $i\in[d]$, where $d\in\mathbb{N}$. Assume that the characteristic polynomial of A_i splits in $\mathbb{F}$: $\det(\lambda I_{m_i} - A_i) = \prod_{j_i\in[m_i]}(\lambda - \alpha_{j_i,i})$, for $i\in[d]$. Prove that $\det(\lambda I_{m_1\cdots m_d} - \otimes_{i\in[d]}A_i) = \prod_{j_i\in[m_i],i\in[d]}(\lambda - \alpha_{j_1,1}\cdots\alpha_{j_d,d})$. Furthermore, show that if each A_i is diagonalizable, then $\otimes_{i\in[d]}A_i$ is diagonalizable. (This statement can be viewed as a generalization of Worked-out Problem 2.3.1-1)

Chapter 3

Canonical forms

In this chapter, we will discuss two different canonical forms for similarity; the Jordan canonical form, which applies only when the base field is algebraically closed and the rational canonical form, which applies in all cases.

3.1 Jordan canonical forms

For a matrix $A \in \mathbb{F}^{n\times n}$, the polynomial $p(z) := \det(zI_n - A)$ is called the *characteristic polynomial* of A. An element $\lambda \in \mathbb{F}$ is called an *eigenvalue* of A if there exists $0 \neq \mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$. This $\mathbf{x} \neq 0$ is called an *eigenvector* of A corresponding to λ. Note that for an eigenvalue λ of, since $A\mathbf{x} = \lambda\mathbf{x}$, (for some $\mathbf{x} \neq 0$), then $\mathbf{x}$ is in the kernel of $\lambda I_n - A$. Using Theorem 1.7.17, $\lambda I_n - A$ is not invertible and $\det(\lambda I_n - A) = 0$. On the other hand, if $\lambda \in \mathbb{F}$ is a root of $p(z)$, then $\lambda I_n - A$ is not invertible and reusing Theorem 1.7.17 yields $\ker(\lambda I_n - A) \neq 0$. If $0 \neq \mathbf{x} \in \ker(\lambda I_n - A)$, then $A\mathbf{x} = \lambda\mathbf{x}$ and then λ is an eigenvalue of A. Therefore, if we call the set of eigenvalues of A as spectrum of A and denote it by spec A, then we have

$$\operatorname{spec} A = \{\lambda \in \mathbb{F};\ p(\lambda) = 0\}.$$

Geometrically, an eigenvector corresponding to a real, nonzero eigenvalue points in a direction that is stretched, and the eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction is reversed. In particular, the eigenvector does not change its direction under $T_A \in L(\mathbf{V})$, where $\mathbf{V}$ is a vector space.

Note that for an algebraically closed field $\mathbb{F}$, every matrix has at least one eigenvalue. However, the definition of eigenvalue does not show how to compute them in practice. To do this, we need to use the fact that eigenvalues are the roots of $\det(zI_n - A)$. If λ is an eigenvalue of A, then eigenvectors for λ are elements of the null space of $A - \lambda I_n$, which can be found via row reduction. Then, the problem of finding λ is solved by investigating the case $A - \lambda I_n$ is singular. Thus, we have reduced the problem of determining eigenvalues to the problem of determining invertibility of a matrix. This can be considered as a motivation to define determinant.

Lemma 3.1.1 *Let $A \in \mathbb{F}^{n\times n}$. Then,* $\det(zI_n - A) = z^n + \sum_{i=1}^n a_i z^{n-i}$ *and $(-1)^i a_i$ is the sum of all $i \times i$ principal minors of A. Assume that* $\det(zI_n - A) = (z - z_1)\dots(z - z_n)$*, and denote* $\mathbf{z} := (z_1, \dots, z_n)^\top \in \mathbb{F}^n$*. Then, $(-1)^i a_i = \sigma_i(\mathbf{z})$, for $i = 1, \dots, n$.*

Proof. Consider $\det(zI - A)$. To obtain the coefficient of z^{n-i}, we need to take the product of some $n-i$ diagonal elements of $zI_n - A$: $(z - a_{j_1 j_1}) \dots (z - a_{j_{n-i} j_{n-i}})$. We take z^{n-i} in this product. Then, this product is multiplied by the $\det(-A[\boldsymbol{\alpha}, \boldsymbol{\alpha}])$, where $\boldsymbol{\alpha}$ is the complement of $\{j_1, \dots, j_{n-i}\}$ in the set $[n]$. This shows that $(-1)^i a_i$ is the sum of all principal minors of A of order i.

Suppose that $\det(zI_n - A)$ splits to linear factors in $\mathbb{F}[z]$. Then, (1.14.6) implies that $(-1)^i a_i = \sigma_i(\mathbf{z})$. □

Corollary 3.1.2 *Let $A \in \mathbb{F}^{n \times n}$ and assume that $\det(zI_n - A) = \prod_{i=1}^n (z - z_i)$. Then*

$$\operatorname{tr} A := \sum_{i=1}^n a_{ii} = \sum_{i=1}^n z_i, \quad \det A = \prod_{i=1}^n z_i.$$

Definition 3.1.3 *Let $\mathrm{GL}(n, \mathbb{F}) \subset \mathbb{F}^{n \times n}$ denote the set (group) of all $n \times n$ invertible matrices with entries in a given field $\mathbb{F}$. Two matrices $A, B \in \mathbb{F}^{n \times n}$ are called* similar*, and this is denoted by $A \sim B$, if $B = PAP^{-1}$, for some $P \in \mathrm{GL}(n, \mathbb{F})$. The set of all $B \in \mathbb{F}^{n \times n}$ similar to a fixed $A \in \mathbb{F}^{n \times n}$ is called the similarity class corresponding to A, or simply a similarity class.*

The following proposition is straightforward:

Proposition 3.1.4 *Let $\mathbb{F}$ be a field. Then, the similarity relation on $\mathbb{F}^{n \times n}$ is an equivalence relation. Furthermore, if $B = UAU^{-1}$ then*

1. *For any integer $m \geq 2$, $B^m = UA^mU^{-1}$.*
2. *If A is invertible, then B is invertible and $B^{-1} = UA^{-1}U^{-1}$.*

Corollary 3.1.5 *Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$. Assume that $T : \mathbf{V} \to \mathbf{V}$ is a linear transformation. Then, the set of all representation matrices of T in different bases is a similarity class. (Use (1.17.2) and (1.17.3), where $m = n$, $\mathbf{x}_i = \mathbf{y}_i, i = 1, \dots, n, X = Y$.) Hence, the characteristic polynomial of T is defined as $\det(zI_n - A) = z^n + \sum_{i=0}^{n-1} a_i z^i$, where A is the representation matrix of T in any basis $[\mathbf{u}_1, \dots, \mathbf{u}_n]$, and this definition is independent of the choice of a basis. In particular, $\det T := \det A$, and $\operatorname{tr} T^m = \operatorname{tr} A^m$, for any non-negative integer. (T^0 is the identity operator, i.e. $T^0(\mathbf{v}) = \mathbf{v}$, for all $\mathbf{v} \in \mathbf{V}$, and $A^0 = I$.)*

An element $\mathbf{v} \in \mathbf{V}$ is called an *eigenvector* of T corresponding to the *eigenvalue* $\lambda \in \mathbb{F}$, if $\mathbf{v} \neq \mathbf{0}$ and $T(\mathbf{v}) = \lambda \mathbf{v}$. This is equivalent to the existence $\mathbf{0} \neq \mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \lambda \mathbf{x}$. Hence, $(\lambda I - A)\mathbf{x} = \mathbf{0}$ which implies that $\det(\lambda I - A) = 0$. Thus, λ is the zero of the characteristic polynomial of A and T. The assumption λ is a zero of the characteristic polynomial yields that the system $(\lambda I - A)\mathbf{x}$ has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$.

Corollary 3.1.6 *Let $A \in \mathbb{F}^{n \times n}$. Then, λ is an eigenvalue of A if and only if λ is a zero of the characteristic polynomial of A $\det(zI - A)$. Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$. Assume that $T : \mathbf{V} \to \mathbf{V}$ is a linear transformation. Then, λ is an eigenvalue of T if and only if λ is a zero of the characteristic polynomial of T.*

Example 3.1.7 *Here, we first show that 5 is an eigenvalue of* $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \in \mathbb{R}^{2\times 2}$. *We must show that there is a non-zero vector* $\mathbf{x} \in \mathbb{R}^2$ *such that* $A\mathbf{x} = 5\mathbf{x}$. *Clearly, this is equivalent to the equation* $(A - 5I)\mathbf{x} = 0$. *Then, we need to compute the null space of the matrix* $A - 5I = \begin{bmatrix} -4 & 2 \\ 4 & -2 \end{bmatrix}$.
Since the rows (or columns) of $A - 5I$ *are clearly linearly dependent, using the fundamental theorem of invertible matrices,* $N(A) > 0$. *Thus,* $A\mathbf{x} = 5\mathbf{x}$ *has a nontrivial solution and this tells us 5 is an eigenvalue of* A. *Next, we find its eigenvectors by computing the null space.*

$$\left[\begin{array}{c|c} A - 5I & 0 \end{array}\right] = \left[\begin{array}{cc|c} -4 & 2 & 0 \\ 4 & -2 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{array}\right].$$

Therefore, if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ *is an eigenvector corresponding to 5, then* $x_1 - \frac{1}{2}x_2 = 0$, *or* $x_1 = \frac{1}{2}x_2$. *Thus, the eigenvectors are of the form* $\mathbf{x} = \begin{bmatrix} \frac{1}{2}x_2 \\ x_2 \end{bmatrix}$, *equivalently the non-zero multiples of* $\begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$.
Note that the set of all eigenvectors corresponding to an eigenvalue λ *of an* $n \times n$ *matrix* A *is the set of non-zero vectors in the null space of* $A - \lambda I$. *This means that the set of eigenvectors of* A *together with the zero vector in* $\mathbb{F}^n$ *is the null space of* $A - \lambda I$.

Definition 3.1.8 *Assume that* A *is an* $n \times n$ *matrix and* λ *is an eigenvalue of* A. *The collection of all eigenvectors corresponding to* λ, *together with the zero vector, is called the eigenspace of* λ *and in denoted by* $\mathbf{E}_\lambda$. *Then, in the above example,* $\mathbf{E}_5 = \left\{ t \begin{bmatrix} 1 \\ 2 \end{bmatrix}; t \in \mathbb{R} \right\}$.

Moreover, for any $m \in \mathbb{N}$ and $\lambda \in \mathbb{F}$, we define E_λ^m as follows:

$$E_\lambda^m = \{\mathbf{x} \in \mathbb{F}; (A - \lambda I)^m \mathbf{x} = 0\}.$$

It can be shown that $E_\lambda^m \neq 0$ for some m if and only if λ is an eigenvalue of A, and $E_\lambda^m \cap E_\mu^m \neq \{0\}$, for some $m, n > 0$ yields $\lambda = \mu$. Furthermore, if $p(\mathbf{x}) = \prod_\lambda (\mathbf{x} - \lambda)^{n_\lambda}$ is the characteristic polynomial of A, then $E_\lambda^m \subseteq E_\lambda^{n_\lambda}$ for all m, $\dim E_\lambda^{n_\lambda} = n_\lambda$ and $\mathbb{F}^n = \bigoplus_\lambda E_\lambda^{n_\lambda}$.
The proof is left as an exercise.

Definition 3.1.9 *A linear operator* $T : \mathbf{V} \to \mathbf{V}$ *is said diagonalizable if it admits a diagonal matrix representation with respect to some basis of* $\mathbf{V}$, *i.e. there is a basis* β *of* $\mathbf{V}$ *such that the matrix* $[T]_\beta$ *is diagonal.*

Note that a basis of $\mathbb{F}^n$ consisting of eigenvectors of A is called an *eigenbasis* for A.
Diagonal matrices are possess a very simple structure and they allow for a very fast computation of determinants and inverses, for instance. Here, we will have a closer look at how to transform matrices into diagonal form. More specifically, we will look at $T \in L(\mathbf{V})$ of finite-dimensional vector spaces, which are similar to a diagonal matrix.

Proposition 3.1.10 *Let* $\mathbf{V}$ *be* n*-dimensional vector space over* $\mathbb{F}$*. Assume that* $T : \mathbf{V} \to \mathbf{V}$ *is a linear transformation. Then, there exists a basis in* V *such that* T *is represented in this basis by a diagonal matrix*

$$\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) := \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix},$$

(T *is diagonalizable) if and only if the characteristic polynomial of* T *is* $(z-\lambda_1)(z-\lambda_2)\ldots(z-\lambda_n)$*, and* $\mathbf{V}$ *has a basis consisting of eigenvectors of* T*.*

Equivalently, $A \in \mathbb{F}^{n\times n}$ *is similar to a diagonal matrix* $\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ *if and only if* $\det(zI - A) = (z-\lambda_1)(z-\lambda_2)\ldots(z-\lambda_n)$*, and* A *has* n*-linearly independent eigenvectors.*

Proof. Assume that there exists a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ in V such that T is represented in this basis by a diagonal matrix $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then, the characteristic polynomial of T is $\det(zI - \Lambda) = \prod_{i=1}^{n}(z-\lambda_i)$. From the definition of the representation matrix of T, $T(\mathbf{u}_i) = \lambda_i\mathbf{u}_i$, for $i = 1, \ldots, n$. Since each $\mathbf{u}_i \neq \mathbf{0}$, we deduce that each $\mathbf{u}_i$ is an eigenvector of T. By our assumption $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is a basis in $\mathbf{V}$.

Conversely, assume now that $\mathbf{V}$ has a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ consisting eigenvectors of T. Thus, $T(\mathbf{u}_i) = \lambda_i\mathbf{u}_i$ for $i = 1, \ldots, n$. Hence, Λ is the representation matrix of T in the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.

To prove the corresponding results for $A \in \mathbb{F}^{n\times n}$, let $\mathbf{V} := \mathbb{F}^n$ and define the linear operator $T(\mathbf{x}) := A\mathbf{x}$, for all $\mathbf{x} \in \mathbb{F}^n$. □

Lemma 3.1.11 *Let* $A \in \mathbb{F}^{n\times n}$ *and assume that* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *are* k *eigenvectors corresponding to* k *distinct eigenvalues* $\lambda_1, \ldots, \lambda_k$*, respectively. Then,* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *are linearly independent.*

Proof. We prove by induction on k. For $k = 1$, $\mathbf{x}_1 \neq \mathbf{0}$. Hence, hence $\mathbf{x}_1$ is linearly independent. Assume that the lemma holds for $k = m - 1$. Suppose that $k = m$. Assume that $\sum_{i=1}^{m} a_i\mathbf{x}_i = \mathbf{0}$. Then

$$\mathbf{0} = A\mathbf{0} = A\sum_{i=1}^{m} a_i\mathbf{x}_i = \sum_{i=1}^{m} a_iA\mathbf{x}_i = \sum_{i=1}^{m} a_i\lambda_i\mathbf{x}_i.$$

Multiply the equality $\sum_{i=1}^{m} a_i\mathbf{x}_i = \mathbf{0}$ by λ_m and subtract it from the above inequality to deduce that $\sum_{i=1}^{m-1} a_i(\lambda_i - \lambda_m)\mathbf{x}_i = \mathbf{0}$. Since $\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}$ are linearly independent, by the induction hypothesis, we deduce that $a_i(\lambda_i - \lambda_m) = 0$, for $i = 1, \ldots, m-1$. As $\lambda_i - \lambda_m \neq 0$, for $i < m$, we get that $a_i = 0$, for $i = 1, \ldots, m-1$. The assumption that $\sum_{i=1}^{m} a_i\mathbf{x}_i = \mathbf{0}$ yields that $a_m\mathbf{x}_m = \mathbf{0}$. Since $\mathbf{x}_m \neq \mathbf{0}$, we obtain that $a_m = 0$. Hence, $a_1 = \ldots = a_m = 0$. □

Theorem 3.1.12 *Let* $\mathbf{V}$ *be an* n*-dimensional vector space over* $\mathbb{F}$*. Assume that* $T : \mathbf{V} \to \mathbf{V}$ *is a linear transformation. Assume that the characteristic polynomial of* T, $p(z)$ *has* n *distinct roots over* $\mathbb{F}$*, i.e.* $p(z) = \prod_{i=1}^{n}(z-\lambda_i)$ *where* $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$*,* *and $\lambda_i \neq \lambda_j$ for each $i \neq j$. Then, there exists a basis in $\mathbf{V}$ in which T is represented by a diagonal matrix.*

Similarly, let $A \in \mathbb{F}^{n\times n}$ and assume that $\det(zI - A)$ has n distinct roots in $\mathbb{F}$. Then, A is similar to a diagonal matrix.

Proof. It is enough to consider the case of the linear transformation T. Recall that each root of the characteristic polynomial of T is an eigenvalue of T (Corollary 3.1.6). Hence, each λ_i corresponds an eigenvector $\mathbf{u}_i$: $T(\mathbf{u}_i) = \lambda_i \mathbf{u}_i$. Then, the proof of the theorem follows from Lemma 3.1.11 and Proposition 3.1.10. □

If $A \in \mathbb{F}^{n\times n}$, it may happen that $\det(zI - A)$ does not have n roots in $\mathbb{F}$. (See for example Problem 3.2.2-2.) Hence, we cannot *diagonalize* A, i.e. A is not similar to a diagonal matrix. If $\mathbb{F}$ is *algebraically closed*, i.e. any $\det(zI - A)$ has n roots in $\mathbb{F}$. We can apply Proposition 3.1.10 in general and Theorem 3.1.12 in particular to see if A is diagonalizable.

Since $\mathbb{R}$ is not algebraically closed and $\mathbb{C}$ is, that is the reason that we sometimes view a real valued matrix $A \in \mathbb{R}^{n\times n}$ as a complex valued matrix $A \in \mathbb{C}^{n\times n}$. (Again see Problem 3.2.2-2)

Corollary 3.1.13 *Let $A \in \mathbb{C}^{n\times n}$ be nondiagonalizable. Then, its characteristic polynomial must have a multiple root.*

Diagonal matrices $A = P^{-1}BP$ exhibit the nice properties that they can be easily raised to a power:

$$B^k = \left(PAP^{-1}\right)^k = PA^kP^{-1}.$$

Computing A^k is easy because we apply this operation individually to any diagonal element. As an example, this allows to compute inverses of A by performing fewer flops.

Definition 3.1.14 *Let $\sim$ be an equivalence relation on the set X. A subset $A \subseteq X$ is said to be a set of canonical form for $\sim$ if for every $x \in X$, there is exactly one $a \in A$ such that $x \sim a$.*

Example 3.1.15 *We have already seen that row equivalence is an equivalence relation on $\mathbb{F}^{m\times n}$. The subset of reduced row echelon form matrices is a set of canonical form for row equivalence as every matrix is row equivalent to a unique matrix in row echelon form.*

Remark 3.1.16 *Recall that $A, B \in \mathbb{F}^{n\times n}$ are called to be similar if there exists an invertible matrix P such that*

$$A = P^{-1}BP.$$

Similarly is an equivalence relation on $\mathbb{F}^{n\times n}$. As we have seen, two matrices are similar if and only if they represent the same linear operators. Hence, similarity is important to study the structure of linear operators. As we mentioned at the beginning, this chapter is devoted to developing canonical forms for similarity.

Definition 3.1.17

1. *Let k be a positive integer and $\lambda \in \mathbb{F}$. Then, $J_k(\lambda) \in \mathbb{F}^{k\times k}$ is a $k \times k$ upper triangular matrix, with λ on the main diagonal, 1 on the next sub-diagonal and other entries are equal to 0 for $k > 1$:*

$$J_k(\lambda) := \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix},$$

and $J_k(\lambda)$ is called a Jordan block associated with the scalar λ.

2. *If A_{ij} are matrices of the appropriate sizes, then by block matrix*

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ \vdots & \vdots & & \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix},$$

we mean the matrix whose left submatrix is A_{11}, and so on. Thus, the A_{ij}'s are submatrices of A and not entries of A.

3. *Let $A_i \in \mathbb{F}^{n_i \times n_i}$, for $i = 1, \dots, l$. Denote by*

$$A = \oplus_{i=1}^{k} A_i = A_1 \oplus A_2 \oplus \ldots \oplus A_k = \operatorname{diag}(A_1, A_2, \ldots, A_k) :=$$
$$\begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & A_k \end{bmatrix} \in \mathbb{F}^{n\times n}, \quad n = n_1 + n_2 + \ldots + n_k,$$

Here, A is said to be a block diagonal matrix, whose blocks are $A_1, A_2, \ldots, A_k$.

Theorem 3.1.18 *(**The Jordan Canonical Form**) Let $A \in \mathbb{C}^{n\times n}$, ($A \in \mathbb{F}^{n\times n}$, where $\mathbb{F}$ is an algebraically closed field.) Then, A is similar to its Jordan canonical form $\oplus_{i=1}^{k} J_{n_i}(\lambda_i)$ for some $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$, ($\lambda_1, \ldots, \lambda_k \in \mathbb{F}$), and positive integers $n_1, \ldots, n_k$. The Jordan canonical form is unique up to the permutations of the Jordan blocks $J_{n_1}(\lambda_1), \ldots, J_{n_k}(\lambda_k)$.*

Equivalently, let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation of an n-dimensional space over $\mathbb{C}$, or any other algebraically closed field. Then, there exists a basis in $\mathbf{V}$, such that $\oplus_{i=1}^{k} J_{n_i}(\lambda_i)$ is the representation matrix of T in this basis. The blocks $J_{n_i}(\lambda_i), i = 1, \ldots, k$ are unique.

Note that $A \in \mathbb{C}^{n\times n}$ is diagonalizable if and only in its Jordan canonical form $k = n$, i.e. $n_1 = \ldots = n_n = 1$. For $k < n$, the *Jordan canonical form* is the simplest form of the similarity class of a nondiagonalizable $A \in \mathbb{C}^{n\times n}$.

We will prove Theorem 3.1.18 in the next several sections.

Example 3.1.19 *Consider the matrix*

$$A = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 3 & 3 \\ -1 & -2 & -2 \end{bmatrix},$$

we see that $\det(\mathbf{x}I - A) = (\mathbf{x} - 3)^3$*. For* $\lambda = 1$*, we have* $A - I = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ -1 & -2 & -3 \end{bmatrix}$ *which has rank 1. Thus,* $\dim N(A - I) = 2$*. Now,* $(A - I)^2 = 0$*, so* $\dim N\left((A - I)^2\right) = 3$*. Next,* $\dim N(A - I) = 2$ *tells us the JCF of* A *has 2 blocks with eigenvalue 1. The* $\dim N\left((A - I)^2\right) - \dim N(A - I) = 1$ *condition tells us one of these blocks has size at least 2, and so the other has size 1. Thus JCF*$(A) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

The following is an algorithm to compute the Jordan canonical form of $A \in \mathbb{F}^{n \times n}$.

Algorithm:

1. Compute and factor the characteristic polynomial of A.

2. Let m_λ be the maximal integer $k \leq n_\lambda$ such that $E_\lambda^k / E_\lambda^{k-1}$ is not a zero subspace.

3. For each λ, compute a basis $\mathcal{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ for $E_\lambda^{m_\lambda} / E_\lambda^{m_\lambda - 1}$ and lift to elements of $E_\lambda^{m_\lambda}$. Add the elements $(A - \lambda I)^m \mathbf{x}_i$ to $\mathcal{A}$, for $1 \leq m < m_\lambda$.

4. Set $i = m_\lambda - 1$.

5. Compute $\mathcal{A} \cap E_\lambda^i$ to a basis for $E_\lambda^i / E_\lambda^{i-1}$. Add the element $(A - \lambda I)^m \mathbf{x}$ to $\mathcal{A}$ for all m and $\mathbf{x} \in \mathcal{A}$.

6. If $i \geq 1$, set $i = i - 1$, and return to the previous step.

7. Output $\mathcal{A}$, the matrix for A with respect to a suitable ordering of $\mathcal{A}$ is in Jordan Canonical Form.

Proof of correctness. To verify that this algorithm works, we need to check that it is always possible to complete $\mathcal{A} \cap E_\lambda^k$ to a basis for $E_\lambda^k / E_\lambda^{k-1}$. Suppose $\mathcal{A} \cap E_\lambda^k$ is linearly dependent. Then, there are $\mathbf{x}_1, \ldots, \mathbf{x}_s \in \mathcal{A} \cap E_\lambda^k$ with $\sum_i c_i \mathbf{x}_i = 0$, and not all $c_i = 0$. By the construction of $\mathcal{A}$, we know that $\mathbf{x}_i = (A - \lambda I)\mathbf{y}_i$, for some $\mathbf{y}_i \in \mathcal{A}$, so consider $\mathbf{y} = \sum_i c_i \mathbf{y}_i$. Then $\mathbf{y} \neq 0$, since the $\mathbf{y}_i$'s are linearly independent, and not all c_i are zero. Indeed, by the construction of the $\mathbf{y}_i$'s, we know $\mathbf{y} \notin E_\lambda^k$. But $(A - \lambda I)\mathbf{w} = 0$, so $\mathbf{y} \in E_\lambda^1$, which is a contradiction, since $k \geq 1$.

Example 3.1.20 *Consider the matrix*

$$A = \begin{bmatrix} 2 & -4 & 2 & 2 \\ -2 & 0 & 1 & 3 \\ -2 & -2 & 3 & 3 \\ -2 & -6 & 3 & 7 \end{bmatrix}.$$

Then, the characteristic polynomial of A is $(x-2)^2(x-4)^2$. A basis for E_2^1 is $\{(2,1,0,2)^\top,(0,1,2,0)^\top\}$, so since there is a two-dimensional 2 blocks, each of size one. To confirm this, check that $E_2^m = E_2^1$ for all $m > 1$. A basis for E_4^1 is $\{(0,1,1,1)^\top\}$, while a basis for E_4^2 is $\{(0,1,1,1)^\top,(1,0,0,1)^\top\}$, so we can take $\{(1,0,0,1)^\top\}$ as a basis for E_4^2/E_4^1. Then $(A-4I)(1,0,0,1)^\top = (0,1,1,1)^\top$, so our basis is then $\{(2,1,0,2)^\top,(0,1,2,0)^\top,(0,1,1,1)^\top,(1,0,01)^\top\}$. The matrix of the transformation with respect to this basis is:

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

3.2 An application of diagonalizable matrices

The sequence of *Fibonacci numbers* F_n

$$0,1,1,2,3,5,8,13,21,34,\ldots$$

is defined recursively as follows:
One begins with $F_0 = 0$ and $F_1 = 1$. After that every number is defined to be sum of its two predecessors:

$$F_n = F_{n-1} + F_{n-2}, \quad \text{for } n > 1.$$

Johannes Kepler (1571-1630) observed that the ratio of consecutive Fibonacci numbers converges to the *golden ratio*.

Theorem 3.2.1 (Kepler) $\lim \frac{F_{n+1}}{F_n} = \frac{1+\sqrt{5}}{2}$.

The purpose of this section is using of linear algebra tools to prove Kepler's theorem. In order to this, we will need to find an explicit formula for Fibonacci numbers. Particularly, we need the following lemma. See also subsection 4.1.1.

Lemma 3.2.2 *For $n > 1$, $F_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}}$.*

Proof. Let T be the linear operator on $\mathbb{R}^2$ represented by the matrix

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

with respect to the standard basis of $\mathbb{R}^2$. Then, for the vector $\mathbf{v}_k$ whose coordinates are two consecutive Fibonacci numbers $(F_k, F_{k-1})^\top$, we have that $T(\mathbf{v}_k) = A\begin{bmatrix} F_k \\ F_{k-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} F_k \\ F_{k-1} \end{bmatrix} = \begin{bmatrix} F_k + F_{k-1} \\ F_k \end{bmatrix} = \begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix} = \mathbf{v}_{k+1}$. Therefore, $\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = A^n \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and this equation helps us to calculate the powers A^n of A by using the diagonalization. We start with finding the eigenvalues of A which are

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \text{ and } \lambda_2 = \frac{1-\sqrt{5}}{2},$$

and we conclude that A is diagonalizable as its eigenvalues are real and distinct. Next, we find the eigenvectors corresponding to λ_1 and λ_2. We solve the equations $A\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \lambda_1 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$, $A\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \lambda_2 \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ and it is obtained that $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} \lambda_2 \\ 1 \end{bmatrix}$. Therefore, the matrix of change of basis between standard basis and the basis of eigenvectors is $p = \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix}$ and so $p^{-1}Ap = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. This means $A^n = p\begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix}p^{-1}$. Totally, we have

$$\begin{aligned}
\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} &= A^n \begin{bmatrix} 1 \\ 0 \end{bmatrix} = p\begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} p^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \frac{1}{\lambda_1 - \lambda_2} \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} \begin{bmatrix} 1 & -\lambda_2 \\ -1 & \lambda_1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \frac{1}{\lambda_1 - \lambda_2} \begin{bmatrix} \lambda_1^{n+1} - \lambda_2^{n+1} \\ \lambda_1^n - \lambda_2^n \end{bmatrix}.
\end{aligned}$$

Equating the entries of the vectors in the last formula we obtain

$$F_n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2} = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}},$$

as claimed. □

We now are ready to prove Kepler's theorem by using the above lemma.

$$\begin{aligned}
\lim_{n\to\infty} \frac{F_{n+1}}{F_n} &= \lim_{n\to\infty} \frac{(1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1}}{2^{n+1}\sqrt{5}} \frac{2^n\sqrt{5}}{(1+\sqrt{5})^n - (1-\sqrt{5})^n} \\
&= \frac{1}{2} \lim_{n\to\infty} \frac{(1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1}}{(1+\sqrt{5})^n - (1-\sqrt{5})^n} \\
&= \frac{1}{2} \lim_{n\to\infty} \frac{(1+\sqrt{5}) - \left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n (1-\sqrt{5})}{1 - \left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n} = \frac{1+\sqrt{5}}{2}.
\end{aligned}$$

Note that

$$\lim_{n\to\infty} \left(\frac{1-\sqrt{5}}{1+\sqrt{5}}\right)^n = 0$$

because

$$\left|\frac{1-\sqrt{5}}{1+\sqrt{5}}\right| < 1.$$

3.2.1 Worked-out Problems

1. Show that the Jordan block $J_k(\lambda) \in \mathbb{F}^{n\times n}$ is similar to $J_k(\lambda)^\top$.
   Solution:

Consider the $k \times k$ permutation matrix $P = \begin{bmatrix} 0 & & \cdots & 1 \\ 0 & & 1 & 0 \\ & \iddots & \vdots & \\ 1 & & \cdots & 0 \end{bmatrix}$.
Then, $P^{-1} J_k(\lambda) P = J_k(\lambda)^\top$ and so $J_k(\lambda) \sim J_k(\lambda)^\top$.

2. Assume that the characteristic polynomial of A splits in $\mathbb{F}$. Show that A is similar to $A^\top$.
Solution:
Denote by $JCF(A)$, a Jordan canonical form of A. Using Theorem 3.1.18, A is similar to $JCF(A) = \bigoplus\limits_i J_{k_i}(\lambda_i)$. Clearly, $A^\top \sim \bigoplus\limits_i J_{k_i}(\lambda_i)^\top$. Using the previous problem, we conclude that $\bigoplus\limits_i J_{k_i}(\lambda_i) \sim \bigoplus\limits_i J_{k_i}(\lambda_i)^\top$ and this yields $A \sim A^\top$.

3.2.2 Problems

1. Let $\mathbf{V}$ be a vector space over $\mathbb{F}$. (You may assume that $\mathbb{F} = \mathbb{C}$). Let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation. Suppose that $\mathbf{u}_i$ is an eigenvector of T with the corresponding eigenvalue λ_i, for $i = 1, \ldots, m$. Show by induction on m that if $\lambda_1, \ldots, \lambda_m$ are m distinct scalars, then $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are linearly independent.

2. Let $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$.

   (a) Show that A is not diagonalizable over the real numbers $\mathbb{R}$.

   (b) Show that A is diagonalizable over the complex numbers $\mathbb{C}$.

   (c) Find $U \in \mathbb{C}^{2\times 2}$ and a diagonal $\Lambda \in \mathbb{C}^{2\times 2}$ such that $A = U \Lambda U^{-1}$.

3. Let $A = \oplus_{i=1}^k J_{n_i}(\lambda_i)$. Show that $\det(zI - A) = \prod_{i=1}^k (z - \lambda_i)^{n_i}$. (You may use the fact that the determinant of an upper triangular matrix is the product of its diagonal entries.)

4. Let $A = \oplus_{i=1}^k A_i$, where $A_i \in \mathbb{C}^{n_i \times n_i}, i = 1, \ldots, k$. Show that $\det(zI_n - A) = \prod_{i=1}^k \det(zI_{n_i} - A_i)$.
(**Hint:** First show the identity for $k = 2$ using the determinant expansion by rows. Then use induction for $k > 2$.)

5. (a) Show that any eigenvector of $J_n(\lambda) \in \mathbb{C}^{n\times n}$ is in the subspace spanned by $\mathbf{e}_1$. Conclude that $J_n(\lambda)$ is not diagonalizable unless $n = 1$.

   (b) What is the rank of $zI_n - J_n(\lambda)$ for a fixed $\lambda \in \mathbb{C}$ and for each $z \in \mathbb{C}$?

   (c) What is the rank of $zI - \oplus_{i=1}^k J_{n_i}(\lambda_i)$ for fixed $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$ and for each $z \in \mathbb{C}$?

6. Let $A \in \mathbb{C}^{n\times n}$ and assume that $\det(zI_n - A) = z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n$ has n distinct complex roots. Show that $A^n + a_1 A^{n-1} + \ldots a_{n-1} A + a_n I_n = \mathbf{0}$, where $\mathbf{0} \in \mathbb{C}^{n\times n}$ denotes the zero matrix, i.e. the matrix whose all entries are 0. (This is a special case of the Cayley-Hamilton theorem, which states that the above identity holds for *any* $A \in \mathbb{C}^{n\times n}$.)
(**Hint:** Use the fact that A is diagonalizable.)

7. If $A, B \in \mathbb{F}^{n \times n}$, show that AB and BA have the same characteristic polynomial.

8. Let $1 \le m \le n \in \mathbb{N}$ and suppose that $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{n \times m}$. Prove that $\det(zI - BA) = z^{n-m} \det(zI - AB)$.
(In particular, if $m = n$, then the characteristic polynomials of AB and BA are the same, see the previous problem.)

9. If $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, show that all of its eigenvalues are real.

10. If $A \in \mathbb{F}^{n \times n}$ is an invertible matrix and $P_A(z)$ and $P_{A^{-1}}(z)$ denote the characteristic polynomial of A and A^{-1}, respectively, show that $P_{A^{-1}}(z) = \frac{z^n}{P_A(0)} P_A(\frac{1}{z})$.

11. If $A \in \mathbb{R}^{n \times n}$ with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, show that $\mathbb{R}^n = \bigoplus_{i=1}^{n} E_{\lambda_i}$.

12. If $A \in \mathbb{F}^{n \times n}$ is a diagonalizable matrix with the characteristic polynomial $p_A(z) = \prod_{i=1}^{k} (z - \lambda_i)^{d_i}$ and $V = \{B \in \mathbb{F}^{n \times n}; AB = BA\}$, prove that $\dim V = \sum_{i=1}^{k} d_i^2$.

13. Prove that a matrix is invertible if and only if it does not have a zero eigenvalue. (See Theorem 1.9.2.)

14. Let A be a square matrix with the block form $A = \begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix}$, where X and Z are square matrices. Show that $\det A = \det X \det Z$.

15. Let $A \in \mathbb{F}^{m \times n}$.

    (a) Prove that A in its reduced row echelon form must have the block form

    $$J_i = \begin{bmatrix} I_i & 0 \\ 0 & 0 \end{bmatrix}.$$

    (b) Show that A is equivalent to exactly one matrix of the form J_i.

    (c) Conclude that the set of these matrices is a set of canonical form for equivalence, (See Lemma 1.7.16.)

16. Assume that A is a doubly stochastic matrix. Show that for each eigenvalue λ of A, $\lambda \le 1$.

17. Show that the matrix $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \in \mathbb{R}^{2 \times 2}$ has complex eigenvalues if θ is not a multiple of π. Give the geometric interpretation of it. (See also Problem 1.10.2-4.)

18. Let $A \in \mathbb{F}^{n \times n}$. Prove that

    (a) if λ is an eigenvalue of A, then λ^k is an eigenvalue of A^k, for any $k \in \mathbb{N}$.

    (b) if λ is an eigenvalue of A and A is invertible, then λ^{-1} is an eigenvalue of A^{-1}.

3.3 Matrix polynomials

Let $P(z) = [p_{ij}(z)]_{i=j=1}^{m,n}$ be an $m \times n$ matrix whose entries are polynomials in $\mathbb{F}[z]$. The set of all such $m \times n$ matrices is denoted by $\mathbb{F}[z]^{m\times n}$. Clearly, $\mathbb{F}[z]^{m\times n}$ is a vector space over $\mathbb{F}$, of infinite dimension. Given $p(z) \in \mathbb{F}[z]$ and $P(z) \in \mathbb{F}[z]^{m\times n}$, one can define $p(z)P(z) := [p(z)p_{ij}(z)] \in \mathbb{F}[z]^{m\times n}$. Again, this product satisfies nice distribution properties. Thus, $\mathbb{F}[z]^{m\times n}$ is a *module* over the ring $\mathbb{F}[z]$. (Note $\mathbb{F}[z]$ is not a field!)

Let $P(z) = [p_{ij}(z)] \in \mathbb{F}[z]^{m\times n}$. Then, $\deg P(z) := \max_{i,j} \deg p_{ij}(z) = l$. Write

$$p_{ij}(z) = \sum_{k=0}^{l} p_{ij,k} z^{l-k},\ P_k := [p_{ij,k}]_{i=j=1}^{m,n} \in \mathbb{F}^{m\times n} \text{ for } k = 0, \dots, l.$$

Then

$$P(z) = P_0 z^l + P_1 z^{l-1} + \dots + P_l, \quad P_i \in \mathbb{F}^{m\times n},\ i = 0, \dots, l, \tag{3.3.1}$$

is a *matrix polynomial* with coefficients in $\mathbb{F}^{m\times n}$.

Assume that $P(z), Q(z) \in \mathbb{F}[z]^{n\times n}$. Then, we can define $P(z)Q(z) \in \mathbb{F}[z]$. Note that in general $P(z)Q(z) \neq Q(z)P(z)$. Hence, $\mathbb{F}[z]^{n\times n}$ is a *noncommutative* ring. For $P(z) \in \mathbb{F}^{n\times n}$ of the form (3.3.1) and any $A \in \mathbb{F}^{n\times n}$, we define

$$P(A) = \sum_{i=0}^{l} P_i A^{l-i} = P_0 A^l + P_1 A^{l-1} + \dots + P_l, \text{ where } A^0 = I_n.$$

Given two polynomials $p, q \in \mathbb{F}[z]$, one can divide p by $q \not\equiv 0$ with the residue r, i.e. $p = tq + r$, for some unique $t, r \in \mathbb{F}[z]$, where $\deg r < \deg q$. One can trivially generalize that to polynomial matrices:

Proposition 3.3.1 *Let $p(z), q(z) \in \mathbb{F}[z]$ and assume that $q(z) \not\equiv 0$. Let $p(z) = t(z)q(z) + r(z)$, where $t(z), r(z) \in \mathbb{F}[z]$ are unique polynomials with $\deg r(z) < \deg q(z)$. Let $n > 1$ be an integer, and define the following scalar polynomials: $P(z) := p(z)I_n, Q(z) := q(z)I_n, T(z) := t(z)I_n, R(z) := r(z)I_n \in \mathbb{F}[z]^{n\times n}$. Then, $P(A) = T(A)Q(A) + R(A)$, for any $A \in \mathbb{F}^{n\times n}$.*

Proof. Since $A^i A^j = A^{i+j}$, for any non-negative integer, with $A^0 = I_n$, the equality $P(A) = T(A)Q(A) + R(A)$ follows trivially from the equality $p(z) = t(z)q(z) + r(z)$. □

Recall that p is divisible by q, denoted as $q|p$, if $p = tq$, i.e. r is the zero polynomial. Note that if $q(z) = (z - a)$, then $p(z) = t(z)(z - a) + p(a)$. Thus, $(z - a)|p$ if and only if $p(a) = 0$. Similar results hold for square polynomial matrices, which are not scalar.

Lemma 3.3.2 *Let $P(z) \in \mathbb{F}[z]^{n\times n}, A \in \mathbb{F}^{n\times n}$. Then, there exists a unique $T_{left}(z)$, of degree degP $- 1$ if $\deg P > 0$ or degree $-\infty$ if $\deg P \le 0$, such that*

$$P(z) = T_{left}(z)(zI - A) + P(A). \tag{3.3.2}$$

In particular, $P(z)$ is divisible from the right by $zI - A$ if and only if $P(A) = 0$.

Proof. We prove the lemma by induction on $\deg P$. If $\deg P \le 0$, i.e. $P(z) = P_0 \in \mathbb{F}^{n\times n}$, then $T_{left} = \mathbf{0}, P(A) = P_0$ and the lemma trivially holds. Suppose that the lemma holds for all P with $\deg P \le l-1$, where $l \ge 1$. Let $P(z)$ be of degree $l \ge 1$ of the form (3.3.1). Then $P(z) = P_0 z^l + \tilde{P}(z)$, where $\tilde{P}(z) = \sum_{i=1}^{l} P_i z^{l-1}$. By the induction assumption $\tilde{P}(z) = \tilde{T}_{left}(z)(zI_n - A) + \tilde{P}(A)$, where $\tilde{T}_{left}(z)$ is unique. A straightforward calculation shows that

$$P_0 z^l = \hat{T}_{left}(z)(zI_n - A) + P_0 A^l, \text{ where } \hat{T}_{left}(z) = \sum_{i=0}^{l-1} P_0 A^i z^{l-i-1},$$

and $\hat{T}_{left}$ is unique. Hence, $T_{left}(z) = \hat{T}_{left}(z) + \tilde{T}_{left}$ is unique, $P(A) = P_0 A^l + \tilde{P}(A)$ and (3.3.2) follows.

Suppose that $P(A) = \mathbf{0}$. Then $P(z) = T_{left}(z)(zI - A)$, i.e. $P(z)$ is divisible by $zI_n - A$ from the right. Assume that $P(z)$ is divisible by $(zI_n - A)$ from the right, i.e. there exists $T(z) \in \mathbb{F}[z]^{n\times n}$ such that $P(z) = T(z)(zI_n - A)$. Subtract (3.3.2) from $P(z) = T(z)(zI_n - A)$ to deduce that $\mathbf{0} = (T(z) - T_{left}(z))(zI_n - A) - P(A)$. Hence, $T(z) = T_{left}(z)$ and $P(A) = 0$. □

The above lemma can be generalized to any $Q(z) = Q_0 z^l + Q_1 z^{l-1} + \ldots + Q_l \in \mathbb{F}[z]$, where $Q_0 \in \mathrm{GL}(n,\mathbb{F})$; there exists unique $T_{left}(z), R_{left}(z) \in \mathbb{F}[z]$ such that

$$P(z) = T_{left}(z)Q(z) + R_{left}(z),\ \deg R_{left} < \deg Q,\ Q(z) = \sum_{i=0}^{l} Q_i z^{l-i},\ Q_0 \in \mathrm{GL}(n,\mathbb{F}). \tag{3.3.3}$$

Here we agree that $(Az^i)(Bz^j) = (AB)z^{i+j}$, for any $A, B \in \mathbb{F}^{n\times n}$ and non-negative integers i, j.

Theorem 3.3.3 (Cayley-Hamilton theorem) *Let $A \in \mathbb{F}^{n\times n}$ and $p(z) = \det(zI_n - A)$ be the characteristic polynomial of A. Let $P(z) = p(z)I_n \in \mathbb{F}[z]^{n\times n}$. Then, $P(A) = \mathbf{0}$.*

Proof. Let $A(z) = zI_n - A$. Fix $z \in \mathbb{F}$ and let $B(z) = [b_{ij}(z)]$ be the adjoint matrix of $A(z)$, whose entries are the cofactors of $A(z)$. That is $b_{ij}(z)$ is $(-1)^{i+j}$ times the determinant of the matrix obtained from $A(z)$ by deleting row j and column i. If one views z as indeterminate, then $B(z) \in \mathbb{F}[z]^{n\times n}$. Consider the identity

$$A(z)B(z) = B(z)A(z) = \det A(z) I_n = p(z)I_n = P(z).$$

Hence, $(zI_n - A)$ divides $P(z)$ from the right. Lemma 3.3.2 yields that $P(A) = \mathbf{0}$. □

Note that the Cayley-Hamilton theorem holds even if $\mathbb{F}$ is not algebraically closed by doing a field extension.

For $p, q \in \mathbb{F}[z]$, let (p, q) denote the *greatest common divisor* of p, q. If p and q are identically zero, then (p, q) is the zero polynomial. Otherwise (p, q) is a polynomial s of the highest degree that divides p and q. Also s is determined up to a multiple of a non-zero scalar and it can be chosen as a unique *monic* polynomial:

$$s(z) = z^l + s_1 z^{l-1} + \ldots + s_l \in \mathbb{F}[z]. \tag{3.3.4}$$

Equality (1.14.3) yields.

Corollary 3.3.4 *Let $p, q \in \mathbb{F}[z]$ be coprime. Then, there exist $u, v \in \mathbb{F}[z]$ such that $1 = up + vq$. Let $n > 1$ be an integer and define $P(z) := p(z)I_n, Q(z) := q(z)I_n, U(z) := u(z)I_n, V(z) := u(z)I_n \in \mathbb{F}[z]^{n\times n}$. Then, for any $A \in \mathbb{F}^{n\times n}$, we have the identity $I_n = U(A)P(A) + V(A)Q(A)$, where $U(A)P(A) = P(A)U(A)$ and $V(A)Q(A) = Q(A)V(A)$.*

Let us consider that case where p and $q \in \mathbb{F}[z]$ are both non-zero polynomials that split (to linear factors) over $\mathbb{F}$. Thus

$$p(z) = p_0(z-\alpha_1)\dots(z-\alpha_i), p_0 \neq 0, \quad q(z) = q_0(z-\beta_1)\dots(z-\beta_j), q_0 \neq 0.$$

In that case $(p, q) = 1$, if p and q do not have a common root. If p and q have a common zero, then (p, q) is a non-zero polynomial that has the maximal number of common roots of p and q counting with multiplicities.

From now on, for any $p \in \mathbb{F}[z]$ and $A \in \mathbb{F}^{n\times n}$ we identify $p(A)$ with $P(A)$, where $P(z) = p(z)I_n$.

Definition 3.3.5 *For an eigenvalue λ, its algebraic multiplicity is the multiplicity of λ as a root of the characteristic polynomial and it is denoted by $m_a(\lambda)$. The geometric multiplicity of λ is the maximal number of linearly independent eigenvectors corresponding to it and it is denoted by $m_g(\lambda)$. Also λ is called an algebraically simple eigenvalue if its algebraic multiplicity is one.*

Example 3.3.6 *Consider the matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Let us find its eigenvalues and eigenvectors. Characteristic polynomial is $(1-x)^2$. So $\lambda = 1$ is a double root. Eigenvectors corresponding to this eigenvalue:*

$$\begin{bmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

So $x_2 = 0$, and x_1 can be anything. There is only one linearly independent vector: $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
Consider another example: the identity matrix,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then, by the same way, $\lambda = 1$ is a double root of the characteristic polynomial and the only eigenvalue. But any non-zero vector can serve as its eigenvector, because for any $x \in \mathbb{R}^2$ we have: $Ax = x = 1\cdot x$. So, it has two linearly independent eigenvector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
Here, for both matrices $\lambda = 1$ is the only eigenvalue with algebraic multiplicity two. But its geometric multiplicity is one for $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and two for $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Theorem 3.3.7 *Let $A \in \mathbb{F}^{n\times n}$ and λ is an eigenvalue of A. Then $m_g(\lambda) \leq m_a(\lambda)$.*

Proof. Set $m = m_g(\lambda)$ and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a basis for E_λ. Extend it to a basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for $\mathbb{F}^n$. Let $P \in \mathbb{F}^{n\times n}$ be the invertible matrix with column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $B = P^{-1}AP$. Then, for each $1 \le j \le m$, we have

$$Be_j = P^{-1}APe_j = P^{-1}A\mathbf{v}_j = P^{-1}\lambda\mathbf{v}_j = \lambda(P^{-1}\mathbf{v}_j) = \lambda e_j,$$

and so B has the block form

$$B = \begin{bmatrix} \lambda I_m & X \\ 0 & Y \end{bmatrix}$$

Using Problem 3.2.2-13 and the fact that similar matrices have the same characteristic polynomials, we have

$$\begin{aligned} P_A(t) = P_B(t) &= \det\begin{bmatrix} (t-\lambda)I_m & -X \\ 0 & tI_{n-m} - Y \end{bmatrix} = \det((t-\lambda)I_m)\det(tI_{n-m} - Y) \\ &= (t-\lambda)^m P_Y(t). \end{aligned}$$

This shows that λ appears at least m times as a root of $P_A(t)$, and then $m_a(t) \ge m$.

3.4 Minimal polynomial and decomposition to invariant subspaces

Recall that $\mathbb{F}^{n\times n}$ is a vector space over $\mathbb{F}$ of dimension n^2. Let $A \in \mathbb{F}^{n\times n}$ and consider the powers $A^0 = I_n, A, A^2, \ldots, A^m$. Let m be the smallest positive integer such that these $m+1$ matrices are linearly dependent as vectors in $\mathbb{F}^{n\times n}$. (*Note* that $A^0 \ne \mathbf{0}$.) Thus, there exists $(b_0, \ldots, b_m)^\top \in \mathbb{F}^n$ for which $(b_0, \ldots, b_m)^\top \ne 0$ and $\sum_{i=0}^m b_i A^{m-i} = 0$. If $b_0 = 0$, then $A^0, \ldots, A^{m-1}$ are linearly dependent, which contradicts the definition of m. Hence, $b_0 \ne 0$. Divide the linear dependence by b_0 to obtain $\sum_{i=0}^m a_i A^{m-i} = 0$, where $a_i = \frac{b_i}{b_0}$, $0 \le i \le m$.

Define $\psi(z) = z^m + \sum_{i=0}^m a_i z^{m-i} \in \mathbb{F}[z]$. Then, $\psi(A) = 0$. Here, ψ is called the *minimal polynomial* of A. In principle $m \le n^2$, but in fact $m \le n$.

Theorem 3.4.1 *Let $A \in \mathbb{F}^{n\times n}$ and $\psi(z)$ be its minimal polynomial. Assume that $p(z) \in \mathbb{F}[z]$ is an annihilating polynomial of A, i.e. $p(A) = 0_{n\times n}$. Then, ψ divides p. In particular, the characteristic polynomial $p(z) = \det(zI_n - A)$ is divisible by $\psi(z)$. Hence,* $\deg\psi \le \deg p = n$.

Proof. Divide the annihilating polynomial p by ψ to obtain $p(z) = t(z)\psi(z) + r(z)$, where $\deg r < \deg\psi = m$. Proposition 3.3.1 yields that $p(A) = t(A)\psi(A) + r(A)$, which implies that $r(A) = 0$. Assume that $l = \deg r(z) \ge 0$, i.e. r is not identically the zero polynomial. Therefore, $A^0, \ldots, A^l$ are linearly dependent, which contradicts the definition of m. Hence $r(z) \equiv 0$.

The Cayley-Hamilton theorem yields that the characteristic polynomial $p(z)$ of A annihilates A. Hence, $\psi | p$ and $\deg\psi \le \deg p = n$. □

Comparing the minimal polynomial of an algebraic element in a field extension with the minimal polynomial of a matrix, we can see their common features. For example, if $\mathbb{E}/\mathbb{F}$ is a field extension and we denote the minimal polynomial of an algebraic element $\alpha \in \mathbb{E}$ by $p(z)$ and if we denote the minimal polynomial of $A \in \mathbb{F}^{n\times n}$

by $\psi(z)$, then both $p(z)$ and $\psi(z)$ are annihilator for α and A, respectively and both are non-zero polynomials of minimum degree with this character. According to the definition, $p(z)$ is irreducible over $\mathbb{F}$. Now, this question can be asked that if $\psi(z)$ is necessarily irreducible over $\mathbb{F}$, too? We will answer to this question in the last Worked-out Problem!

Theorem 3.4.2 *Let $A \in \mathbb{F}^{n\times n}$. Denote by $q(z)$ the g.c.d (monic polynomial), of all the entries of* adj $(zI_n - A) \in \mathbb{F}[z]^{n\times n}$*, which is the g.c.d of all $(n-1)\times(n-1)$ minors of $(zI_n - A)$. Then, $q(z)$ divides the characteristic polynomial $p(z) = \det(zI_n - A)$ of A. Furthermore, the minimal polynomial of A is equal to $\frac{p(z)}{q(z)}$.*

Proof. Expand $\det(zI_n - A)$ by the first column. Since $q(z)$ divides each $(n-1)\times(n-1)$ minor of $zI_n - A$, we deduce that $q(z)$ divides $p(z)$. Let $\phi(z) = \frac{p(z)}{q(z)}$. As $q(z)$ divides each entry of adj $(zI_n - A)$, we deduce that adj $(zI_n - A) = q(z)C(z)$, for some $C(z) \in \mathbb{F}[z]^{n\times n}$. Divide the equality $p(z)I_n =$ adj $(zI_n - A)(zI_n - A)$ by $q(z)$ to deduce that $\phi(z)I_n = C(z)(zI_n - A)$. Lemma 3.3.2 yields that $\phi(A) = 0$.

Let $\psi(z)$ be the minimal polynomial of A. Theorem 3.4.1 yields that ψ divides ϕ. We now show that ϕ divides ψ. Theorem 3.4.1 implies that $p(z) = s(z)\psi(z)$, for some monic polynomial $s(z)$. Since $\psi(A) = 0$, Lemma 3.3.2 yields that $\psi(z)I_n = D(z)(zI_n - A)$, for some $D(z) \in \mathbb{F}^{n\times n}[z]$. Thus, $p(z)I_n = s(z)\psi(z)I_n = s(z)D(z)(zI_n - A)$. Since $p(z)I_n =$ adj $(zI_n - A)(zI_n - A)$, we conclude that $s(z)D(z) =$ adj $(zI_n - A)$. As all the entries of $D(z)$ are polynomials, it follows that $s(z)$ divides all the entries of adj $(zI_n - A)$. Since $q(z)$ is the g.c.d of all entries of adj $(zI_n - A)$, we deduce that $s(z)$ divides $q(z)$. Consider the equality $p(z) = s(z)\psi(z) = q(z)\phi(z)$. Thus, $\psi(z) = \frac{q(z)}{s(z)}\phi(z)$. Hence, $\phi(z)$ divides $\psi(z)$. As $\psi(z)$ and $\phi(z)$ are monic, we deduce that $\psi(z) = \phi(z)$. □

Proposition 3.4.3 *Let $A \in \mathbb{F}^{n\times n}$ and assume that $\lambda \in \mathbb{F}$ is an eigenvalue of A with the corresponding eigenvector $\mathbf{x} \in \mathbb{F}^n$. Then, for any $h(z) \in \mathbb{F}[z]$, $h(A)\mathbf{x} = h(\lambda)\mathbf{x}$. In particular, λ is a root of the minimal polynomial $\psi(z)$ of A, i.e. $\psi(\lambda) = 0$.*

Proof. Clearly, $A^m\mathbf{x} = \lambda^m\mathbf{x}$. Hence, $h(A)\mathbf{x} = h(\lambda)\mathbf{x}$. Assume that $h(A) = 0$. As $\mathbf{x} \neq \mathbf{0}$, we deduce that $h(\lambda) = 0$. Then, $\psi(\lambda) = 0$. □

Combining Theorem 3.4.1 and Proposition 3.4.3, we conclude that the minimal polynomial and the characteristic polynomial of a matrix have the same roots. (probably with the different multiplicities).

Definition 3.4.4 *A matrix $A \in \mathbb{F}^{n\times n}$ is called nonderogatory if the minimal polynomial of A is equal to its characteristic polynomial.*

Definition 3.4.5 *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$, and assume that $\mathbf{V}_1, \ldots, \mathbf{V}_i$ are non-zero subspaces of $\mathbf{V}$. Then, $\mathbf{V}$ is a direct sum of $\mathbf{V}_1, \ldots, \mathbf{V}_i$, denoted as $\mathbf{V} = \oplus_{j=1}^{i}\mathbf{V}_j$ if any vector $\mathbf{v} \in \mathbf{V}$ has a unique representation as $\mathbf{v} = \mathbf{v}_1 + \ldots + \mathbf{v}_i$, where $\mathbf{v}_j \in \mathbf{V}_j$, for $j = 1, \ldots, i$. Equivalently, let $[\mathbf{v}_{j1}, \ldots, \mathbf{v}_{jl_j}]$ be a basis of $\mathbf{V}_j$, for $j = 1, \ldots, i$. Then, $\dim\mathbf{V} = \sum_{j=1}^{i}\dim\mathbf{V}_j = \sum_{j=1}^{i} l_j$ and the $\dim\mathbf{V}$ vectors $\mathbf{v}_{11}, \ldots, \mathbf{v}_{1l_1}, \ldots, \mathbf{v}_{i1}, \ldots, \mathbf{v}_{il_i}$ are linearly independent.*

Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator. A subspace $\mathbf{U}$ of $\mathbf{V}$ is called a *T-invariant subspace*, or simply an invariant subspace when there is no ambiguity about T, if $T(\mathbf{u}) \in \mathbf{U}$, for each $\mathbf{u} \in \mathbf{U}$. We denote this fact by $T\mathbf{U} \subseteq \mathbf{U}$. Denote by $T|_{\mathbf{U}}$ the restriction of T to the invariant subspace of T. Clearly, $T|_{\mathbf{U}}$ is a linear operator on $\mathbf{U}$.

Note that $\mathbf{V}$ and the zero subspace $\{\mathbf{0}\}$ are invariant subspaces. Those are called *trivial* invariant subspaces. The subspace $\mathbf{U}$ is called a *nontrivial* invariant subspace if it is an invariant subspace such that $0 < \dim \mathbf{U} < \dim \mathbf{V}$.

Since the representation matrices of T in different bases form a similarity class, we can define the *minimal polynomial* $\psi(z) \in \mathbb{F}[z]$ of T, as the minimal polynomial of any representation matrix of T. Equivalently, $\psi(z)$ is the monic polynomial of the minimal degree which annihilates T; $\psi(T) = \mathbf{0}$.

Theorem 3.4.6 (**Primary decomposition Theorem**) *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator on a finite non-zero dimensional vector space $\mathbf{V}$. Let $\psi(z)$ be the minimal polynomial of T. Assume that $\psi(z)$ decomposes to $\psi(z) = \psi_1(z)\dots\psi_k(z)$, where each $\psi_i(z)$ is a monic polynomial of degree at least 1. Suppose furthermore that for each pair $i \neq j$, $\psi_i(z)$ and $\psi_j(z)$ are coprime. Then, $\mathbf{V}$ is a direct sum of $\mathbf{V}_1, \dots, \mathbf{V}_k$, where each $\mathbf{V}_i$ is a nontrivial invariant subspace of T. Furthermore, the minimal polynomial of $T|_{\mathbf{V}_i}$ is equal to $\psi_i(z)$, for $i = 1, \dots, k$. Moreover, each $\mathbf{V}_i$ is uniquely determined by $\psi_i(z)$ for $i = 1, \dots, k$.*

Proof. We prove the theorem by induction on $k \geq 2$. Let $k = 2$. Then, $\psi(z) = \psi_1(z)\psi_2(z)$. Let $\mathbf{V}_1 := \psi_2(T)\mathbf{V}$ and $\mathbf{V}_2 = \psi_1(T)\mathbf{V}$ be the images of the operators $\psi_2(T), \psi_1(T)$, respectively. Observe that

$$T\mathbf{V}_1 = T(\psi_2(T)\mathbf{V}) = (T\psi_2(T))\mathbf{V} = (\psi_2(T)T)\mathbf{V} = \psi_2(T)(T\mathbf{V}) \subseteq \psi_2(T)\mathbf{V} = \mathbf{V}_1.$$

Thus, $\mathbf{V}_1$ is a T-invariant subspace. Assume that $\mathbf{V}_1 = \{\mathbf{0}\}$. This is equivalent to $\psi_2(T) = 0$. By Theorem 3.4.1, ψ divides ψ_2 which is impossible since $\deg \psi = \deg \psi_1 + \deg \psi_2 > \deg \psi_1$. Thus, $\dim \mathbf{V}_1 > 0$. Similarly, $\mathbf{V}_2$ is a non-zero T-invariant subspace. Let $T_i = T|_{\mathbf{V}_i}$, for $i = 1, 2$. Clearly

$$\psi_1(T_1)\mathbf{V}_1 = \psi_1(T)\mathbf{V}_1 = \psi_1(T)(\psi_2(T)\mathbf{V}) = (\psi_1(T)\psi_2(T))\mathbf{V} = \{\mathbf{0}\},$$

since ψ is the minimal polynomial of T. Hence, $\psi_1(T_1) = \mathbf{0}$, i.e. ψ_1 is an annihilating polynomial of T_1. Similarly, $\psi_2(T_2) = \mathbf{0}$.

Let $\mathbf{U} = \mathbf{V}_1 \cap \mathbf{V}_2$. Then, $\mathbf{U}$ is an invariant subspace of T. We claim that $\mathbf{U} = \{\mathbf{0}\}$, i.e. $\dim \mathbf{U} = 0$. Assume to the contrary that $\dim \mathbf{U} \geq 1$. Let $Q := T|_{\mathbf{U}}$ and denote by $\phi \in \mathbb{F}[z]$ the minimal polynomial of Q. Clearly, $\deg \phi \geq 1$. Since $\mathbf{U} \subseteq \mathbf{V}_i$, it follows that ψ_i is an annihilating polynomial of Q for $i = 1, 2$. Hence, $\phi | \psi_1$ and $\phi | \psi_2$, i.e. ϕ is a nontrivial factor of ψ_1 and ψ_2. This contradicts the assumption that ψ_1 and ψ_2 are coprime. Hence, $\mathbf{V}_1 \cap \mathbf{V}_2 = \{\mathbf{0}\}$.

Since $(\psi_1, \psi_2) = 1$, there exist polynomials $f, g \in \mathbb{F}[z]$ such that $\psi_1 f + \psi_2 g = 1$. Hence, $I = \psi_1(T)f(T) + \psi_2(T)g(T)$, where I is the identity operator $I\mathbf{v} = \mathbf{v}$ on $\mathbf{V}$. In particular, for any $\mathbf{v} \in \mathbf{V}$ we have $\mathbf{v} = \mathbf{v}_2 + \mathbf{v}_1$, where $\mathbf{v}_1 = \psi_2(T)(g(T)\mathbf{v}) \in \mathbf{V}_1, \mathbf{v}_2 = \psi_1(T)(f(T)\mathbf{v}) \in \mathbf{V}_2$. Since $\mathbf{V}_1 \cap \mathbf{V}_2 = \{\mathbf{0}\}$, it follows that $\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2$. Let $\tilde{\psi}_i$ be the minimal polynomial of T_i. Then, $\tilde{\psi}_i | \psi_i$ for $i = 1, 2$. Hence, $\tilde{\psi}_1\tilde{\psi}_2 | \psi_1\psi_2$. Let $\mathbf{v} \in \mathbf{V}$. Therefore, $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_i \in \mathbf{V}_i, i = 1, 2$. Using the facts that

$\tilde{\psi}_1(T)\tilde{\psi}_2(T) = \tilde{\psi}_2(T)\tilde{\psi}_1(T)$, $\tilde{\psi}_i$ is the minimal polynomial of T_i, and the definition of T_i, we deduce

$$\tilde{\psi}_1(T)\tilde{\psi}_2(T)\mathbf{v} = \tilde{\psi}_2(T)\tilde{\psi}_1(T)\mathbf{v}_1 + \tilde{\psi}_1(T)\tilde{\psi}_2(T)\mathbf{v}_2 = \mathbf{0}.$$

Hence, the monic polynomial $\theta(z) := \tilde{\psi}_1(z)\tilde{\psi}_2(z)$ is an annihilating polynomial of T. Thus, $\psi(z)|\theta(z)$ which implies that $\psi(z) = \theta(z)$, hence $\tilde{\psi}_i = \tilde{\psi}$ for $i = 1, 2$.

It is left to show that $\mathbf{V}_1$ and $\mathbf{V}_2$ are unique. Let $\bar{\mathbf{V}}_i := \{\mathbf{v} \in \mathbf{V} : \psi_i(T)\mathbf{v} = \mathbf{0}\}$, for $i = 1, 2$. Therefore, $\bar{\mathbf{V}}_i$ is a subspace that contains $\mathbf{V}_i$, for $i = 1, 2$. If $\psi_i(T)\mathbf{v} = \mathbf{0}$, then

$$\psi_i(T)(T(\mathbf{v})) = (\psi_i(T)T)\mathbf{v} = (T\psi_i(T)\mathbf{v}) = T(\psi_i(T)\mathbf{v}) = T\mathbf{0} = \mathbf{0}.$$

Hence $\bar{\mathbf{V}}_i$ is T-invariant subspace. We claim that $\bar{\mathbf{V}}_i = \mathbf{V}_i$. Suppose to the contrary that $\dim \bar{\mathbf{V}}_i > \dim \mathbf{V}_i$, for some $i \in \{1, 2\}$. Let $j \in \{1, 2\}$ and $j \neq i$. Then, $\dim(\bar{\mathbf{V}}_i \cap \mathbf{V}_j) \geq 0$. As before we conclude that $\mathbf{U} := \bar{\mathbf{V}}_i \cap \mathbf{V}_j$ is T-invariant subspace. As above, the minimal polynomial of $T|_{\mathbf{U}}$ must divide $\psi_1(z)$ and $\psi_2(z)$, which contradicts the assumption that $(\psi_1, \psi_2) = 1$. This concludes the proof of the theorem for $k = 2$.

Assume that $k \geq 3$. Let $\hat{\psi}_2 := \psi_2 \ldots \psi_k$. Then, $(\psi_1, \hat{\psi}_2) = 1$ and $\psi = \psi_1\hat{\psi}_2$. Then $\mathbf{V} = \mathbf{V}_1 \oplus \hat{\mathbf{V}}_2$, where $T : \mathbf{V}_1 \to \mathbf{V}_1$ has the minimal polynomial ψ_1, and $T : \hat{\mathbf{V}}_2 \to \hat{\mathbf{V}}_2$ has the minimal polynomial $\hat{\psi}_2$. Note that $\mathbf{V}_1$ and $\hat{\mathbf{V}}_2$ are unique. Apply the induction hypothesis for $T|_{\hat{V}_2}$ to deduce the theorem. □

3.4.1 Worked-out Problems

1. Let $A \in \mathbb{F}^{n\times n}$. Show that A and $A^\top$ have the same minimal polynomial.
   Solution:
   If $\psi(z)$ denotes the minimal polynomial of A, since $\psi(A^\top) = \psi(A)^\top$, then $\psi(z)$ is the minimal polynomial of $A^\top$ as well.

2. Let $A \in \mathbb{C}^{n\times n}$ and $A^2 = -A$.

   (a) List all possible eigenvalues of A.

   (b) Show that A is diagonalizable.

   (c) Assume that $B^2 = -A$. Is B diagonalizable?

   Solution:

   (a) If λ is an eigenvalue of A, then $Au = \lambda u$, for some non-zero $u \in \mathbb{C}^n$. Then, $A^2u = \lambda^2 u = -Au = -\lambda u$ and so $(\lambda^2 + \lambda)u = 0$. This means $\lambda(\lambda + 1) = 0$ and then $\lambda_1 = 0$ and $\lambda_2 = -1$ are possible eigenvalues.

   (b) Clearly, $p(z) = z^2 + z$ is an annihilating polynomial of A. Using Theorem 3.3.3, its minimal polynomial is either $z^2 + z$ or z or $z + 1$. Each of them has simple roots. Hence, by the next Worked-out Problem, A is diagonalizable.

   (c) No. Obviously $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ satisfies $A^2 = -A = 0$ and $B^2 = -A = 0$, where $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. But since $P_B(z) = \det(zI - B) = z^2$ and $\psi_B(z)|P_B(z)$, then $\psi_B(z)$ is either z or z^2 and clearly $\psi_B(B) = 0$ if $\psi_B(z) = z$. Then, $\psi_B(z) = z^2$ and by the next worked-out problem B is not diagonalizable.

3. If $A \in \mathbb{F}^{n\times n}$, show that A is diagonalizable over $\mathbb{F}$ if and only if $\det(zI - A)$ splits to linear factors over $\mathbb{F}$ and its minimal polynomial has simple roots.
Solution:
First, assume that $\psi_A(z) = (z - \lambda_1)\cdots(z - \lambda_m)$ is the minimal polynomial of A. Then, λ_i's are the eigenvalues of A which are distinct. For any eigenvalue λ_i, let $\mathbf{E}_{\lambda_i} = \{\mathbf{v} \in \mathbb{F}^n;\ A\mathbf{v} = \lambda_i\mathbf{v}\}$ be the corresponding eigenspaces. We will show that $\mathbb{F}^n = \bigoplus_{i=1}^{m}\mathbf{E}_{\lambda_i}$. According to Worked-out Problem 3.4.1-5, the eigenvectors with different eigenvalues are linearly independent, so it suffices to show $\mathbb{F}^n = \overset{m}{\underset{i=1}{\oplus}}\mathbf{E}_{\lambda_i}$, as the sum will then be direct by linear independence. Now, we want to find polynomials $h_1(z), \ldots, h_m(z)$ in $\mathbb{F}[z]$ such that

$$1 = \sum_{i=1}^{m} h_i(z),\ h_i(z) \equiv 0 \ \Big(\mod \frac{\psi_A}{z - \lambda_i}\Big). \tag{3.4.1}$$

The congruence condition implies the polynomial $(z - \lambda_i)h_i(z)$ is divisible by $\psi_A(z)$. If we substitute the matrix A for z in (3.4.1) and apply both sides to any vector $\mathbf{v} \in \mathbb{F}^n$, we get $\mathbf{v} = \sum_{i=1}^{m} h_i(A)(\mathbf{v})$,

$$(A - \lambda_i I)h_i(A)(\mathbf{v}) = 0. \tag{3.4.2}$$

The second equation says that $h_i(A)(\mathbf{v})$ lies in $\mathbf{E}_{\lambda_i}$ and the first equation says $\mathbf{v}$ is a sum of such eigenvectors, so $\mathbb{F}^n = \bigoplus_{i=1}^{m}\mathbf{E}_{\lambda_i}$. Then, using bases from each $\mathbf{E}_{\lambda_i}$ provides an eigenbasis for A. Now we find $h_i(z)$'s fitting (3.4.2). For $1 < i < m$, let $f_i(z) = \psi_A(z)/(z - \lambda_i) = \prod_{j\neq i}(z - \lambda_i)$. Since λ_i's are distinct, the polynomials $f_1(z), \ldots, f_m(z)$ are relatively prime as an m-tuple, so some $\mathbb{F}[z]$-linear combination of them equals 3.4.1:

$$1 = \sum_{i=1}^{m} g_i(z)f_i(z), \tag{3.4.3}$$

where $g_i(z) \in \mathbb{F}[z]$. Let $h_i(z) = g_i(z)f_i(z)$. By now, we have proved $\mathbb{F}^n = \bigoplus_{i=1}^{m}\mathbf{E}_{\lambda_i}$. Thus, $\mathbb{F}^n$ has a basis of eigenvectors for A and by Proposition 3.1.10, A is diagonalizable.
Now, assume that A is diagonalizable, so all eigenvalues of A are in $\mathbb{F}$ and $\mathbb{F}^n$ is the direct sum of the eigenvectors for A. We want to show the minimal polynomial of A in $\mathbb{F}[z]$ splits and has distinct roots. Let $\lambda_1, \ldots, \lambda_m$ be the different eigenvalues of A, so $\mathbf{V} = \bigoplus_{i=1}^{m}\mathbf{E}_{\lambda_i}$. We will show that $f(z) = \prod_{i=1}^{m}(z - \lambda_i) \in \mathbb{F}[z]$ is the minimal polynomial of A in $F[z]$. By hypothesis, the eigenvectors of A span $\mathbb{F}^n$. Let $\mathbf{v}$ be an eigenvector, say $A\mathbf{v} = \lambda\mathbf{v}$. Then, $A - \lambda$ annihilates $\mathbf{v}$. The matrices $(A - \lambda_i I)$'s commute with each other and one of them annihilates $\mathbf{v}$, so their product $f(A)$ annihilates $\mathbf{v}$. Thus, $f(A)$ annihilates the span of the eigenvectors, which is $\mathbb{F}^n$, so $f(A) = 0$. The minimal polynomial is therefore a factor $f(z)$. At the same time, each root of $f(z)$ is an eigenvalue of A and so is a root of the minimal polynomial of A. Since the roots of $f(z)$ each occur once, $f(z)$ must be the minimal polynomial of A.

4. Let $A \in \mathbb{C}^{n\times n}$ be of finite order, i.e. $A^m = I_n$, for some positive integer m. Then, A is diagonalizable.
   Solution:
   Since $A^m = I_n$, A is annihilated by $z^m - 1$. Then, the minimal polynomial of A is a factor of $z^m - 1$. The polynomial $z^m - 1$ has distinct root in $\mathbb{C}$. The previous problem completes the proof.

5. Show that eigenvectors for distinct eigenvalues are linearly independent.
   Solution:
   Assume that $A \in \mathbb{F}^{n\times n}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are eigenvectors of A with distinct eigenvalues $\lambda_1, \ldots, \lambda_m$, we show that $\mathbf{v}_i$'s are linearly independent by induction on m. The case $m = 1$ is obvious. If $m > 1$, suppose $\sum_{i=1}^m c_i\mathbf{v}_i = 0$, for some $c_i \in \mathbb{F}$. Since $A\mathbf{v}_1 = \lambda_i\mathbf{v}_i$ and $A\sum_{i=1}^m c_i\mathbf{v}_i = 0$, then
   $$\sum_{i=1}^m c_i\lambda_i\mathbf{v}_i = 0. \tag{3.4.4}$$
   Multiplying by linear relation $\sum_{i=1}^m c_i\mathbf{v}_i = 0$ by λ_m, we get
   $$\sum_{i=1}^m c_i\lambda_m\mathbf{v}_i = 0. \tag{3.4.5}$$
   Subtracting (3.4.5) from (3.4.4), we get $\sum_{i=1}^{m-1} c_i(\lambda_i - \lambda_m)\mathbf{v}_i = 0$. By induction hypothesis, $c_i(\lambda_i - \lambda_m) = 0$, for $i = 1, \ldots, m-1$. Since $\lambda_1, \ldots, \lambda_{m-1}, \lambda_m$ are distinct, $\lambda_i - \lambda_m \neq 0$, for $i = 1, \ldots, m-1$. Then, $c_i = 0$, for $i = 1, \ldots, m-1$. Thus the linear relation $\sum_{i=1}^m c_i\mathbf{v}_i = 0$ becomes $c_m\mathbf{v}_m = 0$. The vector $\mathbf{v}_m$ is non-zero, then $c_m = 0$.

6. Show that the minimal polynomial of a matrix $A \in \mathbb{F}^{n\times n}$ is not necessarily irreducible over $\mathbb{F}$.
   Solution:
   If $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$, then $\psi_A(z) = z^2$ which is reducible over $\mathbb{R}$.

3.4.2 Problems

1. Let $A, B \in \mathbb{F}^{n\times n}$ and $p(z) \in \mathbb{F}[z]$. Show the following statements:
   (a) If $B = UAU^{-1}$, for some $U \in \mathrm{GL}(n, \mathbb{F})$, then $p(B) = Up(A)U^{-1}$.
   (b) If $A \sim B$, then A and B have the same minimal polynomial.
   (c) Let $A\mathbf{x} = \lambda\mathbf{x}$. Then, $p(A)\mathbf{x} = p(\lambda)\mathbf{x}$. Deduce that each eigenvalue of A is a root of the minimal polynomial of A.
   (d) Assume that A has n distinct eigenvalues. Then, A is nonderogatory.

2. (a) Show that the Jordan block $J_k(\lambda) \in \mathbb{F}^{k\times k}$ is nonderogatory.
   (b) Let $\lambda_1, \ldots, \lambda_k \in \mathbb{F}$ be k distinct elements. Let
   $$A = \oplus_{i=j=1}^{k,l_i} J_{m_{ij}}(\lambda_i), \text{ where } m_i = m_{i1} \geq \ldots \geq m_{il_i} \geq 1, \text{ for } i = 1, \ldots, k. \tag{3.4.6}$$
   Here m_{ij} and l_i are positive integers be integers. Find the minimal polynomial of A. When A is nonderogatory?

3. Find the characteristic and the minimal polynomials of

$$C := \begin{bmatrix} 2 & 2 & -2 & 4 \\ -4 & -3 & 4 & -6 \\ 1 & 1 & -1 & 2 \\ 2 & 2 & -2 & 4 \end{bmatrix}$$

4. Let $A := \begin{bmatrix} x & y \\ u & v \end{bmatrix}$. Then, A is a point in four dimensional space $\mathbb{R}^4$.

   (a) What is the condition that A has a multiple eigenvalue ($\det(zI_2 - A) = (z - \lambda)^2$) ? Conclude that the set (variety) all 2×2 matrices with a multiple eigenvalue is a quadratic hypersurface in $\mathbb{R}^4$, i.e. it satisfies a polynomial equation in (x, y, u, v) of degree 2. Hence its dimension is 3.

   (b) What is the condition that A has a multiple eigenvalue and it is a diagonalizable matrix, i.e. similar to a diagonal matrix? Show that this is a line in $\mathbb{R}^4$. Hence its dimension is 1.

   (c) Conclude that the set (variety) of 2×2 matrices which have multiple eigenvalues and diagonalizable is "much smaller" than the set of 2×2 matrix with multiple eigenvalue.

   This fact holds for any $n \times n$ matrices in $\mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$.

5. Show that $A \in \mathbb{F}^{n\times n}$ is diagonalizable if and only if $m_g(\lambda) = m_a(\lambda)$, for any eigenvalue λ of A.

6. Let $A \in \mathbb{F}^{n\times n}$ and λ is an eigenvalue of it. Show that the multiplicity of λ is at least the dimension of the eigenspace E_λ.

7. Programming Problem
   Spectrum and pseudo spectrum: Let $A = [a_{ij}]_{i,j=1}^n \in \mathbb{C}^{n\times n}$. Then, $\det(zI_n - A) = (z - \lambda_1)\cdots(z - \lambda_n)$ and the *spectrum* of A is given as spec $A := \{\lambda_1, \ldots, \lambda_n\}$. In computations, the entries of A are known or given up to a certain precision. Say, in regular precision each a_{ij} is known with precision to eight digits: $a_1 \cdot a_2 \cdots a_8 \times 10^m$ for some integer, e.g. $1.2345678 \times 10^{-12}$, in floating point notation. Thus, with a given matrix A, we associate a whole class of matrices $\mathcal{C}(A) \subset \mathbb{C}^{n\times n}$ of matrices $B \in \mathbb{C}^{n\times n}$ that are represented by A. For each $B \in \mathcal{C}(A)$, we have the spectrum spec B. Then, the *pseudo spectrum* of A is the union of all the spectra of $B \in \mathcal{C}(A)$: pspec $A := \cup_{B\in\mathcal{C}(A)}$spec B. Note that spec A and pspec A are subsets of the complex plane $\mathbb{C}$ and can be easily plotted by computer. The shape of pspec A gives an idea of our real knowledge of the spectrum of A, and to changes of the spectrum of A under perturbations.
   The purpose of this programming problems to give the reader a taste of this subject.
   In all the computations use double precision.

   (a) Choose at random $A = [a_{ij}] \in \mathbb{R}^{5\times 5}$ as follows, each entry a_{ij} is chosen at random from the interval $[-1, 1]$, using uniform distribution. Find the spectrum of A and plot the eigenvalues of A on the $X - Y$ axis as complex numbers, marked say as +, where the center of + is at each eigenvalue.

i. For each $\epsilon = 0.1, 0.01, 0.0001, 0.000001$ do the following: For $i = 1, \dots, 100$, choose $B_i \in \mathbb{R}^{5\times 5}$ at random as A in the item (a) and find the spectrum of $A + \epsilon B_i$. Plot these spectra, each eigenvalue of $A + \epsilon B_i$ plotted as on the $X - Y$ axis, together with the plot of the spectrum of A. (Altogether you will have 4 graphs.)

(b) Let $A := \mathrm{diag}[0.1C, [-0.5]]$, i.e. $A \in \mathbb{R}^{5\times 5}$ be a block diagonal matrix where the first 4×4 block is $0.1C$, where the matrix C is given in Problem part (i) of part (a) above with this specific A. (Again you will have 4 graphs.)

(c) Repeat (a) by choosing at random a symmetric matrix $A = [a_{ij}] \in \mathbb{R}^{5\times 5}$. That is choose at random a_{ij} for $1 \le i \le j$, and let $a_{ji} = a_{ij}$ for $i < j$.

i. Repeat the part (i) of (a). (B_j are not symmetric!) You will have 4 graphs.

ii. Repeat part (i) of (a), with the restriction that each B_j is a random symmetric matrix, as explained in (c). You will have 4 graphs.

(d) Can you draw some conclusions about these numerical experiments?

8. Assume that $A, B \in \mathbb{F}^{n\times n}$ are similar and $A = PBP^{-1}$. Let $E_\lambda(A)$ be the λ-eigenspace for A and $E_\lambda(B)$ be the λ-eigenspace for B. Prove that $E_\lambda(B) = P^{-1}E_\lambda(A)$, i.e.

$$E_\lambda(B) = \{P^{-1}x;\ x \in E_\lambda(A)\}.$$

Note that although similar matrices have the same eigenvalues (Worked-out Problem 1.17.1), they do not usually have the same eigenvectors or eigenspaces. The above problem states that there is a precise relation between the eigenspaces of similar matrices.

3.5 Quotient of vector spaces and induced linear operators

We have already defined the concepts of kernel, image, and cokernel of a group homomorphism. We have the similar definitions for a linear transformation. If $\mathbf{V}$ and $\mathbf{W}$ are two vector spaces over the field $\mathbb{F}$ and $T \in L(\mathbf{V}, \mathbf{W})$, then we have: $\ker T = \{\mathbf{x} \in \mathbf{V} : T(\mathbf{x}) = 0\}$, $\mathrm{Im}\, T = \{T(\mathbf{x}) : \mathbf{x} \in \mathbf{V}\}$ and $\mathrm{coker} T = \mathbf{W}/\mathrm{Im}\, T$.

Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Assume that $\mathbf{U}$ is a subspace of $\mathbf{V}$. Then by $\mathbf{V}/\mathbf{U}$ we denote the set of cosets $\mathbf{x}+\mathbf{U} := \{\mathbf{y} \in \mathbf{V},\ \mathbf{y} = \mathbf{x}+\mathbf{u}$ for all $\mathbf{u} \in \mathbf{U}\}$. An element in $\hat{\mathbf{V}} := \mathbf{V}/\mathbf{U}$ is denoted by $\hat{\mathbf{x}} := \mathbf{x} + \mathbf{U}$.

Note that *codimension* of $\mathbf{U}$ is denoted by codim $\mathbf{U}$ and defined as codim $\mathbf{U} = \dim \mathbf{V} - \dim \mathbf{U}$. One of the results in the following lemma is that codim $\mathbf{U} = \dim \mathbf{V}/\mathbf{U}$.

Lemma 3.5.1 *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Assume that $\mathbf{U}$ is a subspace of $\mathbf{V}$. Then, $\mathbf{V}/\mathbf{U}$ is a finite dimensional vectors space over $\mathbb{F}$, where $(\mathbf{x}+\mathbf{U}) + (\mathbf{y}+\mathbf{U}) := (\mathbf{x}+\mathbf{y}+\mathbf{U})$ and $a(\mathbf{x}+\mathbf{U}) := a\mathbf{x}+\mathbf{U}$. Specifically, the neutral element is identified with the coset $\mathbf{0}+\mathbf{U} = \mathbf{U}$. In particular,* $\dim \mathbf{V}/\mathbf{U} =$

$\dim \mathbf{V} - \dim \mathbf{U}$. *More precisely, if* $\mathbf{U}$ *has a basis* $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ *and* $\mathbf{V}$ *has a basis* $\{\mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{u}_{m+1}, \dots, \mathbf{u}_{m+l}\}$, *then* $\hat{\mathbf{V}}$ *has a basis* $\{\hat{\mathbf{u}}_{m+1}, \dots, \hat{\mathbf{u}}_{m+l}\}$. *(Note that if* $\mathbf{U} = \{\mathbf{0}\}$, *then* $\hat{\mathbf{V}} = \mathbf{V}$ *and if* $\mathbf{U} = \mathbf{V}$, *then* $\hat{\mathbf{V}}$ *is the trivial subspace consisting of the zero element.)*

The proof is straightforward and is left as an exercise.

Lemma 3.5.2 *Assume that the assumptions of Lemma 3.5.1 hold. Let* $T \in L(\mathbf{V}) := L(\mathbf{V}, \mathbf{V})$. *Assume furthermore that* $\mathbf{U}$ *is an invariant subspace of* T. *Then,* T *induces a linear operator* $\hat{T} : \hat{\mathbf{V}} \to \hat{\mathbf{V}}$ *by the equality:* $\hat{T}(\mathbf{x} + \mathbf{U}) := T(\mathbf{x}) + \mathbf{U}$, *i.e.* $\hat{T}(\hat{\mathbf{x}}) = \widehat{T(\mathbf{x})}$.

Proof. We first need to show that $\widehat{T(\mathbf{x})}$ is independent of the coset representative. Namely $T(\mathbf{x} + \mathbf{u}) + \mathbf{U} = T(\mathbf{x}) + (T(\mathbf{u}) + \mathbf{U}) = T(\mathbf{x}) + \mathbf{U}$, for each $\mathbf{u} \in \mathbf{U}$. Since $T(\mathbf{u}) \in \mathbf{U}$, we deduce that $T(\mathbf{u}) + \mathbf{U} = \mathbf{U}$. The linearity of $\hat{T}$ follows from the linearity of T. □

3.6 Isomorphism theorems for vector spaces

The notion of isomorphism gives a sense in which several examples of vector spaces that we have seen so far are the same. For example, assume that $\mathbf{V}$ is the vector space of row vectors with two real components and $\mathbf{W}$ is the vector space of column vectors with two real components. Clearly, $\mathbf{V}$ and $\mathbf{W}$ are the same in some sense and we can associate to each row vector $(a\ b)$ the corresponding column vector $\binom{a}{b}$. We see that $T(x\ y) = \binom{x}{y}$ is linear and $R(T) = \mathbf{W}$ and $N(T) = \{0\}$. So T is an isomorphism. The importance of this concept is that any property of $\mathbf{V}$ as a vector space can be translated into an equivalent property of $\mathbf{W}$. For instance, if we have a set of row vectors and want to know if they are linearly independent, it would suffice to answer the same question for the corresponding row vectors.

We already studied the isomorphism theorems for groups in Section 1.1. Here, we investigate the similar theorems for vector spaces.

Theorem 3.6.1 *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *be vector spaces over the field* $\mathbb{F}$ *and* $T \in L(\mathbf{V}, \mathbf{W})$. *If* $\mathbf{V}'$ *and* $\mathbf{W}'$ *are subspaces of* $\mathbf{V}$ *and* $\mathbf{W}$, *respectively, and* $T(\mathbf{V}') \subseteq \mathbf{W}'$, *then* T *induces a linear transformation* $\bar{T} : \mathbf{V}/\mathbf{V}' \to \mathbf{W}/\mathbf{W}'$ *defined by* $T(\mathbf{v} + \mathbf{V}') = T(\mathbf{v}) + \mathbf{W}'$.

Proof. First, we show that $\bar{T}$ is well-defined. If $\mathbf{x} + \mathbf{V}' = \mathbf{y} + \mathbf{V}'$, for $\mathbf{x}, \mathbf{y} \in \mathbf{V}$, then $\mathbf{x} - \mathbf{y} \in \mathbf{V}'$ and so $T(\mathbf{x}) - T(\mathbf{y}) \in T(\mathbf{V}') \subseteq \mathbf{W}'$. Then, $T(\mathbf{x}) + \mathbf{W}' = T(\mathbf{y}) + \mathbf{W}'$, and $\bar{T}(\mathbf{x} + \mathbf{W}) = \bar{T}(\mathbf{y} + \mathbf{W})$. By linearity of T it follows that $\bar{T}$ is also linear. □

The first isomorphism theorem. If $T : \mathbf{V} \to \mathbf{W}$ is a linear operator, then

$$\mathbf{V}/\ker(T) = \mathrm{Im}(T).$$

Proof. Let $\mathbf{V}' = \ker(T)$ and $\mathbf{W}' = \{0\}$. Then, T induces a linear transformation $\bar{T} : \mathbf{V}/\ker T \to \mathrm{Im}(T)$ defined by $\bar{T}(\mathbf{v} + \ker(T)) = T(\mathbf{v})$. If $\mathbf{y} \in \mathrm{Im}(T)$, there exists $\mathbf{x} \in \mathbf{V}$ such that $T(\mathbf{x}) = \mathbf{y}$, and so $\bar{T}(\mathbf{x} + \ker(T)) = \mathbf{y}$. Therefore, $\bar{T}$ is surjective. Now, we check that $\bar{T}$ is injective, for $\mathbf{x} \in \mathbf{V}$, $\mathbf{x} + \ker T \in \ker \bar{T}$ if and only if $T(\mathbf{x}) = 0$, that is, $\mathbf{x} \in \ker T$. Thus, $\bar{T}$ is injective. Using the previous theorem, we conclude that $\bar{T}$ is linear. We are done! □

The first isomorphism theorem implies that up to isomorphism, images of linear operators on $\mathbf{V}$ are the same as quotient space of $\mathbf{V}$. The second and third isomorphism theorems are consequences of the first isomorphism theorem.

The second isomorphism theorem. Let $\mathbf{V}$ be a vector space and let $\mathbf{V}_1$ and $\mathbf{V}_2$ be subspaces of $\mathbf{V}$. Then

$$\frac{\mathbf{V}_1 + \mathbf{V}_2}{\mathbf{V}_2} \cong \frac{\mathbf{V}_1}{\mathbf{V}_1 \cap \mathbf{V}_2}$$

Proof. Let $T : \mathbf{V}_1 + \mathbf{V}_2 \to \frac{\mathbf{V}_1}{\mathbf{V}_1 \cap \mathbf{V}_2}$ be defined by $T(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{v}_1 + (\mathbf{V}_1 \cap \mathbf{V}_2)$. It is easy to show that T is a well-defined surjective linear operator with kernel $\mathbf{V}_2$. An application of the first isomorphism theorem then completes the proof. □

The third isomorphism theorem. Let $\mathbf{V}$ be a vector space, and suppose that $\mathbf{V}_1 \subset \mathbf{V}_2 \subset \mathbf{V}$ are subspaces of $\mathbf{V}$. Then

$$\frac{\mathbf{V}/\mathbf{V}_1}{\mathbf{V}_2/\mathbf{V}_1} \cong \frac{\mathbf{V}}{\mathbf{V}_2}.$$

Proof. Let $T : \mathbf{V}/\mathbf{V}_1 \to \mathbf{V}/\mathbf{V}_2$ be defined by $T(\mathbf{v} + \mathbf{V}_1) = \mathbf{v} + \mathbf{V}_2$. It is easy to show that T is a well-defined surjective linear transformation whose kernel is $\mathbf{V}_2/\mathbf{V}_1$. The rest follows from the first isomorphism theorem. □

Figure 3.1: Kernel and Image of a linear mapping $T : \mathbf{V} \to \mathbf{W}$.

3.7 Existence and uniqueness of the Jordan canonical form

Definition 3.7.1 *A matrix $A \in \mathbb{F}^{n\times n}$ or a linear transformation $T : \mathbf{V} \to \mathbf{V}$ is called nilpotent if $A^m = 0$ or $T^m = 0$, respectively. The minimal $m \geq 1$ for which $A^m = 0$ or $T^m = 0$ is called the index of nilpotency of A and T, respectively and denoted by* index A *or* index T*, respectively.*

Examples

1. *The matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is nilpotent, since $A^2 = 0$.*

2. *Any triangular matrix with zero along the main diagonal is nilpotent. For example, the matrix $A = \begin{bmatrix} 0 & 2 & 1 & 6 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ is nilpotent and its index of nilpotency*

is 4;

$$A^2 = \begin{bmatrix} 0 & 0 & 2 & 7 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \textit{and} \quad A^4 = 0.$$

3. *Though the examples above have a large number of zero entries, a typical nilpotent matrix does not. For example, the matrix* $A = \begin{bmatrix} 5 & -3 & 2 \\ 15 & -9 & 6 \\ 10 & -6 & 4 \end{bmatrix}$ *is nilpotent, though the matrix has no zero entries.*

Assume that A or T are nilpotent, then the s-numbers of A or T are defined as

$$\begin{aligned} s_i(A) &:= \text{rank } A^{i-1} - 2\text{rank } A^i + \text{rank } A^{i+1}, \qquad (3.7.1) \\ s_i(T) &:= \text{rank } T^{i-1} - 2\text{rank } T^i + \text{rank } T^{i+1}, \ i = 1, \ldots \end{aligned}$$

Note that A or T are nilpotent with the index of nilpotency m if and only if z^m is the minimal polynomial of A or T, respectively. Furthermore, if A or T, are nilpotent then the maximal l, for which $s_l > 0$ is equal to the index of nilpotency of A or T, respectively. Moreover, if A is nilpotent, then it is singular but the converse does not satisfy. For example, consider the matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ which is singular but not nilpotent.

Example 3.7.2 *Consider the matrix* $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ *in* $\mathbb{R}^{4\times4}$. *We have*

$$A^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \textit{and} \quad A^3 = 0.$$

Then, A is nilpotent with the index of nilpotency 3. We calculate s-numbers for A;

$$\begin{aligned} s_1(A) &= \text{rank } I - 2\text{rank } A + \text{rank } A^2 = 4 - 2 \times 2 + 1 = 1 \\ s_2(A) &= \text{rank } A - 2\text{rank } A^2 + \text{rank } A^3 = 2 - 2 \times 1 + 0 = 0 \\ s_3(A) &= \text{rank } A^2 - 2\text{rank } A^3 + \text{rank } A^4 = 1 - 2 \times 0 + 0 = 1, \end{aligned}$$

and $s_i(A) = 0$, for $i > 3$.
Therefore, the maximal l, for which $s_l > 0$ is 3 which is equal to the index of nilpotency of A.

Proposition 3.7.3 *Let $T : \mathbf{V} \to \mathbf{V}$ be a nilpotent operator, with the index of nilpotency m, on the finite dimensional vector space $\mathbf{V}$. Then*

$$\text{rank } T^i = \sum_{j=i+1}^{m} (j-i)s_j = (m-i)s_m + (m-i-1)s_{m-1} + \ldots + s_{i+1}, \quad (3.7.2)$$

$i = 0, \ldots, m-1.$

Proof. Since $T^l = \mathbf{0}$, for $l \ge m$, it follows that $s_m(T) = \operatorname{rank} T^{m-1}$ and $s_{m-1} = \operatorname{rank} T^{m-2} - 2\operatorname{rank} T^{m-1}$ if $m > 1$. This proves (3.7.2) for $i = m-1, m-2$. For other values of i, (3.7.2) follows straightforward from (3.7.1) by induction on $m - i \ge 2$. □

Theorem 3.7.4 *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation on a finite dimensional vector space. Assume that T is nilpotent with the index of nilpotency m. Then, $\mathbf{V}$ has a basis of the form*

$$\begin{aligned} &\mathbf{x}_j, T(\mathbf{x}_j), \ldots T^{l_j - 1}(\mathbf{x}_j),\ j = 1, \ldots, i,\ \textit{where } m = l_1 \ge \ldots \ge l_i \ge 1 \\ &\textit{and } T^{l_j}(\mathbf{x}_j) = 0,\ j = 1, \ldots, i. \end{aligned} \tag{3.7.3}$$

More precisely, the number of l_j, which are equal to an integer $l \in [m]$, is equal to $s_l(T)$ given in (3.7.1).

Proof. We first show by induction on the dimension of $\mathbf{V}$ that there exists a basis in $\mathbf{V}$ given by (3.7.3). If $\dim \mathbf{V} = 1$ and T is nilpotent, then any $\mathbf{x} \in \mathbf{V} \setminus \{\mathbf{0}\}$ is basis satisfying (3.7.3). Suppose that for $\dim \mathbf{V} \le N$ any nilpotent T has a basis of the form (3.7.3). Assume now that $\dim \mathbf{V} = N + 1$.

Assume that $T^m = 0$ and $T^{m-1} \ne 0$. Hence, $\operatorname{rank} T^{m-1} = \dim \operatorname{Im} T^{m-1} > 0$. Thus, there exists $\mathbf{0} \ne \mathbf{y}_1 \in \operatorname{Im} T^{m-1}$. Therefore, $T^{m-1}(\mathbf{x}_1) = \mathbf{y}_1$, for some $\mathbf{x}_1 \in \mathbf{V}$. Clearly $T^m(\mathbf{x}_1) = \mathbf{0}$. We claim that $\mathbf{x}_1, T(\mathbf{x}_1), \ldots, T^{m-1}(\mathbf{x}_1)$ are linearly independent. Suppose that

$$\sum_{i=1}^{m} a_i T^{i-1}(\mathbf{x}_i) = \mathbf{0}. \tag{3.7.4}$$

Apply T^{m-1} to this equality to deduce that $\mathbf{0} = a_1 T^{m-1}(\mathbf{x}_1) = a_1 \mathbf{y}_1$. Since $\mathbf{y}_1 \ne \mathbf{0}$, it follows that $a_1 = 0$. Now apply to (3.7.4) T^{m-2} to deduce that $a_2 = 0$. Continue this process to deduce that $a_1 = \ldots = a_m = 0$.

Let $\mathbf{U} = \operatorname{span}\{\mathbf{x}_1, T(\mathbf{x}_1), \ldots, T^{m-1}(\mathbf{x}_1)\}$. Clearly, $T\mathbf{U} \subset \mathbf{U}$. If $\mathbf{U} = \mathbf{V}$, we are done. Now, assume that $m = \dim \mathbf{U} < \dim \mathbf{V}$. We now consider $\hat{\mathbf{V}} := \mathbf{V}/\mathbf{U}$ and the corresponding $\hat{T} : \hat{\mathbf{V}} \to \hat{\mathbf{V}}$. Since $T^m = 0$, it follows that $\hat{T}^m = 0$. Consequently, $\hat{T}$ is nilpotent. Assume that $\hat{T}^{m'} = 0$ and $\hat{T}^{m'-1} \ne 0$. Then, $m \ge m'$. We now apply the induction hypothesis to $\hat{T}$. Thus, $\hat{\mathbf{V}}$ has a basis of the form $\{\hat{\mathbf{x}}_j, \hat{T}(\hat{\mathbf{x}}_j), \ldots \hat{T}^{l_j - 1}(\hat{\mathbf{x}}_j)\}$, for $j = 2, \ldots, i$. Here, $m' = l_2 \ge \ldots \ge l_i \ge 1$. Furthermore, $\hat{T}^{l_j}(\hat{\mathbf{x}}_j) = \mathbf{0}$ in $\hat{\mathbf{V}}$ for $j = 2, \ldots, i$. Assume that $\hat{\mathbf{x}}_j = \mathbf{z}_j + \mathbf{U}$, for $j = 2, \ldots, i$. As $\hat{T}^{l_j}(\hat{\mathbf{x}}_j) = \mathbf{0}$, it follows that $T^{l_j}(\mathbf{z}_j) = \sum_{i=1}^{m} a_i T^{i-1}(\mathbf{x}_1)$. Apply T^{m-l_j} to both sides of this identity. Since $T^m = 0$, we deduce that $\mathbf{0} = \sum_{i=1}^{l_j} a_i T^{m - l_{j-1} + i - 1}(\mathbf{x}_1)$. As $\mathbf{x}_1, \ldots, T^{m-1}(\mathbf{x}_1)$ are linearly independent, we conclude that $a_1 = \ldots = a_{l_j} = 0$. Let $\mathbf{x}_j := \mathbf{z}_j - \sum_{i=1}^{m-l_j} a_{l_j + i} T^{i-1}(\mathbf{x}_1)$. Observe that $T^{l_j}(\mathbf{x}_j) = \mathbf{0}$, for $j = 2, \ldots, i$.

We claim that the vectors in (3.7.3) form a basis in $\mathbf{V}$, as claimed. Assume that there is a linear combination of these vectors that gives zero vector in $\mathbf{V}$:

$$\sum_{j=1}^{i} \sum_{k=1}^{l_j} b_{jk} T^{k-1}(\mathbf{x}_j) = \mathbf{0}. \tag{3.7.5}$$

Consider this linear combination in $\hat{\mathbf{V}}$. As $\hat{\mathbf{x}}_1 = \mathbf{0}$ in $\hat{\mathbf{V}}$, then the above equality reduces to:

$$\sum_{j=2}^{i} \sum_{k=1}^{l_j} b_{jk} \hat{T}^{k-1}(\hat{\mathbf{x}}_j) = \mathbf{0}.$$

The induction hypothesis yield that $b_{jk} = 0$, for each $j \geq 2$ and each corresponding k. Hence, (3.7.5) reduces to $\sum_{k=1}^{m} b_{1k} T^{k-1}(\mathbf{x}_1) = \mathbf{0}$. As $\mathbf{x}_1, \ldots, T^{m-1}(\mathbf{x}_1)$ are linearly independent, we deduce that $b_{1k} = 0$, for $k \in [m]$. Hence the vectors in (3.7.3) are linearly independent. Note that the number of these vectors is $m + \dim \hat{V} = \dim V = N + 1$. Then, the vectors in (3.7.3) form a basis in $\mathbf{V}$.

It is left to show that the number of l_j, which is equal to an integer $l \in [m]$, is equal to $s_l(T)$ given in (3.7.1). Since $T^l = 0$, for $l \geq m$, it follows that $s_m = \operatorname{rank} T^{m-1} = \dim \operatorname{Im} T^{m-1}$. Note that if $l_j < m$, then $T^{m-1}(\mathbf{x}_j) = \mathbf{0}$. Hence, $\operatorname{Im} T^{m-1} = \operatorname{span}\{T^{m-1}(\mathbf{x}_1), \ldots, T^{m-1}(\mathbf{x}_{s_m(T)})\}$. That is, $s_m(T)$ is the number of l_j that are equal to m. Use Lemma 3.7.3 and the basis of $\mathbf{V}$ given by (3.7.3) to deduce that the number of l_j, which are equal to an integer $l \in [m]$, is equal to $s_l(T)$ for $l = m-1, \ldots, 1$. □

Corollary 3.7.5 *Let Tsatisfy the assumption of Theorem 3.7.4. Denote* $\mathbf{V}_j := \operatorname{span}\{T^{l_j-1}(\mathbf{x}_j), \ldots, T(\mathbf{x}_j), \mathbf{x}_j\}$, *for* $j = 1, \ldots, i$. *Then, each* V_j *is a* T*-invariant subspace,* $T|_{\mathbf{V}_j}$ *is represented by* $J_{l_j}(0) \in \mathbb{C}^{l_j \times l_j}$ *in the basis* $\{T^{l_j-1}(\mathbf{x}_j), \ldots, T(\mathbf{x}_j), \mathbf{x}_j\}$, *and* $\mathbf{V} = \oplus_{j=1}^{i} \mathbf{V}_j$. *Each* l_j *is uniquely determined by the sequence* $s_i(T), i = 1, \ldots, .$ *Namely, the index* m *of the nilpotent* T *is the largest* $i \geq 1$ *such that* $s_i(T) \geq 1$. *Let* $k_1 = s_m(T), l_1 = \ldots = l_{k_1} = p_1 = m$ *and define recursively* $k_r := k_{r-1} + s_{p_r}(T)$, $l_{k_{r-1}+1} = \ldots = l_{k_r} = p_r$, *where* $2 \leq r, p_r \in [m-1], s_{p_r}(T) > 0$ *and* $k_{r-1} = \sum_{j=1}^{m-p_r} s_{m-j+1}(T)$.

Definition 3.7.6 $T : \mathbf{V} \to \mathbf{V}$ *be a nilpotent operator. Then, the sequence* $(l_1, \ldots, l_i)$ *defined in Theorem 3.7.4, which gives the lengths of the corresponding Jordan blocks of* T *in a decreasing order is called the Segré characteristic of* T. *The Weyr characteristic of* T *is the dual to Segre's characteristic. That is, consider an* $m \times i$, $0-1$ *matrix* $B = [b_{pq}] \in \{0,1\}^{m \times i}$. *The* j*-th column of* B *has* 1 *in the rows* $1, \ldots, l_j$ *and* 0 *in the rest of the rows. Let* ω_p *be the* p*-th row sum of* B, *for* $p = 1, \ldots, m$. *Then,* $\omega_1 \geq \ldots \geq \omega_m \geq 1$ *is the Weyr characteristic.*

Proof of Theorem 3.1.18 (The Jordan Canonical Form)

Let $p(z) = \det(zI_n - A)$ be the characteristic polynomial of $A \in \mathbb{C}^{n \times n}$. Since $\mathbb{C}$ is algebraically closed , then $p(z) = \prod_{j=1}^{k} (z - \lambda_j)^{n_j}$. Here $\lambda_1, \ldots, \lambda_k$ are k distinct roots, (eigenvalues of A), where $n_j \geq 1$ is the multiplicity of λ_j in $p(z)$. Note that $\sum_{j=1}^{k} n_j = n$. Let $\psi(z)$ be the minimal polynomial of A. By Theorem 3.4.1, $\psi(z) | p(z)$. Using Problem 3.4.2-1.c, we deduce that $\psi(\lambda_j) = 0$, for $j = 1, \ldots, k$. Hence

$$\det(zI_n - A) = \prod_{j=1}^{k} (z - \lambda_j)^{n_j}, \ \psi(z) = \prod_{j=1}^{k} (z - \lambda_j)^{m_j}, \tag{3.7.6}$$
$$1 \leq m_j \leq n_j, \ \lambda_j \neq \lambda_i \text{ for } j \neq i, \ i, j = 1, \ldots, k.$$

Let $\psi_j := (z - \lambda_j)^{m_j}$, for $j = 1, \ldots, k$. Then $(\psi_j, \psi_i) = 1$, for $j \neq i$. Let $\mathbf{V} := \mathbb{C}^n$ and $T : \mathbf{V} \to \mathbf{V}$ be given by $T(\mathbf{x}) := A\mathbf{x}$, for any $\mathbf{x} \in \mathbb{C}^n$. Then, $\det(zI_n - A)$ and $\psi(z)$ are the characteristic and the minimal polynomial of T, respectively. Use Theorem 3.4.6 to obtain the decomposition $\mathbf{V} = \oplus_{i=1}^{k} \mathbf{V}_i$, where each $\mathbf{V}_i$ is a nontrivial T-invariant subspace such that the minimal polynomial of $T_i := T|_{\mathbf{V}_i}$ is ψ_i, for $i = 1, \ldots, k$. That is, $T_i - \lambda_i I_i$, where I_i is the identity operator, i.e. $I_i \mathbf{v} = \mathbf{v}$, for all $\mathbf{v} \in \mathbf{V}_i$, is a nilpotent operator on $\mathbf{V}_i$ and index $(T_i - \lambda_i I_i) = m_i$. Let $Q_i := T_i - \lambda_i I_i$. Then,

Q_i is nilpotent and index $Q_i = m_i$. Apply Theorem 3.7.4 and Corollary 3.7.5 to deduce that $\mathbf{V}_i = \oplus_{j=1}^{q_j} \mathbf{V}_{i,j}$, where each $V_{i,j}$ is Q_i-invariant subspace, and each $V_{i,j}$ has a basis in which Q_i is represented by a Jordan block $J_{m_{ij}}(0)$, for $j = 1, \ldots, q_j$. According to Corollary 3.7.5

$$m_i = m_{i1} \geq \ldots m_{iq_i} \geq 1, \quad i = 1, \ldots, k. \tag{3.7.7}$$

Furthermore, the above sequence is completely determined by rank $Q_i^j, j = 0, 1, \ldots$ for $i = 1, \ldots, k$. Noting that $T_i = Q_i + \lambda_i I_i$, it easily follows that each $V_{i,j}$ is a T_i-invariant subspace, hence T-invariant subspace. Moreover, in the same basis of $\mathbf{V}_{i,j}$ that Q_i represented by $J_{m_{ij}}(0)$, T_i is represented by $J_{m_{ij}}(\lambda_i)$, for $j = 1, \ldots, q_i$ and $i = 1, \ldots, k$. This shows the existence of the Jordan canonical form.

We now show that the Jordan canonical form is unique, up to a permutation of factors. Note that the minimal polynomial of A is completely determined by its Jordan canonical form. Namely, $\psi(z) = \prod_{i=1}^{k}(z - z_i)^{m_{i1}}$, where m_{i1} is the biggest Jordan block with the eigenvalues λ_i. (See Problem 3.4.2-2.) Thus, $m_{i1} = m_i$, for $i = 1, \ldots, k$. Now, Theorem 3.4.6 yields that the subspaces $\mathbf{V}_1, \ldots, \mathbf{V}_k$ are uniquely determined by ψ. Thus, each T_i and $Q_i = T - \lambda_i I_i$ are uniquely determined. Using Theorem 3.7.4, we get that rank $Q_i^j, j = 0, 1, \ldots$ determines the sizes of the Jordan blocks of Q_i. Hence, all the Jordan blocks corresponding to λ_i are uniquely determined for each $i \in [k]$. □

Corollary 3.7.7 *Let $A \in \mathbb{F}^{n \times n}$ and assume that the characteristic polynomial of $p(z) = \det(zI_n - A)$ splits to linear factors, i.e. (3.7.6) holds. Let B be the Jordan canonical form of A. Then*

1. *The multiplicity of the eigenvalue λ_i in the minimal polynomial $\psi(z)$ of A is the size of the biggest Jordan block corresponding to λ_i in B.*

2. *The number of Jordan blocks in B corresponding to λ_i is the nullity of $A - \lambda_i I_n$, i.e. the number of Jordan block in B corresponding to $A - \lambda_i I_n$ is the number of linearly independent eigenvectors of A corresponding to the eigenvalue λ_i.*

3. *Let λ_i be an eigenvalue of A. Then, the number of the Jordan blocks of order i corresponding to λ_i in B is given in (3.7.9).*

Proof.
1. Since $J_n(0)^n = 0$ and $J_n(0)^{n-1} \neq 0$, it follows that the minimal polynomial of $J_n(\lambda)$ is $(z - \lambda_n)^n = \det(zI_n - J_n(\lambda))$. Use Problem 3.7.2-3.a to deduce the first part of the corollary.
2. Since $J_n(0)$ has one independent eigenvector, use Problem 3.7.2-3.b to deduce the second part of the corollary.
3. Use Problem 3.7.2-2.a to establish the last part of the corollary. □

3.7.1 Worked-out Problems

1. Show that the only nilpotent diagonalizable matrix is the zero matrix.
   Solution:
   Assume that $A \in \mathbb{F}^{n\times n}$ is a diagonalizable and nilpotent matrix with the index of nilpotency m. Since A is diagonalizable, then $A = C^{-1}DC$, for a diagonal matrix D and $C \in \mathrm{GL}(n,\mathbb{F})$. Thus, $D^m = (CAC^{-1})^m = CA^M C^{-1} = 0$. However, if the diagonal entries of D are $d_1,\dots,d_n$, then the diagonal entries of D^m are just $d_1^m,\dots,d_n^m$. Since $D^m = 0$, then $d_i^m = 0$, for all i and hence $d_i = 0$, for all i. This means $D = 0$ and it follows that $A = 0$.

2. Show that the index of nilpotency of the nilpotent matrix $A \in \mathbb{F}^{n\times n}$ is less than or equal to n.
   Solution:
   We already discussed that the minimal polynomial of a nilpotent matrix is z^m, for some positive integer m. Since the minimal polynomial and the characteristic polynomial of a matrix have the same roots (probably with the different multiplicities), then the characteristic polynomial of A is z^n and based on the definition of minimal polynomial, $m \le n$. Now, the statement follows from the fact that the index of nilpotency and the degree of the minimal polynomial of A are the same.

3. The matrix $A \in \mathbb{F}^{n\times n}$ is called *idempotent* if $A^2 = A$. Show that if A is idempotent, then $\det A$ is equal 0 or 1.
   Solution:
   Since $A^2 = A$, taking determinant of both sides of this equation, we find
   $$\det A = \det(A^2). \tag{3.7.8}$$
   Then, $\det(A^2) = (\det A)^2 = \det A$. Hence, $\det A(\det A - 1) = 0$ and so $\det A = 0$ or 1.

3.7.2 Problems

1. Let $T : \mathbf{V} \to \mathbf{V}$ be nilpotent with $m = \operatorname{index} T$. Let $(\omega_1,\dots,\omega_m)$ be the Weyr characteristic. Show that $\operatorname{rank} T^j = \sum_{p=j+1}^m \omega_p$, for $j = 1,\dots,m-1$.

2. Let $A \in \mathbb{C}^{n\times n}$ and assume that $\det(zI_n - A) = \prod_{i=1}^k (z-\lambda_i)^{n_i}$, where $\lambda_1,\dots,\lambda_k$ are k distinct eigenvalues of A. Let
   $$s_i(A,\lambda_j) := \operatorname{rank}\,(A-\lambda_j I_n)^{i-1} - 2\operatorname{rank}\,(A-\lambda_j I_n)^{i} + \operatorname{rank}\,(A-\lambda_j I_n)^{i+1} \tag{3.7.9}$$
   $$i = 1,\dots,n_j, j = 1,\dots,k.$$

   (a) Show that $s_i(A,\lambda_j)$ is the number of Jordan blocks of order i corresponding to λ_j for $i = 1,\dots,n_j$.

   (b) Show that in order to find all Jordan blocks of A corresponding to λ_j one can stop computing $s_i(A,\lambda_j)$ at the smallest $i \in [n_j]$ such that $1s_1(A,\lambda_j) + 2s_2(A,\lambda_j)\dots + is_i(A,\lambda_j) = n_j$.

3. Let $C = F \oplus G = \operatorname{diag}(F,G), F \in \mathbb{F}^{l\times l}, G \in \mathbb{F}^{m\times m}$.

(a) Assume that ψ_F, ψ_G are the minimal polynomials of F and G, respectively. Show that ψ_C, the minimal polynomial of C, is equal to $\frac{\psi_F \psi_G}{(\psi_F, \psi_G)}$.

(b) Show that $\operatorname{null} C = \operatorname{null} F + \operatorname{null} G$. In particular, if G is invertible, i.e. 0 is not eigenvalue of G, then $\operatorname{null} C = \operatorname{null} F$.

3.8 Cyclic subspaces and rational canonical forms

In this section, we study the structure of a linear operator on a finite dimensional vector space, using the primary decomposition theorem of section 3.4. Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator and assume that $\mathbf{V}$ is finite dimensional over the field $\mathbb{F}$. Let $\mathbf{0} \ne \mathbf{u} \in \mathbf{V}$. Consider the sequence of vectors $\mathbf{u} = T^0(\mathbf{u}), T(\mathbf{u}), T^2(\mathbf{u}), \ldots$. Since $\dim \mathbf{V} < \infty$, there exists a positive integer $l \ge 1$ such that $\mathbf{u}, T(\mathbf{u}), \ldots, T^{l-1}(\mathbf{u})$ are linearly independent, and $\mathbf{u}, T(\mathbf{v}), \ldots, T^l(\mathbf{u})$ are linearly dependent. Hence, l is the smallest integer such that

$$T^l(\mathbf{u}) = -\sum_{i=1}^{l} a_i T^{l-i}(\mathbf{u}), \tag{3.8.1}$$

for some scalars $a_i \in \mathbb{F}$, $1 \le i \le l$.

Clearly, $l \le \dim \mathbf{V}$. The polynomial $\psi_{\mathbf{u}}(z) := z^l + \sum_{i=1}^{l} a_i z^{l-i}$ is called the *minimal* polynomial of $\mathbf{u}$, with respect to T. It is a monic polynomial and has the minimum degree among annihilating polynomials of $\mathbf{u}$. Its property is similar to the property of the minimal polynomial of T. Namely, if a polynomial $\phi \in \mathbb{F}[z]$ annihilates $\mathbf{u}$, i.e., $\phi(T)(\mathbf{u}) = 0$, then $\psi_{\mathbf{u}} | \phi$. In particular, the minimal polynomial $\psi(z)$ of T is divisible by $\psi_{\mathbf{u}}$, since $\psi(T)(\mathbf{u}) = 0(\mathbf{u}) = \mathbf{0}$. Assume that $\mathbf{U} = \operatorname{span}\{T^i(\mathbf{u}); i = 0, 1, \ldots\}$. Clearly, every vector $\mathbf{w} \in \mathbf{U}$ can be uniquely represented as $\phi(T)(\mathbf{u})$, where $\phi \in \mathbb{F}[z]$ and $\deg \phi \le l-1$. Hence, $\mathbf{U}$ is a T-invariant subspace. The subspace $\mathbf{U}$ is called the *cyclic* subspace, generated by $\mathbf{u}$. Note that in the basis $\{\mathbf{u}_1 = \mathbf{u}, \mathbf{u}_2 = T(\mathbf{u}), \ldots, \mathbf{u}_l = T^{l-1}(\mathbf{u})\}$, the linear transformation $T|\mathbf{U}$ is given by the matrix

$$\begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & -a_l \\ 1 & 0 & 0 & \ldots & 0 & -a_{l-1} \\ 0 & 1 & 0 & \ldots & 0 & -a_{l-2} \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -a_1 \end{bmatrix} \in \mathbb{F}^{l \times l}. \tag{3.8.2}$$

The above matrix is called the *companion matrix*, corresponding to the polynomial $\psi_{\mathbf{u}}(z)$. (Sometimes, the transpose of the above matrix is called the companion matrix.)

Lemma 3.8.1 *Let $\mathbf{V}$ be a finite dimensional vector space and $T \in L(\mathbf{V})$. Suppose $\mathbf{u}$ and $\mathbf{w}$ are two non-zero elements of $\mathbf{V}$ and $\psi_{\mathbf{u}}$ and $\psi_{\mathbf{w}}$ are the minimal polynomials of $\mathbf{u}$ and $\mathbf{w}$ with respect to T, respectively and $(\psi_{\mathbf{u}}, \psi_{\mathbf{w}}) = 1$.Then, $\psi_{\mathbf{u}+\mathbf{w}} = \psi_{\mathbf{u}} \cdot \psi_{\mathbf{w}}$.*

Proof. Let $\mathbf{U}$ and $\mathbf{W}$ be the cyclic invariant subspaces generated by $\mathbf{u}$ and $\mathbf{w}$, respectively. We claim that the T-invariant subspace $\mathbf{X} := \mathbf{U} \cap \mathbf{W}$ is the trivial subspace $\{\mathbf{0}\}$. Assume to the contrary, there exists $\mathbf{0} \ne \mathbf{x} \in \mathbf{X}$. Let $\mathbf{X}_1 \subset \mathbf{X}$ be a nontrivial cyclic subspace generated by $\mathbf{x}$ and $\psi_{\mathbf{x}}$ be the minimal polynomial corresponding to $\mathbf{x}$. Since $\mathbf{X}_1 \subset \mathbf{U}$, it follows that $\mathbf{x} = \phi(T)(\mathbf{u})$, for some $\phi(z) \in \mathbb{F}[z]$. Hence, $\psi_{\mathbf{u}}(T)(\mathbf{x}) = \mathbf{0}$. Thus $\psi_{\mathbf{x}}|\psi_{\mathbf{u}}$. Similarly $\psi_{\mathbf{x}}|\psi_{\mathbf{w}}$. This contradicts the assumption that $(\psi_{\mathbf{u}}, \psi_{\mathbf{w}}) = 1$. Hence, $\mathbf{U} \cap \mathbf{W} = \{\mathbf{0}\}$. Let $\phi = \psi_{\mathbf{u}}\psi_{\mathbf{w}}$. Clearly

$$\phi(T)(\mathbf{u} + \mathbf{w}) = \psi_{\mathbf{w}}(T)(\psi_{\mathbf{u}}(T)(\mathbf{u})) + \psi_{\mathbf{u}}(T)(\psi_{\mathbf{w}}(T)(\mathbf{w}) = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Thus, ϕ is an annihilating polynomial of $\mathbf{u} + \mathbf{w}$. Let $\theta(z)$ be an annihilating polynomial of $\mathbf{u} + \mathbf{w}$. Then, $\mathbf{0} = \theta(T)(\mathbf{u}) + \theta(T)(\mathbf{w})$. Since $\mathbf{U}$ and $\mathbf{V}$ are T-invariant subspaces and $\mathbf{U} \cap \mathbf{W} = \{\mathbf{0}\}$, it follows that $\mathbf{0} = \theta(T)(\mathbf{u}) = \theta(T)(\mathbf{w})$. Hence, $\psi_{\mathbf{u}}|\theta, \psi_{\mathbf{w}}|\theta$. As $(\psi_{\mathbf{u}}, \psi_{\mathbf{w}}) = 1$, it follows that $\phi|\theta$. Hence, $\psi_{\mathbf{u}+\mathbf{w}} = \phi$. □

Theorem 3.8.2 *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator and $1 \le \dim V < \infty$. Let $\psi(z)$ be the minimal polynomial of T. Then, there exists $\mathbf{0} \ne \mathbf{u} \in \mathbf{V}$ such that $\psi_{\mathbf{u}} = \psi$.*

Proof. Assume first that $\psi(z) = (\phi(z))^l$, where $\phi(z)$ is irreducible in $\mathbb{F}[z]$, i.e. $\phi(z)$ is not divisible by any polynomial θ such $1 \le \deg\theta < \deg\phi$, and l is a positive integer. Let $0 \ne \mathbf{w} \in \mathbf{V}$. Recall that $\psi_{\mathbf{w}}|\psi$. Hence, $\psi_{\mathbf{w}} = (\phi)^{l(\mathbf{w})}$, where $l(\mathbf{w}) \in [l]$ is a positive integer. Assume to the contrary that $1 \le l(\mathbf{w}) \le l - 1$, for each $0 \ne \mathbf{w} \in \mathbf{V}$. Then, $(\phi(T))^{l-1}(\mathbf{w}) = \mathbf{0}$. As $(\phi(T))^{l-1}(\mathbf{0}) = \mathbf{0}$, we deduce that $(\phi(z))^{l-1}$ is an annihilating polynomial of T. Therefore, $\psi|\phi^{l-1}$ which is impossible. Hence there exists $\mathbf{u} \ne \mathbf{0}$ such that $l(\mathbf{u}) = l$, i.e. $\psi_{\mathbf{u}} = \psi$.

Consider now the general case

$$\psi(z) = \prod_{i=1}^{k} (\phi_i(z))^{l_i}, \ l_i \in \mathbb{N}, \ \phi_i \text{ irreducible }, (\phi_i, \phi_j) = 1, \text{ for } i \ne j, \tag{3.8.3}$$

where $k > 1$. Theorem 3.4.6 implies that $V = \oplus_{i=1}^{k} \mathbf{V}_i$, where $\mathbf{V}_i = \ker \phi_i^{l_i}(T)$ and $\phi_i^{l_i}$ is the minimal polynomial of $T|\mathbf{V}_i$. The first case of the proof yields the existence of $\mathbf{0} \ne \mathbf{u}_i \in \mathbf{V}_i$ such that $\psi_{\mathbf{u}_i} = \phi_i^{l_i}$ for $i = 1, \ldots, k$. Let $\mathbf{w}_j = \mathbf{u}_1 + \ldots + \mathbf{u}_j$. Use Lemma 3.8.1 to deduce that $\psi_{\mathbf{w}_j} = \prod_{i=1}^{j} \phi_i^{l_i}$. Hence, $\psi_{\mathbf{u}} = \psi$ for $\mathbf{u} := \mathbf{w}_k$. □

Theorem 3.8.3 (**The cyclic decomposition theorem for V**) *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator and $0 < \dim \mathbf{V} < \infty$. Then, there exists a decomposition of V to a direct sum of r T-cyclic subspaces $V = \oplus_{i=1}^{r} \mathbf{U}_i$ with the following properties. Assume that ψ_i is the minimal polynomial of $T|\mathbf{U}_i$, then ψ_1 is the minimal polynomial of T. Furthermore, $\psi_{i+1}|\psi_i$, for $i = 1, \ldots, r - 1$.*

Proof. We prove the theorem by induction on $\dim \mathbf{V}$. For $\dim \mathbf{V} = 1$, any $0 \ne \mathbf{u}$ is an eigenvector of T: $T(\mathbf{u}) = \lambda\mathbf{u}$, so $\mathbf{V}$ is cyclic, and $\psi(z) = \psi_1(z) = z - \lambda$. Assume now that the theorem holds of all non-zero vector spaces of dimension less than $n+1$. Suppose that $\dim \mathbf{V} = n + 1$. Let $\psi(z)$ be the minimal polynomial of T, $m = \deg\psi$ and T is nonderogatory, i.e., $m = n + 1$, which is the degree of the characteristic polynomial of T. Theorem 3.8.2 implies that $\mathbf{V}$ is a cyclic subspace.This yields $r = 1$, and the theorem holds in this case.

Assume now that T is derogatory. Then, $m < n + 1$. Theorem 3.8.2 implies the existence of $\mathbf{0} \ne \mathbf{u}_1$ such that $\psi_{\mathbf{u}_1} = \psi$. Let $\mathbf{U}_1$ be the cyclic subspace generated by $\mathbf{u}_1$. Let $\hat{V} := \mathbf{V}/\mathbf{U}_1$. Thus, $1 \le \dim \hat{V} = \dim \mathbf{V} - m \le n$. We now apply the induction hypothesis to $\hat{T} : \hat{\mathbf{V}} \to \hat{\mathbf{V}}$. Therefore, $\hat{V} = \oplus_{i=2}^{r} \hat{U}_i$, where $\hat{\mathbf{U}}_i$ is a cyclic subspace generated by $\hat{\mathbf{u}}_i, i = 2, \dots, r$. The minimal polynomial of $\hat{\mathbf{u}}_i$ is $\psi_i, i = 2, \dots, r$ and ψ_2 is the minimal polynomial of $\hat{T}$, and $\psi_{i+1}|\psi_i$, for $i = 2, \dots, r-1$. Observe first that for any polynomial $\phi(z)$, we have the identity $\phi(\hat{T})([\mathbf{x}]) = [\phi(T)(\mathbf{x})]$. Since $\psi(T)(\mathbf{x}) = \mathbf{0}$, it follows that $\psi_1 := \psi(z)$ is an annihilating polynomial of $\hat{T}$. Hence, $\psi_2|\psi_1$.

Observe next that since $\psi_{i+1}|\psi_i$, for $i = 1, \dots, r-1$, it follows that $\psi_i|\psi_1 = \psi$. Consequently, $\psi = \theta_i \psi_i$, where θ_i is a monic polynomial and $i = 2, \dots, r$. Since $[\mathbf{0}] = \psi_i(\hat{T})([\hat{\mathbf{u}}_i]) = [\psi_i(T)(\hat{\mathbf{u}}_i)]$, it follows that $\psi_i(T)(\hat{\mathbf{u}}_i) \in \mathbf{U}_1$. Hence, $\psi_i(T)(\hat{\mathbf{u}}_i) = \omega_i(T)(\mathbf{u}_1)$, for some $\omega(z) \in \mathbb{F}[z]$. Hence, $\mathbf{0} = \psi(T)(\hat{\mathbf{u}}_i) = \theta_i(T)(\psi_i(T)(\hat{\mathbf{u}}_i)) = \theta_i(T)(\omega_i(T)(\mathbf{u}_1))$. Therefore, $\psi_{\mathbf{u}_1} = \psi$ divides $\theta_i \omega_i$. Then, $\psi_i|\omega_i$ and so $\omega_i = \psi_i \alpha_i$. Define $\mathbf{u}_i = \hat{\mathbf{u}}_i - \alpha_i(T)(\mathbf{u}_1)$, for $i = 2, \dots, r$. The rest of the proof of theorem follows from Problem 3.8.2-5. □

The cyclic decomposition theorem can be used to determine a set of canonical forms for similarity as follows.

Theorem 3.8.4 (Rational canonical form) *Let $A \in \mathbb{F}^{n \times n}$. Then, there exist r monic polynomials $\psi_1, \dots, \psi_r$, of degree one at least, satisfying the following conditions: First $\psi_{i+1}|\psi_i$, for $i = 1, \dots, r$. Second, $\psi_1 = \psi$ is the minimal polynomial of A. Then, A is similar to $\oplus_{i=1}^{r} C(\psi_i)$, where $C(\psi_i)$ is the companion matrix of the form (3.8.2) corresponding to the polynomial ψ_i.*

Decompose each $\psi_i = \psi_{i,1} \dots \psi_{i,t_i}$ to the product of its irreducible components, as in the decomposition of $\psi = \psi_1$ given in (3.8.3). Then, A is similar to $\oplus_{i=l=1}^{r,t_i} C(\psi_{i,l})$.

Suppose finally that $\psi(z) = \prod_{j=1}^{k}(z - \lambda_j)^{m_j}$, where $\lambda_1, \dots, \lambda_k$ are the k distinct eigenvalues of A. Then, A is similar to the Jordan canonical form given in Theorem 3.1.18.

Proof. Identify A with $T : \mathbb{F}^n \to \mathbb{F}^n$, where $T(\mathbf{x}) = A\mathbf{x}$. Use Theorem 3.8.3 to decompose $\mathbb{F}^n = \oplus_{i=1}^{r} \mathbf{U}_i$, where each $\mathbf{U}_i$ is a cyclic invariant subspace such that $T|\mathbf{U}_i$ has the minimal polynomial ψ_i. Since $\mathbf{U}_i$ is cyclic, then $T|\mathbf{U}_i$ is represented by $C(\psi_i)$ in the appropriate basis of $\mathbf{U}_i$ as shown in the beginning of this section. Hence, A is similar to $\oplus_{i=1}^{r} C(\psi_i)$.

Consider next $T_i := T|\mathbf{U}_i$. Use Theorem 3.4.6 to deduce that $\mathbf{U}_i$ decomposes to a direct sum of T_i invariant subspaces $\oplus_{l=1}^{t_i} \mathbf{U}_{i,l}$, where the minimal polynomial of $T_{i,l} := T_i|\mathbf{U}_{i,l}$ is $\psi_{i,l}$. Since $\mathbf{U}_i$ was cyclic, i.e., T_i was nonderogatory, it follows that each $T_{i,l}$ must be nonderogatory, i.e. $\mathbf{U}_{i,l}$ cyclic. (See Problem 3.8.2-3.) Recall that each $T_{i,l}$ is represented in a corresponding basis by $C(\psi_{i,l})$. Hence, A is similar to $\oplus_{i=l=1}^{r,t_i} C(\psi_{i,l})$.

Suppose finally that $\psi(z)$ splits to linear factors. Hence, $\psi_{i,l} = (z - \lambda_l)^{m_{i,l}}$ and $T_{i,l} - \lambda_l I$ is nilpotent of index $m_{i,l}$. Thus, there exists a basis in $\mathbf{U}_{i,l}$ such that $T_{i,l}$ is represented by the Jordan block $J_{m_{i,l}}(\lambda_l)$. Therefore, A is similar to a sum of corresponding Jordan blocks. □

The polynomials $\psi_1, \dots, \psi_k$ appearing in Theorem 3.8.3 are called *invariant polynomials* of T or its representation matrix A, in any basis. The polynomials $\psi_{i,1}, \dots, \psi_{i,t_i}, i = 1, \dots, k$ appearing in Theorem 3.8.4 are called the *elementary divisors* of A, or the corresponding linear transformation T represented by A. We now show that these polynomials are uniquely defined.

Lemma 3.8.5 *(Cauchy-Binet formula) For two positive integers $1 \le p \le m$, denote by $\mathrm{Q}_{p,m}$ the set of all subsets $\boldsymbol{\alpha}$ of $\{1,\dots,m\}$ of cardinality p. Let $A \in \mathbb{F}^{m\times n}, B \in \mathbb{F}^{n\times l}$ and denote $C = AB \in \mathbb{F}^{m\times l}$. Then, for any integer $p \in [\min(m,n,l)], \boldsymbol{\alpha} \in \mathrm{Q}_{p,m}, \boldsymbol{\beta} \in \mathrm{Q}_{p,l}$ the following identity holds.*

$$\det C[\boldsymbol{\alpha},\boldsymbol{\beta}] = \sum_{\boldsymbol{\gamma}\in\mathrm{Q}_{p,n}} \det A[\boldsymbol{\alpha},\boldsymbol{\gamma}] \det B[\boldsymbol{\gamma},\boldsymbol{\beta}]. \tag{3.8.4}$$

Proof. It is enough to check the case where $\boldsymbol{\alpha} = \boldsymbol{\beta} = \{1,\dots,p\}$. This is equivalent to the assumption that $p = m = l \le n$. For $p = m = n = l$, the Cauchy-Binet formula reduces to $\det AB = (\det A)(\det B)$. Thus, it is sufficient to consider the case $p = m = l < n$. Let $C = [c_{ij}]_{i=j=1}^{p}$. Then, $c_{ij} = \sum_{t_j=1}^{n} a_{it_j} b_{t_j j}$ for $i,j = 1,\dots,p$. Use multilinearity of the determinant to deduce

$$\det C = \det[\sum_{t_j=1}^{n} a_{it_j} b_{t_j j}]_{i=j=1}^{n} = \sum_{t_1,\dots,t_p=1}^{n} \det[a_{it_j} b_{t_j j}]_{i=j=1}^{n} =$$
$$\sum_{t_1,\dots,t_p=1}^{n} \det A[\{1,\dots,p\},\{t_1,\dots,t_p\}] b_{t_1 1} b_{t_2 2}\dots b_{t_p p}.$$

Observe next that $\det A[\{1,\dots,p\},\{t_1,\dots,t_p\}] = 0$ if $t_i = t_j$, for some $1 \le i < j \le p$, since the columns t_i and t_j in $A[\{1,\dots,p\},\{t_1,\dots,t_p\}]$ are equal. Consider the sum

$$\sum_{\{t_1,t_2,\dots,t_p\}=\boldsymbol{\gamma}\in\mathrm{Q}_{p,n}} A[\{1,\dots,p\},\{t_1,\dots,t_p\}] b_{t_1 1} b_{t_2 2}\dots b_{t_p p}.$$

The above arguments yield that this sum is

$$\det(A[\langle p\rangle,\boldsymbol{\gamma}]C[\boldsymbol{\gamma},\langle p\rangle]) = (\det A[\langle p\rangle,\boldsymbol{\gamma}])(\det C[\boldsymbol{\gamma},\langle p\rangle]).$$

This establishes (3.8.4). □

Proposition 3.8.6 *Let $A(z) \in \mathbb{F}^{m\times n}[z]$. Denote by $r = \operatorname{rank} A(z)$ the size of the biggest minor of $A(z)$ which is not a zero polynomial. For an integer $k \in [r]$, let $\delta_k(z)$ be the greatest common divisor of all $k\times k$ minors of $A(z)$, which is assumed to be a monic polynomial. Assume that $\delta_0 = 1$. Then, $\delta_i|\delta_{i+1}$, for $i = 0,\dots,r-1$.*

Proof. Expand a non-zero $(k+1)\times(k+1)$ minor of $A(z)$ to deduce that $\delta_k(z)|\delta_{k+1}(z)$. In particular, $\delta_i|\delta_j$, for $1 \le i < j \le r$. □

Proposition 3.8.7 *Let $A(z) \in \mathbb{F}[z]^{m\times n}$. Assume that $P \in \mathrm{GL}(m,\mathbb{F})$ and $Q \in \mathrm{GL}(n,\mathbb{F})$. Let $B(z) = PA(z)Q$. Then*

1. $\operatorname{rank} A(z) = \operatorname{rank} B(z) = r$.
2. $\delta_k(A(z)) = \delta_k(B(z))$, *for* $k = 0,\dots,r$.

Proof. Use Cauchy-Binet to deduce that $\operatorname{rank} B(z) \le \operatorname{rank} A(z)$. Since $A(z) = P^{-1}B(z)Q^{-1}$, it follows that $\operatorname{rank} A(z) \le \operatorname{rank} B(z)$. Hence, $\operatorname{rank} A(z) = \operatorname{rank} B(z) = r$. Use the Cauchy-Binet to deduce that $\delta_k(A(z))|\delta_k(B(z))$. As $A(z) = P^{-1}B(z)Q^{-1}$, we deduce also that $\delta_k(B(z))|\delta_k(A(z))$. □

Definition 3.8.8 *Let $A \in \mathbb{F}^{n\times n}$ and define $A(z) := zI_n - A$. Then, the polynomials of degree* 1 *at least in the sequence $\phi_i = \frac{\delta_{n-i+1}}{\delta_{n-i}(z)}, i = 1, \dots, n$ are called the invariant polynomials of A.*

Note that the product of all invariant polynomials is $\det(zI_n - A)$. In view of Proposition 3.8.7, we obtain that similar matrices have the same invariant polynomials. Hence, for linear transformation $T : \mathbf{V} \to \mathbf{V}$, we can define its invariant polynomials of T by any representation matrix of T in a basis of $\mathbf{V}$.

Theorem 3.8.9 *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator. Then, the invariant polynomials of T are the polynomials $\psi_1, \dots, \psi_r$ appearing in Theorem 3.8.3.*

To prove this theorem we need the following lemma:

Lemma 3.8.10 *Let $A = \operatorname{diag}(B, C)(= B \oplus C) \in \mathbb{F}^{n\times n}$, where $B \in \mathbb{F}^{l\times l}, C \in \mathbb{F}^{m\times m}, m = n - l$. Let $p \in [n-1]$ and assume that $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_p\}, \boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\} \in Q_{p,n}$:*

$$1 \le \alpha_1 < \dots < \alpha_p \le n, \quad 1 \le \beta_1 < \dots < \beta_p \le n.$$

If $\#\alpha \cap [l] \ne \#\beta \cap [l]$, then $\det A[\boldsymbol{\alpha}, \boldsymbol{\beta}] = 0$.

Proof. Let $k := \#\boldsymbol{\alpha} \cap [l]$, i.e. $\alpha_k \le l, \alpha_{k+1} > l$. Consider a term in $\det A[\boldsymbol{\alpha}, \boldsymbol{\beta}]$. Up to a sign it is $\prod_{j=1}^{p} a_{\alpha_j \beta_{\sigma(j)}}$, for some bijection $\sigma : [p] \to [p]$. Then, this product is zero unless $\sigma([k]) = [k]$ and $\beta_k \le l, \beta_{k+1} > l$. □

Proof of Theorem 3.8.9. From the proof of Theorem 3.8.3, it follows that $\mathbf{V}$ has a basis in which T is represented by $C := \oplus_{i=1}^{r} C(\psi_i)$. Assume that $C(\psi_i) \in \mathbb{F}^{l_i \times l_i}, i \in [r]$, where $n = \sum_{i=1}^{r} l_i$.

Let $\theta := z^l + a_1 z^{m-1} + \dots + a_l$. Then, the companion matrix $C(\theta)$ is given by (3.8.2). Consider the submatrix of $B(z)$ of $zI_l - C(\theta)$ obtained by deleting the first row an the last column. It is an upper triangular matrix with -1 on the diagonal. Hence, $\det B(z) = (-1)^{l-1}$. Therefore, $\delta_{l-1}(zI_m - C(\theta)) = 1$. Thus, $\delta_{l-j}(zI_l - C(\theta)) = 1$, for $j \ge 1$.

We claim that $\delta_{n-j}(zI_n - C) = \prod_{i=j+1}^{r} \psi_j$, for $j \in [r]$. Consider a minor of order $n - j$ of the block diagonal matrix $zI_n - C$ of the form $(zI_n - C)[\boldsymbol{\alpha}, \boldsymbol{\beta}]$, where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in Q_{n-j,n}$. Lemma 3.8.10 claims that $\det(zI_n - C)[\boldsymbol{\alpha}, \boldsymbol{\beta}] = 0$, unless we choose our minor such that in each submatrix $(zI_{l_i} - C(\psi_i))$, we delete the same number of rows and columns. Deleting the first row and last column in $(zI_{l_i} - C(\psi_i))$, we get a minor with value $(-1)^{l_i - 1}$. Recall also that $\delta_{l_i - k}(zI_{l_i} - C(\psi_i)) = 1$, for $k \ge 1$. Since $\psi_{i+1} | \psi_i$, it follows that $\delta_{n-j}(zI_n - C)$ is obtained by deleting the first row and the last column in $(zI_{l_i} - C(\psi_i))$, for $i = 1, \dots, j$ and $j = 1, \dots, r$. ($\psi_{r+1} := 1$.) Thus, $\delta_{n-j}(zI_n - C) = \prod_{i=j+1}^{r} \psi_j$, for $j \in [r]$. Hence, $\psi_i = \frac{\delta_{n-i+1}(zI_n - C)}{\delta_{n-i}(zI_n - C)}$. □

Note that the above theorem implies that the invariant polynomials of A satisfy $\phi_{i+1} | \phi_i$, for $i = 1, \dots, r - 1$.

The irreducible factors $\phi_{i,1}, \dots, \phi_{i,t_i}$ of ψ_i given in Theorem 3.8.4, for $i = 1, \dots, r$ are called the *elementary divisors* of A. The matrices $\oplus_{i=1}^{k} C(\psi_i)$ and $\oplus_{i=l=1}^{k,t_i} C(\psi_{i,l})$ are called the *rational canonical forms of* A. Those are the canonical forms in the case that the characteristic polynomial of A does not split to linear factors over $\mathbb{F}$. Note that the diagonal elements of the Jordan canonical form of a matrix A are the eigenvalues of A, each appearing the number of times equal to its algebraic multiplicity. However, the rational canonical form does not expose the eigenvalues of the matrix, even when these eigenvalues lie in the base field.

Example 3.8.11 *Let $T \in L(\mathbb{R}^7)$ with the minimal polynomial $\psi(z) = (z-1)(z^2+1)^2$. Since its elementary divisors are $z-1$ and $(z^2+1)^2$, then we have the following possibilities for the list of elementary divisors:*

1. $z-1$, $z-1$, $z-1$, $(z^2+1)^2$
2. $z-1$, $(z^2+1)^2$, z^2+1

These correspond to the following rational canonical forms:

1. $$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

2. $$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

3.8.1 Worked-out Problems

1. Consider the following matrix:

$$C = \begin{bmatrix} 2 & 2 & -2 & 4 \\ -4 & -3 & 4 & -6 \\ 1 & 1 & -1 & 2 \\ 2 & 2 & -2 & 4 \end{bmatrix}.$$

Assume that the minimal polynomial of C is $\psi_C(z) = z(z-1)^2$. Find the rational canonical form of C.
Solution:
Since $\frac{p_C(z)}{\psi_1(z)} = z$, then $\psi_2(z) = z$. According to the definition of the companion matrix,

$$C(\psi_1) = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix},$$

and $C(\psi_2) = 0$ and so

$$C \sim \begin{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix} & \\ & [0] \end{bmatrix}.$$

3.8.2 Problems

1. Let $\phi(z) = z^l + \sum_{i=1}^{l} a_i z^{l-i}$. Denote by $C(\phi)$ the matrix (3.8.2). Show the following statements:

   (a) $\phi(z) = \det(zI - C(\phi))$.

   (b) The minimal polynomial of $C(\phi)$ is ϕ, i.e. $C(\phi)$ is nonderogatory.

   (c) Assume that $A \in \mathbb{F}^{n\times n}$ is nonderogatory. Show that A is similar to $C(\phi)$, where ϕ is a characteristic polynomial of A.

2. Let $T : \mathbf{V} \to \mathbf{V}, \dim \mathbf{V} < \infty, \mathbf{0} \ne \mathbf{u}, \mathbf{w}$. $\psi_{\mathbf{u}+\mathbf{w}} = \frac{\psi_{\mathbf{u}}\psi_{\mathbf{w}}}{(\psi_{\mathbf{u}}, \psi_{\mathbf{w}})}$.

3. Let $T : V \to V$ and assume that V is a cyclic space, i.e. T is nondegenerate. Let ψ be the minimal and the characteristic polynomial. Assume that $\psi = \psi_1\psi_2$, where $\deg\psi_1, \deg\psi_2 \ge 1$, and $(\psi_1, \psi_2) = 1$. Show that there exist $\mathbf{0} \ne \mathbf{u}_1, \mathbf{u}_2$ such that $\psi_i = \psi_{\mathbf{u}_i}, i = 1, 2$. Furthermore, $\mathbf{V} = \mathbf{U}_1 \oplus \mathbf{U}_2$, where $\mathbf{U}_i$ is the cyclic subspace generated by $\mathbf{u}_i$.

4. Let T is a linear operator $T : \mathbf{V} \to \mathbf{V}$ and $\mathbf{U}$ is a subspace of $\mathbf{V}$.

   (a) Assume that $T\mathbf{U} \subseteq \mathbf{U}$. Show that T induces a linear transformation $\hat{T} : \mathbf{V}/\mathbf{U} \to \mathbf{V}/\mathbf{U}$, i.e. $\hat{T}([\mathbf{x}]) := [T(\mathbf{x})]$.

   (b) Assume that $T\mathbf{U} \nsubseteq \mathbf{U}$. Show that $\hat{T}([\mathbf{x}]) := [T(\mathbf{x})]$ does not make sense.

5. In this problem we finish the proof of Theorem 3.8.3.

   (a) Show that ψ_i is an annihilating polynomial of $\mathbf{u}_i$, for $i \ge 2$.

   (b) By considering the vector $[\hat{\mathbf{u}}_i] = [\mathbf{u}_i]$, show that $\psi_{\mathbf{u}_i} = \psi_i$ for $i \ge 2$.

   (c) Show that the vectors $\mathbf{u}_i, T(\mathbf{u}_i), \ldots, T^{\deg\psi_i - 1}(\mathbf{u}_i), i = 1, \ldots, r$ are linearly independent.
   (**Hint**: Assume linear dependence of all vectors, and then quotient this dependence by $\mathbf{U}_1$. Then, use the induction hypothesis on $\hat{V}$.)

   (d) Let $\mathbf{U}_i$ be the cyclic subspace generated by $\mathbf{u}_i$, for $i = 2, \ldots, r$. Conclude the proof of Theorem 3.8.3.

6. Let the assumptions of Theorem 3.8.3 hold. Show that the characteristic polynomial of T is equal to $\prod_{i=1}^{r} \psi_i$.

7. Prove that any $A \in \mathbb{F}^{n\times n}$ can be written as the product of two symmetric matrices. (Hint: Use rational canonical form.)

8. Let $A, B \in \mathbb{F}^{n\times n}$. Prove that A is similar to B if and only if they have the same elementary divisors.

Chapter 4

Applications of Jordan canonical form

4.1 Functions of Matrices

Let $A \in \mathbb{C}^{n\times n}$. Consider the iterations

$$\mathbf{x}_l = A\mathbf{x}_{l-1},\ \mathbf{x}_{l-1} \in \mathbb{C}^n, \quad l = 1, \ldots \tag{4.1.1}$$

Clearly, $\mathbf{x}_l = A^l\mathbf{x}_0$. To compute $\mathbf{x}_l$ from $\mathbf{x}_{l-1}$, one need to perform $n(2n-1)$ flops, (operations: n^2 multiplications and $n(n-1)$ additions). If we want to compute $\mathbf{x}_{10^8}$, we need $10^8 n(2n-1)$ operations, if we simply program the iterations (4.1.1). If $n = 10$, it will take us some time to do these iterations, and we will probably run to the roundoff error, which will render our computations meaningless. Is there any better way to find $\mathbf{x}_{10^8}$? The answer is *yes*, and this is the purpose of this section.

For a scalar function f and a matrix $A \in \mathbb{C}^{n\times n}$ we denote by $f(A)$ to be a matrix of the same dimension as A. This provides a generalization of a scalar variable matrix $f(z)$, $z \in \mathbb{C}$. If $f(x)$ is a polynomial or rational function, it is natural to define $f(A)$ by substituting A for x, replacing division by matrix inversion and replacing 1 by the identity matrix. For example if $f(x) = \frac{1+x^3}{1-x}$ then

$$f(A) = (I-A)^{-1}(I+A^3),$$

if 1 is not an eigenvalue of A. As rational functions of a matrix commute, so it does not matter if we write $(I-A)^{-1}(I+A^3)$ or $(I+A^3)(I-A)^{-1}$. In what follows, we are going to provide the formal definition of a matrix function. Let $A \in \mathbb{C}^{n\times n}$. Using Theorem 3.1.18 (The Jordan canonical form) we have

$$P^{-1}AP = \bigoplus_{i=1}^{k} J_{n_i}(\lambda_i),$$

for some $P \in \mathrm{GL}(n,\mathbb{C})$, and λ_i's are the eigenvalues of A each of order n_i. The function f is said to be defined on the spectrum of A if the values

$$f^{(j)}(\lambda_i), \ \ j = 0, 1, \ldots, n_i - 1 \quad i \in [k]$$

exist. These are called the values of the function f on the spectrum of A.

Definition 4.1.1 (Matrix function) *Let f be defined on the spectrum of A. Then, $f(A) = Pf(J)P^{-1} = P\operatorname{diag}(f(J_k))P^{-1}$, where*

$$f(J_k) = \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \cdots & \frac{f^{(n_i-1)}(\lambda_i)}{(n_i-1)!} \\ & f(\lambda_i) & \ddots & \vdots \\ & & \ddots & f'(\lambda_i) \\ & & & f(\lambda_i) \end{bmatrix}.$$

For example for $J = \begin{bmatrix} \frac{1}{4} & 1 \\ 0 & \frac{1}{4} \end{bmatrix}$ and $f(x) = x^2$, we have

$$f(J) = \begin{bmatrix} f(\frac{1}{4}) & f'(\frac{1}{2}) \\ 0 & f(\frac{1}{4}) \end{bmatrix} = \begin{bmatrix} \frac{1}{16} & 1 \\ 0 & \frac{1}{16} \end{bmatrix},$$

which is easily proved to be J^2.

Theorem 4.1.2 *Let $A \in \mathbb{C}^{n\times n}$ and*

$$\det(zI_n - A) = \prod_{i=1}^{k}(z-\lambda_i)^{n_i},\ \psi(z) = \prod_{i=1}^{k}(z-\lambda_i)^{m_i}, \tag{4.1.2}$$

$$1 \le m := \deg\psi = \sum_{i=1}^{k} m_i \le n = \sum_{i=1}^{k} n_i,$$

$$1 \le m_i \le n_i,\ \lambda_i \ne \lambda_j,\ \textit{for } i \ne j,\ i,j = 1,\ldots,k,$$

where $\psi(z)$ is the minimal polynomial of A. Then, there exist unique m linearly independent matrices $Z_{ij} \in \mathbb{C}^{n\times n}$, for $i = 1,\ldots,k$ and $j = 0,\ldots,m_i - 1$, which depend on A, such that for any polynomial $f(z)$ the following identity holds

$$f(A) = \sum_{i=1}^{k}\sum_{j=0}^{m_i-1}\frac{f^{(j)}(\lambda_i)}{j!}Z_{ij}. \tag{4.1.3}$$

($Z_{ij}, i = 1,\ldots,k, j = 1,\ldots,m_i$ are called the A-components.)

Proof. We start first with $A = J_n(\lambda)$. So $J_n(\lambda) = \lambda I_n + H_n$, where $H_n := J_n(0)$. Thus, H_n is a nilpotent matrix, with $H_n^n = \mathbf{0}$ and H_n^j has 1's on the j-th subdiagonal and all other elements are equal 0, for $j = 0,1,\ldots,n-1$. Hence, $I_n = H_n^0, H_n, \ldots, H_n^{n-1}$ are linearly independent.

Let $f(z) = z^l$. Then

$$A^l = (\lambda I_n + H_n)^l = \sum_{j=0}^{l}\binom{l}{j}\lambda^{l-j}H_n^j = \sum_{j=0}^{\min(l,n-1)}\binom{l}{j}\lambda^{l-j}H_n^j.$$

The last equality follows from the equality $H^j = \mathbf{0}$, for $j \ge n$. Note that $\psi(z) = \det(zI_n - J_n(\lambda)) = (z-\lambda)^n$, i.e. $k = 1$ and $m = m_1 = n$. From the above equality we conclude that $Z_{1j} = H_n^j$, for $j = 0,\ldots$ if $f(z) = z^l$ and $l = 0,1,\ldots$. With this definition of Z_{1j}, (4.1.3) holds for $K_l z^l$, where $K_l \in \mathbb{C}$ and $l = 0,1,\ldots$. Hence, (4.1.3) holds for any polynomial $f(z)$ for this choice of A.

Assume now that A is a direct sum of Jordan blocks as in (3.4.6): $A = \oplus_{i=j=1}^{k,l_i} J_{m_{ij}}(\lambda_i)$. Here $m_i = m_{i1} \ge \ldots \ge m_{il_i} \ge 1$ for $i = 1,\ldots,k$, and $\lambda_i \ne \lambda_j$ for $i \ne j$. Thus, (4.1.2) holds with $n_i = \sum_{j=1}^{l_i} m_{ij}$ for $i = 1,\dots,k$. Let $f(z)$ be a polynomial. Then, $f(A) = \oplus_{i=j=1}^{k,l_i} f(J_{m_{ij}}(\lambda_i))$. Use the results for $J_n(\lambda)$ to deduce

$$f(A) = \oplus_{i=j=1}^{k,l_i} \sum_{r=0}^{m_{ij}-1} \frac{f^{(r)}(\lambda_i)}{r!} H_{m_{ij}}^r.$$

Let $Z_{ij} \in \mathbb{C}^{n\times n}$ be a block diagonal matrix of the following form. For each integer $l \in [k]$ with $l \neq i$, all the corresponding blocks to $J_{lr}(\lambda_l)$ are equal to zero. In the block corresponding to $J_{m_{ir}}(\lambda_i)$ Z_{ij} has the block matrix $H_{m_{ir}}^j$, for $j = 0,\dots,m_i - 1$. Note that each Z_{ij} is a non-zero matrix with $0-1$ entries. Furthermore, two different Z_{ij} and $Z_{i'j'}$ do not have a common 1 entry. Hence, $Z_{ij}, i = 1,\dots,k, j = 0,\dots,m_i - 1$ are linearly independent. It is straightforward to deduce (4.1.3) from the above identity.

Let $B \in \mathbb{C}^{n\times n}$. Then, $B = UAU^{-1}$ where A is the Jordan canonical form of B. Recall that A and B have the same characteristic polynomial. Let $f(z) \in \mathbb{C}[z]$. Then (4.1.3) holds. Clearly

$$f(B) = Uf(A)U^{-1} = \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \frac{f^{(j)}(\lambda_i)}{j!} UZ_{ij}U^{-1}.$$

Hence, (4.1.3) holds for B, where $UZ_{ij}U^{-1}, i = 1,\dots,k, j = 0,\dots,m_{ij-1}$ are the B-components.

The uniqueness of the A-components follows from the existence and uniqueness of the Lagrange-Sylvester interpolation polynomial, explained below.

□

Theorem 4.1.3 ***(The Lagrange-Sylvester interpolation polynomial).*** *Let $\lambda_1,\dots,\lambda_k \in \mathbb{C}$ be k-distinct numbers. Let $m_1,\dots,m_k$ be k positive integers and $m = m_1 + \dots + m_k$. Let $s_{ij}, i = 1,\dots,k, j = 0,\dots,m_i - 1$ be any m complex numbers. Then, there exists a unique polynomial $\phi(z)$ of degree at most $m - 1$ satisfying the conditions $\phi^{(j)}(\lambda_i) = s_{ij}$, for $i = 1,\dots,k, j = 0,\dots,m_i - 1$. (For $m_i = 1, i = 1,\dots,k$, ϕ is the Lagrange interpolating polynomial.)*

Proof. The Lagrange interpolating polynomial is given by the formula

$$\phi(z) = \sum_{i=1}^{k} \frac{(z-\lambda_1)\dots(z-\lambda_{i-1})(z-\lambda_{i+1})\dots(z-\lambda_k)}{(\lambda_i-\lambda_1)\dots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\dots(\lambda_i-\lambda_k)} s_{i0}.$$

In the general case, one can determine $\phi(z)$ as follows. Let $\psi(z) := \prod_{i=1}^{k}(z-\lambda_i)^{m_i}$. Then

$$\phi(z) = \psi(z) \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \frac{t_{ij}}{(z-\lambda_i)^{m_i-j}} = \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} t_{ij}(z-\lambda_i)^j \theta_i(z). \tag{4.1.4}$$

Here

$$\theta_i = \frac{\psi(z)}{(z-\lambda_i)^{m_i}} = \prod_{j\neq i}(z-\lambda_j)^{m_j}, \text{ for } i = 1,\dots,k. \tag{4.1.5}$$

Observe

$$\frac{d^l\theta_i}{dz^l}(\lambda_r) = 0, \text{ for } l = 0,\dots,m_r - 1 \text{ and } r \neq i. \tag{4.1.6}$$

Now start to determine $t_{i0}, t_{i1}, \dots, t_{i(m_i-1)}$ recursively for each fixed value of i. This is done by using the values of $\phi(\lambda_i), \phi'(\lambda_i), \dots, \phi^{(m_i-1)}(\lambda_i)$ in the above formula for $\phi(z)$. Note that $\deg\phi \leq m-1$. It is straightforward to show that

$$t_{i0} = \frac{\phi(\lambda_i)}{\theta_i(\lambda_i)},\ t_{i1} = \frac{\phi'(\lambda_i) - t_{i0}\theta_i'(\lambda_i)}{\theta_i(\lambda_i)},\ t_{i2} = \frac{\phi''(\lambda_i) - t_{i0}\theta_i''(\lambda_i) - 2t_{i1}\theta_i'(\lambda_i)}{2\theta_i(\lambda_i)}. \tag{4.1.7}$$

The uniqueness of ϕ is shown as follows. Assume that $\theta(z)$ is another Lagrange-Sylvester polynomial of degree less than m. Then, $\omega(z) := \phi(z) - \theta(z)$ must be divisible by $(z-\lambda_i)^{m_i}$, since $\omega^{(j)}(\lambda_i) = 0$, for $j = 0,\dots,m_i - 1$, for each $i = 1,\dots,k$. Hence, $\psi(z)|\omega(z)$. As $\deg\omega(z) \leq m-1$, it follows that $\omega(z)$ is the zero polynomial, i.e. $\phi(z) = \theta(z)$. □

How to find the components of A:

Assume that the minimal polynomial of A, $\psi(z)$ is given by (4.1.2). Then, we have

$Z_{ij} = \phi_{ij}(A)$, where $\phi_{ij}(z)$ is the Lagrange-Sylvester polynomial of degree m at most satisfying

$$\phi_{ij}(\lambda_p) = 0, \text{ for } p \neq i \text{ and } \phi_{ij}^{(q)}(\lambda_i) = j!\delta_{pj},\ j = 0,\dots,m_i - 1. \tag{4.1.8}$$

Indeed, assume that $\phi_{ij}(z)$ satisfies (4.1.8). Use (4.1.3) to deduce that $\phi_{ij}(A) = Z_{ij}$.

To find ϕ_{ij} do the following steps. Set

$$\phi_{ij}(z) = \frac{\theta_i(z)(z-\lambda_i)^j}{\theta_i(\lambda_i)}(1 + \sum_{l=j+1}^{m_i-1} a_l(z-\lambda_i)^{l-j}), j = 0,\dots,m_i - 2, \tag{4.1.9}$$

$$\phi_{i(m_i-1)}(z) = \frac{\theta_i(z)(z-\lambda_i)^{m_i-1}}{\theta_i(\lambda_i)}. \tag{4.1.10}$$

To find $a_{j+1},\dots,a_{m_i-1}$ in (4.1.9), use the conditions

$$\phi_{ij}^{(q)}(\lambda_i) = 0 \text{ for } q = j+1,\dots,m_i - 1. \tag{4.1.11}$$

Namely, the condition for $q = j+1$ determines a_{q+1}. Next the condition for $q = j+2$ determines a_{q+2}. Continue in this manner to find all $a_{j+1},\dots,a_{m_i-1}$.

We now explain briefly why these formulas hold. Since $\theta_i(z)|\phi_{ij}(z)$, it follows that $\phi_{ij}^{(q)}(\lambda_p) = 0$ for $p \neq i$ and $q = 0,\dots,m_q - 1$. (That is, the first condition of (4.1.8) holds.) Since $(z-\lambda_i)^j|\phi_{ij}(z)$, it follows that $\phi_{ij}^{(q)}(\lambda_i) = 0$, for $q = 0,\dots,j-1$. A straightforward calculation shows that $\phi_{ij}^{(j)}(\lambda_i) = j!$. The values of $a_{j+1},\dots,a_{m_i-1}$ are determined by (4.1.11).

Proof of the uniqueness of A-components. Let $\phi_{ij}(z)$ be the Lagrange-Sylvester polynomial given by (4.1.8). Then, (4.1.3) yields that $Z_{ij} = \phi_{ij}(A)$. □

The following Proposition gives a computer algorithm to find Z_{ij}. (Not recommended for hand calculations!).

Proposition 4.1.4 *Let $A \in \mathbb{C}^{n\times n}$. Assume that the minimal polynomial $\psi(z)$ be given by (4.1.2) and denote $m = \deg\psi$. Then, for each integers $u, v \in [n]$ denote by $a_{uv}^{(l)}$ and $(Z_{ij})_{uv}$ the (u,v) entries of A^l and of the A-component Z_{ij}, respectively. Then, $(Z_{ij})_{uv}, i = 1,\dots,k, j = 0,\dots,m_i - 1$ are the unique solutions of the following system with m unknowns*

$$\sum_{i=1}^{k}\sum_{j=0}^{m_i-1}\binom{l}{j}\lambda_i^{\max(l-j,0)}(Z_{ij})_{uv} = a_{uv}^{(l)}, \quad l = 0,\dots,m-1. \tag{4.1.12}$$

(Note that $\binom{l}{j} = 0$ for $j > l$.)

Proof. Consider the equality (4.1.3) for $f(z) = z^l$, where $l = 0,\dots,m-1$. Restricting these equalities to (u,v) entries, we deduce that $(Z_{ij})_{uv}$ satisfy the system (4.1.12). Thus, the systems (4.1.12) are solvable for each pair $(u,v), u,v = 1,\dots,n$. Let $X_{ij} \in \mathbb{C}^{n\times n}, i = 1,\dots,k, j = 1,\dots,m_i - 1$ such that $(X_{ij})_{uv}$ satisfy the system (4.1.12), for each $u,v \in [n]$. Hence, $f(A) = \sum_{i=1}^{k}\sum_{j=0}^{m_i-1}\frac{f^{(j)}(\lambda_i)}{j!}T_{ij}$, for $f(z) = z^l$ and $l = 0,\dots,m-1$. Hence, the above equality holds for any polynomial $f(z)$ of degree less than m. Apply the above formula to the Lagrange-Sylvester polynomial ϕ_{ij} as given in the proof of the uniqueness of the A-components. Then, $\phi_{ij}(A) = X_{ij}$. So $X_{ij} = Z_{ij}$. Thus, each system (4.1.12) has a unique solution. □

The algorithm for finding the A-components and its complexity.

(a) Set $i = 1$.

(b) Compute and store A^i. Check if $I_n, A, \dots, A^i$ are linearly independent.

(c) $m = i$ and express $A^m = \sum_{i=1}^{m} a_i A^{m-i}$. Then, $\psi(z) = z^m - \sum_{i=1}^{m} a_i z^{m-i}$ is the minimal polynomial.

(d) Find the k roots of $\psi(z)$ and their multiplicities: $\psi(z) = \prod_{i=1}^{k}(z-\lambda_i)^{m_i}$.

(e) Find the A-components by solving n^2 systems (4.1.12).

Complexity of an algorithm is a measure of the amount of time or space required by an algorithm for an input of a given size.
The maximum complexity to find $\psi(z)$ happens when $m = n$. Then, we need to compute and store $I_n, A, A^2, \dots, A^n$. So, we need n^3 storage space. Viewing $I_n, A, \dots, A^i$ as row vectors arranged as $i \times n^2$ matrix $B_i \in \mathbb{C}^{i\times n^2}$, we bring B_i to a row echelon form: $C_i = U_i B_i$, $U_i \in C^{i\times i}$. Note that C_i is essentially upper triangular. Then, we add $i+1$-th row: A^{i+1} to the B_i to obtain $C_{i+1} = U_{i+1}B_{i+1}$. (C_i is $i\times i$ submatrix of C_{i+1}.) To get C_{i+1} from C_i we need $2in^2$ flops. In the case $m = nC_{n^2+1}$ has the last row zero. So to find $\psi(z)$ we need at most Kn^4 flops. ($K \le 2$?). The total storage space is around $2n^3$.
Now to find the roots of $\psi(z)$ with certain precision will take a polynomial time, depending on the precision.
To solve n^2 systems with n variables, given in (4.1.12), use Gauss-Jordan for the

augmented matrix $[S\,T]$. Here, $S \in \mathbb{C}^{n\times n}$ stands for the coefficient of the system (4.1.12), depending on $\lambda_1,\dots,\lambda_k$. Next, $T \in \mathbb{C}^{n\times n^2}$ given the "left-hand side" of n^2 systems of (4.1.12). One needs around n^3 storage space. Bring $[S\,T]$ to $[I_n\,Q]$ using Gauss-Jordan to find A-components. To do that we need about n^4 flops.
In summary, we need storage of $2n^3$ and around $4n^4$ flops. (This would suffice to find the roots of $\psi(z)$ with good enough precision.)

4.1.1 Linear recurrence equation

Consider the linear homogeneous recurrence equation of order n over a field $\mathbb{F}$.

$$u_m = a_1u_{m-1} + a_2u_{m-2} + \cdots + a_nu_{m-n}, \quad m = n, n+1, \dots \tag{4.1.13}$$

The initial conditions are the values $u_0,\dots,u_{n-1}$. Given the initial conditions, then the value of each u_m for $m \geq n$ are determined recursively by (4.1.13). Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ a_n & a_{n-1} & a_{n-2} & a_{n-3} & \dots & a_2 & a_1 \end{bmatrix} \in \mathbb{F}^{n\times n}, \mathbf{x}_k = \begin{bmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+n-1} \end{bmatrix} \in \mathbb{F}^n, \tag{4.1.14}$$

where $k = 0, 1, \dots$. Hence, the linear homogeneous recurrence equation (4.1.13) is equivalent to the homogeneous linear system relation

$$\mathbf{x}_k = A\mathbf{x}_{k-1}, \quad k = 1, 2, \dots. \tag{4.1.15}$$

Then, the formula for u_m, $m \geq n$, can be obtained from the last coordinate of $\mathbf{x}_{m-n+1} = A^{m-n+1}\mathbf{x}_0$.
Thus, we could use the results of this section to find u_m. Note that this can also be used as a tool to compute a formula for Fibonacci numbers as discussed in subsection 3.1.1. See also subsection 6.2.

4.1.2 Worked-out Problems

1. Compute the components of $A = \begin{bmatrix} 2 & 2 & -2 & 4 \\ -4 & -3 & 4 & -6 \\ 1 & 1 & -1 & 2 \\ 2 & 2 & -2 & 4 \end{bmatrix}$.
(Assume that its minimal polynomial is $\psi(z) = z(z-1)^2$.)
Solution:
We have:

$$\lambda_1 = 0,\ \theta_1(z) = (z-1)^2,\ \theta_1(\lambda_1) = 1,\ \ \lambda_2 = 1,\ \theta_2(z) = z,\ \theta_2(\lambda_2) = 1.$$

Next, (4.1.10) yields

$$\phi_{10} = \frac{\theta_1(z)z^0}{\theta_1(0)} = (z-1)^2, \quad \phi_{21} = \frac{\theta_2(z)(z-1)^1}{\theta_2(1)} = z(z-1)$$

Also, (4.1.9) yields that $\phi_{20} = z(1 + a_1(z-1))$. The value of a_1 is determined by the condition $\phi'_{20}(1) = 0$. Hence, $a_1 = -1$. So $\phi_{20} = z(2-z)$. To

$$Z_{10} = (A-I)^2, \quad Z_{20} = A(2I - A), \quad Z_{21} = A(A-I).$$

Use (4.1.3) to deduce

$$f(A) = f(0)(A-I)^2 + f(1)A(A-2I) + f'(1)A(A-I), \tag{4.1.16}$$

for any polynomial $f(z)$.

4.1.3 Problems

1. Let $A \in \mathbb{C}^{n\times n}$ and assume that $\det(zI_n - A) = \prod_{i=1}^{k}(z-\lambda_i)^{n_i}$, and the minimal polynomial $\psi(z) = \prod_{i=1}^{k}(z-\lambda_i)^{m_i}$, where $\lambda_1,\dots,\lambda_k$ are k distinct eigenvalues of A. Let $Z_{ij}, j = 0,\dots,m_i-1, i = 1,\dots,k$ are the A-components.

   (a) Show that $Z_{ij}Z_{pq} = \mathbf{0}$ for $i \neq p$.

   (b) What is the exact formula for $Z_{ij}Z_{ip}$?

4.2 Power stability, convergence and boundedness of matrices

Let $A \in \mathbb{C}^{n\times n}$ Assume that the minimal polynomial $\psi(z)$ is given by (4.1.2) and denote by $Z_{ij}, i = 1,\dots,k, j = 0,\dots,m_j-1$ the A-components. Using Theorem 4.1.2 obtain

$$A^l = \sum_{i=1}^{k}\sum_{j=0}^{m_i-1}\binom{l}{j}\lambda_i^{\max(l-j,0)}Z_{ij}, \tag{4.2.1}$$

for each positive integer l.

If we know the A-components, then to compute A^l we need only around $2mn^2 \le 2n^3$ flops! Thus, we need at most $4n^4$ flops to compute A^l, including the computations of A-components, without dependence on l! (Note that $\lambda_i^j = e^{\log j\lambda_i}$.) So to find $\mathbf{x}_{10^8} = A^{10^8}\mathbf{x}_0$, discussed in the beginning of the previous section, we need about 10^4 flops. So to compute $\mathbf{x}_{10^8}$, we need about $10^4 10^2$ flops compared with $10^8 10^2$ flops using the simple-minded algorithm explained in the beginning of the previous section. There are much simpler algorithms to compute A^l which are roughly of the order $(\log_2 l)^2 n^3$ of computations and $(\log_2 l)^2 n^2$ $(4n^2)$ storage.

Definition 4.2.1 *Let $A \in \mathbb{C}^{n\times n}$. A is called power stable if $\lim_{l\to\infty} A^l = 0$. Also, A is called power convergent if $\lim_{l\to\infty} A^l = B$, for some $B \in \mathbb{C}^{n\times n}$. A is called power bounded if there exists $K > 0$ such that the absolute value of every entry of every $A^l, l = 1,\dots$ is bounded above by K.*

Theorem 4.2.2 *Let $A \in \mathbb{C}^{n\times n}$. Then*

1. *A is power stable if and only if each eigenvalue of A is in the interior of the unit disk: $|z| < 1$.*

2. *A is power convergent if and only if each eigenvalue λ of A satisfies one of the following conditions*

   (a) $|\lambda| < 1$;

   (b) $\lambda = 1$ *and each Jordan block of the JCF of* A *with an eigenvalue* 1 *is of order* 1*, i.e.* 1 *is a simple zero of the minimal polynomial of* A.

3. *A is power bounded if and only if each eigenvalue λ of A satisfies one of the following conditions*

   (a) $|\lambda| < 1$;

   (b) $|\lambda| = 1$ *and each Jordan block of the JCF of* A *with an eigenvalue* λ *is of order* 1*, i.e.* λ *is a simple zero of the minimal polynomial of* A.

(Clearly, power convergence implies power boundness.)

Proof. Consider the formula (4.2.1). Since the A-components $Z_{ij}, i = 1, \ldots, k, j = 0, \ldots, m_i - 1$ are linearly independent, we need to satisfy the conditions of the theorem for each term in (4.2.1), which is $\binom{l}{j}\lambda_i^{l-j}Z_{ij}$ for $l >> 1$. Note that for a fixed j, $\lim_{l\to\infty}\binom{l}{j}\lambda_i^{l-j} = 0$ if and only if $|\lambda_i| < 1$. Hence, we deduce the condition *1* of the theorem.

Note that the sequence $\binom{l}{j}\lambda_i^{l-j}, l = j, j+1, \ldots,$ converges if and only if either $|\lambda_i| < 1$ or $\lambda_i = 1$ and $j = 0$. Hence, we deduce the condition *2* of the theorem.

Note that the sequence $\binom{l}{j}\lambda_i^{l-j}, l = j, j+1, \ldots,$ is bounded if and only if either $|\lambda_i| < 1$ or $|\lambda_i| = 1$ and $j = 0$. Thus, we deduce the condition *3* of the theorem.

□

Corollary 4.2.3 *Let $A \in \mathbb{C}^{n\times n}$ and consider the iterations $\mathbf{x}_l = A\mathbf{x}_{l-1}$ for $l = 1, \ldots$. Then for any $\mathbf{x}_0$*

1. $\lim_{l\to\infty}\mathbf{x}_l = \mathbf{0}$ *if and only if A is power stable.*

2. $\mathbf{x}_l, l = 0, 1, \ldots$ *converges if and only if A is power convergent.*

3. $\mathbf{x}_l, l = 0, 1, \ldots$ *is bounded if and only if A is power bounded.*

Proof. If A satisfies the conditions of an item of Theorem 4.2.2, then the corresponding condition of the corollary clearly holds. Assume that the conditions of an item of the corollary holds. Choose $\mathbf{x}_0 = \mathbf{e}_j = (\delta_{1j}, \ldots, \delta_{nj})^\top$, for $j = 1, \ldots, n$ to deduce the corresponding condition of Theorem 4.2.2. □

Theorem 4.2.4 *Let $A \in \mathbb{C}^{n\times n}$ and consider the nonhomogeneous iterations*

$$\mathbf{x}_l = A\mathbf{x}_{l-1} + \mathbf{b}_l, \quad l = 0, \ldots \tag{4.2.2}$$

Then, we have the following statements:

1. $\lim_{l\to\infty}\mathbf{x}_l = \mathbf{0}$ *for any $\mathbf{x}_0 \in \mathbb{C}^n$ and any sequence $\mathbf{b}_0, \mathbf{b}_1, \ldots$ satisfying the condition $\lim_{l\to\infty}\mathbf{b}_l = 0$ if and only if A is power stable.*

2. *The sequence $\mathbf{x}_l, l = 0, 1, \ldots$ converges for any $\mathbf{x}_0$ and any sequence $\mathbf{b}_0, \mathbf{b}_1, \ldots$ satisfying the condition $\sum_{l=0}^{l}\mathbf{b}_l$ converges if and only if A is power convergent.*

3. *The sequence* $\mathbf{x}_l, l = 0, 1, \ldots$ *is bounded for any* $\mathbf{x}_0$ *and any sequence* $\mathbf{b}_0, \mathbf{b}_1, \ldots$ *satisfying the condition* $\sum_{l=0}^{l} \|\mathbf{b}_l\|_\infty$ *converges if and only if* A *is power bounded. (Here,* $\|(x_1, \ldots, x_n)\|_\infty = \max_{i\in[n]} |x_i|$*.)*

Proof. Assume that $\mathbf{b}_l = 0$. Since $\mathbf{x}_0$ is arbitrary, we deduce the necessity of all the conditions from Theorem 4.2.2. The sufficiency of the above conditions follows from the Jordan canonical form of A:

Let $J = U^{-1}AU$, where U is an invertible matrix and J is the Jordan canonical form of A. Letting $\mathbf{y}_l := U^{-1}\mathbf{x}_l$ and $\mathbf{c}_l = U^{-1}\mathbf{b}_l$, it is enough to prove the sufficiency part of the theorem for the case where A is sum of Jordan blocks. In this case the system (4.2.2) reduces to independent systems of equations for each Jordan block. Thus, it is left to prove the theorem when $A = J_n(\lambda)$.

1. We show that if $A = J_n(\lambda)$ and $|\lambda| < 1$, then $\lim_{l\to\infty} \mathbf{x}_l = \mathbf{0}$, for any $\mathbf{x}_0$ and $\mathbf{b}_l, l = 1, \ldots$ if $\lim_{l\to\infty} \mathbf{b}_l = 0$. We prove this claim by the induction on n. For $n = 1$, (4.2.2) is reduced to

$$x_l = \lambda x_{l-1} + b_l, \quad x_0, x_l, b_l \in \mathbb{C} \text{ for } l = 1, \ldots \tag{4.2.3}$$

It is straightforward to show, e.g. use induction that

$$x_l = \sum_{i=0}^{l} \lambda^i b_{l-i} = b_l + \lambda b_{l-1} + \ldots + \lambda^l b_0 \quad l = 1, \ldots, \text{ were } b_0 := x_0. \tag{4.2.4}$$

Let $\beta_m = \sup_{i\ge m} |b_i|$. Since $\lim_{l\to\infty} b_l = 0$, it follows that each β_m is finite, the sequence $\beta_m, m = 0, 1, \ldots$ decreasing and $\lim_{m\to\infty} \beta_m = 0$. Fix m. Then, for $l > m$ we have:

$$|x_l| \le \sum_{i=0}^{l} |\lambda|^i |b_{l-i}| = \sum_{i=0}^{l-m} |\lambda|^i |b_{l-i}| + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \le$$
$$\beta_m \sum_{i=0}^{l-m} |\lambda|^i + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \le \beta_m \sum_{i=0}^{\infty} |\lambda|^i + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| =$$
$$\frac{\beta_m}{1-|\lambda|} + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \to \frac{\beta_m}{1-|\lambda|} \text{ as } l \to \infty.$$

That is, $\limsup_{l\to\infty} |x_l| \le \frac{\beta_m}{1-|\lambda|}$. As $\lim_{m\to\infty} \beta_m = 0$, it follows that $\limsup_{l\to\infty} |x_l| = 0$, which is equivalent to the statement $\lim_{l\to\infty} x_l = 0$. This proves the case $n = 1$.

Assume that the theorem holds for $n = k$. Let $n = k + 1$. View $\mathbf{x}_l^\top := (x_{1,l}, \mathbf{y}_l^\top,)^\top, \mathbf{b}_l = (b_{1,l}, \mathbf{c}_l^\top,)^\top$, where $\mathbf{y}_l = (x_{2,l}, \ldots, x_{k+1,l})^\top, \mathbf{c}_l \in \mathbb{C}^k$ are the vectors composed of the last k coordinates of $\mathbf{x}_l$ and $\mathbf{b}_l$, respectively. Then (4.2.2) for $A = J_{k+1}(\lambda)$ for the last k coordinates of $\mathbf{x}_l$ is given by the system $\mathbf{y}_l = J_k(\lambda)\mathbf{y}_{l-1} + \mathbf{c}_l$ for $l = 1, 2, \ldots$. Since $\lim_{l\to\infty} \mathbf{c}_l = \mathbf{0}$, the induction hypothesis yields that $\lim_{l\to\infty} \mathbf{y}_l = \mathbf{0}$. The system (4.2.2) for $A = J_{k+1}(\lambda)$ for the first coordinate is $x_{1,l} = \lambda x_{1,l-1} + (x_{2,l-1} + b_{1,l})$, for $l = 1, 2, \ldots$. From induction hypothesis and the assumption that $\lim_{l\to\infty} \mathbf{b}_l = \mathbf{0}$, we deduce that $\lim_{l\to\infty} x_{2,l-1} + b_{1,l} = 0$. Hence, from the case $k = 1$, we deduce that $\lim_{l\to\infty} x_{1,l} = 0$. Therefore, $\lim_{l\to\infty} \mathbf{x}_l = 0$. The proof of this case is concluded.

2. Assume that each eigenvalue λ of A satisfies the following conditions: either $|\lambda| < 1$ or $\lambda = 1$ and each Jordan block corresponding to 1 is of order 1. As we pointed out, we assume that A is a direct sum of its Jordan form. So first we consider $A = J_k(\lambda)$ with $|\lambda| < 1$. Since we assumed that $\sum_{l=1}^{\infty} \mathbf{b}_l$ converges, we deduce that $\lim_{l\to\infty} \mathbf{b}_l = \mathbf{0}$. Thus, by part *1* we get that $\lim_{l\to\infty} \mathbf{x}_l = 0$.

   Assume now that $A = (1) \in \mathbb{C}^{1\times 1}$. Thus, we consider (4.2.3) with $\lambda = 1$. Then, (4.2.4) yields that $x_l = \sum_{i=0}^{l} b_l$. By the assumption of the theorem, $\sum_{i=1}^{\infty} \mathbf{b}_l$ converges, hence the sequence $x_l, l = 1, \ldots$ converges.

3. As in the part *2*, it is enough to consider the case $J_1(\lambda)$ with $|\lambda| = 1$. Note that (4.2.4) yields that $|x_l| \le \sum_{i=0}^{l} |b_i|$. The assumption that $\sum_{i=1}^{\infty} |\mathbf{b}_i|$ converges implies that $|x_l| \le \sum_{i=0}^{\infty} |b_i| < \infty$.

□

Remark 4.2.5 *The stability, convergence and boundedness of the nonhomogeneous systems:*

$$\mathbf{x}_l = A_l \mathbf{x}_{l-1},\ A_l \in \mathbb{C}^{n\times n}, \quad l = 1, \ldots,$$
$$\mathbf{x}_l = A_l \mathbf{x}_{l-1} + \mathbf{b}_l,\ A_l \in \mathbb{C}^{n\times n},\ \mathbf{b}_l \in \mathbb{C}^n \quad l = 1, \ldots,$$

are much harder to analyze.

4.3 e^{At} and stability of certain systems of ODE

The exponential function e^z has the Maclaurin expansion

$$e^z = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \ldots = \sum_{l=0}^{\infty} \frac{z^l}{l!}.$$

Hence, for each $A \in \mathbb{C}^{n\times n}$ one defines

$$e^A := I_n + A + \frac{A^2}{2} + \frac{A^3}{6} + \ldots = \sum_{l=0}^{\infty} \frac{A^l}{l!}.$$

More generally, if $t \in \mathbb{C}$ then

$$e^{At} := I_n + At + \frac{A^2t^2}{2} + \frac{A^3t^3}{6} + \ldots = \sum_{l=0}^{\infty} \frac{A^l t^l}{l!}.$$

Therefore, e^{At} satisfies the matrix differential equation

$$\frac{de^{At}}{dt} = Ae^{At} = e^{At}A. \tag{4.3.1}$$

Also, one has the standard identity $e^{At}e^{Au} = e^{A(t+u)}$ for any complex numbers t, u.

Proposition 4.3.1 *Let $A \in \mathbb{C}^{n\times n}$ and consider the system of linear system of n ordinary differential equations with constant coefficients $\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t)$, where $\mathbf{x}(t)= (x_1(t),\dots,x_n(t))^\top \in \mathbb{C}^n$, satisfying the initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$. Then, $\mathbf{x}(t) = e^{A(t-t_0)}\mathbf{x}_0$ is the unique solution to the above system. More generally, let $\mathbf{b}(t) = (b_1(t),\dots,b_n(t))^\top \in \mathbb{C}^n$ be any continuous vector function on $\mathbb{R}$ and consider the nonhomogeneous system of n ordinary differential equations with the initial condition:*

$$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + \mathbf{b}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{4.3.2}$$

Then, this system has a unique solution of the form

$$\mathbf{x}(t) = e^{A(t-t_0)}\mathbf{x}_0 + \int_{t_0}^{t} e^{A(t-u)}\mathbf{b}(u)du. \tag{4.3.3}$$

Proof. The uniqueness of the solution of (4.3.2) follows from the uniqueness of solutions to system of ODE (Ordinary Differential Equations). The first part of the proposition follows from (4.3.1). To deduce the second part, one does the *variations of parameters*. Namely, one tries a solution $x(t) = e^{A(t-t_0)}\mathbf{y}(t)$, where $\mathbf{y}(t) \in \mathbb{C}^n$ is unknown vector function. Hence

$$\mathbf{x}' = (e^{A(t-t_0)})'\mathbf{y}(t)+e^{A(t-t_0)}\mathbf{y}'(t) = Ae^{A(t-t_0)}\mathbf{y}(t)+e^{A(t-t_0)}\mathbf{y}'(t) = A\mathbf{x}(t)+e^{A(t-t_0)}\mathbf{y}'(t).$$

Substitute this expression of $\mathbf{x}(t)$ to (4.3.2) to deduce the differential equation $\mathbf{y}' = e^{-A(t-t_0)}\mathbf{b}(t)$. Since $\mathbf{y}(t_0) = \mathbf{x}_0$, this simple equation has a unique solution $\mathbf{y}(t) = \mathbf{x}_0 + \int_{t_0}^{u} e^{A(u-t_0)}\mathbf{b}(u)du$. Now, multiply by $e^{A(t-t_0)}$ and use the fact that $e^{At}e^{Au} = e^{A(u+v)}$ to deduce (4.3.3). □

Use (4.1.3) for e^{zt} and the observation that $\frac{d^j e^{zt}}{dz^j} = t^j e^{zt}, j = 0,1,\dots$ to deduce:

$$e^{At} = \sum_{j=1}^{k}\sum_{j=0}^{m_i-1}\frac{t^j e^{\lambda_i t}}{j!}Z_{ij}. \tag{4.3.4}$$

We can substitute this expression for e^{At} in (4.3.3) to get a simple expression of the solution $\mathbf{x}(t)$ of (4.3.2).

Definition 4.3.2 *Let $A \in \mathbb{C}^{n\times n}$. A is called exponentially stable, or simple stable, if $\lim_{t\to\infty} e^{At} = \mathbf{0}$. A is called exponentially convergent if $\lim_{t\to\infty} e^{At} = B$, for some $B \in \mathbb{C}^{n\times n}$. A is called exponentially bounded if there exists $K > 0$ such that the absolute value of every entry of every $e^{At}, t \in [0,\infty)$ is bounded above by K.*

Theorem 4.3.3 *Let $A \in \mathbb{C}^{n\times n}$. Then*

1. *A is stable if and only if each eigenvalue of A is in the left half of the complex plane: $\Re z < 0$.*

2. *A is exponentially convergent if and only if each eigenvalue λ of A satisfies one of the following conditions*

   (a) $\Re\lambda < 0$;

(b) *$\lambda = 2\pi l i$, for some integer l, and each Jordan block of the JCF of A with an eigenvalue λ is of order 1, i.e. λ is a simple zero of the minimal polynomial of A.*

3. *A is exponentially bounded if and only if each eigenvalue λ of A satisfies one of the following conditions*

   (a) *$\Re\lambda < 0$;*

   (b) *$\Re\lambda = 0$ and each Jordan block of the JCF of A with an eigenvalue λ is of order 1, i.e. λ is a simple zero of the minimal polynomial of A.*

Proof. Consider the formula (4.3.4). Since the A-components $Z_{ij}, i = 1, \dots, k, j = 0, \dots, m_i - 1$ are linearly independent, we need to satisfy the conditions of the theorem for each term in (4.3.4), which is $\frac{t^j}{j!} e^{\lambda_i t} Z_{ij}$. Note that for a fixed j, $\lim_{t\to\infty} \frac{t^j}{j!} e^{\lambda_i t} = 0$ if and only if $\Re\lambda_i < 0$. Hence, we deduce the condition *1* of the theorem.
Note that the function $\frac{t^j}{j!} e^{\lambda_i t}$ converges as $t \to \infty$ if and only if either $\Re\lambda_i < 0$ or $e^{\lambda_i} = 1$ and $j = 0$. Hence, we deduce the condition *2* of the theorem.
Note that the function $\frac{t^j}{j!} e^{\lambda_i t}$ is bounded for $t \geq 0$ if and only if either $\Re\lambda_i < 0$ or $|e^{\lambda_i}| = 1$ and $j = 0$. Hence, we deduce the condition *3* of the theorem. □

Corollary 4.3.4 *Let $A \in \mathbb{C}^{n\times n}$ and consider the system of differential equations $\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t), \mathbf{x}(t_0) = \mathbf{x}_0$. Then, for any $\mathbf{x}_0$*

1. *$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$ if and only if A is stable.*

2. *$\mathbf{x}(t)$ converges as $t \to \infty$ if and only if A is exponentially convergent.*

3. *$\mathbf{x}(t), t \in [0, \infty)$ is bounded if and only if A is exponentially bounded.*

Theorem 4.3.5 *Let $A \in \mathbb{C}^{n\times n}$ and consider the system of differential equations (4.3.2). Then, for any $\mathbf{x}_0 \in \mathbb{C}^n$, we have the following statements:*

1. *$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$, for any continuous function $\mathbf{b}(t)$, such that $\lim_{t\to\infty} \mathbf{b}(t) = \mathbf{0}$, if and only if A is stable.*

2. *$\mathbf{x}(t)$ converges as $t \to \infty$, for any continuous function $\mathbf{b}(t)$, such that $\int_{t_0}^{\infty} \mathbf{b}(u)du$ converges, if and only if A is exponentially convergent.*

3. *$\mathbf{x}(t), t \in [0, \infty)$ is bounded for any continuous function $\mathbf{b}(t)$, such that $\int_{t_0}^{\infty} |\mathbf{b}(u)|du$ converges if and only if A is exponentially bounded.*

Proof. It is left as Problem 4.3.2-2.

4.3.1 Worked-out Problems

1. Let $B = [b_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and assume that each entry of B in non-negative. Assume that there exists a vector $\mathbf{u} = (u_1, \dots, u_n)^\top$ with positive coordinates such that $B\mathbf{u} = \mathbf{u}$.

   (a) Show that $B^k = [b_{ij}^{(k)}]_{i,j=1}^n$ has non-negative entries for each $k \in \mathbb{N}$.

(b) Show that $b_{ij}^{(k)} \leq \frac{u_i}{u_j}$, for $i, j = 1, \ldots, n$ and $k \in \mathbb{N}$.

(c) Show that each eigenvalue of B satisfies $|\lambda| \leq 1$.

(d) Suppose that λ is an eigenvalue of B and $|\lambda| = 1$. What is the multiplicity of λ in the minimal polynomial of B?

Solution:

(a) It is immediate as It is clear since B is a non-negative matrix.

(b) Since $B\mathbf{u} = \mathbf{u}$, by induction on k, we can show that $B^k\mathbf{u} = \mathbf{u}$, for any $k \in \mathbb{N}$. Now, we have $u_i = \sum_{p=1}^{n} b_{ip}^{(k)} u_p \geq b_{ij}^{(k)} u_j$, for all $1 \leq i, j \leq n$. Then $b_{ij} \leq \frac{u_i}{u_j}$.

(c) According to part (b), B is power bounded. The statement is immediate from the third part of Theorem 4.2.2.

(d) According to the third part of Theorem 4.2.2, λ would be a simple zero of the minimal polynomial of B.

4.3.2 Problems

1. Consider the nonhomogeneous system $\mathbf{x}_l = A_l \mathbf{x}_{l-1}$, $A_l \in \mathbb{C}^{n\times n}$, $l = 1, \ldots$. Assume that the sequence $A_l, l = 1, \ldots,$ is periodic, i.e. $A_{l+p} = A_l$ for all $l = 1, \ldots,$ and a fixed positive integer p.

   (a) Show that for each $\mathbf{x}_0 \in \mathbb{C}^n$ $\lim_{l\to\infty} \mathbf{x}_l = \mathbf{0}$ if and only if $B := A_p A_{p-1} \ldots A_1$ is power stable.

   (b) Show that for each $\mathbf{x}_0 \in \mathbb{C}^n$ the sequence $\mathbf{x}_l, l = 1, \ldots,$ converges if and only if the following conditions satisfy. First, B is power convergent, i.e. $\lim_{l\to\infty} B^l = C$. Second, $A_i C = C$, for $i = 1, \ldots, p$.

   (c) Find a necessary and sufficient conditions such that for each $\mathbf{x}_0 \in \mathbb{C}^n$ the sequence $\mathbf{x}_l, l = 1, \ldots,,$ is bounded.

2. Prove Theorem 4.3.5.

3. For any $A \in \mathbb{F}^{n\times n}$, prove that $\det e^A = e^{\operatorname{tr} A}$.

4. Let $A \in \mathbb{R}^{n\times n}$ and $P \in \mathrm{GL}(n, \mathbb{R})$. Prove that $e^{P^{-1}AP} = P^{-1}e^A P$.

Chapter 5

Inner product spaces

Through this chapter, we assume that $\mathbb{F}$ is either the real field or complex field and $\mathbf{V}$ is a vector space over $\mathbb{F}$. (Unless stated otherwise.) Assume that $T \in L(\mathbf{U}, \mathbf{V})$. We abbreviate $T(\mathbf{x})$ to $T\mathbf{x}$ and no ambiguity will arise.

5.1 Inner product

In this section, we shall study a certain type of scalar-valued function on pairs of vectors, known as an inner product. The first example of an inner product is the dot product in $\mathbb{R}^3$. The dot product of $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3)$ in $\mathbb{R}^3$ is the scalar $\sum_{i=1}^{3} x_i y_i$. Geometrically, this scalar is the product of the length of $\mathbf{x}$, the length of $\mathbf{y}$ and the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$. Then, we can define length and angle in $\mathbb{R}^3$ algebraically. An inner product on a vector space is the generalization of the dot product. This section deals with inner product and its properties. Then, we may turn to discuss length and angle (orthogonality).

Definition 5.1.1 *The function $\langle \cdot, \cdot \rangle : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ is called an inner product if the following conditions hold for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$ and $a \in \mathbb{F}$:*

(i) Conjugate symmetry:
$\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$,

(ii) Linearity in the first argument:
$\langle a\mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$,

(iii) Positive-definiteness:
$\langle \mathbf{x}, \mathbf{x} \rangle \geqslant 0$
$\langle \mathbf{x}, \mathbf{x} \rangle = 0$ *if and only if* $\mathbf{x} = 0$.

If the second condition in positive-definiteness is dropped, the resulting structure is called a semi-inner product. Other standard properties of inner products are mentioned in Problems 5.7.2-1 and 5.7.2-2.
The vector space $\mathbf{V}$ endowed with the inner product $\langle \cdot, \cdot \rangle$ is called the inner product space. We will use the abbreviation IPS for inner product space.

Example 5.1.2 *On $\mathbb{R}^n$ we have the standard inner product (the dot product) defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} \mathbf{x}_i, \mathbf{y}_i$, where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$.*

Note that inner product generalizes the notion of the dot product of vectors in $\mathbb{R}^n$.

Example 5.1.3 *On $\mathbb{C}^n$ we have the standard inner product defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n \mathbf{x}_i, \overline{\mathbf{y}_i}$, where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{C}^n$ and $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{C}^n$.*

Example 5.1.4 *Let $\mathbf{V} = \mathbb{F}^{n \times n}$. For $A = [a_{ij}]$ and $B = [b_{ij}] \in \mathbf{V}$, define inner product*

$$\langle A, B \rangle = \sum_{i,j} a_{ij} \overline{b_{ij}}.$$

Define conjugate transpose $B^* = (\overline{B})^\top$. Then

$$\langle A, B \rangle = \operatorname{tr} B^* A.$$

Example 5.1.5 (Integration) *Let $\mathbf{V}$ be the vector space of all $\mathbb{F}$-value continuous functions on $[0,1]$. For $f, g \in \mathbf{V}$, define*

$$\langle f, g \rangle = \int_0^1 f \overline{g} dt.$$

This is an inner product on $\mathbf{V}$. In some context, this is called L^2 inner product space. This can be done in any "space" where you have an idea of integration and it comes under Measure Theory.

Matrix of Inner Product. Let $\{e_1, \ldots, e_n\}$ be a basis of $\mathbf{V}$. Let $p_{ij} = \langle e_i, e_j \rangle$ and $P = [p_{ij}] \in \mathbb{F}^{n \times n}$. Then, for $v = x_1 e_1 + \ldots + x_n e_n \in \mathbf{V}$ and $w = y_1 e_1 + \cdots + y_n e_n \in \mathbf{V}$ we have

$$\langle v, w \rangle = \sum \mathbf{x}_i \overline{\mathbf{y}_j} p_{ij} = (x_1, \ldots, x_n) P \begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ \vdots \\ \overline{y_n} \end{pmatrix}.$$

This matrix P is called the *matrix of the inner product* with respect to the basis $\{e_1, \ldots, e_n\}$.

Definition 5.1.6 *A function $\| \ \| : \mathbf{V} \to \mathbb{F}$ is called a norm on $\mathbf{V}$ if it has the following properties for all $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ and $a \in \mathbb{F}$:*

(i) Positive definiteness:
$\|\mathbf{x}\| \geqslant 0$
$\|\mathbf{x}\| = 0$ *if and only if* $\mathbf{x} = 0$,

(ii) Homogeneity:
$\|a\mathbf{x}\| = |a| \|\mathbf{x}\|$,

(iii) Triangle inequality:
$\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

The vector space $\mathbf{V}$ endowed with the norm $\| \ \|$ is called the normed linear space.

Note that the notion of norm generalizes the notion of length of a vector in $\mathbb{R}^n$.

Example 5.1.7 *On $\mathbb{R}^n$ we have the following norms for $\mathbf{x} = (x_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^n$:*

(1) $\|\mathbf{x}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$,

(2) $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$,

(3) $\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$, *(p is a real number, $p \geq 1$).*

The interested reader is encouraged to verify the above properties to see that (1), (2) and (3) define norms.

Note that $\| \, \|_p$ is called the *ℓ_p norm.* The ℓ_2 norm is called the *Euclidean norm.* The Euclidean norm on the matrix space $\mathbb{F}^{m\times n}$ is called the *Frobenius norm*:

$$\|A\|_{\mathbb{F}} = \left(\sum_{i,j} |a_{ij}|^2\right)^{\frac{1}{2}}, \quad A = [a_{ij}] \in \mathbb{F}^{m\times n}.$$

Let $\| \, \|_t$ be a norm on $\mathbb{F}^n$ and $A \in \mathbb{F}^n$. Define $\|A\| = \max\{\|A\mathbf{x}\|_t; \, \|\mathbf{x}\|_t = 1, \, \mathbf{x} \in \mathbb{F}^n\}$. Then $\| \, \|$ is a norm called *operator norm* induced by $\| \, \|_t$.
The operator norm on $\mathbb{F}^{m\times n}$ induced by Euclidean norm is called the *spectral norm*, denoted $\| \, \|_\infty$:

$$\|A\|_\infty = \max\{\|A\mathbf{x}\|_2; \, \|\mathbf{x}\|_2 = 1, \, \mathbf{x} \in \mathbb{F}^n\}.$$

Remark 5.1.8 *If $\langle \, , \, \rangle$ is an inner product, the function $\| \, \| : \mathbf{V} \to \mathbb{F}$ defined as $\|\mathbf{x}\| = (\langle \mathbf{x}, \mathbf{x} \rangle)^{\frac{1}{2}}$ is a norm (Why? Use Theorem 5.1.9 to justify it.). This norm is called the induced norm by the inner product $\langle \, , \, \rangle$. Also, $\|\mathbf{x}\|$ is called the length of* $\mathbf{x}$.

Also, if $\| \, \| : \mathbf{V} \to \mathbb{R}$ is a norm, the metric space $(\mathbf{V}, d)$ defined as $d(x, y) = \|x - y\|$ is called the *induced metric by the norm $\| \, \|$*.

Theorem 5.1.9 (The Cauchy-Schwarz inequality) *Let $\mathbf{V}$ be an IPS and $\mathbf{x}, \mathbf{y} \in \mathbf{V}$. Then we have*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|y\|,$$

with equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. Here $\| \, \|$ is the induced norm by inner product.

Proof. If either $\mathbf{x}$ or $\mathbf{y}$ is zero, the result follows. Assume that $\mathbf{x}, \mathbf{y} \neq 0$. Then, for any real number $r \in \mathbb{R}$,

$$\begin{aligned} 0 \leq \|\mathbf{x} + r\mathbf{y}\|^2 &= \langle \mathbf{x} + r\mathbf{y}, \mathbf{x} + r\mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + r\langle \mathbf{x}, \mathbf{y} \rangle + r\langle \mathbf{y}, \mathbf{x} \rangle + r^2\langle \mathbf{y}, \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + r\overline{\langle \mathbf{y}, \mathbf{x} \rangle} + r\langle \mathbf{x}, \mathbf{y} \rangle + r^2\langle \mathbf{y}, \mathbf{y} \rangle \\ &\leq \langle \mathbf{x}, \mathbf{x} \rangle + 2r|\langle \mathbf{y}, \mathbf{x} \rangle| + r^2\langle \mathbf{y}, \mathbf{y} \rangle = f(r). \end{aligned}$$

This implies that the quadratic polynmial $f(r)$ must have non-positive discriminant, that is,

$$4|\langle \mathbf{y}, \mathbf{x} \rangle|^2 - 4\langle \mathbf{y}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{x} \rangle \leq 0,$$

from which the Cauchy-Schwarz inequality follows. Moreover, if equality holds, then there exists an $r \in \mathbb{F}$ such that $f(r) = 0$, that is, $0 = \|\mathbf{x} + r\mathbf{y}\|^2$, and so $\mathbf{x} + r\mathbf{y} = 0$, which implies that $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$. □

Proposition 5.1.10 *Let $\mathbf{V}$ be a real vector space. Identify $\mathbf{V}_c$ with the set of pairs $(\mathbf{x}, \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbf{V}$. Then, $\mathbf{V}_c$ is a vector space over $\mathbb{C}$ with*

$$(a + ib)(\mathbf{x}, \mathbf{y}) := a(\mathbf{x}, \mathbf{y}) + b(-\mathbf{y}, \mathbf{x}), \quad \text{for all } a, b \in \mathbb{R},\ \mathbf{x}, \mathbf{y} \in \mathbf{V}.$$

If $\mathbf{V}$ has a basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ over $\mathbb{F}$, then $\{(\mathbf{e}_1, \mathbf{0}), ..., (\mathbf{e}_n, \mathbf{0})\}$ is a basis of $\mathbf{V}_c$ over $\mathbb{C}$. Any inner product $\langle \cdot, \cdot \rangle$ on $\mathbf{V}$ over $\mathbb{R}$ induces the following inner product on $\mathbf{V}_c$:

$$\langle (\mathbf{x}, \mathbf{y}), (\mathbf{u}, \mathbf{v}) \rangle = \langle \mathbf{x}, \mathbf{u} \rangle + \langle \mathbf{y}, \mathbf{v} \rangle + i(\langle \mathbf{y}, \mathbf{u} \rangle - \langle \mathbf{x}, \mathbf{v} \rangle), \ \textit{for } \mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in \mathbf{V}.$$

We leave the proof of this proposition to the reader as Problem 5.7.2-3.

Definition 5.1.11 (Angle) *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be non-zero and $\langle \cdot, \cdot \rangle$ denote the standard inner product. Since $|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\| \|\mathbf{y}\|$, we can define the angle between $\mathbf{x}$, $\mathbf{y}$ as follows:*

$$\angle(\mathbf{x}, \mathbf{y}) = \arccos \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

In particular, vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal (denoted by $\mathbf{x} \perp \mathbf{y}$) if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. We can define this notion in any inner product space as follows:

Definition 5.1.12 *Let $\mathbf{V}$ be an IPS. Then*
(i) $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ are called orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.
(ii) $S, T \subset \mathbf{V}$ are called orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, for any $\mathbf{x} \in S$, $\mathbf{y} \in T$.
(iii) For any $S \subset \mathbf{V}$, $S^\perp \subset \mathbf{V}$ denotes the maximal orthogonal set to S, i.e. $S^\perp = \{\mathbf{v} \in \mathbf{V};\ \langle \mathbf{v}, \mathbf{w} \rangle = 0, \text{ for all } \mathbf{w} \in S\}$.
(iv) $\{\mathbf{x}_1, ..., \mathbf{x}_m\} \subset \mathbf{V}$ is called an orthonormal set if

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{ij}, \quad (i, j = 1, ..., m),$$

where δ_{ij} denotes the Kronecker delta.
(v) $\{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbf{V}$ is called an orthonormal basis if it is an orthonormal set which is a basis in $\mathbf{V}$.

Note that $S^\perp$ is always a subspace of $\mathbf{V}$ (even if A was not). Furthermore,

(i) $\{0\}^\perp = \mathbf{V}$,

(ii) $\mathbf{V}^\perp = \{0\}$,

(iii) $S^\perp = (\text{span} S)^\perp$.

Gram-Schmidt algorithm. Let $\mathbf{V}$ be an IPS and $S = \{\mathbf{x}_1, ..., \mathbf{x}_m\} \subset \mathbf{V}$ a finite (possibly empty) set. Then, $\tilde{S} = \{\mathbf{e}_1, ..., \mathbf{e}_p\}$ is the orthonormal set ($p \geq 1$) or the empty set ($p = 0$) obtained from S using the following recursive steps:
(a) If $\mathbf{x}_1 = 0$, remove it from S. Otherwise, replace $\mathbf{x}_1$ by $\|\mathbf{x}_1\|^{-1}\mathbf{x}_1$.
(b) Assume that $\{\mathbf{x}_1, ..., \mathbf{x}_k\}$ is an orthonormal set and $1 \leq k < m$. Let $\mathbf{y}_{k+1} = \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{x}_{k+1}, \mathbf{x}_i \rangle \mathbf{x}_i$. If $\mathbf{y}_{k+1} = 0$, remove $\mathbf{x}_{k+1}$ from S. Otherwise, replace $\mathbf{x}_{k+1}$ by $\|\mathbf{y}_{k+1}\|^{-1}\mathbf{y}_{k+1}$.

Indeed, the Gram-Schmidt process is a method for orthonormalization of a set of vectors in an inner product space.

Corollary 5.1.13 *Let $\mathbf{V}$ be an IPS and $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbf{V}$ be n linearly independent vectors. Then, the Gram-Schmidt algorithm on S is given as follows:*

$$\begin{aligned}
&\mathbf{y}_1 := \mathbf{x}_1,\ r_{11} := \|\mathbf{y}_1\|,\ \mathbf{e}_1 := \frac{\mathbf{y}_1}{r_{11}}, \\
&r_{ji} := \langle \mathbf{x}_i, \mathbf{e}_j \rangle,\ j = 1, ..., i-1, \\
&\mathbf{p}_{i-1} := \sum_{j=1}^{i-1} r_{ji}\mathbf{e}_j, \quad \mathbf{y}_i := \mathbf{x}_i - \mathbf{p}_{i-1}, \\
&r_{ii} := \|\mathbf{y}_i\|,\ \mathbf{e}_i := \frac{\mathbf{y}_i}{r_{ii}},\ i = 2, ..., n.
\end{aligned} \tag{5.1.1}$$

In particular, $\mathbf{e}_i \in S_i$ and $\|\mathbf{y}_i\| = \operatorname{dist}(\mathbf{x}_i, S_{i-1})$, where $S_i = \operatorname{span}\{\mathbf{x}_1, ..., \mathbf{x}_i\}$, for $i = 1, ..., n$ and $S_0 = \{\mathbf{0}\}$. (See Problem 5.7.2-4 for the definition of $\operatorname{dist}(\mathbf{x}_i, S_{i-1})$.)

Corollary 5.1.14 *Any (ordered) basis in a finite dimensional IPS, $\mathbf{V}$ induces an orthonormal basis by the Gram-Schmidt algorithm.*

See Problem 5.7.2-4 for some known properties related to the above notions.

5.2 Explanation of G-S process in standard Euclidean space

First observe that $r_{i(k+1)} := \langle \mathbf{x}_{k+1}, \mathbf{e}_i \rangle$ is the scalar projection of $\mathbf{x}_{k+1}$ on $\mathbf{e}_i$. Next observe that $\mathbf{p}_k$ is the projection of $\mathbf{x}_{k+1}$ on $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\} = \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Hence, $\mathbf{y}_{k+1} = \mathbf{x}_{k+1} - \mathbf{p}_k \perp \operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. Thus, $r_{(k+1)(k+1)} = \|\mathbf{x}_{k+1} - \mathbf{p}_k\|$ is the distance of $\mathbf{x}_{k+1}$ to $\operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\} = \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. The assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent yields that $r_{(k+1)(k+1)} > 0$. Hence, $\mathbf{e}_{k+1} = r_{(k+1)(k+1)}^{-1}(\mathbf{x}_{k+1} - \mathbf{p}_k)$ is a vector of unit length orthogonal to $\mathbf{e}_1, \ldots, \mathbf{e}_k$.

5.3 An example of G-S process

Let $\mathbf{x}_1 = (1,1,1,1)^\top, \mathbf{x}_2 = (-1,4,4,-1)^\top$, $\mathbf{x}_3 = (4,-2,2,0)^\top$. Then

$$r_{11} = \|\mathbf{x}_1\| = 2, \mathbf{e}_1 = \frac{1}{r_{11}}\mathbf{x}_1 = (\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2})^\top,$$
$$r_{12} = \mathbf{e}_1^\top\mathbf{x}_2 = 3, \mathbf{p}_1 = r_{12}\mathbf{e}_1 = 3\mathbf{e}_1 = (\frac{3}{2},\frac{3}{2},\frac{3}{2},\frac{3}{2})^\top,$$
$$\mathbf{y}_2 = \mathbf{x}_2 - \mathbf{p}_1 = (-\frac{5}{2},\frac{5}{2},\frac{5}{2},-\frac{5}{2})^\top, r_{22} = \|\mathbf{y}_2\| = 5,$$
$$\mathbf{e}_2 = \frac{1}{r_{22}}(\mathbf{y}_2) = (-\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2})^\top,$$
$$r_{13} = \mathbf{e}_1^\top\mathbf{x}_3 = 2, r_{23} = \mathbf{e}_2^\top\mathbf{x}_3 = -2,$$
$$\mathbf{p}_2 = r_{13}\mathbf{e}_1 + r_{23}\mathbf{e}_2 = (2,0,0,2)^\top,$$
$$\mathbf{y}_3 = \mathbf{x}_3 - \mathbf{p}_2 = (2,-2,2,-2)^\top, r_{33} = \|\mathbf{y}_3\| = 4,$$
$$\mathbf{e}_3 = \frac{1}{r_{33}}(\mathbf{x}_3 - \mathbf{p}_2) = (\frac{1}{2},-\frac{1}{2},\frac{1}{2},-\frac{1}{2})^\top.$$

5.4 QR Factorization

A QR Factorization of a real matrix A is its decomposition into a product A =QR of unitary matrix Q and an upper triangular matrix R. QR factorization is often used to solve the least square problem. (Theorem 5.5.7)

Let $A = [\mathbf{a}_1\ \mathbf{a}_2 \dots \mathbf{a}_n] \in \mathbb{R}^{m\times n}$ and assume that rank $A = n$, i.e. the columns of A are linearly independent. Perform G-S process with the book keeping as above:

- $r_{11} := \|\mathbf{a}_1\|$, $\mathbf{e}_1 := \frac{1}{r_{11}}\mathbf{a}_1$.
- Assume that $\mathbf{e}_1, \dots, \mathbf{e}_{k-1}$ were computed. Then, $r_{ik} := \mathbf{e}_i^\top\mathbf{a}_k$ for $i = 1, \dots, k-1$, $\mathbf{p}_{k-1} := r_{1k}\mathbf{e}_1 + r_{2k}\mathbf{e}_2 + \dots r_{(k-1)k}\mathbf{e}_{k-1}$ and $r_{kk} := \|\mathbf{a}_k - \mathbf{p}_{k-1}\|$, $\mathbf{e}_k := \frac{1}{r_{kk}}(\mathbf{a}_k - \mathbf{p}_{k-1})$, for $k = 2, ..., n$.

Let $Q = [\mathbf{e}_1\ \mathbf{e}_2 \dots \mathbf{e}_n] \in \mathbb{R}^{m\times n}$ and $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & r_{22} & r_{23} & \dots & r_{2n} \\ 0 & 0 & r_{33} & \dots & r_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & r_{nn} \end{bmatrix}$

Then, $A = QR$, $Q^\top Q = I_n$ and $A^\top A = R^\top R$. The LSS (Least Squares Solution) of $A\mathbf{x} = \mathbf{b}$ is given by the upper triangular system $R\hat{\mathbf{x}} = Q^\top\mathbf{b}$ which can be solved by back substitution. Formally, $\hat{\mathbf{x}} = R^{-1}Q^\top\mathbf{b}$. (See Theorem 5.5.7.)

Proof. $A^\top A\mathbf{x} = R^\top Q^\top QR\mathbf{x} = R^\top R\mathbf{x} = A^\top\mathbf{b} = R^\top Q^\top\mathbf{b}$. Multiply from left by $(R^\top)^{-1}$ to get $R\hat{\mathbf{x}} = Q^\top\mathbf{b}$

Note that $QQ^\top\mathbf{b}$ is the projection of $\mathbf{b}$ on the columns space of A. The matrix $P := QQ^\top$ is called an *orthogonal projection*. It is symmetric and $P^2 = P$, as $(QQ^\top)(QQ^\top) = Q(Q^\top Q)Q^\top = Q(I)Q^\top = QQ^\top$. Note that $QQ^\top : \mathbb{R}^m \to \mathbb{R}^m$ is the orthogonal projection.

The assumption that rank $A = n$ is equivalent to the assumption that $A^\top A$ is invertible. So, the LSS $A^\top A\hat{x} = A^\top\mathbf{b}$ has unique solution $\hat{\mathbf{x}} = (A^\top A)^{-1}\mathbf{b}$. Hence, the projection of $\mathbf{b}$ on the column space of A is $P\mathbf{b} = A\hat{\mathbf{x}} = A(A^\top A)^{-1}A^\top\mathbf{b}$. Therefore, $P = A(A^\top A)^{-1}A^\top$.

5.5 An example of QR algorithm

Let $A = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3] = \begin{bmatrix} 1 & -1 & 4 \\ 1 & 4 & -2 \\ 1 & 4 & 2 \\ 1 & -1 & 0 \end{bmatrix}$ be the matrix corresponding to the Example of G-S algorithm §5.3. Then

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} 2 & 3 & 2 \\ 0 & 5 & -2 \\ 0 & 0 & 4 \end{bmatrix},$$

$$Q = [\mathbf{q}_1\ \mathbf{q}_2\ \mathbf{q}_3] = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

(*Explain why in this example $A = QR$.*) Note that $QQ^\top : \mathbb{R}^4 \to \mathbb{R}^4$ is the projection on $\mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.

Remark 5.5.1 *It is known, e.g. [8] that the Gram-Schmidt process as described in Corollary 5.1.13 is numerically unstable. That is, there is a severe loss of orthogonality of $\mathbf{y}_1, \ldots$ as we proceed to compute $\mathbf{y}_i$. In computations, one uses either a modified GSP or Householder orthogonalization [8].*

Definition 5.5.2 (**Modified Gram-Schmidt algorithm**.) *Let $\mathbf{V}$ be an IPS and $S = \{\mathbf{x}_1, ..., \mathbf{x}_m\} \subset \mathbf{V}$ a finite (possibly empty) set. Then, $\tilde{S} = \{\mathbf{e}_1, ..., \mathbf{e}_p\}$ is either the orthonormal set ($p \geq 1$) or the empty set ($p = 0$) obtained from S using the following recursive steps:*

1. *Initialize $j = 1$ and $p = m$.*
2. *If $\mathbf{x}_j \neq \mathbf{0}$ goto 3. Otherwise, replace p by $p - 1$. If $j > p$ exit. Replace $\mathbf{x}_i$ by $\mathbf{x}_{i+1}$, for $i = j, \ldots, p$. Repeat.*
3. *$\mathbf{e}_j := \frac{1}{\|\mathbf{x}_j\|}\mathbf{x}_j$.*
4. *$j = j + 1$. If $j > p$ exit.*
5. *For $i = 1, \ldots, j - 1$, let $\mathbf{x}_j := \mathbf{x}_j - \langle \mathbf{x}_j, \mathbf{e}_i \rangle \mathbf{e}_i$.*
6. *Goto 2.*

MGS algorithm is stable, needs mn^2 flops, which is more time consuming than GS algorithm.

Lemma 5.5.3 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{R}$. Let $\mathbf{U}$ be a subspace of $\mathbf{V}$. Then*

$$\mathbf{V} = \mathbf{U} \oplus \mathbf{U}^\perp \tag{5.5.1}$$

Proof. If $\mathbf{U} = \mathbf{V}$ or $\mathbf{U} = \{\mathbf{0}\}$, the lemma is trivial. So we assume that $n = \dim \mathbf{V} > m = \dim \mathbf{U} \geq 1$. Choose a basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for $\mathbf{U}$. Complete this basis to a basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for $\mathbf{V}$. Perform the Gram-Schmidt process on $\mathbf{u}_1, \dots, \mathbf{u}_n$ to obtain an orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. Recall that $\{\mathbf{u}_1, \dots, \mathbf{u}_i\} = \mathrm{span}\{\mathbf{v}_1, \dots, \mathbf{v}_i\}$, for $i = 1, \dots, n$. Hence, $\mathbf{U} = \mathrm{span}\{\mathbf{u}_1, \dots, \mathbf{u}_m\} = \mathrm{span}\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, i.e. $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is an orthonormal basis in $\mathbf{U}$. Clearly, $\mathbf{y} = \sum_{i=1}^{n} \langle \mathbf{y}, \mathbf{v}_i \rangle \mathbf{v}_i \in \mathbf{U}^{\perp}$ if and only if $\mathbf{y} = \sum_{i=m+1}^{n} a_i \mathbf{v}_i$, i.e. $\{\mathbf{v}_{m+1}, \dots, \mathbf{v}_n\}$ is a basis in $\mathbf{U}^{\perp}$. So when we write a vector $\mathbf{z} = \sum_{i=1}^{n} z\mathbf{v}_i$, it is of unique form $\mathbf{u} + \mathbf{w}, \mathbf{u} = \sum_{i=1}^{m} z_i \mathbf{v}_i \in \mathbf{U}, \mathbf{w} = \sum_{i=m+1}^{n} z_i \mathbf{v}_i \in \mathbf{U}^{\perp}$. □

Corollary 5.5.4 *Let the assumptions of Lemma 5.5.3 hold. Then,* $(\mathbf{U}^{\perp})^{\perp} = \mathbf{U}$.

Lemma 5.5.5 *For* $A \in \mathbb{F}^{n \times m}$*:*

a) $N(A^{\top}) = R(A)^{\perp}$,

b) $N(A^{\top})^{\perp} = R(A)$.

The proof is left as an exercise.

Lemma 5.5.6 *(Fredholm alternative) Let* $A \in \mathbb{C}^{m \times n}$*. Then, the system*

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \in \mathbb{C}^n, \ \mathbf{b} \in \mathbb{C}^m \tag{5.5.2}$$

is solvable if and only if for each $\mathbf{y} \in \mathbb{C}^m$ *satisfying* $\mathbf{y}^{\top} A = \mathbf{0}$*, the equality* $\mathbf{y}^{\top}\mathbf{b} = 0$ *holds.*

Proof. Suppose that (5.5.2) solvable. Then, $\mathbf{y}^{\top}\mathbf{b} = \mathbf{y}^{\top}A\mathbf{x} = \mathbf{0}^{\top}\mathbf{x} = 0$. Observe next that $\mathbf{y}^{\top}A = \mathbf{0}^{\top}$ if and only if $\mathbf{y}^{\top} \in R(A)^{\perp}$. As $(R(A)^{\perp})^{\perp} = R(A)$, it follows that if $\mathbf{y}^{\top}\mathbf{b} = 0$, for each $\mathbf{y}^{\top}A = \mathbf{0}$, then $\mathbf{b} \in R(A)$, i.e. (5.5.2) is solvable. □

Assume that (5.5.2) has no solution. One may seek the best approximation to a solution. Finding the best approximate solution to an inconsistent linear system (system with no solution) is the basis of a “least square solution”.

Theorem 5.5.7 *(The least squares theorem) Consider the system* (5.5.2)*. Multiply the both sides of this system by* A^* *to obtain the least squares system corresponding to* (5.5.2)*:*

$$A^*A\mathbf{x} = A^*\mathbf{b}, \quad A \in \mathbb{C}^{m \times n}, \mathbf{b} \in \mathbb{C}^m. \tag{5.5.3}$$

Then, (5.5.3) *is always solvable. For each* $\mathbf{x}_0$*, the vector* $A\mathbf{x}_0$ *is the orthogonal projection of* $\mathbf{b}$ *on* $R(A)$*. That is,* $A\mathbf{x}_0 \in R(A)$ *and* $\mathbf{b} - A\mathbf{x}_0 \in R(A)^{\perp}$*. Furthermore,*

$$\|\mathbf{b} - \mathbf{z}\| \geq \|\mathbf{b} - A\mathbf{x}_0\|, \text{ for any } \mathbf{z} \in R(A). \tag{5.5.4}$$

Equality holds if and only if $\mathbf{z} = A\mathbf{x}_0$.

Proof. Suppose that $\mathbf{z}^{\top}A^*A = \mathbf{0}$. Then, $0 = \mathbf{z}^{\top}A^*A\bar{\mathbf{z}} = \|\mathbf{z}^{\top}A^*\|^2$. Hence, $\mathbf{z}^{\top}A^* = \mathbf{0}$. In particular $\mathbf{z}^{\top}A^*\mathbf{b} = 0$. Lemma 5.5.6 yields that (5.5.3) is solvable. Let $\mathbf{x}_0$ be a solution of (5.5.3). So, $A^*(\mathbf{b} - A\mathbf{x}_0) = \mathbf{0}$. Since the columns of A span $R(A)$, it follows that $\mathbf{b} - A\mathbf{x}_0 \in R(A)^{\perp}$. Clearly, $A\mathbf{x}_0 \in R(A)$. Hence, $A\mathbf{x}_0$ is the orthogonal projection on the range of A. Let $\mathbf{z} \in R(A)$. Use the condition that $\mathbf{b} - A\mathbf{x}_0 \in R(A)^\perp$ to deduce

$$\|\mathbf{b}-\mathbf{z}\|^2 = \|(\mathbf{b}-A\mathbf{x}_0)+(A\mathbf{x}_0-\mathbf{z})\|^2 = \|\mathbf{b}-A\mathbf{x}_0\|^2 + \|A\mathbf{x}_0-\mathbf{z}\|^2 \geq \|\mathbf{b}-A\mathbf{x}_0\|^2. \quad (5.5.5)$$

Equality holds if and only if $\mathbf{z} = A\mathbf{x}_0$. □

5.6 The best fit line

A *line of best fit* is a straight line that is the best approximation of the given set of data. We explain this notion by the following problem more precisely.

Problem: Fit a straight line $y = a + bx$ in the $X - Y$ plane through m given points $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$.

Solution: The line should satisfy m conditions:

$$\begin{array}{ccccccccc} 1\cdot a & + & x_1\cdot b & = & y_1 \\ 1\cdot a & + & x_2\cdot b & = & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1\cdot a & + & x_m\cdot b & = & y_m \end{array} \Rightarrow \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \mathbf{y} = \mathbf{c}.$$

$$A \qquad \mathbf{z} \quad = \quad \mathbf{c}, \ \mathbf{z} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

The least squares system $A^\top A\mathbf{z} = A^\top \mathbf{c}$:

$$\begin{bmatrix} m & x_1 + x_2 + \cdots + x_m \\ x_1 + x_2 + \cdots + x_m & x_1^2 + x_2^2 + \cdots + x_m^2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + \cdots + y_m \\ x_1y_1 + x_2y_2 + \cdots + x_my_m \end{bmatrix},$$

$$\det A^\top A = m(x_1^2 + x_2^2 + \cdots + x_m^2) - (x_1 + x_2 + \cdots + x_m)^2,$$

$$\det A^\top A = 0 \ \text{ if and only if } \ x_1 = x_2 = \cdots = x_m.$$

If $\det A^\top A \neq 0$, then

$$a^* = \frac{(\sum_{i=1}^m x_i^2)(\sum_{i=1}^m y_i) - (\sum_{i=1}^m x_i)(\sum_{i=1}^m x_iy_i)}{\det A^\top A},$$

$$b^* = \frac{-(\sum_{i=1}^m x_i)(\sum_{i=1}^m y_i) + m(\sum_{i=1}^m x_iy_i)}{\det A^\top A}.$$

We now explain the solution for the best fit line. We are given m points in the plane $(x_1, y_1), \ldots, (x_m, y_m)$. We are trying to fit a line $y = bx + a$ through these m points. Suppose we chose the parameters $a, b \in \mathbb{R}$. Then, this line passes through the point $(x_i, bx_i + a)$, for $i = 1, \ldots, m$. The square of the distance between the points (x_i, y_i) and $(x_i, bx_i + a)$ is $(y_i - (1\cdot a + x_i \cdot b))^2$. The sum of the squares of distances is $\sum_{i=1}^m (y_i - (1\cdot a + x_i\cdot b))^2$. Note that this sum is $\|\mathbf{y} - A\mathbf{z}\|^2$, where A, $\mathbf{z}$, $\mathbf{y}$ are as above. So $A\mathbf{z} \in R(A)$. Hence, $\min_{\mathbf{z}\in\mathbb{R}^2} \|\mathbf{y} - A\mathbf{z}\|^2$ is achieved for the least square solution $\mathbf{z}^* = (a^*, b^*)$ given as above, (if not all x_i are equal.) So the line $y = a^* + b^*x$ is the best fit line.

Example 5.6.1 *Given three points in* $\mathbb{R}^2$: $(0,1), (3,4), (6,5)$. *Find the best least square fit by a linear function* $y = a + bx$ *to these three points.*

Solution.

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 6 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ 4 \\ 5 \end{bmatrix},$$

$$\mathbf{z} = (A^\top A)^{-1} A^\top \mathbf{c} = \begin{bmatrix} 3 & 9 \\ 9 & 45 \end{bmatrix}^{-1} \begin{bmatrix} 10 \\ 42 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} \\ \frac{2}{3} \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

The best least square fit by a linear function is $y = \frac{4}{3} + \frac{2}{3}x$.

5.7 Geometric interpretation of the determinant (Second encounter)

Definition 5.7.1 *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ be m given vectors. Then, the parallelepiped $\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is defined as follows. The 2^m vertices of $\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ are of the form $\mathbf{v} := \sum_{i=1}^m a_i \mathbf{x}_i$, where $a_i = 0, 1$ for $i = 1, \ldots, m$. Two vertices $\mathbf{v} = \sum_{i=1}^m a_i \mathbf{x}_i$ and $\mathbf{w} = \sum_{i=1}^m b_i \mathbf{x}_i$ of $\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ are adjacent, i.e. connected by an edge in $\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, if $\|(a_1, \ldots, a_m)^\top - (b_1, \ldots, b_m)^\top\| = 1$, i.e. the $0-1$ coordinates of $(a_1, \ldots, a_m)^\top$ and $(b_1, \ldots, b_m)^\top$ differ only at one coordinate k, for some $k \in [m]$.*

Note that if $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the standard basis in $\mathbb{R}^n$, i.e. $\mathbf{e}_i = (\delta_{1i}, \ldots, \delta_{ni})^\top, i = 1, \ldots, n$, then $\mathrm{P}(\mathbf{e}_1, \ldots, \mathbf{e}_m)$ is the m-dimensional unit cube, whose edges are parallel to $\mathbf{e}_1, \ldots, \mathbf{e}_m$ and its center (of gravity) is $\frac{1}{2}(\underbrace{1, \ldots, 1}, 0, \ldots, 0)^\top$, where 1 appears m times for $1 \le m \le n$.

For $m > n$, $\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is "flattened" parallelepiped, since $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are always linearly dependent in $\mathbb{R}^n$ for $m > n$.

Assuming that the volume element generated by $\mathbf{e}_1, \ldots, \mathbf{e}_n$, denoted as $\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n$ is positive, then the volume element of $\mathbf{e}_{\sigma(1)} \wedge \cdots \wedge \mathbf{e}_{\sigma(n)}$ has the sign (orientation) of the permutation σ.

Proposition 5.7.2 *Let $A \in \mathbb{R}^{n \times n}$ and view $A = [\mathbf{c}_1\ \mathbf{c}_2 \ldots \mathbf{c}_n]$ as an ordered set of n vectors, (columns), $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Then $|\det A|$ is the n-dimensional volume of the parallelepiped $\mathrm{P}(\mathbf{c}_1, \ldots, \mathbf{c}_n)$. If $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are linearly independent, then the orientation in $\mathbb{R}^n$ induced by $\mathbf{c}_1, \ldots, \mathbf{c}_n$ is the same as the orientation induced by $\mathbf{e}_1, \ldots, \mathbf{e}_n$ if $\det A > 0$, and is the opposite orientation if $\det A < 0$.*

Proof. $\det A = 0$ if and only if the columns of A are linearly dependent. If $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are linearly dependent, then $\mathrm{P}(\mathbf{c}_1, \ldots, \mathbf{c}_n)$ lies in a subspace of $\mathbb{R}^n$, i.e. some $n-1$ dimensional subspace, and hence the n-dimensional volume of $\mathrm{P}(\mathbf{c}_1, \ldots, \mathbf{c}_n)$ is zero.

Assume now that $\det A \neq 0$, i.e. $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are linearly independent. Perform Gram-Schmidt process. Then, $A = QR$, where $Q = [\mathbf{e}_1\ \mathbf{e}_2 \ldots \mathbf{e}_n]$ is an orthogonal matrix and $R = [r_{ji}] \in \mathbb{R}^{n \times n}$ is an upper diagonal matrix. (See Problem 5.7.2-5.) So $\det A = \det Q \det R$. Since $Q^\top Q = I_n$, we deduce that $1 = \det I_n = \det Q^\top \det Q = \det Q \det Q = (\det Q)^2$. So $\det Q = \pm 1$ and the sign of $\det Q$ is the sign of $\det A$.

Hence, $|\det A| = \det R = r_{11} r_{22} \ldots r_{nn}$. Recall that r_{11} is the length of the vector $\mathbf{c}_1$, and r_{ii} is the distance of the vector $\mathbf{e}_i$ to the subspace spanned by $\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}$, for $i = 2, \ldots, n$. (See Problem 5.7.2-4, parts f, g and i.) Thus, the length of $\mathrm{P}(\mathbf{c}_1)$ is

r_{11}. The distance of $\mathbf{c}_2$ to $\mathrm{P}(\mathbf{c}_1)$ is r_{22}. Thus, the area, i.e. 2-dimensional volume of $\mathrm{P}(\mathbf{c}_1, \mathbf{c}_2)$ is $r_{11}r_{22}$. Continuing in this manner we deduce that the $i-1$ dimensional volume of $\mathrm{P}(\mathbf{c}_1, \dots, \mathbf{c}_{i-1})$ is $r_{11}\dots r_{(i-1)(i-1)}$. As the distance of $\mathbf{c}_i$ to $\mathrm{P}(\mathbf{c}_1, \dots, \mathbf{c}_{i-1})$ is r_{ii}, it follows that the i-dimensional volume of $\mathrm{P}(\mathbf{c}_1, \dots, \mathbf{c}_i)$ is $r_{11}\dots r_{ii}$. For $i = n$, we get that $|\det A| = r_{11}\dots r_{nn}$, which is equal to the n-dimensional volume of $\mathrm{P}(\mathbf{c}_1, \dots, c_n)$.

As we already pointed out, the sign of $\det A$ is equal to the sign of $\det Q = \pm 1$. If $\det Q = 1$, it is possible to "rotate" the standard basis in $\mathbb{R}^n$ to the basis given by the columns of an orthogonal matrix Q with $\det Q = 1$. If $\det Q = -1$, we need one reflection, i.e. replace the standard basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ by the new basis $\{\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3, \dots, \mathbf{e}_n\}$ and rotate the new basis $\{\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3, \dots, \mathbf{e}_n\}$ to the basis consisting of the columns of an orthogonal matrix Q', where $\det Q' = 1$. □

Theorem 5.7.3 ***(The Hadamard determinant inequality)*** *Let $A = [\mathbf{c}_1, \dots, \mathbf{c}_n] \in \mathbb{C}^{n\times n}$. Then, $|\det A| \le \|\mathbf{c}_1\|\,\|\mathbf{c}_2\|\dots\|\mathbf{c}_n\|$. Equality holds if and only if either $\mathbf{c}_i = \mathbf{0}$, for some i or $\langle \mathbf{c}_i, \mathbf{c}_j\rangle = 0$, for all $i \ne j$, i.e. $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ is an orthogonal system.*

Proof. Assume first that $\det A = 0$. Clearly, the Hadamard inequality holds. Equality in Hadamard inequality holds if and only if $\mathbf{c}_i = \mathbf{0}$, for some i.

Assume now that $\det A \ne 0$ and perform the Gram-Schmidt process. From (5.1.1), it follows that $A = QR$, where Q is a unitary matrix, i.e. $Q^*Q = I_n$ and $R = [r_{ji}] \in \mathbb{C}^{n\times n}$ is upper triangular with r_{ii} real and positive numbers. So $\det A = \det Q \det R$. Thus

$$1 = \det I_n = \det Q^*Q = \det Q^* \det Q = \overline{\det Q} \det Q = |\det Q|^2 \Rightarrow |\det Q| = 1.$$

Hence, $|\det A| = \det R = r_{11}r_{22}\dots r_{nn}$. According to Problem 5.7.2-4 and the proof of Proposition 5.7.2, we know that $\|\mathbf{c}_i\| \ge \mathrm{dist}(\mathbf{c}_i, \mathrm{span}\{\mathbf{c}_1, \dots, \mathbf{c}_{i-1}\}) = r_{ii}$, for $i = 2, \dots, n$. Hence, $|\det A| = \det R \le \|\mathbf{c}_1\|\,\|\mathbf{c}_2\|\dots\|\mathbf{c}_n\|$. Equality holds if
$\|\mathbf{c}_i\| = \mathrm{dist}(\mathbf{c}_i, \mathrm{span}\{\mathbf{c}_1, \dots, \mathbf{c}_{i-1}\})$, for $i = 2, \dots, n$. Use Problem 5.7.2-4 to deduce that $\|\mathbf{c}_i\| = \mathrm{dist}(\mathbf{c}_i, \mathrm{span}\{\mathbf{c}_1, \dots, \mathbf{c}_{i-1}\})$ if an only if $\langle \mathbf{c}_i, \mathbf{c}_j\rangle = 0$, for $j = 1, \dots, i-1$. Use these conditions for $i = 2, \dots$ to deduce that equality in Hadamard inequality holds if and only if $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ is an orthogonal system. □

An application of Fredholm alternative

Let $A = [a_{ij}] \in \mathbb{Z}_2^{n\times n}$ be a symmetric matrix. It has been proven in [5] that $\mathrm{diag}\, A \in \mathrm{Im}\, A$. Here, we give Noga Alon's short proof for this statement:

$$\mathbf{x}^\top A\mathbf{x} = \sum_{i,j=1}^n a_{ij}x_ix_j = 2\sum_{1\le i<j\le n} a_{ij}x_ix_j + \mathbf{x}^\top \mathrm{diag}\, A = \mathbf{x}^\top \mathrm{diag}\, A.$$

Then, if $\mathbf{x}^\top A = 0$, so $x^\top \mathrm{diag}\, A = 0$. The Fredholm alternative implies that $\mathrm{diag}\, A$ in the range of A.

5.7.1 Worked-out Problems

1. Let $A = [a_{ij}]_{i,j=1}^n \in \mathbb{C}^{n\times n}$ such that $|a_{ij}| \le 1$, for $i, j = 1, \ldots, n$. Show that $|\det A| = n^{\frac{n}{2}}$ if and only if $A^*A = AA^* = nI_n$.
Solution:
Assume first that $|\det A| = n^{\frac{n}{2}}$ and set $\mathbf{c}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}$. Then, Theorem 5.7.3 and the assumption $|a_{ij}| \le 1$ yield $\|\mathbf{c}_i\| = \sqrt{n}$, for any $1 \le i \le n$. Reusing the assumption $|a_{ij}| \le 1$ follows $|a_{ij}| = 1$, for $i, j = 1, \ldots, n$. Finally, as $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ is an orthogonal system, then $A^*A = AA^* = nI_n$.
Conversely, assume that $A^*A = AA^* = nI_n$. We have $|\det A|^2 = \det A \det A^* = \det AA^* = \det nI_n = n^n$. Then $|\det A| = n^{\frac{n}{2}}$.

2. Let $\mathbf{u} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} 2 \\ 0 \\ -2 \\ 1 \end{bmatrix}$ and $\mathbf{U} = span\{\mathbf{u}, \mathbf{v}\}$. Find an orthogonal projection of the matrix $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ on $\mathbf{U}$ using least squares with a corresponding matrix.
Solution:
Define $A = \begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 1 & -2 \\ -1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$. The least squares theorem implies that the system $A^*A\mathbf{x} = A^*b$ is solvable, $\mathbf{x}_0 = \frac{1}{35}\begin{bmatrix} 8 \\ 2 \end{bmatrix}$ and $A\mathbf{x}_0 \in R(A)$, i.e. $A\mathbf{x}_0 = \frac{1}{35}\begin{bmatrix} 18 \\ -2 \\ -14 \\ -6 \end{bmatrix}$ is an orthogonal projection of b.

3. Find $A = [a_{ij}] \in \mathbb{C}^{3\times 3}$ satisfying $|a_{ij}| \le 1$ and $|\det A|^2 = 27$.
Solution:
Define $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \xi_1 & \xi_2 \\ 1 & \xi_1^2 & \xi_2^2 \end{bmatrix}$, where $\xi_k = e^{\frac{2k\pi}{3}i} = \cos\frac{2k\pi}{3} + i\sin\frac{2k\pi}{3}$, $k = 1, 2$. Clearly, $|a_{ij}| \le 1$ and $AA^* = A^*A = 3I_n$. Using Worked-out Problem 5.7.1-1, we obtain $|\det A| = 3^{\frac{3}{2}} = \sqrt{27}$.

5.7.2 Problems

1. Let $\mathbf{V}$ be an IPS over $\mathbb{F}$. Show that

(a) $\langle \mathbf{0}, \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{0}\rangle = 0$,

(b) for $\mathbb{F} = \mathbb{R}$, $\langle \mathbf{z}, a\mathbf{x} + b\mathbf{y}\rangle = a\langle \mathbf{z}, \mathbf{x}\rangle + b\langle \mathbf{z}, \mathbf{y}\rangle$, for all $a, b \in \mathbb{R}$, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$,

(c) for $\mathbb{F} = \mathbb{C}$, $\langle \mathbf{z}, a\mathbf{x} + b\mathbf{y}\rangle = \bar{a}\langle \mathbf{z}, \mathbf{x}\rangle + \bar{b}\langle \mathbf{z}, \mathbf{y}\rangle$, for all $a, b \in \mathbb{C}$, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$.

2. Let $\mathbf{V}$ be an IPS. Show that

(a) $||a\mathbf{x}|| = |a|\ ||\mathbf{x}||$, for $a \in \mathbb{F}$ and $\mathbf{x} \in \mathbf{V}$.

(b) The *triangle inequality*

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

and equality holds if either $\mathbf{x} = 0$ or $\mathbf{y} = a\mathbf{x}$, for some non-negative $a \in \mathbb{R}$. (Hint: Use the Cauchy-Schwarz inequality.)

(c) Pythagorean law; if $\mathbf{x} \perp \mathbf{y}$, then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|y\|^2.$$

(d) Parallelogram identity;

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2.$$

3. Prove Proposition 5.1.10.

4. Let $\mathbf{V}$ be a finite dimensional IPS of dimension n. Assume that $S \subset \mathbf{V}$.

(a) Show that if $\{\mathbf{x}_1, ..., \mathbf{x}_m\}$ is an orthonormal set, then $\mathbf{x}_1, ..., \mathbf{x}_m$ are linearly independent.

(b) Assume that $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ is an orthonormal basis in $\mathbf{V}$. Show that for any $\mathbf{x} \in \mathbf{V}$, the orthonormal expansion holds

$$\mathbf{x} = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{e}_i\rangle \mathbf{e}_i. \tag{5.7.1}$$

Furthermore, for any $\mathbf{x}, \mathbf{y} \in \mathbf{V}$

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{e}_i\rangle \overline{\langle \mathbf{y}, \mathbf{e}_i\rangle}. \tag{5.7.2}$$

(c) Assume that S is a finite set. Let $\tilde{S}$ be the set obtained by the Gram-Schmidt process. Show that $\tilde{S} = \varnothing$ if and only if span$S = \{\mathbf{0}\}$. Moreover, prove that if $\tilde{S} \ne \varnothing$, then $\{\mathbf{e}_1, ..., \mathbf{e}_p\}$ is an orthonormal basis in span S.

(d) Show that there exists an orthonormal basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ in $\mathbf{V}$ and $0 \le m \le n$ such that

$$\begin{aligned}
&\mathbf{e}_1, ..., \mathbf{e}_m \in S, \quad \text{span } S = \text{span}\{\mathbf{e}_1, ..., \mathbf{e}_m\},\\
&S^\perp = \text{span}\{\mathbf{e}_{m+1}, ..., \mathbf{e}_n\},\\
&(S^\perp)^\perp = \text{span} S.
\end{aligned}$$

(e) Assume from here to the end of the problem that S is a subspace of $\mathbf{V}$. Show $\mathbf{V} = S \oplus S^\perp$.

(f) Let $\mathbf{x} \in \mathbf{V}$ and let $\mathbf{x} = \mathbf{u} + \mathbf{v}$, for unique $\mathbf{u} \in S$, $\mathbf{v} \in S^\perp$. Let $P(\mathbf{x}) := \mathbf{u}$ be the projection of $\mathbf{x}$ on S. Show that $P : \mathbf{V} \to \mathbf{V}$ is a linear transformation satisfying

$$P^2 = P, \quad \text{Range } P = S, \quad \text{Ker } P = S^\perp.$$

(g) Show that

$$\begin{aligned}&\text{dist}(\mathbf{x}, S) := \|\mathbf{x} - P\mathbf{x}\| \le \|\mathbf{x} - \mathbf{w}\| \text{ for any } \mathbf{w} \in S,\\ &\text{and equality holds if and only if } \mathbf{w} = P\mathbf{x}.\end{aligned} \tag{5.7.3}$$

(h) Show in part (g) that equality holds if and only if $\mathbf{w} = P\mathbf{x}$.

(i) Show that $\text{dist}(\mathbf{x}, S) = \|\mathbf{x} - \mathbf{w}\|$, for some $\mathbf{w} \in S$ if and only if $\mathbf{x} - \mathbf{w}$ is orthogonal to S.

(j) Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be an orthonormal basis of S. Show that for each $\mathbf{x} \in \mathbf{V}$, $P\mathbf{x} = \sum_{i=1}^p \langle \mathbf{y}, \mathbf{e}_i \rangle \mathbf{e}_i$.

(Note that $P\mathbf{x}$ is called *the least square approximation* to $\mathbf{x}$ in the subspace S.)

5. Let $X \in \mathbb{C}^{m\times n}$ and assume that $m \ge n$ and rank $X = n$. Let $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{C}^m$ be the columns of X, i.e. $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$. Assume that $\mathbb{C}^m$ is an IPS with the standard inner product $< \mathbf{x}, \mathbf{y} >= \mathbf{y}^*\mathbf{x}$. Perform the Gram-Schmidt algorithm (1.1) to obtain the matrix $Q = [\mathbf{e}_1, ..., \mathbf{e}_n] \in \mathbb{C}^{m\times n}$. Let $R = [r_{ji}]_1^n \in \mathbb{C}^{n\times n}$ be the upper triangular matrix with r_{ji}, $j \le i$ given by (5.1.1). Show that $\bar{Q}^T Q = I_n$ and $X = QR$. (This is the QR algorithm.) Show that if in addition $X \in \mathbb{R}^{m\times n}$, then Q and R are real valued matrices.

6. Let $C \in \mathbb{C}^{n\times n}$ and assume that $\lambda_1, ..., \lambda_n$ are n eigenvalues of C counted with their multiplicities. View C as an operator $C : \mathbb{C}^n \to \mathbb{C}^n$. View $\mathbb{C}^n$ as $2n$-dimensional vector space over $\mathbb{R}$. Let $C = A + \sqrt{-1}B$, $A, B \in \mathbb{R}^{n\times n}$.

a. Then, $\hat{C} := \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$ represents the operator $C : \mathbb{C}^n \to \mathbb{C}^n$ as an operator over $\mathbb{R}$ in suitably chosen basis.

b. Show that $\lambda_1, \bar{\lambda}_1, ..., \lambda_n, \bar{\lambda}_n$ are the $2n$ eigenvalues of $\hat{C}$ counting with multiplicities.

c. Show that the Jordan canonical form of $\hat{C}$ is obtained by replacing each Jordan block $\lambda I + H$ in C by two Jordan blocks $\lambda I + H$ and $\bar{\lambda} I + H$.

7. Let $A = [a_{ij}]_{i,j} \in \mathbb{C}^{n\times n}$. Assume that $|a_{ij}| \le K$, for all $i, j = 1, \ldots, n$. Show that $|\det A| \le K^n n^{\frac{n}{2}}$.

8. Show that for each n, there exists a matrix $A = [a_{ij}]_{i,j=1}^n \in \mathbb{C}^{n\times n}$ such that $|a_{ij}| = 1$, for $i, j = 1, \ldots, n$ and $|\det A| = n^{\frac{n}{2}}$.

9. Let $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ and assume that $a_{ij} = \pm 1, i, j = 1, \ldots, n$. Show that if $n > 2$, then the assumption that $|\det A| = n^{\frac{n}{2}}$ yields that n is divisible by 4.

10. Show that for any $n = 2^m$, $m = 0, 1, \dots$ there exists $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ such that $a_{ij} = \pm 1, i, j = 1, \dots, n$ and $|\det A| = n^{\frac{n}{2}}$. (*Hint*: Try to prove by induction on m that $A \in \mathbb{R}^{2^m \times 2^m}$ can be chosen symmetric, and then construct $B \in \mathbb{R}^{2^{m+1}\times 2^{m+1}}$ using A.)

    Note: A matrix $H = [a_{ij}]_{i,j} \in \mathbb{R}^{n\times n}$ such that $a_{ij} = \pm 1$, for $i, j = 1, \dots, n$ and $|\det H| = n^{\frac{n}{2}}$ is called a *Hadamard* matrix. Hadamard matrices admit several other characterizations; an equivalent definition states that a Hadamard matrix H is an $n \times n$ matrix satisfying the identity $HH^\top = nI_n$. (See Worked-out Problem 5.7.1-1.) Also, another equivalent definition for a Hadamard matrix H is an $n \times n$ matrix with entries in $\{-1, 1\}$ such that two distinct rows or columns have inner product zero. It is conjectured that for each n divisible by 4, there exists a Hadamard matrix. It is known that a necessary condition for the existence of an $n \times n$ Hadamard matrix is that $n = 1, 2, 4k$, for some k, i.e. there exists no $n \times n$ Hadamard matrices for $n \notin \{1, 2, 4k, k \in \mathbb{N}\}$. That this condition is also sufficient is known as the Hadamard conjecture, and has been the subject of a vast amount of literature in recent decades.

11. Let $\mathbf{V}$ be an IPS and $T \in L(\mathbf{V})$. Assume that $\mathbf{v}, \mathbf{w} \in \mathbf{V}$ are two eigenvectors of T with distinct eigenvalues. Prove that $\langle \mathbf{v}, \mathbf{w}\rangle = 0$.

12. Consider rotation matrices $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \in \mathbb{R}^{2\times 2}$, $\theta \in [0, 2\pi)$. Assume that $T \in L(\mathbf{V})$ is orthogonal, where $\mathbf{V}$ is a finite dimensional real vector space. Show that there exists a basis β of $\mathbf{V}$ such that $[T]_\beta$ is block diagonal, and the blocks are either 2×2 rotation matrices or 1×1 matrices consisting of 1 or -1.
    (See Problem 1.10.2-4 and Problem 3.2.2-17).

5.8 Special transformations in IPS

Proposition 5.8.1 *Let $\mathbf{V}$ be an IPS and $T : \mathbf{V} \to \mathbf{V}$ a linear transformation. Then, there exists a unique linear transformation $T^* : \mathbf{V} \to \mathbf{V}$ such that $\langle T\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, T^*\mathbf{y}\rangle$, for all $\mathbf{x}, \mathbf{y} \in \mathbf{V}$.*

We leave its proof as Problems 5.10.2-1 and 5.10.2-2.

Definition 5.8.2 *Let $\mathbf{V}$ be an IPS and $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation. Then*
(a) T is called self-adjoint if $T^* = T$;
(b) T is called anti self-adjoint if $T^* = -T$;
(c) T is called unitary if $T^*T = TT^* = I$;
(d) T is called normal if $T^*T = TT^*$.

Note that a self-adjoint transformation (matrix) is also called Hermitian.

Denote by $\mathbf{S}(\mathbf{V})$, $\mathbf{AS}(\mathbf{V})$, $\mathbf{U}(\mathbf{V})$ and $\mathbf{N}(\mathbf{V})$ the sets of self-adjoint, anti self-adjoint, unitary and normal operators on $\mathbf{V}$, respectively.

Proposition 5.8.3 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{F}$ *with an orthonormal basis* $E = \{\mathbf{e}_1,...,\mathbf{e}_n\}$*. Let* $T : \mathbf{V} \to \mathbf{V}$ *be a linear transformation and* $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ *be the representation matrix of* T *in the basis* E*:*

$$a_{ij} = \langle T\mathbf{e}_j, \mathbf{e}_i\rangle, \quad i, j = 1,...,n. \tag{5.8.1}$$

Then, for $\mathbb{F} = \mathbb{R}$*:*

$$\begin{aligned}
&(a) \quad T^* \text{ is represented by } A^\top,\\
&(b) \quad T \text{ is self−adjoint if and only if } A = A^\top,\\
&(c) \quad T \text{ is anti self−adjoint if and only if } A = -A^\top,\\
&(d) \quad T \text{ is unitary if and only if } A \text{ is orthogonal , i.e. } AA^\top = A^\top A = I,\\
&(e) \quad T \text{ is normal if and only if } A \text{ is normal , i.e. } AA^\top = A^\top A,
\end{aligned}$$

and for $\mathbb{F} = \mathbb{C}$*:*

$$\begin{aligned}
&(a) \quad T^* \text{ is represented by } A^* \ (:= \bar{A}^\top),\\
&(b) \quad T \text{ is self−adjoint if and only if } A = A^*,\\
&(c) \quad T \text{ is anti self−adjoint if and only if } A \text{ is anti hermitian , i.e. } A = -A^*,\\
&(d) \quad T \text{ is unitary if and only if } A \text{ is unitary , i.e. } AA^* = A^*A = I,\\
&(e) \quad T \text{ is normal if and only if } A \text{ is normal , i.e. } AA^* = A^*A.
\end{aligned}$$

We leave the proof as Problem 5.10.2-3.
Let $\mathbf{V}$ be a real vector space. The *complexification* of $\mathbf{V}$ is defined by taking the tensor product of $\mathbf{V}$ with the complex field and it is denoted by $\mathbf{V}_c$.

Proposition 5.8.4 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{R}$ *and* $T \in L(\mathbf{V})$*. Let* $\mathbf{V}_c$ *be the complexification of* $\mathbf{V}$*. Then, that there exists a unique* $T_c \in L(\mathbf{V}_c)$ *such that* $T_c|\mathbf{V} = T$*. Furthermore,* T *is self-adjoint, unitary or normal if and only if* T_c *is self-adjoint, unitary or normal, respectively.*

We leave the proof as Problem 5.10.2-4.

Definition 5.8.5 *For a field* $\mathbb{F}$*, let*

$$\begin{aligned}
&\mathbf{S}(n,\mathbb{F}) := \{A \in \mathbb{F}^{n\times n} : \quad A = A^\top\},\\
&\mathbf{AS}(n,\mathbb{F}) := \{A \in \mathbb{F}^{n\times n} : \quad A = -A^\top\},\\
&\mathbf{O}(n,\mathbb{F}) := \{A \in \mathbb{F}^{n\times n} : \quad AA^\top = A^\top A = I\},\\
&\mathbf{SO}(n,\mathbb{F}) := \{A \in \mathbf{O}(n,\mathbb{F}) : \quad \det A = 1\},\\
&\mathbf{DO}(n,\mathbb{F}) := \mathbf{D}(n,\mathbb{F}) \cap \mathbf{O}(n,\mathbb{F}),\\
&\mathbf{N}(n,\mathbb{R}) := \{A \in \mathbb{R}^{n\times n} : \quad AA^\top = A^\top A\},\\
&\mathbf{N}(n,\mathbb{C}) := \{A \in \mathbb{C}^{n\times n} : \quad AA^* = A^*A\},\\
&\mathbf{H}_n := \{A \in \mathbb{C}^{n\times n} : \quad A = A^*\},\\
&\mathbf{AH}_n := \{A \in \mathbb{C}^{n\times n} : \quad A = -A^*\},\\
&\mathbf{U}_n := \{A \in \mathbb{C}^{n\times n} : \quad AA^* = A^*A = I\},\\
&\mathbf{SU}_n := \{A \in \mathbf{U}_n : \quad \det A = 1\},\\
&\mathbf{DU}_n := \mathbf{D}(n,\mathbb{C}) \cap \mathbf{U}_n.
\end{aligned}$$

See Problem 5.10.2-5 for relations between these classes.

Theorem 5.8.6 *Let $\mathbf{V}$ be an IPS over $\mathbb{C}$ of dimension n. Then, a linear transformation $T : \mathbf{V} \to \mathbf{V}$ is normal if and only if $\mathbf{V}$ has an orthonormal basis consisting of eigenvectors of T.*

Proof. Suppose first that $\mathbf{V}$ has an orthonormal basis $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ such that $T\mathbf{e}_i = \lambda_i \mathbf{e}_i$, $i = 1, ..., n$. From the definition of T^*, it follows that $T^*\mathbf{e}_i = \bar{\lambda}_i \mathbf{e}_i$, $i = 1, ..., n$. Hence, $TT^* = T^*T$.

Assume now T is normal. Since $\mathbb{C}$ is algebraically closed, T has an eigenvalue λ_1. Let $\mathbf{V}_1$ be the subspace of $\mathbf{V}$ spanned by all eigenvectors of T corresponding to the eigenvalue λ_1. Clearly, $T\mathbf{V}_1 \subset \mathbf{V}_1$. Let $\mathbf{x} \in \mathbf{V}_1$. Then, $T\mathbf{x} = \lambda_1 \mathbf{x}$. Thus

$$TT^*\mathbf{x} = (TT^*)\mathbf{x} = (T^*T)\mathbf{x} = T^*T\mathbf{x} = \lambda_1 T^*\mathbf{x} \Rightarrow T^*\mathbf{V}_1 \subset \mathbf{V}_1.$$

Hence, $T\mathbf{V}_1^\perp, T^*\mathbf{V}_1^\perp \subset \mathbf{V}_1^\perp$. Since $\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_1^\perp$, it is enough to prove the theorem for $T|\mathbf{V}_1$ and $T|\mathbf{V}_1^\perp$.

As $T|\mathbf{V}_1 = \lambda_1 I_{\mathbf{V}_1}$, it is straightforward to show $T^*|\mathbf{V}_1 = \bar{\lambda}_1 I_{\mathbf{V}_1}$ (see Problem 5.10.2-2). Hence, for $T|\mathbf{V}_1$ the theorem trivially holds. For $T|\mathbf{V}_1^\perp$ the theorem follows by induction. □

Theorem 5.8.7 is an important result of linear algebra, called *spectral theorem for normal matrices* which states that normal matrices are diagonal with respect to an orthonormal basis. See [17] for its proof and more details.

Theorem 5.8.7 *Assume that $A \in \mathbb{C}^{n\times n}$ is a normal matrix. Then, A is unitarily similar to a diagonal matrix. That is, there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$ such that $A = U\Lambda U^* = U\Lambda U^{-1}$. This is called the spectral decomposition of A. (The columns of U is an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A, and the diagonal entries of Λ are the corresponding eigenvalues of A.)*

The proof of Theorem 5.8.6 yields the following corollary.

Corollary 5.8.8 *Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ of dimension n. Then, the linear transformation $T : \mathbf{V} \to \mathbf{V}$ with a real spectrum is normal if and only if $\mathbf{V}$ has an orthonormal basis consisting of eigenvectors of T.*

Definition 5.8.9 *If T is a linear transformation, the set of all distinct eigenvalues of T is called the spectrum of T and it is denoted by* $\operatorname{spec} T$.

Proposition 5.8.10 *Let $\mathbf{V}$ be an IPS over $\mathbb{C}$ and $T \in \mathbf{N}(\mathbf{V})$. Then*

(a) T is self − adjoint *if and only if* $\operatorname{spec} T \subset \mathbb{R}$,

(b) T is unitary *if and only if* $\operatorname{spec} T \subset S^1 = \{z \in \mathbb{C} : \quad |z| = 1\}$.

Proof. Since T is normal, there exists an orthonormal basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ such that $T\mathbf{e}_i = \lambda_i \mathbf{e}_i$, $i = 1, \ldots, n$. Hence, $T^*\mathbf{e}_i = \bar{\lambda}_i \mathbf{e}_i$. Then

$$T = T^* \text{ if and only if } \lambda_i = \bar{\lambda}_i,\ i = 1, \ldots, n,$$
$$TT^* = T^*T = I \text{ if and only if } |\lambda_i| = 1,\ i = 1, \ldots, n.$$

□

Combine Proposition 5.8.4 and Corollary 5.8.8 with the above proposition to deduce the following corollary:

Corollary 5.8.11 *Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ and $T \in \mathbf{S}(\mathbf{V})$. Then,* $\operatorname{spec} T \subset \mathbb{R}$ *and $\mathbf{V}$ has an orthonormal basis consisting of the eigenvectors of T.*

Corollary 5.8.11 gives another key factor about hermitian matrices and states that all eigenvalues of a hermitian matrix must be real. Note that a matrix $A \in \mathbb{R}^{n\times n}$ with all real eigenvalues need not be hermitian. For instance, $A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}$ has eigenvalues 1 and 2 but $A \neq A^*$.

Proposition 5.8.12 *Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ and let $T \in \mathbf{U}(\mathbf{V})$. Then, $\mathbf{V} = \oplus_{i\in\{-1,1,2,\ldots,k\}} \mathbf{V}_i$, where $k \geq 1$, $\mathbf{V}_i$ and $\mathbf{V}_j$ are orthogonal for $i \neq j$, such that*
(a) $T|\mathbf{V}_{-1} = -I_{\mathbf{V}_{-1}}$ $\dim \mathbf{V}_{-1} \geq 0$,
(b) $T|\mathbf{V}_1 = I_{\mathbf{V}_1}$, $\dim \mathbf{V}_1 \geq 0$,
(c) $T\mathbf{V}_i = \mathbf{V}_i$, $\dim \mathbf{V}_i = 2$, $\operatorname{spec}(T|\mathbf{V}_i) \subset S^1 \backslash \{-1, 1\}$, *for $i = 2, \ldots, k$.*

We leave the proof as Problem 5.10.2-7.

Proposition 5.8.13 *Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ and $T \in \mathbf{AS}(\mathbf{V})$. Then, $\mathbf{V} = \oplus_{i\in\{1,2,\ldots,k\}} \mathbf{V}_i$, where $k \geq 1$, $\mathbf{V}_i$ and $\mathbf{V}_j$ are orthogonal, for $i \neq j$, such that*
(a) $T|\mathbf{V}_1 = 0_{\mathbf{V}_1}$ $\dim \mathbf{V}_0 \geq 0$,
(b) $T\mathbf{V}_i = \mathbf{V}_i$, $\dim \mathbf{V}_i = 2$, $\operatorname{spec}(T|\mathbf{V}_i) \subset \sqrt{-1}\mathbb{R} \backslash \{0\}$, *for $i = 2, \ldots, k$.*

We leave the proof as Problem 5.10.2-8.

Theorem 5.8.14 *Let $\mathbf{V}$ be an IPS over $\mathbb{C}$ of dimension n and $T \in L(\mathbf{V})$. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be n eigenvalues of T counted with their multiplicities. Then, there exists an orthonormal basis $\{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$ of $\mathbf{V}$ with the following properties:*

$$T\operatorname{span}\{\mathbf{g}_1, \ldots, \mathbf{g}_i\} \subset \operatorname{span}\{\mathbf{g}_1, \ldots, \mathbf{g}_i\},\ \langle T\mathbf{g}_i, \mathbf{g}_i\rangle = \lambda_i,\ i = 1, \ldots, n. \tag{5.8.2}$$

Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ of dimension n and $T \in L(\mathbf{V})$ and assume that $\operatorname{spec} T \subset \mathbb{R}$. *Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be n eigenvalues of T counted with their multiplicities. Then, there exists an orthonormal basis $\{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$ of $\mathbf{V}$ such that (5.8.2) holds.*

Proof. Assume first that $\mathbf{V}$ is IPS over $\mathbb{C}$ of dimension n. The proof is by induction on n. For $n = 1$, the theorem is trivial. Assume that $n > 1$. Since $\lambda_1 \in \operatorname{spec} T$, it follows that there exists $\mathbf{g}_1 \in \mathbf{V}$, $\langle \mathbf{g}_1, \mathbf{g}_1\rangle = 1$, such that $T\mathbf{g}_1 = \lambda_1 \mathbf{g}_1$. Let $\mathbf{U} := \operatorname{span}(\mathbf{g}_1)^\perp$. Let P be the orthogonal projection on $\mathbf{U}$. Observe that $P\mathbf{v} = \mathbf{v} - \langle \mathbf{v}, \mathbf{g}_1\rangle \mathbf{g}_1$, for any vector $\mathbf{v} \in \mathbf{V}$. Let $T_1 := PT|_{\mathbf{U}}$. Clearly, $T_1 \in L(\mathbf{V})$. Let

$\tilde{\lambda}_2, ..., \tilde{\lambda}_n$ be the eigenvalues of T_1 counted with their multiplicities. The induction hypothesis yields the existence of an orthonormal basis $\{\mathbf{g}_2, ..., \mathbf{g}_n\}$ of $\mathbf{U}$ such that

$$T_1 \mathrm{span}\{\mathbf{g}_2, ..., \mathbf{g}_i\} \subset \mathrm{span}\{\mathbf{g}_2, ..., \mathbf{g}_i\}, \ \langle T_1\mathbf{g}_i, \mathbf{g}_i\rangle = \tilde{\lambda}_i, \ i = 1, ..., n.$$

As $T_1\mathbf{u} = T\mathbf{u} - \langle T\mathbf{u}, \mathbf{g}_1\rangle \mathbf{g}_1$, it follows that $T\mathrm{span}\{\mathbf{g}_1, ..., \mathbf{g}_i\} \subset \mathrm{span}\{\mathbf{g}_1, ..., \mathbf{g}_i\}$, for $i = 1, ..., n$. Hence, in the orthonormal basis $\{\mathbf{g}_1, ..., \mathbf{g}_n\}$, T is presented by an upper diagonal matrix $B = [b_{ij}]_1^n$, with $b_{11} = \lambda_1$ and $b_{ii} = \tilde{\lambda}_i$, $i = 2, ..., n$. Therefore, $\lambda_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_n$ are the eigenvalues of T counted with their multiplicities. This establishes the theorem in this case. The real case is treated similarly. □

Combine the above results with Problems 5.10.2-6 and 5.10.2-12 to deduce the following corollary:

Corollary 5.8.15 *Let $A \in \mathbb{C}^{n\times n}$ and $\lambda_1, ..., \lambda_n \in \mathbb{C}$ be n eigenvalues of A counted with their multiplicities. Then, there exists an upper triangular matrix $B = [b_{ij}]_1^n \in \mathbb{C}^{n\times n}$, such that $b_{ii} = \lambda_i$, $i = 1, ..., n$, and a unitary matrix $U \in \mathbf{U}_n$ such that $A = UBU^{-1}$. If $A \in \mathbf{N}(n, \mathbb{C})$, then B is a diagonal matrix.*

Let $A \in \mathbb{R}^{n\times n}$ and assume that $\mathrm{spec}\, T \subset \mathbb{R}$. *Then, $A = UBU^{-1}$, where U can be chosen a real orthogonal matrix and B a real upper triangular matrix. If $A \in \mathbf{N}(n, \mathbb{R})$ and* $\mathrm{spec}\, A \subset \mathbb{R}$, *then B is a diagonal matrix.*

It is easy to show that U in the above corollary can be chosen in $\mathbf{SU}_n$ or $\mathbf{SO}(n, \mathbb{R})$, respectively (Problem 5.10.2-11).

Definition 5.8.16 *Let $\mathbf{V}$ be a vector space and assume that $T : \mathbf{V} \to \mathbf{V}$ is a linear operator. Let $0 \neq \mathbf{v} \in \mathbf{V}$. Then, $\mathbf{W} = \mathrm{span}\{\mathbf{v}, T\mathbf{v}, T^2\mathbf{v}, \dots\}$ is called a cyclic invariant subspace of T generated by $\mathbf{v}$. (It is also referred as a Krylov subspace of T generated by $\mathbf{v}$.) Sometimes, we will call $\mathbf{W}$ just a cyclic subspace, or Krylov subspace.*

Theorem 5.8.17 *Let $\mathbf{V}$ be a finite dimensional IPS and $T : \mathbf{V} \to \mathbf{V}$ be a linear operator. For $0 \neq \mathbf{v} \in \mathbf{V}$, let $\mathbf{W} = \mathrm{span}\{\mathbf{v}, T\mathbf{v}, ..., T^{r-1}\mathbf{v}\}$ be a cyclic T-invariant subspace of dimension r generated by $\mathbf{v}$. Let $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ be an orthonormal basis of $\mathbf{W}$ obtained by the Gram-Schmidt process from the basis $\{\mathbf{v}, T\mathbf{v}, \dots, T^{r-1}\mathbf{v}\}$ of $\mathbf{W}$. Then, $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = 0$, for $1 \le i \le j - 2$, i.e. the representation matrix of $T|\mathbf{W}$ in the basis $\{\mathbf{u}_1, \dots, \mathbf{u}_{\}}$ is upper Hessenberg. If T is self-adjoint, then the representation matrix of $T|\mathbf{W}$ in the basis $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is a tridiagonal hermitian matrix.*

Proof. Let $\mathbf{W}_j = \mathrm{span}\{\mathbf{v}, \dots, T^{j-1}\mathbf{v}\}$, for $j = 1, ..., r + 1$. Clearly, $T\mathbf{W}_j \subset \mathbf{W}_{j+1}$, for $j = 1, ..., r$. The assumption that $\mathbf{W}$ is T-invariant subspace yields $\mathbf{W} = \mathbf{W}_r = \mathbf{W}_{r+1}$. Since $\dim \mathbf{W} = r$, it follows that $\mathbf{v}, ..., T^{r-1}\mathbf{v}$ are linearly independent. Hence, $\{\mathbf{v}, \dots, T^{r-1}\mathbf{v}\}$ is a basis for $\mathbf{W}$. Recall that $\mathrm{span}\{\mathbf{u}_1, ..., \mathbf{u}_j\} = \mathbf{W}_j$, for $j = 1, \dots, r$. Let $r \ge j \ge i+2$. Then, $T\mathbf{u}_i \in T\mathbf{W}_i \subset \mathbf{W}_{i+1}$. As $\mathbf{u}_j \perp \mathbf{W}_{i+1}$, it follows that $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = 0$. Assume that $T^* = T$. Let $r \ge i \ge j + 2$. Then, $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = \langle \mathbf{u}_i, T\mathbf{u}_j\rangle = 0$. Hence, the representation matrix of $T|\mathbf{W}$ in the basis $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is a tridiagonal hermitian matrix. □

5.9 Symmetric bilinear and hermitian forms

Definition 5.9.1 *Let $\mathbf{V}$ be a vector space over $\mathbb{F}$ and $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$. Q is called a symmetric bilinear form (on $\mathbf{V}$) if the following conditions are satisfied:*
(i) $Q(\mathbf{x},\mathbf{y}) = Q(\mathbf{y},\mathbf{x})$*, for all* $\mathbf{x},\mathbf{y} \in \mathbf{V}$ *(symmetricity);*
(ii) $Q(a\mathbf{x} + b\mathbf{z}, y) = aQ(\mathbf{x},\mathbf{y}) + bQ(\mathbf{z},\mathbf{y})$*, for all* $a,b \in \mathbb{F}$ *and* $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathbf{V}$ *(bilinearity).*

For $\mathbb{F} = \mathbb{C}$, Q it is called hermitian form or sesquilinear (on $\mathbf{V}$) if Q satisfies the conditions (iii) and (ii) where
(iii) $Q(\mathbf{x},\mathbf{y}) = \overline{Q(\mathbf{y},\mathbf{x})}$, for all $\mathbf{x},\mathbf{y} \in \mathbf{V}$ (barsymmetricity).

Something to notice about the definition of a bilinear form is the similarity it has to an inner product. In essence, a bilinear form is a generalization of an inner product.

Example 5.9.2 *The dot product on $\mathbb{R}^n$ is a symmetric bilinear form.*

Example 5.9.3 *On $\mathbb{C}^n$, let $Q((x_1,\dots,x_n),(y_1,\dots,y_n)) = \sum_{i=1}^n x_i\bar{y}_i$. Regarding $\mathbb{C}^n$ as a real vector space, Q is bilinear. But veiwing $\mathbb{C}^n$ as a complex vector space, Q is not bilinear (it is not linear in its second component). Moreover, $Q(\mathbf{x},\mathbf{y}) = Q(\mathbf{y},\mathbf{x})$. Then Q is a hermitian form.*

The following results are elementary, we leave their proofs as Problems 5.10.2-14 and 5.10.2-15.

Proposition 5.9.4 *Let $\mathbf{V}$ be a vector space over $\mathbb{F}$ with a basis $E = \{\mathbf{e}_1,\dots,\mathbf{e}_n\}$. Then, there is a $1-1$ correspondence between a symmetric bilinear form Q on $\mathbf{V}$ and $A \in \mathbf{S}(n,\mathbb{F})$:*

$$\begin{aligned}&Q(\mathbf{x},\mathbf{y}) = \eta^\top A\xi,\\ &\mathbf{x} = \sum_{i=1}^n \xi_i\mathbf{e}_i,\ \mathbf{y} = \sum_{i=1}^n \eta_i\mathbf{e}_i,\ \xi = (\xi_1,\dots,\xi_n)^\top, \eta = (\eta_1,\dots,\eta_n)^\top \in \mathbb{F}^n.\end{aligned}$$

Let $\mathbf{V}$ be a vector space over $\mathbb{C}$ with a basis $E = \{\mathbf{e}_1,\dots,\mathbf{e}_n\}$. Then, there is $1-1$ correspondence between a hermitian form Q on $\mathbf{V}$ and $A \in \mathbf{H}_n$:

$$\begin{aligned}&Q(\mathbf{x},\mathbf{y}) = \eta^* A\xi,\\ &\mathbf{x} = \sum_{i=1}^n \xi_i\mathbf{e}_i,\ \mathbf{y} = \sum_{i=1}^n \eta_i\mathbf{e}_i,\ \xi = (\xi_1,\dots,\xi_n)^\top, \eta = (\eta_1,\dots,\eta_n)^\top \in \mathbb{C}^n.\end{aligned}$$

Definition 5.9.5 *Let the assumptions of Proposition 5.9.4 hold. Then, A is called the representation matrix of Q in the basis E.*

Proposition 5.9.6 *Let the assumptions of Proposition 5.9.4 hold and assume that $F = \{\mathbf{f_1},\dots,\mathbf{f_n}\}$ is another basis of $\mathbf{V}$. Then, the symmetric bilinear form Q is represented by $B \in \mathbf{S}(n,\mathbb{F})$ in the basis F, where B is congruent to A:*

$$B = U^\top AU, \quad U \in \mathrm{GL}(n,\mathbb{F}),$$

and U is the matrix corresponding to the basis change from F to E. For $\mathbb{F} = \mathbb{C}$ the hermitian form Q is presented by $B \in \mathbf{H}_n$ in the basis F, where B hermite congruent to A:

$$B = U^* AU, \quad U \in \mathrm{GL}(n,\mathbb{C}),$$

and U is the matrix corresponding to the basis change from F to E.

Proposition 5.9.7 *Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{R}$. Let $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ be a symmetric bilinear form and $A \in \mathbf{S}(n, \mathbb{R})$ the representation matrix of Q with respect to a basis E in $\mathbf{V}$ and $\mathbf{V}_c$ be the extension of $\mathbf{V}$ over $\mathbb{C}$. Then, there exists a unique hermitian form $Q_c : \mathbf{V}_c \times \mathbf{V}_c \to \mathbb{C}$ such that $Q_c|_{\mathbf{V}\times\mathbf{V}} = Q$ and Q_c is presented by A with respect to the basis E in $\mathbf{V}_c$.*

We leave its proof as Problem 5.10.2-16.

Convention 5.9.8 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F}$. Let $Q : \mathbf{V}\times\mathbf{V} \to \mathbb{F}$ be either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$. Then, a representation matrix A of Q is chosen with respect to an orthonormal basis E.*

The following proposition is straightforward.

Proposition 5.9.9 *Let $\mathbf{V}$ is an n-dimensional IPS over $\mathbb{F}$ and $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ be either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$. Then, there exists a unique $T \in \mathbf{S}(\mathbf{V})$ such that $Q(\mathbf{x}, \mathbf{y}) = \langle T\mathbf{x}, \mathbf{y}\rangle$, for any $\mathbf{x}, \mathbf{y} \in \mathbf{V}$. In any orthonormal basis of $\mathbf{V}$, Q and T represented by the same matrix A. In particular, the characteristic polynomial $p(\lambda)$ of T is called the characteristic polynomial of Q. Here, Q has only real roots:*

$$\lambda_1(Q) \geq \ldots \geq \lambda_n(Q),$$

which are called the eigenvalues of Q. Furthermore, there exists an orthonormal basis $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ in $\mathbf{V}$ such that $D = \operatorname{diag}(\lambda_1(Q), \ldots, \lambda_n(Q))$ is the representation matrix of Q in F.

Vice versa, for any $T \in \mathbf{S}(\mathbf{V})$ and any subspace $\mathbf{U} \subset \mathbf{V}$, the form $Q(T, \mathbf{U})$ defined by

$$Q(T, \mathbf{U})(\mathbf{x}, \mathbf{y}) := \langle T\mathbf{x}, \mathbf{y}\rangle, \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathbf{U}$$

is either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$.

In the rest of the book, we use the following normalization unless stated otherwise.

Normalization 5.9.10 *Let $\mathbf{V}$ is an n-dimensional IPS over $\mathbb{F}$. Assume that $T \in \mathbf{S}(\mathbf{V})$. Then, arrange the eigenvalues of T counted with their multiplicities in the decreasing order*

$$\lambda_1(T) \geq \ldots \geq \lambda_n(T).$$

The same normalization applies to real symmetric matrices and complex hermitian matrices.

5.10 Max-min characterizations of eigenvalues

Given a hermitian matrix (linear transformation), we can obtain its largest (resp. smallest) eigenvalue by maximizing (resp. minimizing) the corresponding quadratic form over all the unit vectors. This section is devoted to the max-min characterization of eigenvalues of hermitian matrices.

First, we recall the Grassmannian of given dimensional in a fixed vector space.

Definition 5.10.1 *Let* $\mathbf{V}$ *be a finite dimensional space over the field* $\mathbb{F}$. *Denote by* $\mathrm{Gr}(m,\mathbf{V})$ *the set of all* m*-dimensional subspaces in* $\mathbf{V}$ *of dimension* $m \in [n] \cup \{0\}$.

Theorem 5.10.2 *(The convoy principle) Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $T \in \mathbf{S}(\mathbf{V})$. *Then*

$$\lambda_k(T) = \max_{\mathbf{U}\in \mathrm{Gr}(k,\mathbf{V})} \min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{U}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} = \max_{\mathbf{U}\in \mathrm{Gr}(k,\mathbf{V})} \lambda_k(Q(T,\mathbf{U})), \quad k = 1,\ldots,n, \tag{5.10.1}$$

where the form $Q(T,\mathbf{U})$ *is defined in Proposition 5.9.9. For* $k \in [n]$, *let* $\mathbf{U}$ *be an invariant subspace of* T *spanned by eigenvectors* $\mathbf{e}_1,\ldots,\mathbf{e}_k$ *corresponding to the eigenvalues* $\lambda_1(T),\ldots,\lambda_k(T)$. *Then,* $\lambda_k(T) = \lambda_k(Q(T,\mathbf{U}))$. *Let* $\mathbf{U} \in \mathrm{Gr}(k,\mathbf{V})$ *and assume that* $\lambda_k(T) = \lambda_k(Q(T,\mathbf{U}))$. *Then, the subspace* $\mathbf{U}$ *contains an eigenvector of* T *corresponding to* $\lambda_k(T)$.

In particular

$$\lambda_1(T) = \max_{\mathbf{0}\neq\mathbf{x}\in\mathbf{V}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle}, \quad \lambda_n(T) = \min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{V}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} \tag{5.10.2}$$

Moreover for any $\mathbf{x} \neq \mathbf{0}$:

$$\lambda_1(T) = \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} \text{ if and only if } T\mathbf{x} = \lambda_1(T)\mathbf{x},$$
$$\lambda_n(T) = \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} \text{ if and only if } T\mathbf{x} = \lambda_n(T)\mathbf{x},$$

The quotient $\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle}$, $\mathbf{0} \neq \mathbf{x} \in \mathbf{V}$ is called *Rayleigh quotient.* The characterization (5.10.2) is called *convoy principle.*

Proof. Choose an orthonormal basis $E = \{\mathbf{e}_1,\ldots,\mathbf{e}_n\}$ such that

$$T\mathbf{e}_i = \lambda_i(T)\mathbf{e}_i, \; < \mathbf{e}_i, \mathbf{e}_j >= \delta_{ij} \quad i,j = 1,\ldots,n. \tag{5.10.3}$$

Then

$$\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} = \frac{\sum_{i=1}^n \lambda_i(T)|x_i|^2}{\sum_{i=1}^n |x_i|^2}, \quad \mathbf{x} = \sum_{i=1}^n x_i\mathbf{e}_i \neq \mathbf{0}. \tag{5.10.4}$$

The above equality yields straightforward (5.10.2) and the equality cases in these characterizations. Let $\mathbf{U} \in \mathrm{Gr}(k,\mathbf{V})$. Then, the minimal characterization of $\lambda_k(Q(T,\mathbf{U}))$ yields the equality

$$\lambda_k(Q(T,\mathbf{U})) = \min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{U}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle}, \quad \text{for any } \mathbf{U} \in \mathrm{Gr}(k,\mathbf{U}). \tag{5.10.5}$$

Next, there exists $\mathbf{0} \neq \mathbf{x} \in \mathbf{U}$ such that $\langle \mathbf{x}, \mathbf{e}_i\rangle = 0$, for $i = 1,\ldots,k-1$. (For $k = 1$ this condition is void.) Hence

$$\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} = \frac{\sum_{i=k}^n \lambda_i(T)|x_i|^2}{\sum_{i=k}^n |x_i|^2} \leq \lambda_k(T) \Rightarrow \lambda_k(T) \geq \lambda_k(Q(T,\mathbf{U})).$$

Let

$$\begin{aligned}&\lambda_1(T) = \ldots = \lambda_{n_1}(T) > \lambda(T)_{n_1+1}(T) = \ldots = \lambda_{n_2}(T) > \ldots >\\ &\lambda_{n_{r-1}+1}(T) = \ldots = \lambda_{n_r}(T) = \lambda_n(T), \quad n_0 = 0 < n_1 < \ldots < n_r = n. \quad (5.10.6)\end{aligned}$$

Assume that $n_{j-1} < k \le n_j$ and $\lambda_k(Q(T,\mathbf{U})) = \lambda_k(T)$. Then, for $\mathbf{x} \in \mathbf{U}$ with $\langle \mathbf{x}, \mathbf{e}_i \rangle = 0$, we have equality $\lambda_k(Q(T,\mathbf{U})) = \lambda_k(T)$ if and only if $\mathbf{x} = \sum_{i=k}^{n_j} x_i \mathbf{e}_i$. Thus, $T\mathbf{x} = \lambda_k(T)\mathbf{x}$.

Let $\mathbf{U}_k = \mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ and $\mathbf{0} \ne \mathbf{x} = \sum_{i=1}^k x_i \mathbf{e}_i \in \mathbf{U}_k$. Then

$$\frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\sum_{i=1}^k \lambda_i(T)|x_i|^2}{\sum_{i=1}^k |x_i|^2} \ge \lambda_k(T) \Rightarrow \lambda_k(Q(T,\mathbf{U}_k)) \ge \lambda_k(T).$$

Hence, $\lambda_k(Q(T,\mathbf{U}_k)) = \lambda_k(T)$. □

Note that (5.10.1) can be stated as

$$\max\{\min\{\langle T\mathbf{x}, \mathbf{x}\rangle, \mathbf{x} \in \mathbf{U}, \|\mathbf{x}\| = 1 \text{ and } \mathbf{U} \in Gr(k, \mathbf{V})\}\}.$$

It can be shown that for $k > 1$ and $\lambda_1(T) > \lambda_k(T)$, there exists $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ such that $\lambda_k(T) = \lambda_k(T, \mathbf{U})$ and $\mathbf{U}$ is not an invariant subspace of T, in particular $\mathbf{U}$ does not contain all $\mathbf{e}_1, \ldots, \mathbf{e}_k$ satisfying (5.10.3). (See Problem 5.10.2-18.)

Corollary 5.10.3 *Let the assumptions of Theorem 5.10.2 hold. Let $1 \le \ell \le n$. Then*

$$\lambda_k(T) = \max_{\mathbf{W} \in \mathrm{Gr}(\ell, \mathbf{V})} \lambda_k(Q(T, \mathbf{W})), \quad k = 1, \ldots, \ell. \quad (5.10.7)$$

Proof. For $k \le \ell$, apply Theorem 5.10.2 to $\lambda_k(Q(T,\mathbf{W}))$ to deduce that $\lambda_k(Q(T,\mathbf{W})) \le \lambda_k(T)$. Let $\mathbf{U}_\ell = \mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_\ell\}$. Then

$$\lambda_k(Q(T, \mathbf{U}_\ell)) = \lambda_k(T), \quad k = 1, \ldots, \ell.$$

□

Theorem 5.10.4 *(Courant-Fischer principle) Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Then*

$$\lambda_k(T) = \min_{\mathbf{W} \in \mathrm{Gr}(k-1, \mathbf{V})} \max_{\mathbf{0} \ne \mathbf{x} \in \mathbf{W}^\perp} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}, \quad k = 1, \ldots, n.$$

See Problem 5.10.2-19 for the proof of the theorem and the following corollary.

Corollary 5.10.5 *Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Let $k, \ell \in [n]$ be integers satisfying $k \le \ell$. Then*

$$\lambda_{n-\ell+k}(T) \le \lambda_k(Q(T, \mathbf{W})) \le \lambda_k(T), \quad \text{for any } \mathbf{W} \in \mathrm{Gr}(\ell, \mathbf{V}).$$

The following theorem, by Weyl, allows us to obtain an upper bound for the k-th eigenvalue of $S + T$.

Theorem 5.10.6 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $S, T \in \mathbf{S}(\mathbf{V})$*. Then, for any* $i, j \in \mathbb{N}, i + j - 1 \le n$ *the inequality* $\lambda_{i+j-1}(S+T) \le \lambda_i(S) + \lambda_j(T)$ *holds.(This inequality is well-known as Weyl inequality.)*

Proof. Let $\mathbf{U}_{i-1}, \mathbf{V}_{j-1} \subset \mathbf{V}$ be eigenspaces of S and T spanned by the first $i-1, j-1$ eigenvectors of S and T, respectively. So

$$\langle S\mathbf{x}, \mathbf{x}\rangle \le \lambda_i(S)\langle \mathbf{x}, \mathbf{x}\rangle,\ \langle T\mathbf{y}, \mathbf{y}\rangle \le \lambda_j(T)\langle \mathbf{y}, \mathbf{y}\rangle, \text{ for all } \mathbf{x} \in \mathbf{U}_{i-1}^{\perp}, \mathbf{y} \in \mathbf{V}_{j-1}^{\perp}.$$

Note that $\dim \mathbf{U}_{i-1} = i-1, \dim \mathbf{V}_{j-1} = j-1$.. Let $\mathbf{W} = \mathbf{U}_{i-1} + \mathbf{V}_{j-1}$. Then, $\dim \mathbf{W} = l - 1 \le i + j - 2$. Assume that $\mathbf{z} \in \mathbf{W}^{\perp}$. Then, $\langle (S+T)\mathbf{z}, \mathbf{z}\rangle = \langle S\mathbf{z}, \mathbf{z}\rangle + \langle T\mathbf{z}, \mathbf{z}\rangle \le (\lambda_i(S) + \lambda_j(T))\langle \mathbf{z}, \mathbf{z}\rangle$. Hence, $\max_{\mathbf{0} \ne \mathbf{z} \in \mathbf{W}^{\perp}} \frac{\langle (S+T)\mathbf{z}, \mathbf{z}\rangle}{\langle \mathbf{z}, \mathbf{z}\rangle} \le \lambda_i(S) + \lambda_j(T)$. Use Theorem 5.10.4 to deduce that $\lambda_{i+j-1}(S+T) \le \lambda_l(S+T) \le \lambda_i(S) + \lambda_j(T)$. □

Definition 5.10.7 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS. Fix an integer* $k \in [n]$*. Then,* $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\}$ *is called an orthonormal* k*-frame if* $< \mathbf{f}_i, \mathbf{f}_j >= \delta_{ij}$*, for* $i, j = 1, ..., k$*. Denote by* $\mathrm{Fr}(k, \mathbf{V})$ *the set of all orthonormal* k*-frames in* $\mathbf{V}$*.*

Note that each $F_k \in \mathrm{Fr}(k, \mathbf{V})$ induces $\mathbf{U} = \mathrm{span} F_k \in \mathrm{Gr}(k, \mathbf{V})$. Vice versa, any $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ induces the set $\mathrm{Fr}(k, \mathbf{U})$ of all orthonormal k-frames which span $\mathbf{U}$.

Theorem 5.10.8 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $T \in \mathbf{S}(\mathbf{V})$*. Then, for any integer* $k \in [n]$

$$\sum_{i=1}^{k} \lambda_i(T) = \max_{\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V})} \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle.$$

Furthermore

$$\sum_{i=1}^{k} \lambda_i(T) = \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle,$$

for some k*-orthonormal frame* $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\}$ *if and only if* $\mathrm{span} F_k$ *is spanned by* $\mathbf{e}_1, ..., \mathbf{e}_k$ *satisfying (5.10.3).*

Proof. Define

$$\begin{aligned} \mathrm{tr}\, Q(T, \mathbf{U}) &:= \sum_{i=1}^{k} \lambda_i(Q(T, \mathbf{U})), \quad \text{for } \mathbf{U} \in \mathrm{Gr}(k, \mathbf{V}), \\ \mathrm{tr}_k\, T &:= \sum_{i=1}^{k} \lambda_i(T). \end{aligned} \tag{5.10.8}$$

Let $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V})$. Set $\mathbf{U} = \mathrm{span} F_k$. Then, in view of Corollary 5.10.3

$$\sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle = \mathrm{tr}\, Q(T, \mathbf{U}) \le \sum_{i=1}^{k} \lambda_i(T).$$

Let $E_k := \{\mathbf{e}_1, ..., \mathbf{e}_k\}$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ are given by (5.10.3). Clearly, $\mathrm{tr}_k\, T = \mathrm{tr}\, Q(T, \mathrm{span} E_k)$. This shows the maximal characterization of $\mathrm{tr}_k\, T$.

Let $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ and assume that $\mathrm{tr}_k T = \mathrm{tr}\, Q(T, \mathbf{U})$. Hence, $\lambda_i(T) = \lambda_i(Q(T, \mathbf{U}))$, for $i = 1, ..., k$. Then, there exists $G_k = \{\mathbf{g}_1, ..., \mathbf{g}_k\} \in \mathrm{Fr}(k, \mathbf{U}))$ such that

$$\min_{\mathbf{0} \neq \mathbf{x} \in \mathrm{span}\{\mathbf{g}_1, ..., \mathbf{g}_i\}} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \lambda_i(Q(T, \mathbf{U})) = \lambda_i(T),\ i = 1, ..., k.$$

Use Theorem 5.10.2 to deduce that $T\mathbf{g}_i = \lambda_i(T)\mathbf{g}_i$, for $i = 1, ..., k$. □

Theorem 5.10.9 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $T \in \mathbf{S}(\mathbf{V})$*. Then, for any integer* $k, l \in [n]$*, such that* $k + l \leq n$

$$\sum_{i=l+1}^{l+k} \lambda_i(T) = \min_{\mathbf{W} \in \mathrm{Gr}(l, \mathbf{V})} \max_{\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V} \cap \mathbf{W}^\perp)} \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle.$$

Proof. Let $\mathbf{W}_j := \mathrm{span}\{\mathbf{e}_1, ..., \mathbf{e}_j\}, j = 1, \ldots, n$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ are given by (5.10.3). Then, $\mathbf{V}_1 := \mathbf{V} \cap \mathbf{W}_l$ is an invariant subspace of T. Let $T_1 := T|\mathbf{V}_1$. Then, $\lambda_i(T_1) = \lambda_{l+i}(T)$, for $i = 1, \ldots, n - l$. Theorem 5.10.8 for T_1 yields

$$\max_{\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V} \cap \mathbf{W}_l^\perp)} \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle = \sum_{i=l+1}^{l+k} \lambda_i(T).$$

Let $T_2 := T|\mathbf{W}_{l+k}$ and $\mathbf{W} \in \mathrm{Gr}(l, \mathbf{V})$. Set $\mathbf{U} := \mathbf{W}_{l+k} \cap \mathbf{W}^\perp$. Then, $\dim \mathbf{U} \geq k$. Apply Theorem 5.10.8 to $-T_2$ to deduce

$$\sum_{i=1}^{k} \lambda_i(-T_2) \geq \sum_{i=1}^{k} \langle -T\mathbf{f}_i, \mathbf{f}_i\rangle, \text{ for } \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}).$$

The above inequality is equivalent to the inequality

$$\sum_{i=l+1}^{l+k} \lambda_i(T) \leq \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle, \text{ for } \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}) \leq$$
$$\max_{\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V} \cap \mathbf{W}^\perp)} \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle.$$

The above inequalities yield the theorem. □

5.10.1 Worked-out Problems

1. Let $A, B \in H_n$.

   (a) Show that $\lambda_i(A + B) \leq \lambda_i(A) + \lambda_i(B)$, for any $i \in [n]$.

   (b) Show that $\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A + B)$.

   (c) Give the necessary and sufficient condition to have the equality $\lambda_1(A + B) = \lambda_1(A) + \lambda_1(B)$.

   Solution:

   (a) Apply the Weyl inequality for $j = 1$.

(b) Replace A and B by $-A$ and $-B$ and i by j in the previous part:

$$\lambda_j(-A-B) \le \lambda_j(-A) + \lambda_j(-B).$$

We have $\lambda_j(-A-B) = -\lambda_{n-j+1}(A+B)$, $\lambda_j(-A) = -\lambda_{n-j+1}(A)$ and $\lambda_1(-B) = -\lambda_n(B)$.
Then, $\lambda_{n-j+1}(A+B) \ge \lambda_{n-j+1}(A) + \lambda_n(B)$. Setting $i = n-j+1$, we obtain the desired inequality.

(c) Assume that $\lambda_1(A+B)\mathbf{x} = (A+B)\mathbf{x}$, where $\|\mathbf{x}\| = 1$. Then, $\lambda_1(A+B) = \mathbf{x}^*(A+B)\mathbf{x} = \mathbf{x}^*A\mathbf{x} + \mathbf{x}^*B\mathbf{x} \le \lambda_1(A) + \lambda_1(B)$. Equality holds if and only if $\mathbf{x}$ is an eigenvector of A and B corresponding to $\lambda_1(A)$ and $\lambda_1(B)$, i.e. equality holds if and only if A and B have a common eigenvector $\mathbf{x}$ corresponding to $\lambda_1(A)$ and $\lambda_1(B)$.

2. Let $S, T \in S(\mathbf{V})$. We say $T \succ S$ if $\langle T\mathbf{x}, \mathbf{x}\rangle > \langle S\mathbf{x}, \mathbf{x}\rangle$, for all $0 \ne \mathbf{x} \in \mathbf{V}$. T is called *positive definite* if $T \succ 0$, where 0 is the zero operator in $L(\mathbf{V})$.
Assume that $K \in S(n, R)$ is a positive definite matrix. Define in $\mathbb{R}^n$ an inner product $\langle \mathbf{x}, \mathbf{y}\rangle := \mathbf{y}^\top K\mathbf{x}$. Let $A \in \mathbb{R}^{n\times n}$ and view AK as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ by $\mathbf{x} \mapsto (AK)\mathbf{x}$. Show that AK is self-adjoint with respect to the above inner product if and only if $A \in S(n, \mathbb{R})$.
Solution:
Assume first that AK is self-adjoint, then we have:

$$\langle AK\mathbf{x}, \mathbf{y}\rangle = \mathbf{y}^\top KAK\mathbf{x} = \mathbf{y}^\top K^\top A^\top K\mathbf{x} = \mathbf{y}^\top (AK)^\top K = \langle \mathbf{x}, AK\mathbf{y}\rangle.$$

On the other hand, since KA is self-adjoint (as AK is), then we have:

$$\mathbf{y}^\top KAK\mathbf{x} = \mathbf{y}^\top K^\top A^\top K\mathbf{x}.$$

Since this is true for all $\mathbf{x}$ and $\mathbf{y}$, then $KAK = KA^\top K$ and as K is positive definite, then $A \in S(n, \mathbb{R})$.
Conversely, assume that $A \in S(n, \mathbb{R})$. Then

$$\begin{aligned} &\langle \mathbf{x}, A\mathbf{y}\rangle = \langle A\mathbf{x}, \mathbf{y}\rangle \Rightarrow \\ &\overline{A\mathbf{y}}K\mathbf{x} = \bar{\mathbf{y}}KA\mathbf{x} \Rightarrow \\ &\bar{\mathbf{y}}\bar{A}K\mathbf{x} = \bar{\mathbf{y}}KA\mathbf{x}. \end{aligned} \tag{5.10.9}$$

As $K \in S(n, \mathbb{R})$, then (5.10.9) yields $\bar{\mathbf{y}}\overline{AK}\mathbf{x} = \bar{\mathbf{y}}KA\mathbf{x}$. Since this is true for all $\mathbf{x}$ and $\mathbf{y}$, then $KA = \overline{KA} = (KA)^\top$. This implies that AK is self-adjoint.

3. Prove that every real square matrix is the sum of a symmetric matrix and a skew-symmetric matrix.
Solution:
For any matrix $A = [a_{ij}]$, consider the symmetric matrix $B = [b_{ij}]$ and skew-symmetric matrix $C = [c_{ij}]$ with $b_{ij} = \frac{a_{ij}+a_{ji}}{2}$ and $c_{ij} = \frac{a_{ij}-a_{ji}}{2}$, for $i \le j$. Then, note that for $i \le j$, we have $b_{ij} + c_{ij} = a_{ij}$ and $b_{ji} + c_{ji} = b_{ij} - c_{ij} = a_{ji}$, so $A = B + C$ as desired.

4. Let $\langle , \rangle$ be a bilinear form on a real vector space $\mathbf{V}$. Show that there is a symmetric form $(,)$ and a skew-symmetric form $[,]$ so that $\langle , \rangle = (,) + [,]$.

Solution:
Note that a bilinear form is symmetric or skew-symmetric if and only if its corresponding matrix is symmetric or skew-symmetric. Then, let A be the matrix for $\langle,\rangle$; by the previous problem, we can find B symmetric and C skew-symmetric such that $A = B + C$. Then, take $(u,v) = u^\top Bv$ and $[u,v] = u^\top Cv$. Note that $\langle u,v\rangle = (u,v) + [u,v]$ and $(,)$ and $[,]$ are symmetric and skew-symmetric, as desired.

5.10.2 Problems

1. Prove Proposition 5.8.1.

2. Let $P, Q \in L(\mathbf{V}), a, b \in \mathbb{F}$. Show that $(aP + bQ)^* = \bar{a}P^* + \bar{b}Q^*$.

3. Prove Proposition 5.8.3.

4. Prove Proposition 5.8.4 for finite dimensional $\mathbf{V}$. (*Hint*: Choose an orthonormal basis in $\mathbf{V}$.)

5. Show the following statements:

$$\begin{aligned}
&\mathbf{SO}(n,\mathbb{F}) \subset \mathbf{O}(n,\mathbb{F}) \subset \mathrm{GL}(n,\mathbb{F}),\\
&\mathbf{S}(n,\mathbb{R}) \subset \mathbf{H}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{AS}(n,\mathbb{R}) \subset \mathbf{AH}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{S}(n,\mathbb{R}), \mathbf{AS}(n,\mathbb{R}) \subset \mathbf{N}(n,\mathbb{R}) \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{O}(n,\mathbb{R}) \subset \mathbf{U}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{SO}(n,\mathbb{F}),\ \mathbf{O}(n,\mathbb{F}),\ \mathbf{SU}_n,\ \mathbf{U}_n \ \text{ are groups}\\
&\mathbf{S}(n,\mathbb{F}) \ \text{ is an } \mathbb{F}\text{–vector space of dimension } \binom{n+1}{2},\\
&\mathbf{AS}(n,\mathbb{F}) \ \text{ is an } \mathbb{F}\text{–vector space of dimension } \binom{n}{2},\\
&\mathbf{H}_n \ \text{ is an } \mathbb{R}\text{–vector space of dimension } n^2,\\
&\mathbf{AH}_n = i\mathbf{H}_n,
\end{aligned}$$

6. Let $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$ be an orthonormal basis in IPS $\mathbf{V}$ over $\mathbb{F}$. Let $G = \{\mathbf{g}_1, ..., \mathbf{g}_n\}$ be another basis in $\mathbf{V}$. Show that F is an orthonormal basis if and only if the the matrix of change of bases either from E to G or from G to E is a unitary matrix.

7. Prove Proposition 5.8.12

8. Prove Proposition 5.8.13

9. a. Show that $A \in \mathbf{SO}(2,\mathbb{R})$ is of the form $A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \theta \in \mathbb{R}$.

   b. Show that $\mathbf{SO}(2,\mathbb{R}) = e^{\mathbf{AS}(2,\mathbb{R})}$. That is, for any $B \in \mathbf{AS}(2,\mathbb{R})$, $e^B \in \mathbf{SO}(2,\mathbb{R})$, and for any $A \in \mathbf{SO}(n,\mathbb{R})$ one can find $B \in \mathbf{AS}(2,\mathbb{R})$ such that $A = e^B$. (*Hint*: Consider the power series for e^B, $B = \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}$.)

c. Show that $\mathbf{SO}(n,\mathbb{R}) = e^{\mathbf{AS}(n,\mathbb{R})}$. (*Hint*: Use Propositions 5.8.12 and 5.8.13 and part b.)

d. Show that $\mathbf{SO}(n,\mathbb{R})$ is a path connected space. (See part e.)

e. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}$ with $n > 1$. Let $p \in [n-1]$. Assume that $\{\mathbf{x}_1, ..., \mathbf{x}_p\}$ and $\{\mathbf{y}_1, ..., \mathbf{y}_p\}$ are two orthonormal systems in $\mathbf{V}$. Show that these two orthonormal systems are path connected. That is, there are p continuous mappings $\mathbf{z}_i(t) : [0,1] \to \mathbf{V}$, $i = 1, ..., p$, such that for each $t \in [0,1]$, $\{\mathbf{z}_1(t), ..., \mathbf{z}_p(t)\}$ is an orthonormal system and $\mathbf{z}_i(0) = \mathbf{x}_i, \mathbf{z}_i(1) = \mathbf{y}_i, i = 1, ..., p$.
(See also Problem 1.10.2-4.)

10. a. Show that $\mathbf{U}_n = e^{\mathbf{AH}_n}$. (*Hint*: Use Proposition 5.8.10 and its proof.)

    b. Show that $\mathbf{U}_n$ is path connected.

    c. Prove Problem 9e for $\mathbb{F} = \mathbb{C}$.

11. Show that

    (a) $D_1 D D_1^* = D$, for any $D \in \mathbf{D}(n,\mathbb{C})$, $D_1 \in \mathbf{DU}_n$.

    (b) $A \in \mathbf{N}(n,\mathbb{C})$ if and only if $A = UDU^*$, $U \in \mathbf{SU}_n$, $D \in \mathbf{D}(n,\mathbb{C})$.

    (c) $A \in \mathbf{N}(n,\mathbb{R})$, $\sigma(A) \subset \mathbb{R}$ if and only if $A = UDU^\top$, $U \in \mathbf{SO}_n$, $D \in \mathbf{D}(n,\mathbb{R})$.

12. Show that an upper triangular or a lower triangular matrix $B \in \mathbb{C}^{n\times n}$ is normal if and only if B is diagonal. (**Hint**: consider the equality $(BB^*)_{11} = (B^*B)_{11}$.)

13. Let the assumptions of Theorem 5.8.17 hold. Show that instead of performing the Gram-Schmidt process on $\mathbf{v}, T\mathbf{v}, ..., T^{r-1}\mathbf{v}$, one can perform the following process. Let $\mathbf{w}_1 := \frac{1}{\|\mathbf{v}\|}\mathbf{v}$. Assume that one already obtained i orthonormal vectors $\mathbf{w}_1, ..., \mathbf{w}_i$. Let $\tilde{\mathbf{w}}_{i+1} := T\mathbf{w}_i - \sum_{j=1}^{i}\langle T\mathbf{w}_i, \mathbf{w}_j\rangle \mathbf{w}_j$. If $\tilde{\mathbf{w}}_{i+1} = 0$, then stop the process, i.e. one is left with i orthonormal vectors. If $\mathbf{w}_{i+1} \neq 0$ then $\mathbf{w}_{i+1} := \frac{1}{\|\tilde{\mathbf{w}}_{i+1}\|}\tilde{\mathbf{w}}_{i+1}$ and continue the process. Show that the process ends after obtaining r orthonormal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ and $\mathbf{u}_i = \mathbf{w}_i$, for $i = 1, ..., r$. (This is a version of *Lanczos tridiagonalization* process.)

14. Prove Proposition 5.9.4.

15. Prove Proposition 5.9.6.

16. Prove Proposition 5.9.7.

17. Prove Proposition 5.9.9.

18. Let $\mathbf{V}$ be a 3-dimensional IPS and $T \in L(\mathbf{V})$ be self-adjoint. Assume that

$$\lambda_1(T) > \lambda_2(T) > \lambda_3(T), \quad T\mathbf{e}_i = \lambda_i(T)\mathbf{e}_i,\ i = 1, 2, 3.$$

    Let $\mathbf{W} = \mathrm{span}\{\mathbf{e}_1, \mathbf{e}_3\}$.

    (a) Show that for each $t \in (\lambda_3(T), \lambda_1(T))$, there exists $\mathbf{W}(t) \in \mathrm{Gr}(1,\mathbf{W})$ such that $\lambda_1(Q(T,\mathbf{W}(t))) = t$.

    (b) Let $t \in [\lambda_2(T), \lambda_1(T)]$ and set $\mathbf{U}(t) = \mathrm{span}\{\mathbf{W}(t), \mathbf{e}_2\} \in \mathrm{Gr}(2,\mathbf{V})$. Show that $\lambda_2(T) = \lambda_2(Q(T,\mathbf{U}(t))$.

19. (a) Let the assumptions of Theorem 5.10.4 hold and $\mathbf{W} \in \mathrm{Gr}(k-1,\mathbf{V})$. Show that there exists $\mathbf{0} \neq \mathbf{x} \in \mathbf{W}^\perp$ such that $\langle \mathbf{x}, \mathbf{e}_i \rangle = 0$, for $k+1,...,n$, where $\mathbf{e}_1,...,\mathbf{e}_n$ satisfy (5.10.3). Conclude that $\lambda_1(Q(T,\mathbf{W}^\perp)) \geq \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} \geq \lambda_k(T)$.

    (b) Let $\mathbf{U}_\ell = \mathrm{span}\{\mathbf{e}_1,...,\mathbf{e}_\ell\}$. Show that $\lambda_1(Q(T,\mathbf{U}_\ell^\perp)) = \lambda_{\ell+1}(T)$, for $\ell = 1,...,n-1$.

    (c) Prove Theorem 5.10.4.

    (d) Prove Corollary 5.10.5. (**Hint**: Choose $\mathbf{U} \in \mathrm{Gr}(k,\mathbf{W})$ such that $\mathbf{U} \subset \mathbf{W} \cap \mathrm{span}\{\mathbf{e}_{n-\ell+k+1},...,\mathbf{e}_n\}^\perp$. Then, $\lambda_{n-\ell+k}(T) \leq \lambda_k(Q(T,\mathbf{U})) \leq \lambda_k(Q(T,\mathbf{W}))$.)

20. Let $B = [b_{ij}]_{i,j=1}^n \in \mathbf{H}_n$ and denote by $A \in \mathbf{H}_{n-1}$ the matrix obtained from B by deleting the $j-th$ row and column.

    (a) Show the Cauchy interlacing inequalities

    $$\lambda_i(B) \geq \lambda_i(A) \geq \lambda_{i+1}(B), \text{ for } i = 1,...,n-1.$$

    (b) Prove the inequality $\lambda_1(B) + \lambda_n(B) \leq \lambda_1(A) + b_{ii}$.
    (**Hint**. Express the traces of B and A respectively in terms of eigenvalues to obtain

    $$\lambda_1(B) + \lambda_n(B) = b_{ii} + \lambda_1(A) + \sum_{i=2}^{n-1}(\lambda_i(A) - \lambda_i(B)).$$

    Then use the Cauchy interlacing inequalities.)

21. Let $B \in \mathbf{H}_n$ be the following 2×2 block matrix $B = \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^* & B_{22} \end{bmatrix}$. Show that

    $$\lambda_1(B) + \lambda_n(B) \leq \lambda_1(B_{11}) + \lambda_1(B_{22}).$$

    (**Hint**. Assume that $B\mathbf{x} = \lambda_1(B)\mathbf{x}, \mathbf{x}^\top = \{\mathbf{x}_1^\top, \mathbf{x}_2^\top\}$, partitioned as B. Consider $\mathbf{U} = \mathrm{span}\{(\mathbf{x}_1^\top, \mathbf{0})^\top, (\mathbf{0}, \mathbf{x}_2^\top)^\top\}$. Analyze $\lambda_1(Q(T,\mathbf{U})) + \lambda_2(Q(T,\mathbf{U}))$.)

22. Let $T \in \mathbf{S}(\mathbf{V})$. Denote by $\iota_+(T), \iota_0(T), \iota_-(T)$ the number of positive, zero and negative eigenvalues among $\lambda_1(T) \geq ... \geq \lambda_n(T)$. The triple $\iota(T) := (\iota_+(T), \iota_0(T), \iota_-(T))$ is called the *inertia* of T. For $B \in \mathbf{H}_n$, let $\iota(B) := (\iota_+(B), \iota_0(B), \iota_-(B))$ be the inertia of B, where $\iota_+(B), \iota_0(B), \iota_-(B)$ is the number of positive, zero and negative eigenvalues of B, respectively. Let $\mathbf{U} \in \mathrm{Gr}(k,\mathbf{V})$. Prove the statements (a), (b), (c) and (d).

    (a) Assume that $\lambda_k(Q(T,\mathbf{U})) > 0$, i.e. $Q(T,\mathbf{U}) > 0$. Then, $k \leq \iota_+(T)$. If $k = \iota_+(T)$, then one can choose $\mathbf{U}$ to be an invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to positive eigenvalues of T. (Usually such a subspace is not unique.)

    (b) Assume that $\lambda_k(Q(T,\mathbf{U})) \geq 0$, i.e. $Q(T,\mathbf{U}) \geq 0$. Then, $k \leq \iota_+(T) + \iota_0(T)$. If $k = \iota_+(T) + \iota_0(T)$, then $\mathbf{U}$ is the unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to non-negative eigenvalues of T.

    (c) Assume that $\lambda_1(Q(T,\mathbf{U})) < 0$, i.e. $Q(T,\mathbf{U}) < 0$. Then, $k \leq \iota_-(T)$. If $k = \iota_-(T)$, then $\mathbf{U}$ can be chosen to be an invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T, corresponding to negative eigenvalues of T. (Usually such a subspace may not be unique.)

(d) Assume that $\lambda_1(Q(T,\mathbf{U})) \le 0$, i.e. $Q(T,\mathbf{U}) \le 0$. Then, $k \le \iota_-(T) + \iota_0(T)$. If $k = \iota_-(T) + \iota_0(T)$, then $\mathbf{U}$ is a unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to non-positive eigenvalues of T.

23. Let $B \in \mathbf{H}_n$ and assume that $A = PBP^*$, for some $P \in \mathrm{GL}(n,\mathbb{C})$. Show that $\iota(A) = \iota(B)$.

24. Prove that for symmetric real matrices the signs of pivots are the signs of eigenvalues.

25. Show the following statements:

    (a) The set of hermitian matrices is not a subspace of $\mathbb{C}^{n\times n}$ over $\mathbb{C}$, (with the usual addition and scalar multiplication)

    (b) The set of hermitian matrices is subspace of $\mathbb{C}^{n\times n}$ over $\mathbb{R}$, (with the usual addition and scalar multiplication.)

26. Let $\mathbf{V}$ be a complex vector space, $T \in N(\mathbf{V})$ and $f \in \mathbb{C}[z]$. Show the following statements:

    (a) $f(T) \in N(\mathbf{V})$,

    (b) The minimal polynomial of T has distinct roots.

27. Let $A \in \mathbb{C}^{n\times n}$ be a hermitian matrix. Prove that rank A equals the number of non-zero eigenvalues of A.

5.11 Positive definite operators and matrices

To find the matrix analogues of positive (non-negative) real numbers, we introduce positive (non-negative) definite matrices.

Definition 5.11.1 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F}$. Let S and $T \in \mathbf{S}(\mathbf{V})$. Then, $T{>}S$, $(T{\ge}S)$ if $\langle T\mathbf{x},\mathbf{x}\rangle > \langle S\mathbf{x},\mathbf{x}\rangle$, $(\langle T\mathbf{x},\mathbf{x}\rangle \ge \langle S\mathbf{x},\mathbf{x}\rangle)$, for all $\mathbf{0} \ne \mathbf{x} \in \mathbf{V}$. Also, T is called positive (non-negative) definite if $T{>}0$ $(T{\ge}0)$, where 0 is the zero operator in $L(\mathbf{V})$.*

Denote by $\mathbf{S}_+(\mathbf{V})^o$ and $\mathbf{S}_+(\mathbf{V})$ the open set of positive definite self-adjoint operators and the closed set of non-negative self-adjoint operators, respectively.
$(\mathbf{S}_+(\mathbf{V})^o \subset \mathbf{S}_+(\mathbf{V}) \subset \mathbf{S}(\mathbf{V}))$.

Definition 5.11.2 *Let P and Q be either symmetric bilinear forms or hermitian forms. Then, $Q{\succ}P$, $(Q{\ge}P)$ if $Q(\mathbf{x},\mathbf{x}) > P(\mathbf{x},\mathbf{x})$, $(Q(\mathbf{x},\mathbf{x}) \ge P(\mathbf{x},\mathbf{x}))$, for all $\mathbf{0} \ne \mathbf{x} \in \mathbf{V}$. Also, Q is called positive (non-negative) definite if $Q{>}0$ $(Q{\ge}0)$, where 0 is the zero operator in $L(\mathbf{V})$.*

Note that for $A \in \mathbb{C}^{n\times n}$, A is hermitian if and only if $\mathbf{x}^*A\mathbf{x}$ is real for all $\mathbf{x} \in \mathbb{C}^n$. This suggests the following definition of positive (non-negative) definiteness for $A \in \mathbf{H}_n$.

Definition 5.11.3 *For A and $B \in \mathbf{H}_n$, $B{>}A$ ($B{\geq}A$) if $\mathbf{x}^*B\mathbf{x} > \mathbf{x}^*A\mathbf{x}$ ($\mathbf{x}^*B\mathbf{x} \geq \mathbf{x}^*A\mathbf{x}$), for all $\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n$. Moreover, $B\in \mathbf{H}_n$ is called positive (non-negative) definite if $B{>}0$ ($B{\geq}0$). Denote by $\mathbf{H}_{n,+}^o \subset \mathbf{H}_{n,+} \subset \mathbf{H}_n$ the open set of positive definite $n \times n$ hermitian matrices and the closed set of $n \times n$ non-negative hermitian matrices, respectively. Let $\mathbf{S}_+(n,\mathbb{R}) := \mathbf{S}(n,\mathbb{R}) \cap \mathbf{H}_{n,+}$, $\mathbf{S}_+(n,\mathbb{R})^o := \mathbf{S}(n,\mathbb{R}) \cap \mathbf{H}_{n,+}^o$.*

Use Theorem 5.10.2 to deduce the following corollary which states that all eignevalues of a positive definite matrix (or linear operator) are strictly positive real numbers.

Corollary 5.11.4 *Let $\mathbf{V}$ be n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Then, $T{>}0$ ($T{\geq}0$) if and only if $\lambda_n(T) > 0$ ($\lambda_n(T) \geq 0$). Let $S \in \mathbf{S}(\mathbf{V})$ and assume that $T{>}S$ ($T{\geq}S$). Then, $\lambda_i(T) > \lambda_i(S)$ ($\lambda_i(T) \geq \lambda_i(S)$), for $i = 1, \dots, n$.*

We can apply Corollary 5.11.4 to derive an equivalent definition for positive (non-negative) definite operators as follows.
An operator is positive (non-negative) definite if it is symmetric and all its eigenvalues are positive (non-negative). Moreover, since the trace of a matrix is the sum of its eigenvalues then the trace of a positive definite matrix is a positive real number. Also, since the determinant of a matrix is the product of its eigenvalues, then the determinant of a positive definite matrix is a positive real number. In particular, positive definite matrices are always invertible.

Proposition 5.11.5 *Let $\mathbf{V}$ be a finite dimensional IPS. Assume that $T \in \mathbf{S}(\mathbf{V})$. Then, $T{\geq}0$ if and only if there exists $S \in \mathbf{S}(\mathbf{V})$ such that $T = S^2$. Furthermore, $T{>}0$ if and only if S is invertible. For $T \in \mathbf{S}(\mathbf{V})$ with $T{\geq}0$, there exists a unique $S{\geq}0$ in $\mathbf{S}(\mathbf{V})$ such that $T = S^2$. This S is called the square root of T and is denoted by $T^{\frac{1}{2}}$.*

Proof. Assume first that $T{\geq}0$. Let $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ be an orthonormal basis consisting of eigenvectors of T as in (5.10.3). Since $\lambda_i(T) \geq 0$, $i = 1, \dots, n$, we can define $P \in L(\mathbf{V})$ as follows

$$P\mathbf{e}_i = \sqrt{\lambda_i(T)}\mathbf{e}_i, \quad i = 1, \dots, n.$$

Clearly, P is self-adjoint non-negative and $T = P^2$.

Suppose now that $T = S^2$, for some $S \in \mathbf{S}(\mathbf{V})$. Then, $T \in \mathbf{S}(\mathbf{V})$ and $\langle T\mathbf{x}, \mathbf{x}\rangle = \langle S\mathbf{x}, S\mathbf{x}\rangle \geq 0$. Hence, $T{\geq}0$. Clearly, $\langle T\mathbf{x}, \mathbf{x}\rangle = 0$ if and only if $S\mathbf{x} = 0$. Hence, $T{>}0$ if and only if $S \in \mathrm{GL}(\mathbf{V})$. Suppose that $S{\geq}0$. Then, $\lambda_i(S) = \sqrt{\lambda_i(T)}$, $i = 1, \dots, n$. Furthermore, each eigenvector of S is an eigenvector of T. It is straightforward to show that $S = P$, where P is defined above. Clearly, $T{>}0$ if and only if $\sqrt{\lambda_n(T)} > 0$, i.e. if and only if S is invertible. □

Corollary 5.11.6 *Let $B \in \mathbf{H}_n$ (or $\mathbf{S}(n,\mathbb{R})$). Then, $B{\geq}0$ if and only there exists $A \in \mathbf{H}_n$ (or $\mathbf{S}(n,\mathbb{R})$) such that $B = A^2$. Furthermore, $B{>}0$ if and only if A is invertible. For $B{\geq}0$, there exists a unique $A{\geq}0$ such that $B = A^2$. This A is denoted by $B^{\frac{1}{2}}$.*

Definition 5.11.7 *Let* $\mathbf{V}$ *be an IPS. Given a list of vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{V}$*, the matrix*

$$\begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_n \rangle \\ \vdots & \cdots & \vdots \\ \langle \mathbf{x}_n, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{x}_n, \mathbf{x}_n \rangle \end{bmatrix}$$

is denoted by $G(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *and is called the Gramian matrix.*

Theorem 5.11.8 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{F}$*. Let* $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbf{V}$*. Then, the Gramian matrix* $G(\mathbf{x}_1, ..., \mathbf{x}_n) := (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_1^n$ *is a hermitian non-negative definite matrix. (If* $\mathbb{F} = \mathbb{R}$*, then* $G(\mathbf{x}_1, ..., \mathbf{x}_n)$ *is real symmetric non-negative definite.) Also,* $G(\mathbf{x}_1, ..., \mathbf{x}_n) > 0$ *if and only* $\mathbf{x}_1, ..., \mathbf{x}_n$ *are linearly independent. Furthermore, for any integer* $k \in [n-1]$*,*

$$\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \le \det G(\mathbf{x}_1, ..., \mathbf{x}_k)\ \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n). \tag{5.11.1}$$

Equality holds if and only if either $\det G(\mathbf{x}_1, ..., \mathbf{x}_k)\ \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n) = 0$ *or* $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$*, for* $i = 1, ..., k$ *and* $j = k+1, ..., n$*.*

Proof. Clearly, $G(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbf{H}_n$. If $\mathbf{V}$ is an IPS over $\mathbb{R}$ then $G(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbf{S}(n, \mathbb{R})$. Let $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{F}^n$. Then

$$\mathbf{a}^* G(\mathbf{x}_1, ..., \mathbf{x}_n)\mathbf{a} = \langle \sum_{i=1}^n a_i \mathbf{x}_i, \sum_{j=1}^n a_j \mathbf{x}_j \rangle \ge 0.$$

Equality holds if and only if $\sum_{i=1}^n a_i \mathbf{x}_i = 0$. Hence, $G(\mathbf{x}_1, ..., \mathbf{x}_n) \ge 0$ and $G(\mathbf{x}_1, ..., \mathbf{x}_n) > 0$ if and only if $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. In particular, $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \ge 0$ and $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) > 0$ if and only if $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent.

We now prove the inequality (5.11.1). Assume first that the right-hand side of (5.11.1) is zero. Then, either $\mathbf{x}_1, ..., \mathbf{x}_k$ or $\mathbf{x}_{k+1}, ..., \mathbf{x}_n$ are linearly dependent. Hence, $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly dependent and $\det G = 0$.

Assume now that the right-hand side of (5.11.1) is positive. Hence, $\mathbf{x}_1, ..., \mathbf{x}_k$ and $\mathbf{x}_{k+1}, ..., \mathbf{x}_n$ are linearly independent. If $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly dependent, then $\det G = 0$ and strict inequality holds in (5.11.1). It is left to show the inequality (5.11.1) and the equality case when $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. Perform the Gram-Schmidt algorithm on $\mathbf{x}_1, ..., \mathbf{x}_n$ as given in (5.1.1). Let $S_j = \operatorname{span}\{\mathbf{x}_1, ..., \mathbf{x}_j\}$, for $j = 1, ..., n$. Corollary 5.1.1 yields that $\operatorname{span}\{\mathbf{e}_1, ..., \mathbf{e}_{n-1}\} = S_{n-1}$. Hence, $\mathbf{y}_n = \mathbf{x}_n - \sum_{j=1}^{n-1} b_j \mathbf{x}_j$, for some $b_1, ..., b_{n-1} \in \mathbb{F}$. Let G' be the matrix obtained from $G(\mathbf{x}_1, ..., \mathbf{x}_n)$ by subtracting from the n-th row b_j times j-th row. Thus, the last row of G' is $(\langle \mathbf{y}_n, \mathbf{x}_1 \rangle, ..., \langle \mathbf{y}_n, \mathbf{x}_n \rangle) = (0, ..., 0, \|\mathbf{y}_n\|^2)$. Clearly, $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) = \det G'$. Expand $\det G'$ by the last row to deduce

$$\begin{aligned} &\det G(\mathbf{x}_1, ..., \mathbf{x}_n) = \det G(\mathbf{x}_i, ..., \mathbf{x}_{n-1})\ \|\mathbf{y}_n\|^2 = ... = \\ &\det G(\mathbf{x}_1, ..., \mathbf{x}_k)\ \prod_{i=k+1}^n \|\mathbf{y}_i\|^2 = \\ &\det G(\mathbf{x}_1, ..., \mathbf{x}_k)\ \prod_{i=k+1}^n \operatorname{dist}(\mathbf{x}_i, S_{i-1})^2, \quad k = n-1, ..., 1. \end{aligned} \tag{5.11.2}$$

Perform the Gram-Schmidt process on $\mathbf{x}_{k+1},...,\mathbf{x}_n$ to obtain the orthogonal set of vectors $\hat{\mathbf{y}}_{k+1},...,\hat{\mathbf{y}}_n$ such that

$$\hat{S}_j := \mathrm{span}\{\mathbf{x}_{k+1},...,\mathbf{x}_j\} = \mathrm{span}\{\hat{\mathbf{y}}_{k+1},...,\hat{\mathbf{y}}_j\},\ \mathrm{dist}(\mathbf{x}_j,\hat{S}_{j-1}) = \|\hat{\mathbf{y}}_j\|,$$

for $j = k+1,...,n$, where $\hat{S}_k = \{\mathbf{0}\}$. Use (5.11.2) to deduce that $\det G(\mathbf{x}_{k+1},...,\mathbf{x}_n) = \prod_{j=k+1}^{n}\|\hat{\mathbf{y}}_j\|^2$. As $\hat{S}_{j-1} \subset S_{j-1}$, for $j > k$, it follows that

$$\|\mathbf{y}_j\| = \mathrm{dist}(\mathbf{x}_j, S_{j-1}) \le \mathrm{dist}(\mathbf{x}_j,\hat{S}_{j-1}) = \|\hat{\mathbf{y}}_j\|,\ j = k+1,...,n.$$

This shows (5.11.1). Assume now equality holds in (5.11.1). Then, $\|\mathbf{y}_j\| = \|\hat{\mathbf{y}}_j\|$, for $j = k+1,...,n$. Since $\hat{S}_{j-1} \subset S_{j-1}$ and $\hat{\mathbf{y}}_j - \mathbf{x}_j \in \hat{S}_{j-1} \subset S_{j-1}$, it follows that $\mathrm{dist}(\mathbf{x}_j, S_{j-1}) = \mathrm{dist}(\hat{\mathbf{y}}_j, S_{j-1}) = \|\mathbf{y}_j\|$. Hence, $\|\hat{\mathbf{y}}_j\| = \mathrm{dist}(\hat{\mathbf{y}}_j, S_{j-1})$. Part (h) of Problem 5.1.4 yields that $\hat{\mathbf{y}}_j$ is orthogonal on S_{j-1}. In particular each $\hat{\mathbf{y}}_j$ is orthogonal to S_k, for $j = k+1,...,n$. Hence, $\mathbf{x}_j \perp S_k$, for $j = k+1,...,n$, i.e. $\langle \mathbf{x}_j, \mathbf{x}_i\rangle = 0$, for $j > k$ and $i \le k$. Clearly, if the last condition holds, then $\det G(\mathbf{x}_1,...,\mathbf{x}_n) = \det G(\mathbf{x}_1,...,\mathbf{x}_k)\ \det G(\mathbf{x}_{k+1},...,\mathbf{x}_n)$. □

Note that $\det G(\mathbf{x}_1,...,\mathbf{x}_n)$ has the following geometric meaning. Consider a parallelepiped Π in $\mathbf{V}$ spanned by $\mathbf{x}_1,...,\mathbf{x}_n$ starting from the origin $\mathbf{0}$. That is, Π is a convex hull spanned by the vectors $\mathbf{0}$ and $\sum_{i\in S}\mathbf{x}_i$ for all nonempty subsets $S \subset \{1,...,n\}$. Then, $\sqrt{\det G(\mathbf{x}_1,...,\mathbf{x}_n)}$ is the n-volume of Π. The inequality (5.11.1) and equalities (5.11.2) are "obvious" from this geometrical point of view.

Corollary 5.11.9 *Let $0 \le B = [b_{ij}]_1^n \in \mathbf{H}_{n,+}$. Then*

$$\det B \le \det [b_{ij}]_1^k\ \det [b_{ij}]_{k+1}^n,\ \text{for } k = 1,...,n-1.$$

For a fixed k, equality holds if and only if either the right-hand side of the above inequality is zero or $b_{ij} = 0$, for $i = 1,...,k$ and $j = k+1,...,n$.

Proof. From Corollary 5.11.6, it follows that $B = X^2$, for some $X \in \mathbf{H}_n$. Let $\mathbf{x}_1,...,\mathbf{x}_n \in \mathbb{C}^n$ be the n-columns of $X^T = (\mathbf{x}_1,...,\mathbf{x}_n)$. Let $\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{y}^*\mathbf{x}$. Since $X \in \mathbf{H}_n$, we deduce that $B = G(\mathbf{x}_1,...,\mathbf{x}_n)$. □

Theorem 5.11.10 *Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$, then the following statements are equivalent*
(a) $T \succ 0$.
(b) *Let $\{\mathbf{g}_1,...,\mathbf{g}_n\}$ be a basis of $\mathbf{V}$. Then, $\det(\langle T\mathbf{g}_i,\mathbf{g}_j\rangle)_{i,j=1}^k > 0,\ k = 1,...,n$.*

Proof. (a) ⇒ (b). According to Proposition 5.11.5, $T = S^2$, for some $S \in \mathbf{S}(\mathbf{V}) \cap \mathrm{GL}(\mathbf{V})$. Then, $\langle T\mathbf{g}_i,\mathbf{g}_j\rangle = \langle S\mathbf{g}_i, S\mathbf{g}_j\rangle$. Hence, $\det(\langle T\mathbf{g}_i,\mathbf{g}_j\rangle)_{i,j=1}^k = \det G(S\mathbf{g}_1,...,S\mathbf{g}_k)$. Since, S is invertible and $\mathbf{g}_1,...,\mathbf{g}_k$ linearly independent, it follows that $S\mathbf{g}_1,...,S\mathbf{g}_k$ are linearly independent. Theorem 5.11.1 implies that $\det G(S\mathbf{g}_1,...,S\mathbf{g}_k) > 0$, for $k = 1,...,n$.
(b) ⇒ (a). The proof is by induction on n. For $n = 1$, (a) is obvious. Assume that (a) holds for $n = m-1$. Let $\mathbf{U} := \mathrm{span}\{\mathbf{g}_1,...,\mathbf{g}_{n-1}\}$ and $Q := Q(T,\mathbf{U})$. Then, there exists $P \in \mathbf{S}(\mathbf{U})$ such that $< P\mathbf{x},\mathbf{y} >= Q(\mathbf{x},\mathbf{y}) =< T\mathbf{x},\mathbf{y} >$, for any $\mathbf{x},\mathbf{y} \in \mathbf{U}$. By induction $P \succ 0$. Corollary 5.10.3 yields that $\lambda_{n-1}(T) \ge \lambda_{n-1}(P) > 0$. Hence,

T has at least $n-1$ positive eigenvalues. Let $\mathbf{e}_1,\dots,\mathbf{e}_n$ be given by (5.10.3). Then, $\det(\langle T\mathbf{e}_i,\mathbf{e}_j\rangle)_{i,j=1}^n = \prod_{i=1}^n \lambda_i(T) > 0$. Let $A=[a_{pq}]_1^n \in \mathrm{GL}(n,\mathbb{C})$ be the transformation matrix from the basis $\mathbf{g}_1,\dots,\mathbf{g}_n$ to $\mathbf{e}_1,\dots,\mathbf{e}_n$, i.e.

$$\mathbf{g}_i = \sum_{p=1}^n a_{pi}\mathbf{e}_p,\ i=1,\dots,n.$$

It is straightforward to show that

$$\begin{aligned}&(\langle T\mathbf{g}_i,\mathbf{g}_j\rangle)_1^n = A^T(\langle T\mathbf{e}_p,\mathbf{e}_q\rangle)\bar{A} \Rightarrow \\ &\det(\langle T\mathbf{g}_i,\mathbf{g}_j\rangle)_1^n = \det(\langle T\mathbf{e}_i,\mathbf{e}_j\rangle)_1^n|\det A|^2 = |\det A|^2\prod_{i=1}^n \lambda_i(T).\end{aligned} \tag{5.11.3}$$

Since, $\det(\langle T\mathbf{g}_i,\mathbf{g}_j\rangle)_1^n > 0$ and $\lambda_1(T)\ge\dots\ge\lambda_{n-1}(T)>0$, it follows that $\lambda_n(T)>0$. □

The following result is straightforward, its proof is left as Problem 5.12.2-1.

Proposition 5.11.11 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F}$ with the inner product $\langle\cdot,\cdot\rangle$. Assume that $T\in\mathbf{S}(\mathbf{V})$. Then, $T\succ 0$ if and only if $(\mathbf{x},\mathbf{y}) := \langle T\mathbf{x},\mathbf{y}\rangle$ is an inner product on $\mathbf{V}$. Vice versa any inner product $(\cdot,\cdot):\mathbf{V}\times\mathbf{V}\to\mathbb{R}$ is of the form $(\mathbf{x},\mathbf{y}) =< T\mathbf{x},\mathbf{y} >$ for a unique self-adjoint positive definite operator $T\in L(\mathbf{V})$.*

Example 5.11.12 *Each $B\in\mathbf{H}_n$ with $B\succ 0$ induces an inner product on $\mathbb{C}^n$: $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{y}^*B\mathbf{x}$. Each $B\in\mathbf{S}(n,\mathbb{R})$ with $B\succ 0$ induces an inner product on $\mathbb{R}^n$: $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{y}^TB\mathbf{x}$. Furthermore, any inner product on $\mathbb{C}^n$ or $\mathbb{R}^n$ is of the above form. In particular, the standard inner products on $\mathbb{C}^n$ and $\mathbb{R}^n$ are induced by the identity matrix I.*

5.12 Inequalities for traces

Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F}$. Let $T:\mathbf{V}\to\mathbf{V}$ be a linear operator. Then, $\operatorname{tr} T$ is the trace of the representation matrix A of with respect to any orthonormal basis of $\mathbf{V}$. See Problem 5.12.2-3.

Theorem 5.12.1 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F}$. Assume that $S,T\in\mathbf{S}(\mathbf{V})$. Then, $\operatorname{tr} ST$ is bounded from below and above by $\sum_{i=1}^n\lambda_i(S)\lambda_{n-i+1}(T)$ and $\sum_{i=1}^n\lambda_i(S)\lambda_i(T)$, respectively. Namely, we have:*

$$\sum_{i=1}^n \lambda_i(S)\lambda_{n-i+1}(T) \le \operatorname{tr} ST \le \sum_{i=1}^n \lambda_i(S)\lambda_i(T). \tag{5.12.1}$$

Equality for the upper bound holds if and only if $ST=TS$ and there exists an orthonormal basis $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ in $\mathbf{V}$ such that

$$S\mathbf{x}_i = \lambda_i(S)\mathbf{x}_i,\quad T\mathbf{x}_i = \lambda_i(T)\mathbf{x}_i,\quad i=1,\dots,n. \tag{5.12.2}$$

Equality for the lower bound holds if and only if $ST=TS$ and there exists an orthonormal basis $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ of $\mathbf{V}$ such that

$$S\mathbf{x}_i = \lambda_i(S)\mathbf{x}_i,\quad T\mathbf{x}_i = \lambda_{n-i+1}(T)\mathbf{x}_i,\quad i=1,\dots,n. \tag{5.12.3}$$

Proof. Let $\{\mathbf{y}_1,...,\mathbf{y}_n\}$ be an orthonormal basis of $\mathbf{V}$ such that

$$\begin{aligned}
&T\mathbf{y}_i = \lambda_i(T)\mathbf{y}_i, \quad i = 1,...,n,\\
&\lambda_1(T) = ... = \lambda_{i_1}(T) > \lambda_{i_1+1}(T) = ... = \lambda_{i_2}(T) > ... >\\
&\lambda_{i_{k-1}+1}(T) = ... = \lambda_{i_k}(T) = \lambda_n(T),\ 1 \le i_1 < ... < i_k = n.
\end{aligned}$$

If $k = 1$, then $i_1 = n$ and it follows that $T = \lambda_1 I$. Then, the theorem is trivial in this case. Assume that $k > 1$. Then

$$\begin{aligned}
&\operatorname{tr} ST = \sum_{i=1}^{n} \lambda_i(T)\langle S\mathbf{y}_i, \mathbf{y}_i\rangle =\\
&\sum_{i=1}^{n-1}(\lambda_i(T) - \lambda_{i+1}(T))(\sum_{l=1}^{i}\langle S\mathbf{y}_l, \mathbf{y}_l\rangle) + \lambda_n(T)(\sum_{l=1}^{n}\langle S\mathbf{y}_l, \mathbf{y}_l\rangle) =\\
&\sum_{j=1}^{k-1}(\lambda_{i_j}(T) - \lambda_{i_{j+1}}(T))\sum_{l=1}^{i_j}\langle S\mathbf{y}_l, \mathbf{y}_l\rangle + \lambda_n(T)\operatorname{tr} S.
\end{aligned}$$

Theorem 5.10.8 yields that $\sum_{l=1}^{i_j}\langle S\mathbf{y}_l, \mathbf{y}_l\rangle \le \sum_{l=1}^{i_j}\lambda_l(S)$. Substitute these inequalities for $j = 1,...,k-1$ in the above identity to deduce the upper bound in (5.12.1). Clearly the condition (5.12.2) implies that $\operatorname{tr} ST$ is equal to the upper bound in (5.12.1). Assume now that $\operatorname{tr} ST$ is equal to the upper bound in (5.12.1). Then, $\sum_{l=1}^{i_j}\langle S\mathbf{y}_l, \mathbf{y}_l\rangle = \sum_{l=1}^{i_j}\lambda_l(S)$ for $j = 1,...,k-1$. Theorem 5.10.8 yields that $\operatorname{span}\{\mathbf{y}_1,...,\mathbf{y}_{i_j}\}$ is spanned by some i_j eigenvectors of S corresponding to the first i_j eigenvalues of S, for $j = 1,...,k-1$. Let $\{\mathbf{x}_1,...,\mathbf{x}_{i_1}\}$ be an orthonormal basis of $\operatorname{span}\{\mathbf{y}_1,...,\mathbf{y}_{i_1}\}$ consisting of the eigenvectors of S corresponding to the eigenvalues of $\lambda_1(S),...,\lambda_{i_1}(S)$. Since any $0 \ne \mathbf{x} \in \operatorname{span}\{\mathbf{y}_1,...,\mathbf{y}_{i_1}\}$ is an eigenvector of T corresponding to the eigenvalue $\lambda_{i_1}(T)$, it follows that (5.12.2) holds, for $i = 1,...,i_1$. Consider $\operatorname{span}\{\mathbf{y}_1,...,\mathbf{y}_{i_2}\}$. The above arguments imply that this subspace contains i_2 eigenvectors of S and T corresponding to the first i_2 eigenvalues of S and T. Hence, $\mathbf{U}_2$ is the orthogonal complement of $\operatorname{span}\{\mathbf{x}_1,...,\mathbf{x}_{i_1}\}$ in $\operatorname{span}\{\mathbf{y}_1,...,\mathbf{y}_{i_2}\}$, spanned by $\mathbf{x}_{i_1+1},...,\mathbf{x}_{i_2}$, which are $i_2 - i_1$ orthonormal eigenvectors of S corresponding to the eigenvalues $\lambda_{i_1+}(S),...,\lambda_{i_2}(S)$. Since any non-zero vector in $\mathbf{U}_2$ is an eigenvector of T corresponding to the eigenvalue $\lambda_{i_2}(T)$, we deduce that (5.12.2) holds for $i = 1,...,i_2$. Continuing in the same manner we obtain (5.12.2).

To prove the equality case in the lower bound, consider the equality in the upper bound for $\operatorname{tr} S(-T)$. □

Corollary 5.12.2 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS over* $\mathbb{F}$*. Assume that* S *and* $T \in \mathbf{S}(\mathbf{V})$*. Then*

$$\sum_{i=1}^{n}(\lambda_i(S) - \lambda_i(T))^2 \le \operatorname{tr}(S - T)^2. \tag{5.12.4}$$

Equality holds if and only if $ST = TS$ *and* $\mathbf{V}$ *has an orthonormal basis* $\{\mathbf{x}_1,...,\mathbf{x}_n\}$ *satisfying (5.12.2).*

Proof. Note that

$$\sum_{i=1}^{n}(\lambda_i(S) - \lambda_i(T))^2 = \operatorname{tr} S^2 + \operatorname{tr} T^2 - 2\sum_{i=1}^{n}\lambda_i(S)\lambda_i(T).$$

□

Corollary 5.12.3 *Let $S, T \in \mathbf{H}_n$. Then, the inequalities (5.12.1) and (5.12.4) hold. Equality in the upper bounds holds if and only if there exists $U \in \mathbf{U}_n$ such that $S = U \operatorname{diag} \lambda(S) U^*, T = U \operatorname{diag} \lambda(T) U^*$. Equality in the lower bound of (5.12.1) holds if and only if there exists $V \in \mathbf{U}_n$ such that $S = V \operatorname{diag} \lambda(S) V^*, -T = V \operatorname{diag} \lambda(-T) V^*$.*

5.12.1 Worked-out Problems

1. Let $\mathbf{V}$ denote the vector space of real $n \times n$ matrices. Prove that $\langle A, B \rangle = \operatorname{tr} A^\top B$ is a positive definite bilinear form on $\mathbf{V}$. Find an orthonormal basis for this form.
Solution:
First note that $\langle A_1 + A_2, B \rangle = \operatorname{tr}((A_1 + A_2)^\top B) = \operatorname{tr}(A_1^\top B + A_2^\top B) = \operatorname{tr} A_1^\top B + \operatorname{tr} A_2^\top B = \langle A_1, B \rangle + \langle A_2, B \rangle$ and $\langle cA, B \rangle = \operatorname{tr} cA^\top B = c \operatorname{tr} A^\top B = c\langle A, B \rangle$, so the form is bilinear in A. The exact same proof establishes bilinearity in B. Now, note that $\langle 0, 0 \rangle = 0$, and
$$\langle A, A \rangle = \operatorname{tr} A^\top A = \sum_{i=1}^n (A^\top A)_{ii} = \sum_{i=1}^n \sum_{j=1}^n (A^\top)_{ij} A_{ji} = \sum_{i=1}^n \sum_{j=1}^n A_{ij}^2 > 0,$$
for $A \neq 0$, so the form is positive definite and bilinear.
For the next part, let M_{ij} be the matrix with a 1 in the (i, j) entry and 0's elsewhere. We claim that M_{ij} for $1 \le i, j \le n$ forms the desired orthonormal basis; it is clear that these matrices form a basis for $\mathbf{V}$. Then, note that for $(a, b) \neq (x, y)$ we have
$$\langle M_{ab}, M_{xy} \rangle = \operatorname{tr}(M_{ab}^\top M_{xy}) = \sum_{i=1}^n (M_{ab}^\top M_{xy})_{ii} =$$
$$\sum_{i=1}^n \sum_{j=1}^n (M_{ab}^\top)_{ij} (M_{xy})_{ji} = \sum_{i=1}^n \sum_{j=1}^n \delta_{aj}\delta_{bi}\delta_{xj}\delta_{yi} = 0,$$
where δ is the Kronecker delta function and where we've used the fact that a term in the sum is non-zero only when $a = x = j$ and $b = y = i$, which never occurs. Thus, this basis is orthogonal. Furthermore, we find
$$\langle M_{ab}, M_{ab} \rangle = \operatorname{tr}(M_{ab}^\top M_{ab}) = \operatorname{tr} M_{bb} = 1,$$
since $M_{ab}^\top M_{ab} = M_{bb}$. Therefore, M_{ij}'s form the desired orthonormal basis.

5.12.2 Problems

1. Show Proposition 5.11.11.

2. Consider the Hölder inequality
$$\sum_{l=1}^n x_l y_l a_l \le (\sum_{l=1}^n x_l^p a_l)^{\frac{1}{p}} (\sum_{l=1}^n y_l^q a_l)^{\frac{1}{q}}, \tag{5.12.5}$$
for any $\mathbf{x} = (x_1, \ldots, x_n)^\top, \mathbf{y} = (y_1, \ldots, y_n)^\top, \mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}_+^n$ and $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

(a) Let $A \in \mathbf{H}_{n,+}, \mathbf{x} \in \mathbb{C}^n$ and $0 \le i < j < k$ be three integers. Show that

$$\mathbf{x}^* A^j \mathbf{x} \le (\mathbf{x}^* A^i \mathbf{x})^{\frac{k-j}{k-i}} (\mathbf{x}^* A^k \mathbf{x})^{\frac{j-i}{k-i}}. \tag{5.12.6}$$

(**Hint**: Diagonalize A.)

(b) Assume that $A = e^B$, for some $B \in \mathbf{H}_n$. Show that (5.12.6) holds for any three real numbers $i < j < k$.

3. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F}$.

(a) Assume that $T : \mathbf{V} \to \mathbf{V}$ is a linear transformation. Show that for any orthonormal basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$,

$$\operatorname{tr} T = \sum_{i=1}^{n} \langle T\mathbf{x}_i, \mathbf{x}_i \rangle.$$

Furthermore, if $\mathbb{F} = \mathbb{C}$, then $\operatorname{tr} T$ is the sum of the n eigenvalues of T.

(b) Let $S, T \in \mathbf{S}(\mathbf{V})$. Show that $\operatorname{tr} ST = \operatorname{tr} TS \in \mathbb{R}$.

4. Let $A \in \mathbb{R}^{m\times n}$ and rank $A = n$. Show that $A^\top A$ is positive definite.

5. Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Show that e^A is symmetric and positive definite.

6. Show that a matrix $A \in \mathbb{F}^{m\times n}$ is positive definite if it is symmetric and all its pivots are positive.
(Hint: Use Problem 5.10.2-24.)

7. Determine whether the following matrices are positive definite.

(a) $A = \begin{bmatrix} -2 & 1 & 0 \\ -1 & 2 & -1 \\ 0 & 1 & 2 \end{bmatrix}$

(b) $B = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

(Hint: Use Problem 5.10.2-24 and Problem 5.12.2-4.)

5.13 Schur's Unitary Triangularization

Schur's unitary triangularization theorem says that every matrix is unitarily equivalent to a triangular matrix. Precisely, it reals as follows. (See Theorem 5.8.7)

Theorem 5.13.1 *Let $A \in \mathbb{F}^{n\times n}$. Then, there exists a unitary $U \in \mathbb{F}^{n\times n}$ and an upper triangular $\Lambda \in \mathbb{F}^{n\times n}$ such that $A = U\Lambda U^*$. The diagonal entries of Λ are the eigenvalues of A. (Here $\mathbb{F} = \mathbb{C}$.)*

Proof. by induction on n: The results holds for $n = 1$. Assume that $n > 1$ and the statement holds for matrices of order $n-1$. Let $A \in \mathbb{F}^{n\times n}$ and λ be an eigenvalue of it, and $\mathbf{x}_1 \in \mathbb{F}^n$ be a corresponding eigenvector with $\|\mathbf{x}_1\|_2 = 1$. Extend $\mathbf{x}_1$ to an orthonormal basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of $\mathbb{F}^n$. Set $U_1 = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. Then U_1 is unitary and

$$U_1^* A U_1 = \begin{bmatrix} \lambda & y^\top \\ 0 & A_1 \end{bmatrix},$$

where $A_1 \in \mathbb{F}^{(n-1)\times(n-1)}$. By induction hypothesis, there exists a unitary $U_2 \in \mathbb{F}^{(n-1)\times(n-1)}$ such that $U_2^* A_1 U_2$ is upper triangular. Set $U = U_1 \operatorname{diag}(1, U_2) \in \mathbb{F}^{n\times n}$. Then, U is unitary and

$$U^* A U = \begin{bmatrix} \lambda & y^\top U_2 \\ 0 & U_2^* A_1 U_2 \end{bmatrix} = \Lambda,$$

is upper triangular. The diagonal entries of the upper triangular matrix Λ are the eigenvalues of A. □

5.14 Singular Value Decomposition

Singular value decomposition is based on a theorem which says that a rectangular matrix A can be broken down into the product of three matrices; a unitary matrix U, a diagonal matrix Σ, and the transpose of a unitary matrix $\mathbf{V}$.

Let $\mathbf{U}$ and $\mathbf{V}$ be two finite dimensional IPS over $\mathbb{F}$, with the inner products $\langle\cdot,\cdot\rangle_{\mathbf{U}}$ and $\langle\cdot,\cdot\rangle_{\mathbf{V}}$, respectively. Let $\{\mathbf{u}_1, ..., \mathbf{u}_m\}$ and $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ be bases in $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $T : \mathbf{V} \to \mathbf{U}$ be a linear operator. In these bases, T is represented by a matrix $A = [a_{ij}] \in \mathbb{F}^{m\times n}$ as given by

$$T\mathbf{v}_j = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i, \quad j = 1, \ldots, n.$$

Let $T^* : \mathbf{U}^* = \mathbf{U} \to \mathbf{V}^* = \mathbf{V}$. Then, $T^*T : \mathbf{V} \to \mathbf{V}$ and $TT^* : \mathbf{U} \to \mathbf{U}$ are self-adjoint operators. As

$$\langle T^*T\mathbf{v}, \mathbf{v}\rangle_{\mathbf{V}} = \langle T\mathbf{v}, T\mathbf{v}\rangle_{\mathbf{V}} \ge 0, \quad \langle TT^*\mathbf{u}, \mathbf{u}\rangle_{\mathbf{U}} = \langle T^*\mathbf{u}, T^*\mathbf{u}\rangle_{\mathbf{U}} \ge 0,$$

it follows that $T^*T \ge 0, TT^* \ge 0$. Assume that:

$$T^*T\mathbf{c}_i = \lambda_i(T^*T)\mathbf{c}_i,\ \langle \mathbf{c}_i, \mathbf{c}_k\rangle_{\mathbf{V}} = \delta_{ik},\ i,k = 1, ..., n, \tag{5.14.1}$$
$$\lambda_1(T^*T) \ge ... \ge \lambda_n(T^*T) \ge 0,$$
$$TT^*\mathbf{d}_j = \lambda_j(TT^*)\mathbf{d}_j,\ \langle \mathbf{d}_j, \mathbf{d}_l\rangle_{\mathbf{U}} = \delta_{jl},\ j,l = 1, ..., m, \tag{5.14.2}$$
$$\lambda_1(TT^*) \ge ... \ge \lambda_m(TT^*) \ge 0,$$

Proposition 5.14.1 *Let $\mathbf{U}$ and $\mathbf{V}$ be two finite dimensional IPS over $\mathbb{F}$ and $T : \mathbf{V} \to \mathbf{U}$ be a linear transformation. Then,* $\operatorname{rank} T = \operatorname{rank} T^* = \operatorname{rank} T^*T = \operatorname{rank} TT^* = r$. *Furthermore, the self-adjoint non-negative definite operators T^*T and TT^* have exactly r positive eigenvalues, and*

$$\lambda_i(T^*T) = \lambda_i(TT^*) > 0, \quad i = 1, ..., \operatorname{rank} T. \tag{5.14.3}$$

Moreover, for $i \in [r]$, $T\mathbf{c}_i$ and $T^*\mathbf{d}_i$ are eigenvectors of TT^* and T^*T corresponding to the eigenvalue $\lambda_i(TT^*) = \lambda_i(T^*T)$, respectively. Furthermore, if $\mathbf{c}_1, ..., \mathbf{c}_r$ satisfy (5.14.1), then $\tilde{\mathbf{d}}_i := \frac{T\mathbf{c}_i}{\|T\mathbf{c}_i\|}, i = 1, ..., r$ satisfies (5.14.2), for $i = 1, ..., r$. Similar result holds for $\mathbf{d}_1, ..., \mathbf{d}_r$.

Proof. Clearly, $T\mathbf{x} = 0$ if and only if $\langle T\mathbf{x}, T\mathbf{x}\rangle = 0$ if and only if $T^*T\mathbf{x} = 0$. Hence

$$\text{rank}\, T^*T = \text{rank}\, T = \text{rank}\, T^* = \text{rank}\, TT^* = r.$$

Thus, T^*T and TT^* have exactly r positive eigenvalues. Let $i \in [r]$. Then, $T^*T\mathbf{c}_i \neq 0$. Hence, $T\mathbf{c}_i \neq 0$. (5.14.1) yields that $TT^*(T\mathbf{c}_i) = \lambda_i(T^*T)(T\mathbf{c}_i)$. Similarly, $T^*T(T^*\mathbf{d}_i) = \lambda_i(TT^*)(T^*\mathbf{d}_i) \neq 0$. Hence, (5.14.3) holds. Assume that $\mathbf{c}_1, ..., \mathbf{c}_r$ satisfy (5.14.1). Let $\tilde{\mathbf{d}}_1, ..., \tilde{\mathbf{d}}_r$ be defined as above. By the definition, $\|\tilde{\mathbf{d}}_i\| = 1, i = 1, ..., r$. Assume that $1 \le i < j \le r$. Then

$$0 = \langle \mathbf{c}_i, \mathbf{c}_j\rangle = \lambda_i(T^*T)\langle \mathbf{c}_i, \mathbf{c}_j\rangle = \langle T^*T\mathbf{c}_i, \mathbf{c}_j\rangle = \langle T\mathbf{c}_i, T\mathbf{c}_j\rangle \Rightarrow \langle \tilde{\mathbf{d}}_i, \tilde{\mathbf{d}}_j\rangle = 0.$$

Hence, $\{\tilde{\mathbf{d}}_1, ..., \tilde{\mathbf{d}}_r\}$ is an orthonormal system. □

Definition 5.14.2 *Let*

$$\mathbb{R}^n_{\searrow} := \{\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n; x_1 \ge x_2 \ge \cdots \ge x_n\}.$$

For $\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$, let $\underline{\mathbf{x}} = (\underline{x}_1, \dots, \underline{x}_n)^\top \in \mathbb{R}^n_{\searrow}$ be the unique rearrangement of the coordinates of $\mathbf{x}$ in a decreasing order. That is, there exists a permutation π in $[n]$ such that $\underline{x}_i = x_{\pi(i)}$, $i = 1, \dots, n$.
Let $\mathbf{x} = (x_1, \dots, x_n)^\top$, $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$. Then $\mathbf{x}$ is weakly majorized by $\mathbf{y}$ (or $\mathbf{y}$ weakly majorizes $\mathbf{x}$), denoted by $\mathbf{x} \preceq$, if

$$\sum_{i=1}^{k} x_i \le \sum_{i=1}^{k} y_i, \quad k = 1, \dots, n.$$

Also, $\mathbf{x}$ is majorized by $\mathbf{y}$, denoted by $\mathbf{x} \prec \mathbf{y}$, $\mathbf{x} \preceq \mathbf{y}$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

Theorem 5.14.3 (Hardy-Littlewood-Pólya) *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then $\mathbf{x} \prec \mathbf{y}$ if and only if there exists $A \in \Omega_n$ such that $\mathbf{x} = A\mathbf{y}$, [7].*

Let

$$\sigma_i(T) = \sqrt{\lambda_i(T^*T)}, \text{ for } i = 1, ...r, \quad \sigma_i(T) = 0 \text{ for } i > r,$$
$$\sigma_{(p)}(T) := (\sigma_1(T), ..., \sigma_p(T))^\top \in \mathbb{R}^p_{\searrow},\ p \in \mathbb{N}. \tag{5.14.4}$$

Then, $\sigma_i(T) = \sigma_i(T^*), i = 1, ..., \min(m, n)$ are called the *singular values of T and T^**, respectively. Note that the singular values are arranged in a decreasing order. The positive singular values are called *principal singular values of T and T^**, respectively. Note that

$$\begin{aligned}
&\|T\mathbf{c}_i\|^2 = \langle T\mathbf{c}_i, T\mathbf{c}_i\rangle = \langle T^*T\mathbf{c}_i, \mathbf{c}_i\rangle = \lambda_i(T^*T) = \sigma_i^2 \Rightarrow \\
&\|T\mathbf{c}_i\| = \sigma_i,\ i = 1, ..., n, \\
&\|T^*\mathbf{d}_j\|^2 = \langle T^*\mathbf{d}_j, T^*\mathbf{d}_j\rangle = \langle TT^*\mathbf{d}_j, \mathbf{d}_i\rangle = \lambda_i(TT^*) = \sigma_j^2 \Rightarrow \\
&\|T\mathbf{d}_j\| = \sigma_j,\ j = 1, ..., m.
\end{aligned}$$

Let $\{\mathbf{c}_1, ...\mathbf{c}_n\}$ be an orthonormal basis of $\mathbf{V}$ satisfying (5.14.1). Choose an orthonormal basis $\{\mathbf{d}_1, ..., \mathbf{d}_m\}$ as follows:
Set $\mathbf{d}_i := \frac{T\mathbf{c}_i}{\sigma_i}, i = 1, ..., r$. Then, complete the orthonormal set $\{\mathbf{d}_1, ..., \mathbf{d}_r\}$ to an orthonormal basis of $\mathbf{U}$. Since $\operatorname{span}\{\mathbf{d}_1, ..., \mathbf{d}_r\}$ is spanned by all eigenvectors of TT^* corresponding to non-zero eigenvalues of TT^*, it follows that $\ker T^* = \operatorname{span}\{\mathbf{d}_{r+1}, ..., \mathbf{d}_m\}$. Hence, (5.14.2) holds. In these orthonormal bases of $\mathbf{U}$ and $\mathbf{V}$, the operators T and T^* are represented quite simply:

$$\begin{aligned} T\mathbf{c}_i &= \sigma_i(T)\mathbf{d}_i,\ i = 1, ..., n, \quad \text{where } \mathbf{d}_i = 0, \text{ for } i > m, \\ T^*\mathbf{d}_j &= \sigma_j(T)\mathbf{c}_j,\ j = 1, ..., m, \quad \text{where } \mathbf{c}_j = 0, \text{ for } j > n. \end{aligned} \tag{5.14.5}$$

Let

$$\Sigma = [s_{ij}]_{i,j=1}^{m,n},\ s_{ij} = 0, \text{for } i \neq j,\ s_{ii} = \sigma_i, \text{for } i = 1, ..., \min(m.n). \tag{5.14.6}$$

In the case $m \neq n$, we call Σ a diagonal matrix with the diagonal $\sigma_1, ..., \sigma_{\min(m,n)}$. Then, in the bases $[\mathbf{d}_1, ..., \mathbf{d}_m]$ and $[\mathbf{c}_1, ..., \mathbf{c}_n]$, T and T^* are represented by the matrices Σ and $\Sigma^\top$, respectively.

Lemma 5.14.4 *Let $\mathbf{U}$, $\mathbf{V}$, T, Σ, $[\mathbf{c}_1, ..., \mathbf{c}_n]$ and $[\mathbf{d}_1, ..., \mathbf{d}_m]$ be as above. Assume that*

(i) T is presented by the matrix $A \in \mathbb{F}^{m\times n}$ (then T^* is presented by A^*.)

(ii) $U \in \mathbf{U}(m)$ is the unitary matrix representing the change of basis $[\mathbf{d}_1, ..., \mathbf{d}_m]$ to $[\mathbf{c}_1, ..., \mathbf{c}_n]$.

(iii) $V \in \mathbf{U}(n)$ is the unitary matrix representing the change of basis $[\mathbf{u}_1, ..., \mathbf{u}_m]$ to $[\mathbf{v}_1, ..., \mathbf{v}_n]$.

(iv) $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is an orthonormal set.

(v) $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal set.

Then

$$A = U\Sigma V^* \in \mathbb{F}^{m\times n}. \tag{5.14.7}$$

Proof. By the definition, $T\mathbf{v}_j = \sum_{i=1}^m a_{ij}\mathbf{u}_i$. Let $U = [u_{ip}]_{i,p=1}^m, V = [v_{jq}]_{j,q=1}^n$. Then

$$T\mathbf{c}_q = \sum_{j=1}^n v_{jq}T\mathbf{v}_j = \sum_{j=1}^n v_{jq}\sum_{i=1}^m a_{ij}\mathbf{u}_i = \sum_{j=1}^n v_{jq}\sum_{i=1}^m a_{ij}\sum_{p=1}^m \bar{u}_{ip}\mathbf{d}_p.$$

Use the first equality of (5.14.5) to deduce that $U^*AV = \Sigma$. □

Definition 5.14.5 *(5.14.7) is called the singular value decomposition (SVD) of A.*

Theorem 5.14.6 (Singular value decomposition) *Let $A \in \mathbb{F}^{m\times n}$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Then there exists a factorization, called singular value decomposition of A of the form*

$$A = U\Sigma V^*,$$

where $U \in \mathbf{U}(m)$, $V \in \mathbf{U}(n)$ and $\Sigma \in \mathbb{F}^{m\times n}$ is a matrix with non-negative real entries in its diagonal.

Proof. It is immediate from Lemma 5.14.4. □

Proposition 5.14.7 *For the field $\mathbb{F}$, denote by $\mathcal{R}_{m,n,k}(\mathbb{F}) \subset \mathbb{F}^{m\times n}$ the set of all matrices of rank $k \in [\min(m,n)]$ at most. Then, $A \in \mathcal{R}_{m,n,k}(\mathbb{F})$ if and only if A can be expressed as a sum of at most k matrices of rank 1. Furthermore, $\mathcal{R}_{m,n,k}(\mathbb{F})$ is a variety in $\mathbb{F}^{m\times n}$ given by the polynomial conditions: Each $(k+1)\times(k+1)$ minor of A is equal to zero. (By variety one means a set of points in $\mathbb{F}^n$ satisfying a finite number of polynomial equation.)*

For the proof see Problem 5.14.2-2.

Definition 5.14.8 *Let $A \in \mathbb{C}^{m\times n}$ and assume that A has the SVD given by (5.14.7), where $U = [\mathbf{u}_1,\dots,\mathbf{u}_m]$ and $V = [\mathbf{v}_1,\dots,\mathbf{v}_n]$. Denote by $A_k := \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^* \in \mathbb{C}^{m\times n}$, for $k = 1,\dots,\operatorname{rank} A$. For $k > \operatorname{rank} A$, we define $A_k := A\ (= A_{\operatorname{rank} A})$.*

Note that for $1 \le k < \operatorname{rank} A$, the matrix A_k is uniquely defined if and only if $\sigma_k > \sigma_{k+1}$. (See Problem 5.14.2-1.)

An $m\times m$ matrix B is called an $m\times m$ principal submatrix of an $n\times n$ matrix A, if B is obtained from A by removing $n-m$ rows and the same $n-m$ columns. Let B be a square submatrix of A. Then, $\det B$ is called a *minor* of A. Moreover, $\det B$ is called a *principal minor of order m* if B is an $m\times m$ principal submatrix of A.

Theorem 5.14.9 *For the field $\mathbb{F}$ and $A = [a_{ij}] \in \mathbb{F}^{m\times n}$, the following conditions hold:*

$$||A||_F := \sqrt{\operatorname{tr} A^* A} = \sqrt{\operatorname{tr} AA^*} = \sqrt{\sum_{i=1}^{\operatorname{rank} A} \sigma_i(A)^2}. \tag{5.14.8}$$

$$||A||_2 := \max_{\mathbf{x}\in\mathbb{F}^n, ||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 = \sigma_1(A). \tag{5.14.9}$$

$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} ||A-B||_2 = ||A-A_k|| = \sigma_{k+1}(A), k = 1, ..., \operatorname{rank} A - 1. \tag{5.14.10}$$

$$\begin{gathered}\sigma_i(A) \ge \sigma_i((a_{i_p j_q})_{p=1,q=1}^{m',n'}) \ge \sigma_{i+(m-m')+(n-n')}(A), \\ m' \in [m],\ n' \in [n],\ 1 \le i_1 < ... < i_{m'} \le m,\ 1 \le j_1 < ... < j_{n'} \le n.\end{gathered} \tag{5.14.11}$$

Proof. The proof of (5.14.8) is left as Problem 5.14.2-7. We now show the equality in (5.14.9). View A as an operator $A : \mathbb{C}^n \to \mathbb{C}^m$. From the definition of $||A||_2$, it follows

$$||A||_2^2 = \max_{0\ne\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{x}^* A^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \lambda_1(A^*A) = \sigma_1(A)^2,$$

which proves (5.14.9).

We now prove (5.14.10). In the SVD of A, assume that $U = \{\mathbf{u}_1, ..., \mathbf{u}_m\}$ and $V = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$. Then, (5.14.7) is equivalent to the following representation of A:

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,\ \mathbf{u}_1, ..., \mathbf{u}_r \in \mathbb{R}^m,\ \mathbf{v}_1, ..., \mathbf{v}_r \in \mathbb{R}^n,\ \mathbf{u}_i^* \mathbf{u}_j = \mathbf{v}_i^* \mathbf{v}_j = \delta_{ij},\ i, j = 1, ..., r, \tag{5.14.12}$$

where $r = \operatorname{rank} A$. Let $B = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* \in \mathcal{R}_{m,n,k}$. Then, in view of (5.14.9)

$$\|A - B\|_2 = \|\sum_{k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*\|_2 = \sigma_{k+1}.$$

Let $B \in \mathcal{R}_{m,n,k}$. To show (5.14.10), it is enough to show that $\|A - B\|_2 \geq \sigma_{k+1}$. Let

$$\mathbf{W} := \{\mathbf{x} \in \mathbb{R}^n : \quad B\mathbf{x} = 0\}.$$

Then codim $\mathbf{W} \geq k$. Furthermore

$$\|A - B\|_2^2 \geq \max_{\|\mathbf{x}\|_2 = 1, \mathbf{x} \in \mathbf{W}} \|(A - B)\mathbf{x}\|^2 = \max_{\|\mathbf{x}\|_2 = 1, \mathbf{x} \in \mathbf{W}} \mathbf{x}^* A^* A \mathbf{x} \geq \lambda_{k+1}(A^* A) = \sigma_{k+1}^2,$$

where the last inequality follows from the min-max characterization of $\lambda_{k+1}(A^* A)$.

Let $C = [a_{ij_q}]_{i,q=1}^{m,n'}$. Then, $C^* C$ is a principal submatrix of $A^* A$ of dimension n'. The interlacing inequalities between the eigenvalues of $A^* A$ and $C^* C$ yield (5.14.11), for $m' = m$. Let $D = (a_{i_p j_q})_{p,q=1}^{m',n'}$. Then, DD^* is a principle submatrix of CC^*. Use the interlacing properties of the eigenvalues of CC^* and DD^* to deduce (5.14.11). □

We now restate the above materials for linear operators.

Definition 5.14.10 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be finite dimensional vector spaces over* $\mathbb{F}$. *For* $k \in \mathbb{Z}_+$, *denote* $L_k(\mathbf{V}, \mathbf{U}) := \{T \in L(\mathbf{V}, \mathbf{U}) : \operatorname{rank} T \leq k\}$. *Assume furthermore that* $\mathbf{U}$ *and* $\mathbf{V}$ *are IPS. Let* $T \in L(\mathbf{V}, \mathbf{U})$ *and assume that the orthonormal bases of* $[\mathbf{d}_1, \ldots, \mathbf{d}_m]$ *and* $[\mathbf{c}_1, \ldots, \mathbf{c}_n]$ *of* $\mathbf{U}$ *and* $\mathbf{V}$, *respectively satisfy (5.14.5). Define* $T_0 := \mathbf{0}$ *and* $T_k := T$, *for an integer* $k \geq \operatorname{rank} T$. *Let* $k \in [\operatorname{rank} T - 1]$. *Define* $T_k \in L(\mathbf{V}, \mathbf{U})$ *by the equality* $T_k(\mathbf{v}) = \sum_{i=1}^{k} \sigma_i(T) \langle \mathbf{v}, \mathbf{c}_i \rangle \mathbf{d}_i$, *for any* $\mathbf{v} \in \mathbf{V}$.

It is straightforward to show that $T_k \in L_k(\mathbf{V}, \mathbf{U})$ and T_k is unique if and only if $\sigma_k(T) > \sigma_{k+1}(T)$. See Problem 5.14.2-8. Theorem 5.14.9 yields the following corollary:

Corollary 5.14.11 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be finite dimensional IPS over* $\mathbb{F}$ *and* $T : \mathbf{V} \to \mathbf{U}$ *be a linear operator. Then*

$$\|T\|_F := \sqrt{\operatorname{tr} T^* T} = \sqrt{\operatorname{tr} T T^*} = \sqrt{\sum_{i=1}^{\operatorname{rank} T} \sigma_i(T)^2}. \tag{5.14.13}$$

$$\|T\|_2 := \max_{\mathbf{x} \in \mathbf{V}, \|\mathbf{x}\|_2 = 1} \|T\mathbf{x}\|_2 = \sigma_1(T). \tag{5.14.14}$$

$$\min_{Q \in L_k(\mathbf{V}, \mathbf{U})} \|T - Q\|_2 = \sigma_{k+1}(T), \quad k = 1, ..., \operatorname{rank} T - 1. \tag{5.14.15}$$

5.15 Characterizations of singular values

Theorem 5.15.1 *For the field $\mathbb{F}$, assume that $A \in \mathbb{F}^{m\times n}$. Define*

$$H(A) = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in \mathrm{H}_{m+n}(\mathbb{F}). \tag{5.15.1}$$

Then

$$\begin{aligned} &\lambda_i(H(A)) = \sigma_i(A),\ \lambda_{m+n+1-i}(H(A)) = -\sigma_i(A),\ i = 1, ..., \operatorname{rank} A, \\ & \hfill (5.15.2) \\ &\lambda_j(H(A)) = 0,\ j = \operatorname{rank} A + 1, ..., n + m - \operatorname{rank} A. \end{aligned}$$

View A as an operator $A : \mathbb{F}^n \to \mathbb{F}^m$. Choose orthonormal bases $[\mathbf{d}_1, ..., \mathbf{d}_m]$ and $[\mathbf{c}_1, ..., \mathbf{c}_n]$ in $\mathbb{F}^m$ and $\mathbb{F}^n$, respectively satisfying (5.14.5). Then

$$\begin{aligned} &\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix} = \sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix}, \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix} = -\sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix}, \\ &\qquad i = 1, ..., \operatorname{rank} A, \qquad\qquad (5.15.3) \\ &\ker H(A) = \operatorname{span}\{(\mathbf{d}_{r+1}^*, 0)^*, ..., (\mathbf{d}_m^*, 0)^*, (0, \mathbf{c}_{r+1}^*)*, ..., (0, \mathbf{c}_n^*)^*\},\ r = \operatorname{rank} A. \end{aligned}$$

Proof. It is straightforward to show the equalities (5.15.3). Since all the eigenvectors appearing in (5.15.3) are linearly independent, we deduce (5.15.2). □

Corollary 5.15.2 *For the field $\mathbb{F}$, assume that $A \in \mathbb{F}^{m\times n}$. Let $\hat{A} := A[\alpha, \beta] \in \mathbb{F}^{p\times q}$ be a submatrix of A, formed by the set of rows and columns $\alpha \in Q_{p,m}, \beta \in Q_{q,n}$, respectively. Then*

$$\sigma_i(\hat{A}) \le \sigma_i(A) \ \textit{for}\ i = 1, \ldots. \tag{5.15.4}$$

For $l \in [\operatorname{rank} A]$, the equalities $\sigma_i(\hat{A}) = \sigma_i(A), i = 1, \ldots, l$ hold if and only if there exist two orthonormal systems of l right and left singular vectors $\mathbf{c}_1, \ldots, \mathbf{c}_l \in \mathbb{F}^n$, $\mathbf{d}_1, \ldots, \mathbf{d}_l \in \mathbb{F}^n$ satisfying (5.15.3), for $i = 1, \ldots, l$ such that the non-zero coordinates vectors $\mathbf{c}_1, \ldots, \mathbf{c}_l$ and $\mathbf{d}_1, \ldots, \mathbf{d}_l$ are located at the indices β and α, respectively. (See Problem 5.14.2-10.)

Corollary 5.15.3 *Let $\mathbf{V}$ and $\mathbf{U}$ be two IPS over $\mathbb{F}$. Assume that $\mathbf{W}$ is a subspace of $\mathbf{V}$, $T \in L(\mathbf{V}, \mathbf{U})$ and denote by $\hat{T} \in L(\mathbf{W}, \mathbf{U})$ the restriction of T to $\mathbf{W}$. Then, $\sigma_i(\hat{T}) \le \sigma_i(T)$, for any $i \in \mathbb{N}$. Furthermore, $\sigma_i(\hat{T}) = \sigma_i(T)$, for $i = 1, \ldots, l \le \operatorname{rank} T$ if and only if $\mathbf{U}$ contains a subspace spanned by the first l right singular vectors of T. (See Problem 5.14.2-11.)*

We now translate Theorem 5.15.1 to the operator setting.

Lemma 5.15.4 *Let $\mathbf{U}$ and $\mathbf{V}$ be two finite dimensional IPS spaces with the inner products $\langle\cdot,\cdot\rangle_{\mathbf{U}}, \langle\cdot,\cdot\rangle_{\mathbf{V}}$, respectively. Define $\mathbf{W} := \mathbf{V} \oplus \mathbf{U}$ as the induced IPS by*

$$\langle(\mathbf{y}, \mathbf{x}), (\mathbf{v}, \mathbf{u})\rangle_{\mathbf{W}} := \langle\mathbf{y}, \mathbf{v}\rangle_{\mathbf{V}} + \langle\mathbf{x}, \mathbf{u}\rangle_{\mathbf{U}}.$$

Let $T : \mathbf{V} \to \mathbf{U}$ be a linear operator and $T^* : \mathbf{U} \to \mathbf{V}$ be the adjoint of T. Define the operator

$$\hat{T} : \mathbf{W} \to \mathbf{W}, \quad \hat{T}(\mathbf{y}, \mathbf{x}) := (T^*\mathbf{x}, T\mathbf{y}). \tag{5.15.5}$$

Then, $\hat{T}$ is self-adjoint operator and $\hat{T}^2 = T^*T \oplus TT^*$. Hence, the spectrum of $\hat{T}$ is symmetric with respect to the origin and $\hat{T}$ has exactly $2\operatorname{rank} T$ *non-zero eigenvalues. More precisely, if* $\dim \mathbf{U} = m, \dim \mathbf{V} = n$ *then:*

$$\lambda_i(\hat{T}) = -\lambda_{m+n-i+1}(\hat{T}) = \sigma_i(T), \text{ for } i = 1, \dots, \operatorname{rank} T, \tag{5.15.6}$$
$$\lambda_j(\hat{T}) = 0, \text{ for } j = \operatorname{rank} T + 1, \dots, n + m - \operatorname{rank} T.$$

Let $\{\mathbf{d}_1, \dots, \mathbf{d}_{\min(m,n)}\} \in \operatorname{Fr}(\min(m,n), \mathbf{U})$ *and* $\{\mathbf{c}_1, \dots, \mathbf{c}_{\min(m,n)}\} \in \operatorname{Fr}(\min(m,n), \mathbf{V})$ *be the set of vectors satisfying (5.14.5). Define*

$$\mathbf{z}_i := \frac{1}{\sqrt{2}}(\mathbf{c}_i, \mathbf{d}_i), \mathbf{z}_{m+n-i+1} := \frac{1}{\sqrt{2}}(\mathbf{c}_i, -\mathbf{d}_i), i = 1, \dots, \min(m,n). \tag{5.15.7}$$

Then, $\{\mathbf{z}_1, \mathbf{z}_{m+n}, \dots, \mathbf{z}_{\min(m,n)}, \mathbf{z}_{m+n-\min(m,n)+1}\} \in \operatorname{Fr}(2\min(m,n), \mathbf{W})$. *Furthermore,* $\hat{T}\mathbf{z}_i = \sigma_i(T)\mathbf{z}_i$ *and* $\hat{T}\mathbf{z}_{m+n-i+1} = -\sigma_i(T)\mathbf{z}_{m+n-i+1}$*, for* $i = 1, \dots, \min(m,n)$.

The proof is left as Problems 5.14.2-12.

Theorem 5.15.5 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be finite dimensional IPS over* $\mathbb{C}$*, with* $\dim \mathbf{U} = m$, $\dim \mathbf{V} = n$ *and* $T : \mathbf{V} \to \mathbf{U}$ *be a linear operator. Then, for each* $k \in [\min(m,n)]$*, we have*

$$\sum_{i=1}^{k} \sigma_i(T) = \max_{\{\mathbf{f}_1,\dots,\mathbf{f}_k\} \in \operatorname{Fr}(k,\mathbf{U}), \{\mathbf{g}_1,\dots,\mathbf{g}_k\} \in \operatorname{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}} = \tag{5.15.8}$$
$$\max_{\{\mathbf{f}_1,\dots,\mathbf{f}_k\} \in \operatorname{Fr}(k,\mathbf{U}), \{\mathbf{g}_1,\dots,\mathbf{g}_k\} \in \operatorname{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} |\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}|.$$

$\Re$ *stands for the real part of a complex number*
Furthermore, $\sum_{i=1}^k \sigma_i(T) = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$*, for some two k-orthonormal frames* $F_k = \{\mathbf{f}_1, \dots, \mathbf{f}_k\}, G_k = \{\mathbf{g}_1, \dots, \mathbf{g}_k\}$ *if and only* $\operatorname{span}\{(\mathbf{g}_1, \mathbf{f}_1), \dots, (\mathbf{g}_k, \mathbf{f}_k)\}$ *is spanned by k eigenvectors of* $\hat{T}$ *corresponding to the first k eigenvalues of* $\hat{T}$.

Proof. Assume that $\{\mathbf{f}_1, \dots, \mathbf{f}_k\} \in \operatorname{Fr}(k, \mathbf{U})$ and $\{\mathbf{g}_1, \dots, \mathbf{g}_k\} \in \operatorname{Fr}(k, \mathbf{V})$. Let $\mathbf{w}_i := \frac{1}{\sqrt{2}}(\mathbf{g}_i, \mathbf{f}_i), i = 1, \dots, k$. Then, $\{\mathbf{w}_1, \dots, \mathbf{w}_k\} \in \operatorname{Fr}(k, \mathbf{W})$. A straightforward calculation shows $\sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$. The maximal characterization of $\sum_{i=1}^k \lambda_i(\hat{T})$, (Theorem 5.10.8), and (5.15.6) yield the inequality $\sum_{i=1}^k \sigma_i(\hat{T}) \ge \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$ for $k \in [\min(m,n)]$. Let $\mathbf{c}_1, \dots, \mathbf{c}_{\min(m,n)}$ and $\mathbf{d}_1, \dots, \mathbf{d}_{\min(m,n)}$ satisfy (5.14.5). Then, Lemma 5.15.4 yields that $\sum_{i=1}^k \sigma_i(\hat{T}) = \sum_{i=1}^k \Re\langle T\mathbf{c}_i, \mathbf{d}_i\rangle_{\mathbf{U}}$, for $k \in [\min(m,n)]$. This proves the first equality of (5.15.8). The second equality of (5.15.8) is straightforward. (See Problem 5.14.2-13).
Assume now that $\sum_{i=1}^k \sigma_i(T) = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$, for some two k-orthonormal frames $F_k = \{\mathbf{f}_1, \dots, \mathbf{f}_k\}, G_k = \{\mathbf{g}_1, \dots, \mathbf{g}_k\}$. Define $\mathbf{w}_1, \dots, \mathbf{w}_k$ as above. The above arguments yield that $\sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \lambda_i(\hat{T})$. Theorem 5.10.8 implies that $\operatorname{span}\{(\mathbf{g}_1, \mathbf{f}_1), \dots, (\mathbf{g}_k, \mathbf{f}_k)\}$ is spanned by k eigenvectors of $\hat{T}$ corresponding to the first k eigenvalues of $\hat{T}$. Vice versa, assume that $\{\mathbf{f}_1, \dots, \mathbf{f}_k\} \in \operatorname{Fr}(k, \mathbf{U}), \{\mathbf{g}_1, \dots, \mathbf{g}_k\} \in \operatorname{Fr}(k, \mathbf{V})$ and $\operatorname{span}\{(\mathbf{g}_1, \mathbf{f}_1), \dots, (\mathbf{g}_k, \mathbf{f}_k)\}$ is spanned by k eigenvectors of $\hat{T}$ corresponding to the first k eigenvalues of $\hat{T}$. Define $\{\mathbf{w}_1, \dots, \mathbf{w}_k\} \in \operatorname{Fr}(\mathbf{W})$ as above. Then, $\operatorname{span}\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ contains k linearly independent eigenvectors corresponding to the first k eigenvalues of $\hat{T}$. Theorem 5.10.8 and Lemma 5.15.4 conclude that $\sigma_i(T) = \sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$. □

Theorem 5.15.6 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be* m *and* n *dimensional IPS spaces with* $\dim\mathbf{U} = m$ *and* $\dim\mathbf{V} = n$. *Assume that* S *and* $T:\mathbf{V}\to\mathbf{U}$ *be two linear operators. Then*

$$\Re\operatorname{tr}(S^*T) \le \sum_{i=1}^{\min(m,n)} \sigma_i(S)\sigma_i(T). \tag{5.15.9}$$

Equality holds if and only if there exist two orthonormal sets $\{\mathbf{d}_1,\dots,\mathbf{d}_{\min(m,n)}\}\in \operatorname{Fr}(\min(m,n),\mathbf{U})$ *and* $\{\mathbf{c}_1,\dots,\mathbf{c}_{\min(m,n)}\}\in\operatorname{Fr}(\min(m,n),\mathbf{V})$ *such that*

$$S\mathbf{c}_i = \sigma_i(S)\mathbf{d}_i, T\mathbf{c}_i = \sigma_i(T)\mathbf{d}_i, S^*\mathbf{d}_i = \sigma_i(S)\mathbf{c}_i, T^*\mathbf{d}_i = \sigma_i(T)\mathbf{c}_i, i = 1,\dots,\min(m,n). \tag{5.15.10}$$

Proof. Let $A, B\in\mathbb{C}^{n\times m}$. Then
$\operatorname{tr}B^*A = \overline{\operatorname{tr}AB^*}$. Hence, $2\Re\operatorname{tr}AB^* = \operatorname{tr}H(A)H(B)$. Therefore, $2\Re\operatorname{tr}S^*T = \operatorname{tr}\hat S\hat T$. Use Theorem 5.12.1, for $\hat S,\hat T$ and Lemma 5.15.4 to deduce (5.15.9). Equality in (5.15.9) holds if and only if $\operatorname{tr}\hat S\hat T = \sum_{i=1}^{m+n}\lambda_i(\hat S)\lambda_i(\hat T)$.
Clearly, the assumptions that $\{\mathbf{d}_1,\dots,\mathbf{d}_{\min(m,n)}\}\in\operatorname{Fr}(\min(m,n),\mathbf{U})$, $\{\mathbf{c}_1,\dots,\mathbf{c}_{\min(m,n)}\}\in\operatorname{Fr}(\min(m,n),\mathbf{V})$ and the equalities (5.15.10) imply equality in (5.15.9).
Assume equality in (5.15.9). Theorem 5.12.1 and the definitions of $\hat S$ and $hatT$ yield the existence $\{\mathbf{d}_1,\dots,\mathbf{d}_{\min(m,n)}\}\in\operatorname{Fr}(\min(m,n),\mathbf{U})$ and $\{\mathbf{c}_1,\dots,\mathbf{c}_{\min(m,n)}\}\in\operatorname{Fr}(\min(m,n),\mathbf{V})$, such that (5.15.10) hold. □

Theorem 5.15.7 *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be finite dimensional IPS over* $\mathbb{F}$ *and* $T:\mathbf{V}\to\mathbf{U}$ *be a linear operator. Then*

$$\min_{Q\in L_k(\mathbf{V},\mathbf{U})}||T-Q||_F = \sqrt{\sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T)},\quad k = 1,...,\operatorname{rank}T-1. \tag{5.15.11}$$

Furthermore, $||T-Q||_F = \sqrt{\sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T)}$, *for some* $Q\in L_k(\mathbf{V},\mathbf{U}), k<\operatorname{rank}T$, *if and only there* $Q = T_k$, *where* T_k *is defined in Definition 5.14.10.*

Proof. Use Theorem 5.15.6 to deduce that for any $Q\in L(\mathbf{V},\mathbf{U})$ one has

$$||T-Q||_F^2 = \operatorname{tr}T^*T - 2\Re\operatorname{tr}Q^*T + \operatorname{tr}Q^*Q \ge$$
$$\sum_{i=1}^{\operatorname{rank}T}\sigma_i^2(T) - 2\sum_{i=1}^{k}\sigma_i(T)\sigma_i(Q) + \sum_{i=1}^{k}\sigma_i^2(Q) =$$
$$\sum_{i=1}^{k}(\sigma_i(T)-\sigma_i(Q))^2 + \sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T) \ge \sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T).$$

Clearly, $||T-T_k||_F^2 = \sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T)$. Hence, (5.15.11) holds. Vice versa if $Q\in L_k(\mathbf{V},\mathbf{U})$ and $||T-Q||_F^2 = \sum_{i=k+1}^{\operatorname{rank}T}\sigma_i^2(T)$, then the equality case in Theorem 5.15.6 implies that $Q = T_k$. □

Corollary 5.15.8 *For the field* $\mathbb{F}$, *let* $A\in\mathbb{F}^{m\times n}$. *Then*

$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})}||A-B||_F = \sqrt{\sum_{i=k+1}^{\operatorname{rank}A}\sigma_i^2(A)},\quad k = 1,...,\operatorname{rank}A-1. \tag{5.15.12}$$

Furthermore, $||A-B||_F = \sqrt{\sum_{i=k+1}^{\operatorname{rank} A} \sigma_i^2(A)}$, *for some* $B \in \mathcal{R}_{m,n,k}(\mathbb{F}), k < \operatorname{rank} A$, *if and only if* $B = A_k$, *where* A_k *is defined in Definition 5.14.8.*

Theorem 5.15.9 *For the matrix* $A \in \mathbb{F}^{m\times n}$, *we have*

$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} \sum_{i=1}^{j} \sigma_i(A-B) = \sum_{i=k+1}^{k+j} \sigma_i(A), \quad j = 1,...,\min(m,n)-k,\ k = 1,\dots,\min(m,n)-1. \tag{5.15.13}$$

Proof. Clearly, for $B = A_k$, we have the equality $\sum_{i=1}^{j}\sigma_i(A-B) = \sum_{i=k+1}^{k+j}\sigma_i(A)$. Let $B \in \mathcal{R}_{m,n,k}(\mathbb{F})$. Assume that $\mathbf{X} \in \operatorname{Gr}(k,\mathbb{C}^n)$ is a subspace which contains the columns of B. Let $\mathbf{W} = \{(\mathbf{0}^\top, \mathbf{x}^\top)^\top \in \mathbb{F}^{m+n}, \mathbf{x} \in \mathbf{X}\}$. Observe that for any $\mathbf{z} \in \mathbf{W}^\perp$, one has the equality $\mathbf{z}^* H((A-B))\mathbf{z} = \mathbf{z}^* H(A)\mathbf{z}$. Combine Theorems 5.10.9 and 5.15.1 to deduce $\sum_{i=1}^{j}\sigma_i(B-A) \geq \sum_{i=k+1}^{k+j}\sigma_i(A)$. □

Theorem 5.15.10 *Let* $\mathbf{V}$ *be an n-dimensional IPS over* $\mathbb{C}$ *and* $T : \mathbf{V} \to \mathbf{V}$ *be a linear operator. Assume the n eigenvalues of* T $\lambda_1(T),\dots,\lambda_n(T)$ *are arranged the order* $|\lambda_1(T)| \geq \dots \geq |\lambda_n(T)|$. *Let* $\boldsymbol{\lambda}_a(T) := (|\lambda_1(T)|,\dots,|\lambda_n(T)|), \boldsymbol{\sigma}(T) := (\sigma_1(T),\dots,\sigma_n(T))$. *Then,* $\boldsymbol{\lambda}_a(T) \preceq \boldsymbol{\sigma}(T)$. *That is*

$$\sum_{i=1}^{k} |\lambda_i(T)| \leq \sum_{i=1}^{k}\sigma_i(T), \quad i = 1,\dots,n. \tag{5.15.14}$$

Furthermore, $\sum_{i=1}^{k}|\lambda_i(T)| = \sum_{i=1}^{k}\sigma_i(T)$, *for some* $k \in [n]$ *if and only if the following conditions are satisfied. There exists an orthonormal basis* $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ *of* $\mathbf{V}$ *such that:*

1. $T\mathbf{x}_i = \lambda_i(T)\mathbf{x}_i, T^*\mathbf{x}_i = \overline{\lambda_i(T)}\mathbf{x}_i$, *for* $i = 1,\dots,k$.

2. *Denote by* $S : \mathbf{U} \to \mathbf{U}$ *the restriction of* T *to the invariant subspace* $\mathbf{U} = \operatorname{span}\{\mathbf{x}_{k+1},\dots,\mathbf{x}_n\}$. *Then,* $||S||_2 \leq |\lambda_k(T)|$.

Proof. Use Theorem 5.8.14 to choose an orthonormal basis $\{\mathbf{g}_1,\dots,\mathbf{g}_n\}$ of $\mathbf{V}$ such that T is represented by an upper diagonal matrix $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ such that $a_{ii} = \lambda_i(T), i = 1,\dots,n$. Let $\epsilon_i \in \mathbb{C}, |\epsilon_i| = 1$ such that $\bar{\epsilon}_i\lambda_i(T) = |\lambda_i(T)|$, for $i = 1,\dots,n$. Let $S \in L(\mathbf{V})$ be presented in the basis $\{\mathbf{g}_1,\dots,\mathbf{g}_n\}$ by a diagonal matrix $\operatorname{diag}(\epsilon_1,\dots,\epsilon_k,0,\dots,0)$. Clearly, $\sigma_i(S) = 1$, for $i = 1,\dots,k$ and $\sigma_i(S) = 0$, for $i = k+1,\dots,n$. Furthermore, $\Re \operatorname{tr} S^*C = \sum_{i=1}^{k}|\lambda_i(T)|$. Hence, Theorem 5.15.6 yields (5.15.14).

Assume now that $\sum_{i=1}^{k}|\lambda_i(T)| = \sum_{i=1}^{k}\sigma_i(T)$. Hence, equality sign holds in (5.15.9). Hence, there exist two orthonormal bases $\{\mathbf{c}_1,\dots,\mathbf{c}_n\}$ and $\{\mathbf{d}_1,\dots,\mathbf{d}_n\}$ in $\mathbf{V}$ such that (5.15.10) holds. It easily follows that $\{\mathbf{c}_1,\dots,\mathbf{c}_k\}$ and $\{\mathbf{d}_1,\dots,\mathbf{d}_k\}$ are orthonormal bases of $\mathbf{W} := \operatorname{span}\{\mathbf{g}_1,\dots,\mathbf{g}_k\}$. Hence, $\mathbf{W}$ is an invariant subspace of T and T^*. Then, $A = A_1 \oplus A_2$, i.e. A is a block diagonal matrix. Thus, $A_1 = [a_{ij}]_{i,j=1}^{k} \in \mathbb{C}^{k\times k}, A_2 = [a_{ij}]_{i,j=k+1}^{n} \in \mathbb{C}^{(n-k)\times(n-k)}$ represent the restriction of T to $\mathbf{W}, \mathbf{U} := \mathbf{W}^\perp$, denoted by T_1 and T_2, respectively. Hence, $\sigma_i(T_1) = \sigma_i(T)$, for $i = 1,\dots,k$. Note that the restriction of S to $\mathbf{W}$, denoted by S_1 is given by the diagonal matrix $D_1 := \operatorname{diag}(\epsilon_1,\dots,\epsilon_k) \in \mathbf{U}(k)$. Next, (5.15.10) yields that $S_1^{-1}T_1\mathbf{c}_i = \sigma_i(T)\mathbf{c}_i$, for $i = 1,\dots,k$, i.e. $\sigma_1(T),\dots,\sigma_k(T)$ are the eigenvalues of $S_1^{-1}T_1$.

Clearly, $S_1^{-1}T_1$ is presented in the basis $[\mathbf{g}_1, \dots, \mathbf{g}_k]$ by the matrix $D_1^{-1}A_1$, which is a diagonal matrix with $|\lambda_1(T)|, \dots, |\lambda_k(T)|$ on the main diagonal. That is, $S_1^{-1}T_1$ has eigenvalues $|\lambda_1(T)|, \dots, |\lambda_k(T)|$. Therefore, $\sigma_i(T) = |\lambda_i(T)|$, for $i = 1, \dots, k$. Theorem 5.14.9 yields that

$$\operatorname{tr} A_1^* A_1 = \sum_{i,j=1}^{k} |a_{ij}|^2 = \sum_{i=1}^{k} \sigma_i^2(A_1) = \sum_{i=1}^{k} \sigma_i^2(T_1) = \sum_{i=1}^{k} |\lambda_i(T)|^2.$$

As $\lambda_1(T), \dots, \lambda_k(T)$ are the diagonal elements of A_1, it follows from the above equality that A_1, is a diagonal matrix. Hence, we can choose $\mathbf{x}_i = \mathbf{g}_i$, for $i = 1, \dots, n$ to obtain the part *1* of the equality case.
Let $T\mathbf{x} = \lambda\mathbf{x}$, where $\|\mathbf{x}\| = 1$ and $\rho(T) = |\lambda|$. Recall that $\|T\|_2 = \sigma_1(T)$, where $\sigma_1(T)^2 = \lambda_1(T^*T)$ is the maximal eigenvalue of the self-adjoint operator T^*T. The maximum characterization of $\lambda_1(T^*T)$ yields that $|\lambda|^2 = \langle T\mathbf{x}, T\mathbf{x}\rangle = \langle T^*T\mathbf{x}, \mathbf{x}\rangle \le \lambda_1(T^*T) = \|T\|_2^2$. Hence, $\rho(T) \le \|T\|_2$.
Assume now that $\rho(T) = \|T\|_2$. If $\rho(T) = 0$, then $\|T\|_2 = 0$. This means $T = 0$ and theorem holds trivially in this case. Assume that $\rho(T) > 0$. Hence, the eigenvector $\mathbf{x}_1 := \mathbf{x}$ is also the eigenvector of T^*T corresponding to $\lambda_1(T^*T) = |\lambda|^2$. Hence, $|\lambda|^2\mathbf{x} = T^*T\mathbf{x} = T^*(\lambda\mathbf{x})$, which implies that $T^*\mathbf{x} = \bar{\lambda}\mathbf{x}$. Let $\mathbf{U} = \operatorname{span}(\mathbf{x})^\perp$ be the orthogonal complement of $\operatorname{span}(\mathbf{x})$. Since $T\operatorname{span}(\mathbf{x}) = \operatorname{span}(\mathbf{x})$, it follows that $T^*\mathbf{U} \subseteq \mathbf{U}$. Similarly, since $T^*\operatorname{span}(\mathbf{x}) = \operatorname{span}(\mathbf{x})$, then $T\mathbf{U} \subseteq \mathbf{U}$. Thus, $\mathbf{V} = \operatorname{span}(\mathbf{x}) \oplus \mathbf{U}$ and $\operatorname{span}(\mathbf{x})$ and $\mathbf{U}$ are invariant subspaces of T and T^*. Hence, $\operatorname{span}(\mathbf{x})$ and $\mathbf{U}$ are invariant subspaces of T^*T and TT^*. Let T_1 be the restriction of T to $\mathbf{U}$. Then, $T_1^*T_1$ is the restriction of T^*T. Therefore, $\|T_1\|_2^2 \ge \lambda_1(T^*T) = \|T\|_2^2$. The above result implies the equality $\rho(T) = \|T\|_2$. □

Corollary 5.15.11 *Let $\mathbf{U}$ be an n-dimensional IPS over $\mathbb{C}$ and $T : \mathbf{U} \to \mathbf{U}$ be a linear operator. Denote by $|\lambda(T)| = (|\lambda_1(T)|, ..., |\lambda_n(T)|)^T$ the absolute eigenvalues of T, (counting with their multiplicities), arranged in a decreasing order. Then, $|\lambda(T)| = (\sigma_1(T), ..., \sigma_n(T))^\top$ if and only if T is a normal operator.*

5.15.1 Worked-out Problems

1. Let $A \in \mathbb{C}^{n\times n}$ and $\sigma_1(A) \ge \dots \ge \sigma_n(A) \ge 0$ be the singular values of A. Let $\lambda_1(A), \dots, \lambda_n(A)$ be the eigenvalues of A listed with their multiplicities and $|\lambda_1(A)| \ge \dots \ge |\lambda_n(A)|$.

   (a) Show that $|\lambda_1(A)| \le \sigma_1(A)$.

   (b) What is a necessary and sufficient condition for the equality $|\lambda_1(A)| = \sigma_1(A)$?

   (c) Is it true that $\sigma_n(A) \le |\lambda_n(A)|$?

   (d) Show that $\sum_{i=1}^n \sigma_i(A)^2 \ge \sum_{i=1}^n |\lambda_i(A)|^2$.

   Solution:

   (a) Let $A\mathbf{x} = \lambda_1(A)\mathbf{x}$, for $\|\mathbf{x}\| = 1$. Then, $\|A\mathbf{x}\| = |\lambda_1(A)|\|\mathbf{x}\| = |\lambda_1(A)| \le \max_{\|\mathbf{y}\|=1}(\|A\mathbf{y}\|) = \sigma_1(A)$.

(b) Equality holds if and only if $A^*A\mathbf{x} = \lambda_1(A^*A)\mathbf{x}$, i.e. $A^*\mathbf{x} = \overline{\lambda}_1\mathbf{x}$ and $|\lambda_1(A)| = \sigma_1(A)$.

(c) Yes. We have $\sigma_n(A)^2 = \min_{\|\mathbf{y}\|=1} \mathbf{y}^*A^*A\mathbf{y}$. Now, $A\mathbf{y} = \lambda_n(A)\mathbf{y}$, $\|\mathbf{y}\| = 1$. Hence, $\sigma_n(A)^2 \le |\lambda_n(A)|^2$.

(d) Let $B = UAU^*$, where U is unitary and B is upper-triangular. If $B = \begin{bmatrix} \lambda_1 & b_{12} & \dots & & b_{1n} \\ 0 & \lambda_2 & b_{23} & \dots & b_{2n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & \dots & & 0 & \lambda_n \end{bmatrix} = [b_{ij}]$, then $\operatorname{tr} BB^* = \sum_{i,j=0}^n |b_{ij}|^2 = \sum_{i=1}^n |\lambda_i(B)|^2 + \sum_{i=2}^n \sum_{j=i+1}^n |b_{ij}|^2 = \sum_{i=1}^n \sigma_i(B)^2$. Then, $\sum_{i=1}^n |\lambda_i(B)|^2 \le \sum_{i=1}^n \sigma_i(B)^2$.

2. Find the SVD of $A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$.

Solution:
First, we compute the singular values σ_i by finding the eigenvalues of $AA^\top$.

$$AA^\top = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}.$$

The characteristic polynomial is $\det(AA^\top - zI) = z^2 - 34z + 225 = (z-25)(z-9)$, so the singular values are $\sigma_1 = \sqrt{25} = 5$ and $\sigma_2 = \sqrt{9} = 3$.
Now we find an orthonormal set of eigenvectors of $A^\top A$. The eigenvalues of $A^\top A$ are 25, 9, and 0, and since $A^\top A$ is symmetric we know that the eigenvectors will be orthogonal.
For $\lambda = 25$, we have

$$A^\top A - 25I = \begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix},$$

which has the row reduced form $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. A unit-length vector in the kernel of that matrix is $v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$.

For $\lambda = 9$, we have $A^\top A - 9I = \begin{bmatrix} 4 & 12 & 2 \\ 12 & 4 & -2 \\ 2 & -2 & -1 \end{bmatrix}$, which row-reduces to $\begin{bmatrix} 1 & 0 & -\frac{1}{4} \\ 0 & 1 & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}$.

A unit-length vector in the kernel is $v_2 = \begin{bmatrix} \frac{1}{\sqrt{18}} \\ -\frac{1}{\sqrt{18}} \\ \frac{4}{\sqrt{18}} \end{bmatrix}$.

For $\lambda = 0$, we have $A^\top A - 0I = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix}$, which row-reduces to $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$.

A unit-length vector in its kernel is $\begin{bmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}$.

So at this point, we know that

$$A = U\Sigma V^\top = U\begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{18}} & -\frac{1}{\sqrt{18}} & \frac{4}{\sqrt{18}} \\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

Finally, we can compute U by the formula $\sigma u_i = Av_i$. This gives $U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$.

So the SVD is:

$$A = U\Sigma V^\top = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{18}} & -\frac{1}{\sqrt{18}} & \frac{4}{\sqrt{18}} \\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

5.15.2 Problems

1. Let $\mathbf{U}$ and $\mathbf{V}$ be two finite dimensional inner product spaces. Assume that $T \in L(\mathbf{U}, \mathbf{V})$. Show that for any complex number $t \in \mathbb{C}$, $\sigma_i(tT) = |t|\sigma_i(T)$, for all i.

2. Prove Proposition 5.14.7. (Use SVD to prove the nontrivial part of the Proposition.)

3. Let $A \in \mathbb{C}^{m\times n}$ and assume that $U \in \mathbf{U}_m$ and $V \in \mathbf{V}_n$. Show that $\sigma_i(UAV) = \sigma_i(A)$, for all i.

4. Let $A \in \mathrm{GL}(n, \mathbb{C})$. Show that $\sigma_1(A^{-1}) = \sigma_n(A)^{-1}$.

5. Let $\mathbf{U}$ and $\mathbf{V}$ be two IPS of dimensions m and n, respectively. Assume that

$$\mathbf{U} = \mathbf{U}_1 \oplus \mathbf{U}_2, \dim \mathbf{U}_1 = m_1, \dim \mathbf{U}_2 = m_2, \mathbf{U}_1 \perp \mathbf{U}_2,$$
$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2, \dim \mathbf{V}_1 = n_1, \dim \mathbf{V}_2 = n_2, \mathbf{V}_1 \perp \mathbf{V}_2.$$

   Suppose furthermore that $T \in L(\mathbf{V}, \mathbf{U})$, $T\mathbf{V}_1 \subseteq \mathbf{U}_1$ and $T\mathbf{V}_2 \subseteq \mathbf{U}_2$. Let $T_i \in L(\mathbf{V}_i, \mathbf{U}_i)$ be the restriction of T to $\mathbf{V}_i$, for $i = 1, 2$. Then, $\operatorname{rank} T = \operatorname{rank} T_1 + \operatorname{rank} T_2$ and $\{\sigma_1(T), \dots, \sigma_{\operatorname{rank} T}(T)\} = \{\sigma_1(T_1), \dots, \sigma_{\operatorname{rank} T_1}(T_1)\} \cup \{\sigma_1(T_2), \dots, \sigma_{\operatorname{rank} T_2}(T_2)\}$.

6. Let the assumptions of the Definition 5.14.8 hold. Show that for $1 \le k < \operatorname{rank} A$, A_k is uniquely defined if and only if $\sigma_k > \sigma_{k+1}$.

7. Prove the equalities in (5.14.8).

8. Let the assumptions of Definition 5.14.10 hold. Show that for $k \in [\operatorname{rank} T - 1]$, $\operatorname{rank} T_k = k$ and T_k is unique if and only if $\sigma_k(T) > \sigma_{k+1}(T)$.

9. Let $\mathbf{V}$ be an n-dimensional IPS. Assume that $T \in L(\mathbf{V})$ is a normal operator. Let $\lambda_1(T), \dots, \lambda_n(T)$ be the eigenvalues of T arranged in the order $|\lambda_1(T)| \geq \dots \geq |\lambda_n(T)|$. Show that $\sigma_i(T) = |\lambda_i(T)|$, for $i = 1, \dots, n$.

10. Let the assumptions of Corollary 5.15.2 hold.

    (a) Since $\operatorname{rank} \hat{A} \leq \operatorname{rank} A$, show that the inequalities (5.15.4) reduce to $\sigma_i(\hat{A}) = \sigma_i(A) = 0$ for $i > \operatorname{rank} A$.

    (b) Since $H(\hat{A})$ is a submatrix of $H(A)$, use the Cauchy interlacing principle to deduce the inequalities (5.15.4) for $i = 1, \dots, \operatorname{rank} A$. Furthermore, if $p' := m - \#\alpha, q' = n - \#\beta$, then the Cauchy interlacing principle gives the complementary inequalities $\sigma_i(\hat{A}) \geq \sigma_{i+p'+q'}(A)$ for any $i \in \mathbb{N}$.

    (c) Assume that $\sigma_i(\hat{A}) = \sigma_i(A)$, for $i = 1, \dots, l \leq \operatorname{rank} A$. Compare the maximal characterization of the sum of the first k eigenvalues of $H(\hat{A})$ and $H(A)$ given by Theorem 5.10.8, for $k = 1, \dots, l$ to deduce the last part of Corollary (5.15.2).

11. Prove Corollary 5.15.3 by choosing any orthonormal basis in $\mathbf{U}$, an orthonormal basis in $\mathbf{V}$ whose first $\dim \mathbf{W}$ elements span $\mathbf{W}$, and using Problem 10.

12. Prove Lemma 5.15.4.

13. Under the assumptions of Theorem 5.15.5, show the equalities.

$$\max_{\{\mathbf{f}_1,\dots,\mathbf{f}_k\} \in \mathrm{Fr}(k,\mathbf{U}), \{\mathbf{g}_1,\dots,\mathbf{g}_k\} \in \mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} \Re \langle T\mathbf{g}_i, \mathbf{f}_i \rangle_{\mathbf{U}} =$$
$$\max_{\{\mathbf{f}_1,\dots,\mathbf{f}_k\} \in \mathrm{Fr}(k,\mathbf{U}), \{\mathbf{g}_1,\dots,\mathbf{g}_k\} \in \mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} |\langle T\mathbf{g}_i, \mathbf{f}_i \rangle_{\mathbf{U}}|.$$

14. Let $\mathbf{U}$ and $\mathbf{V}$ be and finite dimensional IPS. Assume that P and $T \in L(\mathbf{U}, \mathbf{V})$. Show that $\Re \operatorname{tr}(P^*T) \geq -\sum_{i=1}^{\min(m,n)} \sigma_i(S)\sigma_i(T)$. Equality holds if and only if $S = -P$ and T satisfies the conditions of Theorem 5.15.6.

15. Show that if $A, B \in \mathbb{F}^{m \times n}$ and B is a submatrix of A, then $\|B\|_\infty \leq \|A\|_\infty$.

16. Let $A \in \mathbb{F}^{m \times n}$ be with singular value decomposition $U\Sigma V^*$.

    (a) Show that $\|A\|_2 = \sigma_1$ (the largest singular value).

    (b) If A is invertible, show that

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}.$$

17. Show that the singular values of a positive definite hermitian matrix are the same as its eigenvalues.

5.16 Moore-Penrose generalized inverse

The purpose of constructing a generalized inverse is to obtain a matrix that can serve as the inverse in some sense for a wider class of matrices than invertible ones. A generalized inverse exists for an arbitrary matrix $A \in \mathbb{C}^{m\times n}$.
Let $A \in \mathbb{C}^{m\times n}$. Then, (5.14.12) is called the *reduced* SVD of A. It can be written as

$$A = U_r\Sigma_r V_r^*, \quad r = \operatorname{rank} A, \quad \Sigma_r := \operatorname{diag}(\sigma_1(A),\dots,\sigma_r(A)) \in S_r(\mathbb{R}), \tag{5.16.1}$$
$$U_r = [\mathbf{u}_1,\dots,\mathbf{u}_r] \in \mathbb{C}^{m\times r}, V_r = [\mathbf{v}_1,\dots,\mathbf{v}_r] \in \mathbb{C}^{n\times r}, U_r^*U_r = V_r^*V_r = I_r,.$$

Note that

$$AA^*\mathbf{u}_i = \sigma_i(A)^2\mathbf{u}_i, A^*A\mathbf{v}_i = \sigma_i(A)^2\mathbf{v}_i,$$
$$\mathbf{v}_i = \frac{1}{\sigma_i(A)}A^*\mathbf{u}_i, \mathbf{u}_i = \frac{1}{\sigma_i(A)}A\mathbf{v}_i, i = 1,\dots,r.$$

Then
$$A^\dagger := V_r\Sigma_r^{-1}U_r^* \in \mathbb{C}^{n\times m}, \tag{5.16.2}$$
is the *Moore-Penrose* generalized inverse of A. If $A \in \mathbb{R}^{m\times n}$, then we assume that $U \in \mathbb{R}^{m\times r}$ and $V \in R^{n\times r}$, i.e. U and V are real values matrices over the real numbers $\mathbb{R}$.

Theorem 5.16.1 *Let $A \in \mathbb{C}^{m\times n}$. Then, the Moore-Penrose generalized inverse $A^\dagger \in \mathbb{C}^{n\times m}$ satisfies the following properties.*

1. $\operatorname{rank} A = \operatorname{rank} A^\dagger$.
2. $A^\dagger AA^\dagger = A^\dagger$, $AA^\dagger A = A$, $A^*AA^\dagger = A^\dagger AA^* = A^*$.
3. *$A^\dagger A$ and $AA^\dagger$ are Hermitian non-negative definite idempotent matrices, i.e. $(A^\dagger A)^2 = A^\dagger A$ and $(AA^\dagger)^2 = AA^\dagger$, having the same rank as A.*
4. *The least square solution of $A\mathbf{x} = \mathbf{b}$, i.e. the solution of the system $A^*A\mathbf{x} = A^*\mathbf{b}$, has a solution $\mathbf{y} = A^\dagger\mathbf{b}$. This solution has the minimal norm $||\mathbf{y}||$, for all possible solutions of $A^*A\mathbf{x} = A^*\mathbf{b}$.*
5. *If $\operatorname{rank} A = n$, then $A^\dagger = (A^*A)^{-1}A^*$. In particular, if $A \in \mathbb{C}^{n\times n}$ is invertible then $A^\dagger = A^{-1}$.*

To prove the above theorem we need the following proposition.

Proposition 5.16.2 *Let $E \in \mathbb{C}^{l\times m}$ and $G \in \mathbb{C}^{m\times n}$. Then* $\operatorname{rank} EG \le \min(\operatorname{rank} E, \operatorname{rank} G)$. *If $l = m$ and E is invertible, then* $\operatorname{rank} EG = \operatorname{rank} G$. *If $m = n$ and G is invertible, then* $\operatorname{rank} EG = \operatorname{rank} E$.

Proof. Let $\mathbf{e}_1,\dots,\mathbf{e}_m \in \mathbb{C}^l, \mathbf{g}_1,\dots,\mathbf{g}_n \in \mathbb{C}^m$ be the columns of E and G, respectively. Then, $\operatorname{rank} E = \dim\operatorname{span}\{\mathbf{e}_1,\dots,\mathbf{e}_l\}$. Observe that $EG = [E\mathbf{g}_1,\dots,E\mathbf{g}_n] \in \mathbb{C}^{l\times n}$. Clearly $E\mathbf{g}_i$ is a linear combination of the columns of E. Hence, $E\mathbf{g}_i \in \operatorname{span}\{\mathbf{e}_1,\dots,\mathbf{e}_l\}$. Therefore, $\operatorname{span}\{E\mathbf{g}_1,\dots,E\mathbf{g}_n\} \subseteq \operatorname{span}\{\mathbf{e}_1,\dots,\mathbf{e}_l\}$, which implies that $\operatorname{rank} EG \le \operatorname{rank} E$. Note that $(EG)^T = G^TE^T$. Hence

rank EG = rank $(EG)^T \le$ rank G^T = rank G. Thus
rank $EG \le \min(\text{rank } E, \text{rank } G)$. Suppose E is invertible. Then, rank $EG \le$ rank G = rank $E^{-1}(EG) \le$ rank EG. It follows rank EG = rank G. Similarly, rank EG = rank E if G is invertible. □

Proof of Theorem 5.16.1.

1. Proposition 5.16.2 yields that rank $A^\dagger$ = rank $V_r\Sigma_r^{-1}U_r^* \le$ rank $\Sigma_r^{-1}U_r^* \le$ rank $\Sigma_r^{-1} = r$ = rank A. Since $\Sigma_r = V_r^* A^\dagger U_r$, Proposition 5.16.2 implies that rank $A^\dagger \ge$ rank $\Sigma_r^{-1} = r$. Hence, rank A = rank $A^\dagger$.

2. $AA^\dagger = (U_r\Sigma_r V_r^*)(V_r\Sigma_r^{-1}U_r^*) = U_r\Sigma_r\Sigma_r^{-1}U_r^* = U_rU_r^*$. Hence
$$AA^\dagger A = (U_rU_r^*)(U_r\Sigma_r V_r^*) = U_r\Sigma V_r^* = A.$$
Hence $A^*AA^\dagger = (V_r\Sigma_r U_r^*)(U_rU_r^*) = A^*$. Similarly $A^\dagger A = V_rV_r^*$ and $A^\dagger AA^\dagger = A^\dagger, A^\dagger AA^* = A^*$.

3. Since $AA^\dagger = U_rU_r^*$, we deduce that $(AA^\dagger)^* = (U_rU_r^*)^* = (U_r^*)^*U_r^* = AA^\dagger$, i.e. $AA^\dagger$ is Hermitian. Next, $(AA^\dagger)^2 = (U_rU_r^*)^2 = (U_rU_r^*)(U_rU_r^*) = (U_rU_r^*) = AA^\dagger$, i.e. $AA^\dagger$ is idempotent. Thus, $AA^\dagger$ is non-negative definite. As $AA^\dagger = U_rI_rU_r^*$, the arguments of part *1* yield that rank $AA^\dagger = r$. Similar arguments apply to $A^\dagger A = V_rV_r^*$.

4. Since $A^*AA^\dagger = A^*$, it follows that $A^*A(A^\dagger\mathbf{b}) = A^*\mathbf{b}$, i.e. $\mathbf{y} = A^\dagger\mathbf{b}$ is a least square solution. It is left to show that if $A^*A\mathbf{x} = A^*\mathbf{b}$, then $||\mathbf{x}|| \ge ||A^\dagger\mathbf{b}||$ and equality holds if and only if $\mathbf{x} = A^\dagger\mathbf{b}$.
We now consider the system $A^*A\mathbf{x} = A^*\mathbf{b}$. To analyze this system, we use the full form of SVD given in (5.14.7). It is equivalent to
$(V\Sigma^TU^*)(U\Sigma V^*)\mathbf{x} = V\Sigma^TU^*\mathbf{b}$. Multiplying by V^*, we obtain the system $\Sigma^T\Sigma(V^*\mathbf{x}) = \Sigma^T(U^*\mathbf{b})$. Let $\mathbf{z} = (z_1,\dots,z_n)^T := V^*\mathbf{x}$,
$\mathbf{c} = (c_1,\dots,c_m)^T := U^*\mathbf{b}$. Note that $\mathbf{z}^*\mathbf{z} = \mathbf{x}^*VV\mathbf{x} = \mathbf{x}^*\mathbf{x}$, i.e. $||\mathbf{z}|| = ||\mathbf{x}||$. After these substitutions, the least square system in $z_1,\dots,z_n$ variables is given in the form $\sigma_i(A)^2z_i = \sigma_i(A)c_i$, for $i = 1,\dots,n$. Since $\sigma_i(A) = 0$, for $i > r$, we obtain that $z_i = \frac{1}{\sigma_i(A)}c_i$, for $i = 1,\dots,r$ while $z_{r+1},\dots,z_n$ are free variables. Thus, $||\mathbf{z}||^2 = \sum_{i=1}^r \frac{1}{\sigma_i(A)^2} + \sum_{i=r+1}^n |z_i|^2$. Hence, the least square solution with the minimal length $||\mathbf{z}||$ is the solution with $z_i = 0$, for $i = r+1,\dots,n$. This solution corresponds to the $\mathbf{x} = A^\dagger\mathbf{b}$.

5. Since rank A^*A = rank $A = n$, it follows that A^*A is an invertible matrix. Hence, the least square solution is unique and is given by $\mathbf{x} = (A^*A)^{-1}A^*\mathbf{b}$. Thus, for each $\mathbf{b}$ one has $(A^*A)^{-1}A^*\mathbf{b} = A^\dagger\mathbf{b}$, hence $A^\dagger = (A^*A)^{-1}A^*$.
If A is an $n \times n$ invertible matrix, it follows that $(A^*A)^{-1}A^* = A^{-1}(A^*)^{-1}A^* = A^{-1}$. □

5.16.1 Worked-out Problems

1. A linear transformation T on a vector space $\mathbf{V}$ is called a *projection* if $T^2 = T$.

(a) Show that if T is a projection on $\mathbf{V}$, then so is $I - T$, where I is the identity operator on $\mathbf{V}$.

(b) Show that if T is a projection on $\mathbf{V}$, then $\mathbf{V} = \ker T \oplus \mathrm{Im} T$.

(c) Show that if T is a projection on T, its only possible eigenvalues are 0 and 1.

(d) Show that if T is a projection on $\mathbf{V}$ and $\mathbf{V}$ is finite-dimensional, then T is diagonalizable.

Solution:

(a) For any $v \in \mathbf{V}$, $(I-T)^2 v = (I-T)(I-T)v = (I-T)(v-Tv) = v-Tv-T(v-Tv) = v - Tv - Tv + T^2 v$. Using $T^2 = T$, this reduces to $v - Tv = (I-T)v$. So, $(I-T)^2 = I - T$ and $I - T$ is a projection, too.

(b) For any $v \in \mathbf{V}$, of course $Tv \in \mathrm{Im}(T)$. Now, $T(v - Tv) = Tv - T^2 v = 0$, so $v - Tv \in \ker(T)$. As $v = (v - Tv) + Tv$, we have $v \in \ker(T) + \mathrm{Im}(T)$ and so $\mathbf{V} = \ker(T) + \mathrm{Im}(T)$.

To show the sum is direct, suppose that $v \in \ker(T) \cap \mathrm{Im}(T)$. Then, $Tv = 0$ and also $v = Tw$, for some $w \in \mathbf{V}$. Then, $0 = Tv = T(Tw) = T^2 w = Tw = v$. So $\ker(T) \cap \mathrm{Im}(T) = \{0\}$, and thus $\mathbf{V} = \ker(T) \oplus \mathrm{Im}(T)$.

(c) One can do this directly, using the definition of eigenvalues. Supposing that λ is an eigenvalue of the projection T, we have $Tv = \lambda v$, for some non-zero vector v. Then, $T^2 v = T(Tv) = T(\lambda v) = \lambda T v = \lambda^2 v$. But as $T^2 = T$, this is also $Tv = \lambda v$; since $v \neq 0$, $\lambda^2 = \lambda$ and so λ is either 0 or 1.

(d) Since $\mathbf{V} = \ker(T) \oplus \mathrm{Im}(T)$, if we choose a basis for $\ker(T)$ and a basis for $\mathrm{Im}(T)$, together they form a basis for $\mathbf{V}$. Clearly, each element of the basis for $\ker(T)$ is an eigenvector corresponding to 0. Also, any vector in the basis for $\mathrm{Im}(T)$ is $v = Tw$, for some $w \in \mathbf{V}$; as such $Tv = T^2 w = Tw = 1v$ and so an eigenvector corresponding to 1. We have a basis of $\mathbf{V}$ consisting of eigenvectors, so T is diagonalizable.

5.16.2 Problems

1. A matrix $P \in \mathbb{R}^{n\times n}$ is called an *orthogonal projection* if P is a projection and a symmetric matrix. Let $\mathbf{V} \subseteq \mathbb{R}^n$ be the subspace spanned by the columns of P. Show that for any $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in P\mathbf{V}$ the following inequality holds: $\|\mathbf{a} - \mathbf{b}\| \geq \|\mathbf{a} - P\mathbf{a}\|$. Furthermore, equality holds if and only if $\mathbf{b} = P\mathbf{a}$. That is, $P\mathbf{a}$ is the orthogonal projection of $\mathbf{a}$ on the column space of P.

2. Let $A \in \mathbb{R}^{m\times n}$ and assume that the SVD of A is given by (5.14.7), where $U \in \mathbf{O}(m,\mathbb{R}), V \in \mathbf{O}(n,\mathbb{R})$.

   (a) What is the SVD of A^T?

   (b) Show that $(A^T)^\dagger = (A^\dagger)^T$.

   (c) Suppose that $B \in \mathbb{R}^{l\times m}$. Is it true that $(BA)^\dagger = A^\dagger B^\dagger$? Justify!

3. Let $A \in \mathbb{C}^{m\times n}$. Show that

   (a) $N(A^*) = N(A^\dagger)$.

   (b) $R(A^*) = R(A^\dagger)$.

Chapter 6

Perron-Frobenius theorem

The Perron-Frobenius theorem provides a simple characterization of the eigenvectors and eigenvalues of certain types of matrices called non-negative matrices in particular irreducible and primitive matrices. The importance of this theorem lies in the fact the eigenvalue problems on these types of matrices aries in many different fields of mathematics. As an example of applications of Perron-Frobenius theorem, at the end of this chapter, we will show that it can be used as a tool to give a generalization of Kepler's Theorem.

6.1 Perron-Frobenius theorem

Denote by $\mathbb{R}_+^{m\times n}$, the set of matrices $A = [a_{ij}]_{i,j=1}^{m,n}$, where each entry $a_{ij} \in \mathbb{R}$, $a_{ij} \geq 0$. We denote a non-negative matrix by $A \geq 0$. We say that A is a non-zero non-negative if $A \geq 0$ and $A \neq 0$. This is denoted by $A \gneq 0$. We say that A is a positive matrix if all the entries of A are positive, i.e. $a_{ij} > 0$, for all i, j. We denote that by $A > 0$. Similar notation is for vectors, i.e. $n = 1$. From now on, we assume that all matrices are square matrices unless stated otherwise. A non-negative square matrix $A \in \mathbb{R}_+^{n\times n}$ is called *irreducible*, if $(I + A)^{n-1} > 0$. A non-negative matrix is called *primitive* if $A^p > 0$, for some positive integer p.

Theorem 6.1.1 (Perron-Frobenius) *Let $A \in \mathbb{R}_+^{n\times n}$. Denote by $\rho(A)$, the spectral radius of A, i.e. the maximum of the absolute values of the eigenvalues of A. Then, the following conditions hold.*

1. *$\rho(A)$ is an eigenvalue of A.*
2. *There exists $\mathbf{x} \gneq \mathbf{0}$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$.*
3. *Assume that λ is an eigenvalue of A, $\lambda \neq \rho(A)$ and $|\lambda| = \rho(A)$. Then,* $\mathrm{index}\,(\lambda) \leq \mathrm{index}\,(\rho(A))$. *Furthermore, $\zeta = \frac{\lambda}{\rho(A)}$ is a root of unity of order n at most, i.e. $\zeta^m = 1$, for some m with $m \leq n$.*
4. *Suppose that A is primitive. Then, $\rho(A) > 0$ and $\rho(A)$ is a simple root of the characteristic polynomial, and there exists $\mathbf{x} > 0$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$. Furthermore, if λ is an eigenvalue of A different from $\rho(A)$, then $|\lambda| < \rho(A)$.*
5. *Suppose that A is irreducible. Then, $\rho(A) > 0$ and $\rho(A)$ is a simple root of the characteristic polynomial and there exists $\mathbf{x} > 0$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$. Let*

$\lambda_0 := \rho(A), \lambda_1, \dots, \lambda_{m-1}$ be all distinct eigenvalues of A satisfying $|\lambda| = \rho(A)$. Then, all these eigenvalues are simple roots of the characteristic polynomial of A. Furthermore, $\lambda_j = \rho(A)\zeta^j$, for $j = 0, \dots, m-1$, where $\zeta = e^{\frac{2\pi i}{m}}$ is a primitive $m-th$ root of 1. Finally, the characteristic polynomials of ζA and A are equal.

We first prove this result for symmetric non-negative matrices.

Proposition 6.1.2 *Let $A = A^\top \geq 0$ be a symmetric matrix with non-negative entries. Then, the conditions of Theorem 6.1.1 hold.*

Proof. Recall that A is diagonalizable by an orthogonal matrix. Hence, the index of each eigenvalue of A is 1. Next, recall that each eigenvalue is given by the Rayleigh quotient $\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$. For $\mathbf{x} = (x_1, \dots, x_n)^\top$, let $|\mathbf{x}| := (|x_1|, \dots, |x_n|)^\top$. Note that $\mathbf{x}^\top\mathbf{x} = |\mathbf{x}|^\top|\mathbf{x}|$. Also $|\mathbf{x}^\top A\mathbf{x}| \leq |\mathbf{x}|^\top A|\mathbf{x}|$. Hence, $|\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}| \leq \frac{|\mathbf{x}|^\top A|\mathbf{x}|}{|\mathbf{x}|^\top|\mathbf{x}|}$. Thus, the maximum of the Rayleigh quotient is achieved for some $\mathbf{x} \gneqq \mathbf{0}$. This shows that $\rho(A)$ is an eigenvalue of A, and there exists an eigenvector $\mathbf{x} \gneqq \mathbf{0}$ corresponding to A.

Suppose that $A > 0$. Then, $A\mathbf{x} = \rho(A)\mathbf{x}, \mathbf{x} \gneqq 0$. So, $A\mathbf{x} > \mathbf{0}$. Hence, $\rho(A) > 0$ and $\mathbf{x} > 0$. Assume that there exists another eigenvector $\mathbf{y}$, which is not collinear with $\mathbf{x}$, corresponding to $\rho(A)$. Then, $\mathbf{y}$ can be chosen to be orthogonal to the positive eigenvector $\mathbf{x}$ corresponding to $\rho(A)$, i.e. $\mathbf{x}^\top\mathbf{y} = 0$. Hence, $\mathbf{y}$ has positive and negative coordinates. Therefore

$$|\frac{\mathbf{y}^\top A\mathbf{y}}{\mathbf{y}^\top\mathbf{y}}| < \frac{|\mathbf{y}|^\top A|\mathbf{y}|}{|\mathbf{y}|^\top|\mathbf{y}|} \leq \rho(A).$$

This gives the contradiction $\rho(A) < \rho(A)$. Therefore, $\rho(A)$ is a simple root of the characteristic polynomial of A. In a similar way, we deduce that each eigenvalue λ of A, different from $\rho(A)$, satisfies $|\lambda| < \rho(A)$.

Since the eigenvectors of A^p are the same as A, we deduce the same results if A is a non-negative symmetric matrix which is primitive. Suppose that A is irreducible. Then, $(tI+A)^{n-1} > 0$, for any $t > 0$. Clearly, $\rho(tI+A) = \rho(A)+t$. Hence, the eigenvector corresponding to $\rho(A)$ is positive. Therefore, $\rho(A) > 0$. Each eigenvalue of $tI+A$ is of the form $\lambda + t$, for some eigenvalue λ of A. Since $(tI + A)$ is primitive for any $t > 0$, it follows that $|\lambda + t| < \rho(A) + t$. Let $t \to 0^+$ to deduce that $|\lambda| \leq \rho(A)$. Since all the eigenvalues of A are real, we can have an equality for some eigenvalue $\lambda \neq \rho(A)$ if and only if $\lambda = -\rho(A)$. This will be the case if A has the form $C = \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}$, for some $B \in \mathbb{R}_+^{m\times l}$. It can be shown that if A is a symmetric irreducible matrix such that $-\rho(A)$ is an eigenvalue of A, then $A = PCP^\top$, for some permutation matrix P. We already showed that C and $-C$ have the same eigenvalues, counted with their multiplicities. Hence, $-A$ and A have the same characteristic polynomial. □

Recall the ℓ_∞ norm on $\mathbb{C}^n$ and the corresponding operator norm on $\mathbb{C}^{n\times n}$:

$$\|(x_1, \dots, x_n)^\top\|_\infty := \max_{i\in[n]} |x_i|, \quad \|A\|_\infty := \max_{\|\mathbf{x}\|_\infty \leq 1} \|A\mathbf{x}\|_\infty, \ A \in \mathbb{C}^{n\times n}. \tag{6.1.1}$$

Lemma 6.1.3 *Let $A = [a_{ij}]_{i,j\in[n]} \in \mathbb{C}^{n\times n}$. Then*

$$\|A\|_\infty = \max_{i\in[n]} \sum_{j\in[n]} |a_{ij}|. \tag{6.1.2}$$

Proof. Let $a := \max_{i\in[n]} \sum_{j\in[n]} |a_{ij}|$. Assume that $\|(x_1,\dots,x_n)^\top\| \le 1$. Then

$$|(A\mathbf{x})_i| = |\sum_{j\in[n]} a_{ij}x_j| \le \sum_{j\in[n]} |a_{ij}|\,|x_j| \le \sum_{j\in[n]} |a_{ij}| \le a,$$

for each $i \in [n]$. Hence, $\|A\|_\infty \le a$.

From the definition of a, it follows that $a = \sum_{j\in[n]} |a_{kj}|$, for some $k \in [n]$. Let x_j be a complex number of modulus 1, i.e. $x_j = 1$, such that $|a_{kj}| = a_{kj}x_j$, for each $j \in [n]$. Let $\mathbf{x} = (x_1,\dots,x_n)^\top$. Clearly $\|\mathbf{x}\|_\infty = 1$. Furthermore, $a = \sum_{j\in[n]} a_{kj}x_j$. Hence, $\max_{i\in[n]} \|A\mathbf{x}\|_\infty \ge a$. Therefore $a = \|A\|_\infty$. □

Definition 6.1.4 *Let $A = [a_{ij}] \in \mathbb{C}^{n\times n}$. Denote $|A| := [|a_{ij}|] \in \mathbb{R}_+^{n\times n}$. For $\mathbf{x} = (x_1,\dots,x_n)^\top \ge \mathbf{0}$, define $r(A,\mathbf{x}) := \min\{t \ge 0, |A|\mathbf{x} \le t\mathbf{x}\}$.*

(Note that $r(A,\mathbf{x})$ may take a value ∞.) It is straightforward to show that

$$r(A,\mathbf{x}) := \max_{i\in[n]} \frac{\sum_{j\in[n]} |a_{ij}|x_j}{x_i}, \quad \text{for } \mathbf{x} > \mathbf{0}. \tag{6.1.3}$$

Lemma 6.1.5 *Let $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ and $\mathbf{1} := (1,\dots,1)^\top \in \mathbb{C}^n$. Then*

1. *$\|A\|_\infty = r(A,\mathbf{1})$.*

2. *Assume that $\mathbf{x} = (x_1,\dots,x_n)^\top > \mathbf{0}$. Then, each eigenvalue λ of A satisfies $|\lambda| \le r(A,\mathbf{x})$. That is, $\rho(A) \le r(A,\mathbf{x})$. In particular*

$$\rho(A) \le \|A\|_\infty. \tag{6.1.4}$$

Proof. The statement $\|A\|_\infty = r(A,1)$ follows straightforward from the definition of $\|A\|_\infty$ and $r(A,\mathbf{1})$.

We now show (6.1.4). Assume that $A\mathbf{z} = \lambda\mathbf{z}$, where $\mathbf{z} = (z_1,\dots,z_n)^\top \in \mathbb{C}^n \setminus \{\mathbf{0}\}$. Without loss of generality, we may assume that $|z_k| = \|\mathbf{z}\|_\infty = 1$, for some $k \in [n]$. Then

$$|\lambda| = |\lambda z_k| = |(A\mathbf{z})_k| \le \sum_{j=1}^n |a_{kj}|\,|z_j| \le \sum_{j=1}^n |a_{kj}| \le \|A\|_\infty.$$

Assume that $\mathbf{x} = (x_1,\dots,x_n)^\top > 0$. Let $D = \operatorname{diag}(x_1,\dots,x_n)$. Define $A_1 := D^{-1}AD$. A straightforward calculation shows that $r(A,\mathbf{x}) = r(A_1,\mathbf{1}) = \|A_1\|_\infty$. Clearly, A and A_1 are similar. Use (6.1.4) for A_1 to deduce that $|\lambda| \le r(A,\mathbf{x})$. □

Lemma 6.1.6 *Let $A \in \mathbb{R}_+^{n\times n}$. Then, $r(A,\mathbf{x}) \ge r(A,(I+A)\mathbf{x})$.*

Proof. As $A\mathbf{x} \le r(A,\mathbf{x})\mathbf{x}$, we deduce

$$A((I+A)\mathbf{x}) = (I+A)(A\mathbf{x}) \le (I+A)(r(A,\mathbf{x})\mathbf{x}) = r(A,\mathbf{x})A((I+A)\mathbf{x})).$$

The definition of $r(A,(I+A)\mathbf{x})$ yields the lemma. □

The next result is a basic step in the proof of Theorem 6.1.1 and is due to Wielandt, e.g., [7].

Theorem 6.1.7 *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Then*

$$\rho(A) = \min_{\mathbf{x}>\mathbf{0}} r(A,\mathbf{x}) > 0. \tag{6.1.5}$$

Equality holds if and only if $A\mathbf{x} = \rho(A)\mathbf{x}$. There is a unique positive eigenvector $\mathbf{x} > \mathbf{0}$ corresponding to $\rho(A)$, up to a multiple by a constant. (This eigenvector is called Perron-Frobenius eigenvector) That is, rank $(\rho(A)I - A) = n-1$. *Any other real eigenvector of A corresponding to $\lambda \neq \rho(A)$ has two coordinates with opposite signs.*

Proof. Since for any $a > 0$ and $\mathbf{x} > \mathbf{0}$, we have that $r(A,a\mathbf{x}) = r(A,\mathbf{x})$, it is enough to assume that $\mathbf{x}$ is a probability vector, i.e. $\|\mathbf{x}\|_1 = \sum_{i=1}^n x_i = 1$. Clearly, Π_n, the set of all probability vectors in $\mathbb{R}^n$, is a compact set. (Problem 1.13.2-7) We now consider the extremal problem

$$r(A) := \inf_{\mathbf{x}\in\Pi_n} r(A,\mathbf{x}). \tag{6.1.6}$$

We claim that this infimum is achieved for some $\mathbf{y} \in \Pi_n$. As $r(A,\mathbf{x}) \geq 0$, for $\mathbf{x} \geq \mathbf{0}$, it follows that there exists a sequence $\mathbf{x}_m \in \Pi_n$ such $\lim_{m\to\infty} r(A,\mathbf{x}_m) = r(A)$. Clearly, $r(A,\mathbf{x}_m) \geq r(A)$. Hence, by considering a subsequence if necessary, we can assume that $r(A,\mathbf{x}_m)$ is a nonincreasing sequence, i.e., $r(A,\mathbf{x}_l) \geq r(A,\mathbf{x}_m)$, for $m > l$. Thus, $r(A,\mathbf{x}_l)\mathbf{x}_m \geq r(A,\mathbf{x}_m)\mathbf{x}_m \geq A\mathbf{x}_m$. As $\lim_{m\to\infty}\mathbf{x}_m = \mathbf{y}$, it follows that $r(A,\mathbf{x}_l)\mathbf{y} \geq A\mathbf{y}$. Hence, $r(A,\mathbf{x}_l) \geq r(A,\mathbf{y}) \Rightarrow r(A) \geq r(A,\mathbf{y})$. From the definition of $r(A)$, we deduce that $r(A) \leq r(A,\mathbf{y})$. Hence, $r(A) = r(A,\mathbf{y})$.

Assume first that $\mathbf{y}$ is not an eigenvector of A. Let $\mathbf{z} := r(A)\mathbf{y} - A\mathbf{y}$. So $\mathbf{z} \geq \mathbf{0}$ and $\mathbf{z} \neq 0$. Since A is irreducible, then $(I+A)^{n-1} > 0$. Hence, $(I+A)^{n-1}\mathbf{z} > \mathbf{0}$. This is equivalent to

$$r(A)((I+A)^{n-1}\mathbf{y}) > (I+A)^{n-1}(A\mathbf{y}) = A((I+A)^{n-1}\mathbf{y}).$$

Let $\mathbf{u} := (I+A)^{n-1}\mathbf{y}$. Then, the above inequality yields that $r(A,\mathbf{u}) < r(A)$. Let $\mathbf{v} = b\mathbf{u} \in \Pi_n$, for a corresponding $b > 0$. So $r(A,\mathbf{v}) = r(A,\mathbf{u}) < r(A)$. This contradicts the definition of $r(A)$. Hence, $A\mathbf{y} = r(A)\mathbf{y}$. Clearly, $(I+A)^{n-1}\mathbf{y} = (1+r(A))^{n-1}\mathbf{y}$. As $\mathbf{y} \in \Pi_n$ and $(I+A)^{n-1} > 0$, it follows that $(I+A)^{n-1}\mathbf{y} > \mathbf{0}$. Hence, $\mathbf{y} > 0$. Part *2* of Lemma 6.1.5 yields that $\rho(A) \leq r(A,\mathbf{y}) = r(A)$. As $r(A)$ is an eigenvalue of A, it follows that $r(A) = \rho(A)$. Since A is irreducible, $A \neq 0$. Hence $r(A) = r(A,\mathbf{y}) > 0$. Combine the definition of $r(A)$ with the fact that $A\mathbf{y} = r(A)\mathbf{y}$ to deduce (6.1.5).

We next claim that rank $(\rho(A)I - A) = n-1$, i.e. the dimension of the eigenspace of A corresponding to A is one. Assume to the contrary that $\mathbf{w} \in \mathbb{R}^n$ is an eigenvector of A corresponding to $\rho(A)$ which is not collinear with $\mathbf{y}$, i.e. $w \neq t\mathbf{y}$, for any $t \in \mathbb{R}$. Consider $\mathbf{w} + t\mathbf{y}$. Since $\mathbf{y} > 0$ for $t \gg 0$ and $-t \gg 0$, we have that $w + t\mathbf{y} > 0$ and $w + t\mathbf{y} < 0$, respectively. Hence, there exists $t_0 \in \mathbb{R}$ such that $\mathbf{u} := w + t_0\mathbf{y} \geq \mathbf{0}$ and at least one of the coordinates of $\mathbf{u}$ equals zero. As $\mathbf{w}$ and $\mathbf{y}$ are linearly independent, it follows that $\mathbf{u} \neq \mathbf{0}$. Also, $A\mathbf{u} = \rho(A)\mathbf{u}$. Since A is irreducible, we obtain that $(1+\rho(A))^{n-1}\mathbf{u} = (I+A)^{n-1}\mathbf{u} > \mathbf{0}$. This contradicts the assumption that $\mathbf{u}$ has at least one zero coordinates.

Suppose finally that $A\mathbf{v} = \lambda\mathbf{v}$, for some $\lambda \neq \rho(A)$ and $\mathbf{v} \neq \mathbf{0}$. Observe that $A^\top \geq 0$ and $(I+A^\top)^{n-1} = ((I+A)^{n-1})^\top > 0$. That is, $A^\top$ is irreducible. Hence, there exists $\mathbf{u} > 0$ such that $A^\top\mathbf{u} = \rho(A^\top)\mathbf{u}$. Recall that $\rho(A^\top) = \rho(A)$. Consider now

$\mathbf{u}^\top A\mathbf{v} = \lambda\mathbf{u}^\top\mathbf{v} = \rho(A)\mathbf{u}^\top\mathbf{v}$. As $\rho(A) \neq \lambda$, we deduce that $\mathbf{u}^\top\mathbf{v} = 0$. As $\mathbf{u} > \mathbf{0}$, it follows that $\mathbf{v}$ has two coordinates with opposite signs. □

We now give the full proof of Theorem 6.1.1. We use the following well-known result for the roots of monic polynomials of complex variables [15].

Lemma 6.1.8 (Continuity of the roots of a polynomial) *Let $p(z) = z^n + a_1 z^{n-1} + \ldots + a_n$ be a monic polynomial of degree n with complex coefficients. Assume that $p(z) = \prod_{j=1}^n (z - z_j(p))$. Given $\epsilon > 0$, then there exists $\delta(\epsilon) > 0$ such that for each coefficient vector $\mathbf{b} = (b_1, \ldots, b_n)$ satisfying $\sum_{j=1}^n |b_j - a_j|^2 < \delta(\epsilon)^2$ one can rearrange the roots of $q(z) = z^n + b_1 z^{n-1} + \ldots + b_n = \prod_{j=1}^n (z - z_j(q))$ such that $|z_j(q) - z_j(p)| < \epsilon$, for each $j \in [n]$.*

Corollary 6.1.9 *Assume that $A_m \in \mathbb{C}^{n\times n}$ converges to $A \in \mathbb{C}^{n\times n}$, $m \in \mathbb{N}$. Then, $\lim_{m\to\infty} \rho(A_m) = \rho(A)$. That is, the function $\rho(\cdot) : \mathbb{C}^{n\times n} \to [0, \infty)$, that assigns to each A its spectral radius, is a continuous function.*

Proof of *1.* and *2.* of Theorem 6.1.1. Assume that $A \in \mathbb{R}_+^{n\times n}$, i.e. A has non-negative entries. Let $J_n \in \mathbb{R}_+^{n\times n}$ be a matrix whose all entries are 1. Let $A_m = A + \frac{1}{m}J_n$, for $m \in \mathbb{N}$. Then, each A_m is a positive matrix and $\lim_{m\to\infty} A_m = A$. Corollary 6.1.9 yields that $\lim_{m\to\infty} \rho(A_m) = \rho(A)$. Clearly, each A_m is irreducible. Hence, by Theorem 6.1.7, there exists a probability vector $\mathbf{x}_m \in \Pi_n$ such that $A_m\mathbf{x}_m = \rho(A_m)\mathbf{x}_m$. As Π_n is compact, there exists a subsequence $m_k, k \in \mathbb{N}$ such that $\lim_{k\to\infty} \mathbf{x}_{m_k} = \mathbf{x} \in \Pi_n$. Consider the equalities $A_{m_k}\mathbf{x}_{m_k} = \rho(A_{m_k})\mathbf{x}_{m_k}$. Let $k \to \infty$ to deduce that $A\mathbf{x} = \rho(A)\mathbf{x}$. So $\rho(A)$ is an eigenvector of A with a corresponding probability vector $\mathbf{x}$. □

In what follows, we need the following lemma:

Lemma 6.1.10 *Let $B \in \mathbb{C}^{n\times n}$ and assume that* rank $B = n - 1$. *Then,* adj $B = \mathbf{u}\mathbf{v}^\top \neq 0$, *where $B\mathbf{u} = B^\top\mathbf{v} = \mathbf{0}$. Suppose furthermore that $A \in \mathbb{R}_+^{n\times n}$ and $\rho(A)$ is a geometrically simple eigenvalue, i.e.* rank $B = n - 1$, *where $B = \rho(A)I_n - A$. Then,* adj $B = \mathbf{u}\mathbf{v}^\top$ *and $\mathbf{u}, \mathbf{v} \gneq \mathbf{0}$ are the eigenvectors of A and $A^\top$ corresponding to the eigenvalue $\rho(A)$.*

Proof. Recall that the entries of adj B are all $n-1$ minors of B. Since rank $B = n-1$, we deduce that adj $B \neq 0$. Recall next that $B(\text{adj } B) = (\text{adj } B)B = \det B\, I_n = 0$, as rank $B = n - 1$. Let adj $B = [\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n]$. As $B(\text{adj } B) = 0$, we deduce that $B\mathbf{u}_i = \mathbf{0}$, for $i \in [n]$. Since rank $B = n - 1$, the dimension of the null space is one. So the null space of B is spanned by $\mathbf{u} \neq \mathbf{0}$. Hence, $\mathbf{u}_i = v_i\mathbf{u}$, for some $v_i, i \in [n]$. Let $\mathbf{v} = (v_1, \ldots, v_n)^\top$. Then, adj $B = \mathbf{u}\mathbf{v}^\top$. Since adj $B \neq 0$, it follows that $\mathbf{v} \neq \mathbf{0}$. As $0 = (\text{adj } B)B = \mathbf{u}\mathbf{v}^\top B = \mathbf{u}(B^\top\mathbf{v})$ and $\mathbf{u} \neq \mathbf{0}$, it follows that $B^\top\mathbf{v} = \mathbf{0}$.

Assume now that $A \in \mathbb{R}_+^{n\times n}$. Then, part *1* of Theorem 6.1.1 yields that $\rho(A)$ is an eigenvalue of A. Hence, $B = \rho(A)I_n - A$ is singular. Suppose that rank $B = n - 1$. We claim that adj $B \geq 0$. We first observe that for $t > \rho(A)$, we have $\det(tI_n - A) = \prod_{j=1}^n (t - \lambda_j(A)) > 0$. Indeed, if $\lambda_j(A)$ is real, then $t - \lambda_j(A) > 0$. If $\lambda_j(A)$ is not real, i.e. a complex number, then $\bar{\lambda}_j(A)$ is also an eigenvalue of A. Recall that $t > \rho(A) \geq |\lambda_j(A)|$. Hence, $(t - \lambda_j(A))(t - \bar{\lambda}_j(A)) = |t - \lambda_j(A)|^2 > 0$. So

$\det(tI_n - A) > 0$. Furthermore,

$$\text{adj }(tI_n - A) = \det(tI_n - A)(tI_n - A)^{-1} =$$
$$\det(tI_n - A)t^{-1}(I_n - t^{-1}A)^{-1} = \det(tI_n - A)t^{-1}\sum_{j=0}^{\infty}(t^{-1}A)^j.$$

As $A^j \geq 0$, for each non-negative integer j, it follows that adj $(tI_n - A) \geq 0$, for $t > \rho(A)$. Let $t \searrow \rho(A)$, (t converges to $\rho(A)$ from above), to deduce that adj $B \gneq 0$. As $B = \mathbf{u}\mathbf{v}^\top$, we can assume that $\mathbf{u}, \mathbf{v} \gneq \mathbf{0}$. □

Lemma 6.1.11 *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Then, $\rho(A)$ is an algebraically simple eigenvalue.*

Proof. We may assume that $n > 1$. Theorem 6.1.7 implies that $\rho(A)$ is geometrically simple, i.e. null$(\rho(A)I - A) = 1$. Hence, rank $(\rho(A)I - A) = n-1$. Lemma 6.1.10 yields that adj $(\rho(A)I - A) = \mathbf{u}\mathbf{v}^\top$, where $A\mathbf{u} = \rho(A)\mathbf{u}$, $A^\top\mathbf{v} = \rho(A)\mathbf{v}, \mathbf{u}, \mathbf{v} \ngeq \mathbf{0}$. Theorem 6.1.7 implies that $\mathbf{u}, \mathbf{v} > \mathbf{0}$. Hence, $\mathbf{u}\mathbf{v}^\top$ is a positive matrix. In particular, $\operatorname{tr}\mathbf{u}\mathbf{v}^\top = \mathbf{v}^\top\mathbf{u} > 0$. Since

$$(\det(\lambda I - A))'(\lambda = \rho(A)) = \operatorname{tr}\text{adj }(\rho(A)I - A) = \mathbf{v}^\top\mathbf{u} > 0,$$

we deduce that $\rho(A)$ is a simple root of the characteristic polynomial of A. □
For $A \in \mathbb{R}_+^{n\times n}$ and $\mathbf{x} \ngeq \mathbf{0}$ denote $s(A,\mathbf{x}) = \max\{t \geq 0, A\mathbf{x} \geq t\mathbf{x}\}$. From the definition of $r(A,\mathbf{x})$, we immediately conclude that $r(A,\mathbf{x}) \geq s(A,\mathbf{x})$. We now give a complementary part of Theorem 6.1.7.

Lemma 6.1.12 *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Then*

$$\rho(A) = \max_{\mathbf{x}>\mathbf{0}} s(A,\mathbf{x}) > 0. \tag{6.1.7}$$

Equality holds if and only if $A\mathbf{x} = \rho(A)\mathbf{x}$.

The proof of this lemma is similar to the proof of Theorem 6.1.7, and we skip it.

As usual, denote by $\mathrm{S}^1 := \{z \in \mathbb{C}, |z| = 1\}$ the unit circle in the complex plane.

Lemma 6.1.13 *Let $A \in \mathbb{R}_+^{n\times n}$ be irreducible, $C \in \mathbb{C}^{n\times n}$. Assume that $|C| \leq A$. Then, $\rho(C) \leq \rho(A)$. Equality holds, i.e. there exists $\lambda \in \operatorname{spec} C$, such that $\lambda = \zeta\rho(A)$, for some $\zeta \in \mathrm{S}^1$, if and only if there exists a complex diagonal matrix $D \in \mathbb{C}^{n\times n}$, whose diagonal entries are equal to 1, such that $C = \zeta DAD^{-1}$. The matrix D is unique up to a multiplication by $t \in \mathrm{S}^1$.*

Proof. We can assume that $n > 1$. Suppose that $A = [a_{ij}], C = [c_{ij}]$. Theorem 6.1.7 yields that there exists $\mathbf{u} > \mathbf{0}$ such that $A\mathbf{u} = \rho(A)\mathbf{u}$. Hence, $r(A,\mathbf{u}) = \rho(A)$. Since $|C| \leq A$, it follows that $r(C,\mathbf{u}) \leq r(A,\mathbf{u}) = \rho(A)$. Lemma 6.1.5 yields that $\rho(C) \leq r(C,\mathbf{u})$. Hence, $\rho(C) \leq \rho(A)$.

Suppose that $\rho(C) = \rho(A)$. Hence, we have equalities $\rho(C) = r(C,\mathbf{u}) = \rho(A) = r(A,\mathbf{u})$. So, $\lambda = \zeta\rho(A)$ is an eigenvalue of C, for some $\zeta \in \mathrm{S}^1$. Furthermore, for the corresponding eigenvector $\mathbf{z}$ of C we have

$$|\lambda|\,|\mathbf{z}| = |C\mathbf{z}| \leq |C|\,|\mathbf{z}| \leq A|\mathbf{z}|.$$

Hence, $\rho(A)|z| \le A|z|$. Therefore $s(A,|z|) \ge \rho(A)$. Lemma 6.1.12 yields that $s(A,|z|) = \rho(A)$ and $|z|$ is the corresponding non-negative eigenvector. Therefore, $|C\mathbf{z}| = |C||\mathbf{z}| = A|\mathbf{z}|$. Theorem 6.1.7 yields that $|\mathbf{z}| = \mathbf{u}$.

Let $z_i = d_i u_i, |d_i| = 1$, for $i = 1,\dots,n$. The equality $|C\mathbf{z}| = |C|\,|\mathbf{z}| = A|\mathbf{z}|$ combined with the triangle inequality and $|C| \le A$, yields first that $|C| = A$. Furthermore, for each fixed i, the non-zero complex numbers $c_{i1}z_1,\dots,c_{in}z_n$ have the same argument, i.e. $c_{ij} = \zeta_i a_{ij}\bar{d}_j$, for $j = 1,\dots,n$ and some complex number ζ_j, where $|\zeta_i| = 1$. Recall that $\lambda z_i = (C\mathbf{z})_i$. Hence, $\zeta_i = \zeta d_i$, for $i = 1,\dots,n$. Thus, $C = \zeta DAD^{-1}$, where $D = \mathrm{diag}(d_1,\dots,d_n)$. It is straightforward to see that D is unique up to a multiplication by t, for any $t \in \mathrm{S}^1$.

Suppose now that for $D = \mathrm{diag}(d_1,\dots,d_n)$, where $|d_1| = \dots = |d_n| = 1$ and $|\zeta| = 1$, we have that $C = \zeta DAD^{-1}$. Then, $\lambda_i(C) = \zeta\lambda_i(A)$, for $i = 1,\dots,n$. So $\rho(C) = \rho(A)$. Furthermore, $c_{ij} = \zeta d_i c_{ij}\bar{d}_j, i,j = 1,\dots,n$. Then, $|C| = A$. □

Lemma 6.1.14 *Let $\zeta_1,\dots,\zeta_h \in \mathrm{S}^1$ be h distinct complex numbers which form a multiplicative semi-group, i.e. for any integers $i,j \in [h]$, $\zeta_i\zeta_j \in \{\zeta_1,\dots,\zeta_h\}$. Then, the set $\{\zeta_1,\dots,\zeta_h\}$ is the set, (the group), of all h roots of 1: $e^{\frac{2\pi i\sqrt{-1}}{h}}, i = 1,\dots,h$.*

(Here, we denote the complex number i by $\sqrt{-1}$ to simplify notations.)

Proof. Let $\zeta \in \mathcal{T} := \{\zeta_1,\dots\zeta_h\}$. Consider the sequence $\zeta^i, i = 1,\dots$. Since $\zeta^{i+1} = \zeta\zeta^i$, for $i = 1,\dots,$ and $\mathcal{T}$ is a semigroup, it follows that each ζ^i is in $\mathcal{T}$. As $\mathcal{T}$ is a finite set, we must have two positive integers such that $\zeta^k = \zeta^l$, for $k < l$. Assume that k and l are the smallest possible positive integers. So $\zeta^p = 1$, where $p = l - k \ge 1$, and $\mathcal{T}_p := \{\zeta, \zeta^2, \dots, \zeta^{p-1}, \zeta^p = 1\}$ are all p roots of 1. Here, ζ is called a *p-primitive root* of 1, i.e. $\zeta = e^{\frac{2\pi p_1\sqrt{-1}}{p}}$, where p_1 is an positive integer less than p. Furthermore, p_1 and p are *coprime*, which is denoted by $(p_1,p) = 1$. Note that $\zeta^i \in \mathcal{T}$, for any integer i.

Next, we choose $\zeta \in \mathcal{T}$, such that ζ is a primitive p-root of 1 of the maximal possible order. We claim that $p = h$, which is equivalent to the equality $\mathcal{T} = \mathcal{T}_p$. Assume to the contrary that $\mathcal{T}_p \subsetneq \mathcal{T}$. Let $\eta \in \mathcal{T}\backslash\mathcal{T}_p$. The previous arguments show that η is a q-primitive root of 1. Therefore, $\mathcal{T}_q \subset \mathcal{T}$, and $\mathcal{T}_q \subsetneq \mathcal{T}_p$. Thus, q cannot divide p. Also, the maximality of p yields that $q \le p$. Let $(p,q) = r$ be the g.c.d., the greatest common divisor of p and q. So $1 \le r < q$. Recall that Euclid's algorithm, which is applied to the division of p by q with a residue, yields that there exist two integers i,j such that $ip + jq = r$. Let $l := \frac{pq}{r} > p$ be the least common multiplier of p and q. Observe that $\zeta' = e^{\frac{2\pi\sqrt{-1}}{p}} \in \mathcal{T}_p, \eta' = e^{\frac{2\pi\sqrt{-1}}{q}} \in \mathcal{T}_q$. So

$$\xi := (\eta')^i(\zeta')^j = e^{\frac{2\pi(ip+jq)\sqrt{-1}}{pq}} = e^{\frac{2\pi\sqrt{-1}}{l}} \in \mathcal{T}.$$

As ξ is an l-primitive root of 1, we obtain a contradiction to the maximality of p. So $p = h$ and $\mathcal{T}$ is the set of all h-roots of unity. □

Theorem 6.1.15 *Let $A \in \mathbb{R}_+^{n\times n}$ be irreducible and assume that for a positive integer $h \ge 2$, A has $h-1$ distinct eigenvalues $\lambda_1,\dots,\lambda_{h-1}$, which are distinct from $\rho(A)$, such that $|\lambda_1| = \dots = |\lambda_{h-1}| = \rho(A)$. Then, the following conditions hold*

1. *Assume that A is imprimitive, i.e. not primitive. Then, there exist exactly $h-1\ge 1$ distinct eigenvalues $\lambda_1,\dots,\lambda_{h-1}$ different from $\rho(A)$ and satisfying $|\lambda_i|=\rho(A)$. Furthermore, the following conditions hold.*

   (a) *λ_i is an algebraically simple eigenvalue of A, for $i=1,\dots,h-1$.*

   (b) *The complex numbers $\frac{\lambda_i}{\rho(A)}, i=1,\dots,h-1$ and 1 are all h roots of unity, i.e. $\lambda_i=\rho(A)e^{\frac{2\pi\sqrt{-1}i}{h}}$, for $i=1,\dots,h-1$. Furthermore, if $A\mathbf{z}_i=\lambda_i\mathbf{z}_i, \mathbf{z}_i\ne\mathbf{0}$, then $|\mathbf{z}_i|=\mathbf{u}>0$, the Perron-Frobenius eigenvector $\mathbf{u}$ given in Theorem 6.1.7.*

   (c) *Let ζ be any h-root of 1, i.e. $\zeta^h=1$. Then, the matrix ζA is similar to A. Hence, if λ is an eigenvalue of A, then $\zeta\lambda$ is an eigenvalue of A having the same algebraic and geometric multiplicity as λ.*

   (d) *There exists a permutation matrix $P\in\mathcal{P}_n$ such that $P^\top AP=B$ has a block h-circulate form*

$$B=\begin{bmatrix} 0 & B_{12} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & B_{23} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & 0 & B_{(h-1)h} \\ B_{h1} & 0 & 0 & 0 & \vdots & 0 & 0 \end{bmatrix},$$

$$B_{i(i+1)}\in\mathbb{R}^{n_i\times n_{i+1}}, i=1,\dots,h, B_{h(h+1)}=B_{h1},$$
$$n_{h+1}=n_1, n_1+\dots+n_h=n.$$

   Furthermore, the diagonal blocks of B^h are all irreducible primitive matrices, i.e.

$$C_i:=B_{i(i+1)}\dots B_{(h-1)h}B_{h1}\dots B_{(i-1)i}\in\mathbb{R}_+^{n_i\times n_i},\ i=1,\dots,h, \tag{6.1.8}$$

   are irreducible and primitive.

(In the whole theorem we denote the complex number i by $\sqrt{-1}$ to simplify notations.)

Proof. Assume that $\zeta_i:=\frac{\lambda_i}{\rho(A)}\in\mathrm{S}^1$, for $i=1,\dots,h-1$ and $\zeta_h=1$. Apply Lemma 6.1.13 to $C=A$ and $\lambda=\zeta_i\rho(A)$ to deduce that $A=\zeta_iD_iAD_i^{-1}$, where D_i is a diagonal matrix such that $|D|=I$, for $i=1,\dots,h$. Hence, if λ is an eigenvalue of A, then $\zeta_i\lambda$ is an eigenvalue of A, with the same algebraic and geometric multiplicities as λ. In particular, since $\rho(A)$ is an algebraically simple eigenvalue of A, $\lambda_i=\zeta_i\rho(A)$ is an algebraically simple eigenvalue of A, for $i=1,\dots,h-1$. This establishes (1a).

Let $\mathcal{T}=\{\zeta_1,\dots,\zeta_h\}$. Note that

$$A=\zeta_iD_iAD_i^{-1}=\zeta_iD_i(\zeta_jD_jAD_j^{-1})D_i^{-1}=(\zeta_i\zeta_j)(D_iD_j)A(D_iD_j)^{-1}. \tag{6.1.9}$$

Therefore, $\zeta_i\zeta_j\rho(A)$ is an eigenvalue of A. Hence, $\zeta_i\zeta_j\in\mathcal{T}$, i.e. $\mathcal{T}$ is a semigroup. Lemma 6.1.14 yields that $\zeta_1,\dots,\zeta_n$ are all h roots of 1. Note that if $A\mathbf{z}_i=\lambda_i\mathbf{z}_i, \mathbf{z}_i\ne\mathbf{0}$, then $\mathbf{z}_i=tD_i\mathbf{u}$, for some $0\ne t\in\mathbb{C}$, where $\mathbf{u}>\mathbf{0}$ is the Perron-Frobenius vector given in Theorem 6.1.7. This establishes (1b).

Let $\zeta=e^{\frac{2\pi\sqrt{-1}}{h}}\in\mathcal{T}$. Then, $A=\zeta DAD^{-1}$, where D is a diagonal matrix $D=(d_1,\dots,d_n), |D|=I$. Since D can be replaced by $\bar{d}_1D$, we can assume that $d_1=1$.

(6.1.9) yields that $A = \zeta^h D^h A D^{-h} = IAI^{-1}$. Lemma 6.1.13 implies that $D^h = \mathrm{diag}(d_1^h, \ldots, d_n^h) = tI$. Since $d_1 = 1$, it follows that $D^h = I$. So all the diagonal entries of D are h-roots of unity. Let $P \in \mathcal{P}_n$ be a permutation matrix such that the diagonal matrix $E = P^\top DP$ is of the following block diagonal form

$$E = I_{n_1} \oplus \mu_1 I_{n_2} \oplus \ldots \oplus \mu_{l-1} I_{n_l},\ \mu_i = e^{\frac{2\pi k_i \sqrt{-1}}{h}},$$
$$i = 1, \ldots, l-1,\ 1 \le k_1 < k_2 < \ldots < k_{l-1} \le h-1.$$

Note that $l \le h$ and equality holds if and only if $k_i = i$. Let $\mu_0 = 1$.

Let $B = P^\top AP$. Partition B to a block matrix $[B_{ij}]_{i=j=1}^l$, where $B_{ij} \in \mathbb{R}_+^{n_i \times n_j}$, for $i, j = 1, \ldots, l$. Then, the equality $A = \zeta DAD^{-1}$ yields $B = \zeta EBE^{-1}$. The structure of B and E implies the equalities

$$B_{ij} = \zeta \frac{\mu_{i-1}}{\mu_{j-1}} B_{ij}, \quad i, j = 1, \ldots, l.$$

Since all the entries of B_{ij} are non-negative, we obtain that $B_{ij} = 0$ if $\zeta \frac{\mu_{i-1}}{\mu_{j-1}} \ne 1$. Hence, $B_{ii} = 0$, for $i = 1, \ldots, l$. Since B is irreducible, it follows that not all $B_{i1}, \ldots, B_{il}$ are zero matrices for each $i = 1, \ldots, l$. First, start with $i = 1$. Since $\mu_0 = 1$ and $j_1 \ge 1$, it follows that $\mu_j \ne \zeta$, for $j > 1$. Then, $B_{1j} = 0$ for $j = 3, \ldots, l$. Hence, $B_{12} \ne 0$, which implies that $\mu_1 = \zeta$, i.e. $k_1 = 1$. Now, let $i = 2$ and consider $j = 1, \ldots, l$. As $k_i \in \{k_1 + 1, k_1 + 2, \ldots, h-1\}$, for $i > 1$, it follows that $B_{2j} = 0$, for $j \ne 3$. Hence, $B_{23} \ne 0$ which yields that $k_2 = 2$. Applying these arguments for $i = 3, \ldots, l-1$, we deduce that $B_{ij} = 0$ for $j \ne i+1$, $B_{i(i+1)} \ne 0, k_i = i$ for $i = 1, \ldots, l-1$. It is left to consider $i = l$. Note that

$$\frac{\zeta \mu_{l-1}}{\mu_{j-1}} = \frac{\zeta^l}{\zeta^{j-1}} = \zeta^{l-(j-1)}, \text{ which is different from 1, for } j \in [\ell] \smallsetminus \{1\}.$$

Hence, $B_{lj} = 0$ for $j > 1$. Since B is irreducible, $B_{11} \ne 0$. So $\zeta^l = 1$. As $l \le h$ we deduce that $l = h$. Hence, B has the block form given in (1d). □

6.2 Irreducible matrices

Denote by $\mathbb{R}_+ \supset \mathbb{Z}_+$ the set of non-negative real numbers and non-negative integers, respectively. Let $\mathcal{S} \subset \mathbb{C}$. By $S_n(\mathcal{S}) \subset \mathcal{S}^{n \times n}$ denote the set of all symmetric matrices $A = [a_{ij}]$ with entries in $\mathcal{S}$. Assume that $0 \in \mathcal{S}$. Then, by $S_{n,0}(\mathcal{S}) \subset S_n(\mathcal{S})$ the subset of all symmetric matrices with entries in $\mathcal{S}$ and zero diagonal. Denote by $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ the vector of length n whose all coordinates are $\mathbf{1}$. For any $t \in \mathbb{R}$, we let $\mathrm{sign}t = 0$ if $t = 0$ and $\mathrm{sign}t = \frac{t}{|t|}$ if $t \ne 0$.

Let $D = (V, E)$ be a multidigraph. Assume that $\#V = n$ and label the vertices of V as $1, \ldots, n$. We have a bijection $\phi_1 : V \to [n]$. This bijection induces an isomorphic graph $D_1 = ([n], E_1)$. With D_1 we associate the following matrix $A(D_1) = [a_{ij}]_{i,j=1}^n \in \mathbb{Z}_+^{n \times n}$. Then, a_{ij} is the number of directed edges from the vertex i to the vertex j. (If $a_{ij} = 0$ then there no diedges from i to j.) When no confusion arises, we let $A(D) := A(D_1)$, and we call $A(D)$ the *adjacency matrix* of D. Note that a different bijection $\phi_2 : V \to [n]$ gives rise to a different $A(D_2)$, where $A(D_2) = P^\top A(D_1)P$,

for some permutation matrix $P \in \mathcal{P}_n$.
If D is a simple digraph then $A(D) \in \{0,1\}^{n\times n}$. If D is a multidigraph, then $a_{ij} \in \mathbb{Z}_+$ is the number of diedges from i to j. Hence $A(G) \in \mathbb{Z}_+^{n\times n}$. If G is a multigraph, then $A(G) = A(D(G)) \in S_n(\mathbb{Z}_+)$. If G is a simple graph, then $A(G) \in S_{n,0}(\{0,1\})$.

Proposition 6.2.1 *Let $D = (V,E)$ be a multidigraph on n vertices. Let $A(D)$ be a representation matrix of D. For an integer $k \ge 1$ let $A(D)^k = [a_{ij}^{(k)}] \in \mathbb{Z}_+^{n\times n}$. Then, $a_{ij}^{(k)}$ is the number of walks of length k from the vertex i to the vertex j. In particular, $\mathbf{1}^\top A\mathbf{1}$ and $\operatorname{tr} A$ are the total number of walks and the toral number of closed walks of length k in D.*

Proof. For $k = 1$ the proposition is obvious. Assume that $k > 1$. Recall that

$$a_{ij}^{(k)} = \sum_{i_1,\ldots,i_{k-1}\in[n]} a_{ii_1}a_{i_1i_2}\cdots a_{i_{k-1}j}. \tag{6.2.1}$$

The summand $a_{ii_1}a_{i_1i_2}\cdots a_{i_{k-1}j}$ gives the number of walks of the form $i_0i_1i_2\cdots i_{k-1}i_k$, where $i_0 = i$, $i_k = j$. Indeed, if one of the terms in this product is zero, i.e. there is no diedge (i_p, i_{p+1}), then the product is zero. Otherwise each positive integer $a_{i_pi_{p+1}}$ counts the number of diedges (i_p, i_{p+1}). Hence, $a_{ii_1}a_{i_1i_2}\cdots a_{i_{k-1}j}$ is the number of walks of the form $i_0i_1i_2\cdots i_{k-1}i_k$. The total number of walks from $i = i_0$ to $j = i_k$ of length k is the sum given by (6.2.1). To find out the total number of walks in D of length k is $\sum_{i=j=1}^n a_{ij}^{(k)} = \mathbf{1}^\top A\mathbf{1}$. The total number of closed walks in D of length k is $\sum_{i=1}^k a_{ii}^{(k)} = \operatorname{tr} A(D)^k$. □

With a multipartite graph $G = (V_1 \cup V_2, E)$, where $\#V_1 = m$, $\#V_2 = n$, we associate a representation matrix $B(G) = [b_{ij}]_{i=j=1}^{m,n}$ as follows. Let $\psi_1 : V_1 \to [m]$, $\phi_1 : V_2 \to [m]$ be bijections. This bijection induces an isomorphic graph $D_1 = ([m] \cup [n], E_1)$. Then, b_{ij} is the number of edges connecting $i \in [m]$ to $j \in [n]$ in D_1.
A non-negative matrix $A = [a_{ij}]_{i=j=1}^n \in \mathbb{R}_+^{n\times n}$ induces the following digraph $D(A) = ([n], E)$. The diedge (i,j) is in E if and only if $a_{ij} > 0$. Note that of $A(D(A)) = [\mathrm{sign} a_{ij}] \in \{0,1\}^{n\times n}$.

Theorem 6.2.2 *Let $D = ([n], E)$ be a multidigraph. Then, D is strongly connected if and only if $(I + A(D))^{n-1} > 0$, in particular, a non-negative matrix $A \in \mathbb{R}_+^{n\times n}$ is irreducible if and only if $(I + A)^{n-1} > 0$.*

Proof. Apply the Newton binomial theorem for $(1+t)^{n-1}$ to the matrix $(I + A(D))^{n-1}$

$$(I + A(D))^{n-1} = \sum_{p=0}^{n-1} \binom{n-1}{p} A(D)^p.$$

Recall that all the binomial coefficients $\binom{n-1}{p}$ are positive for $p = 0, \ldots, n-1$. Assume first that $(I + A(D))^{n-1}$ is positive. Hence the (i,j) entry of $A(D)^p$ is positive for some $p = p(i,j)$. Let $i \neq j$. Since $A(D)^0 = I$, we deduce that $p(i,j) > 0$. Use Proposition 6.2.1 to deduce that there is a walk of length p from the vertex i to the vertex j.
Suppose that D is strongly connected. Then, for each $i \neq j$ we must have a path pf length $p \in [1, n-1]$ which connects i and j. Hence, all off-diagonal entries of

$(I+A(D))^{n-1}$ are positive. Clearly, $(I+A(D))^{n-1} \geq I$. Hence, $(I+A(D))^{n-1} > 0$. Let $A \in \mathbb{R}_+^{n\times n}$. Then, the (i,j) entry of $(I+A)^{n-1}$ is positive if and only if the (i,j) entry of $(I+A(D(A)))^{n-1}$ is positive. Hence, A is irreducible if and only if $(I+A)^{n-1} > 0$. □

6.3 Recurrence equation with non-negative coefficients

Recall that the matrix $A = [a_{ij}] \in \mathbb{R}_+^{n\times n}$ is called irreducible if $(I+A)^{n-1} > 0$. We now correspond a digraph $D = D(A)$ of order n to A as follows. The vertex set is $\mathbf{V} = \{a_1, \ldots, a_n\}$. There is an arc $\alpha = (a_i, a_j)$ from a_i to a_j if and only if $a_{ij} > 0$, $(i, j \in [n])$. It is shown that A is irreducible if and only if D is strongly connected. (Theorem 6.2.2.)

We have already seen that $A \in \mathbb{R}_+^{n\times n}$ is called primitive if $A^p > 0$, for some positive integer p. It is proved that if A is an irreducible matrix such that the g.c.d. of the lengths of all cycles in $D(A)$ is 1, then A is primitive. See [7] for more details on primitive matrices.

Consider the following homogeneous recurrence equation:

$$\mathbf{x}_k = A\mathbf{x}_{k-1}, \quad k = 1, 2, \ldots \tag{6.3.1}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ a_n & a_{n-1} & a_{n-2} & \ldots & a_2 & a_1 \end{bmatrix} \in \mathbb{R}^{n\times n}, \ \mathbf{x}_k = \begin{bmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+n-1} \end{bmatrix} \in \mathbb{R}^n$$

(See subsection 4.1.1.)

Assume that $a_1, \ldots, a_n \geq 0$. Note that from the form of A, we have $1 \to 2 \to 3 \to \cdots \to n$. Thus, the condition for A to be irreducible is that $a_1 > 0$, i.e. $n \to 1$. Note that the condition $a_1 > 0$ implies that $D(A)$ has a Hamiltonian cycle. For example, if $a_2 > 0$ (in addition to $a_1 > 0$), then $D(A)$ has a cycle of length $n-1 : 2 \to 3 \to \cdots \to n \to 2$. Since A is assumed to be irreducible $(a_1 > 0)$ and the g.c.d. of the cycles in D_A is 1 (as the g.c.d. of n and $n-1$ is 1), then A is primitive.

Now, we are ready to give a generalization of Kepler's Theorem for Fibonacci numbers which can be considered as an application of Perron-Frobenius theorem.

Theorem 6.3.1 *Consider the homogeneous recurrence equation* (6.3.1). *Assume that $a_1 > 0$ and $a_2, \ldots, a_n \geq 0$. Suppose furthermore that $a_{n_1}, \ldots, a_{n_i} > 0$ where $1 < n_1 < \cdots < n_i \leq n$. Assume that the g.c.d. of $n, n-n_i+1$ is 1. Suppose that $(u_0, \ldots, u_{n-1})^\top \gneq \mathbf{0}$. Then*

$$\lim_{m\to\infty} \frac{u_{m+1}}{u_m} = \rho(A). \tag{6.3.2}$$

More precisely:

$$\lim_{k\to\infty} \frac{1}{\rho(A)^k}\mathbf{x}_k = (\mathbf{v}^\top \mathbf{x}_0)\mathbf{u}, \tag{6.3.3}$$

where $\mathbf{u}$, $\mathbf{v}$ *are defined below.*

Proof. The assumptions of the theorem yield that A is primitive. Then, Perron-Frobenius theorem yields that $\rho(A) > 0$, is a simple eigenvalue of A, and all other eigenvalues of A satisfy $|\lambda| < \rho(A)$. The component of A corresponding to $\rho(A)$ is of the form $\mathbf{u}\mathbf{v}^\top$, where $\mathbf{u}, \mathbf{v} > \mathbf{0}$ are the right and the left eigenvector of A satisfying

$$A\mathbf{u} = \rho(A)\mathbf{u},\ A^\top\mathbf{v} = \rho(A)\mathbf{v},\ \mathbf{v}^\top\mathbf{u} = 1.$$

Hence, writing down A^m in terms of its components and taking $A^m\mathbf{x}_0$, we see that the leading term is

$$\mathbf{x}_k = \rho(A)^k\mathbf{u}\mathbf{v}^\top\mathbf{x}_0 + O(t^k),\ t = \max\{|\lambda_j(A)|, |\lambda_j(A)| < \rho(A)\}.$$

As $\mathbf{v}^\top\mathbf{x}_0 > 0$, we deduce the theorem. □

6.3.1 Worked-out Problems

1. Let $A \in \mathbb{R}_+^{n\times n}$ be irreducible. Show that A is primitive if and only if one of the following conditions hold:

   (a) $n = 1$ and $A > 0$.

   (b) $n > 1$ and each eigenvalue λ of A different from $\rho(A)$ satisfies the inequality $|\lambda| < \rho(A)$.

   Solution:
   If $n = 1$, then A is primitive if and only if $A > 0$. Assume now that $n > 1$. So $\rho(A) > 0$. Considering $B = \frac{1}{\rho(A)}A$, it is enough to consider the case $\rho(A) = 1$. Assume first that if $\lambda \neq 1$ is an eigenvalue of A, then $|\lambda| < 1$. Theorem 6.1.7 implies $A\mathbf{u} = \mathbf{u}, A^\top\mathbf{w} = \mathbf{w}$ for some $\mathbf{u}, \mathbf{w} > \mathbf{0}$. So $\mathbf{w}^\top\mathbf{u} > 0$. Let $\mathbf{v} := (\mathbf{w}^\top\mathbf{u})^{-1}\mathbf{w}$. Then, $A^\top\mathbf{v} = \mathbf{v}$ and $\mathbf{v}^\top\mathbf{u} = 1$. Part *2.* of Theorem 4.6 yields $\lim_{k\to\infty} A^k = Z_{10} \geq 0$, where Z_{10} is the component of A corresponding to an algebraically simple eigenvalue $1 = \rho(A)$. Problem 4.1.3-1.a yields that $Z_{10} = \mathbf{u}\mathbf{v}^\top > 0$. So there exists an integer $k_0 \geq 1$, such that $A^k > 0$, for $k \geq k_0$, i.e. A is primitive.
   Assume now A has exactly $h > 1$ distinct eigenvalues λ satisfying $|\lambda| = 1$. Lemma 6.1.15 implies that there exists a permutation matrix P such that $B = P^\top AP$ is of the form (1d) of Theorem 6.1.15 . Note that B^h is a block diagonal matrix. Hence, $B^{hj} = (B^h)^j$ is a block diagonal matrix for $j = 1, \ldots, \ldots$. Hence, B^{hj} is never a positive matrix, so A^{hj} is never a positive matrix. Hence A is not primitive.

6.3.2 Problems

1. Let $B \in \mathbb{R}_+^{n\times n}$ be an irreducible, imprimitive matrix, having $h > 1$ distinct eigenvalues λ satisfying $|\lambda| = \rho(B)$. Suppose furthermore that B has the form (1d) of Theorem 6.1.15 . Show that B^h is a block diagonal matrix, where each diagonal block is an irreducible primitive matrix whose spectral radius is $\rho(B)^h$. In particular, show that the last claim of (1d) of Theorem 6.1.15 holds.

Bibliography

[1] J.L. Arocha, B.L. Iano, M. Takane. The Theorem of Philip Hall for Vector Spaces, An. Inst. Mat. Univ. Nac. Autónoma México 32 (1992), 1-8 (1993).

[2] A. Barvinok, A course in convexity, GSM 54, Amer. Math. Soc., Providence, RI, 2002.

[3] J.A. Bondy, U.S.R. Murty, Graph Theory, Graduate Texts in Mathematics, Springer, 2008.

[4] L.M. Bregman. Some Properties of Nonnegative Matrices and Their Permanents. Soviet Math. Dokl. 14 (1973), 945-949 [Dokl. Akad. Nauk SSSR 211 (1973), 27-30.

[5] Y. Filmus. Range of Symmetric Matrices over GF(2). Unpublished note, University of Toronto, 2010.

[6] B. Fine, G. Rosenberger. The Fundamental Theorem of Algebra. Undergraduate Texts n Mathematics, Springer-Verlag, New York, 1997.

[7] S. Friedland, Matrices: Algebra, Analysis and Applications, World Scientific, Singapore, 2016.

[8] G.H. Golub and C.F. Van Loan. Matrix Computation, *John Hopkins Univ. Press, 4th Ed.*, Baltimore, 2013.

[9] R. J. Gregorac, A Note on Finitely Generated Groups. Proc. Amer. Math. Soc. 18. 1967, pp. 756-758.

[10] J. Hannah. A Geometric Approach to Determinant. The American Mathematical Monthly. Vol. 103, No. 5 (May, 1996), pp. 401-409.

[11] I.N. Herstein. Abstract Algebra. 3rd Ed. Wiley, New York, 1996.

[12] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 2013.

[13] B. Mendelson. Introduction to Topology, 3rd Ed. Dover Publications, Inc, New York, 1990.

[14] M.A. Nielsen and I.L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2010.

[15] A. Ostrowski, Solution of Equations in Euclidean and Banach Spaces, Academic Press, 1973.

[16] R. Rado. A Theorem on Independence Relations, Quart. J. Math, Oxford Ser. 13 (1942), 83-89.

[17] S. Roman, Advanced Linear Algebra, Springer-Verlag, Graduate Texts in Mathematics Vol. 135, 1992.

[18] W. Rudin. Principles of Mathematical Analysis, 3rd Ed. International Series in Pure and Applied Mathematics. New York, 1974.

[19] R. Schindler. Set Theory: Exploring Independence and Truth. Universitext. Springer, Cham (2014).

[20] A. Schrijver. A Short Proof of Minc's Conjecture. Journal of Combinational Theory, Series A 25, 80-83 (1978).

[21] J. Shipman. Improving of Fundamental Theorem of Algebra. Math. Interlligencer 29 (2009), no. 4, 9-14. 00-01 (12D05).

[22] G. Toth. Measures of Symmetry for Convex Sets and Stability. Universitext. Springer, Cham (2015).

[23] J.H. Van Lint, R.M. Wilson, A Course in Combinatorics, Cambridge University Press, Cambridge, 1992.

Index