MATRICES
Summary of Lectures

MATH 494 Special Topics in Mathematics, Fall 2005

CRN: 23662 - undergrad, 23663 - grad, MWF 3:00-3:50, 302 AH

Instructor: Shmuel Friedland
Department of Mathematics, Statistics and Computer Science
email: friedlan@uic.edu

Last update February 6, 2006

1 Introduction

Theory of matrices related to many fields of biology, business, engineering, medicine, science and social sciences. Let us give a few examples. Recall that

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

is called an $m \times n$ matrix and briefly denoted by $A = (a_{ij})_{i,j=1}^{m,n}$ or just $A = (a_{ij})$. If the entries a_{ij} are in some given set $\mathcal{S}$ we *denote* by $\mathcal{S}^{m\times n}$ the set of all $m \times n$ with entries in $\mathcal{S}$. In some other books $\mathcal{S}^{n\times n}$ is denoted by $\mathrm{M}_{mn}(\mathcal{S})$ and $\mathrm{M}_n(\mathcal{S})$ stands for $\mathrm{M}_{nn}(\mathcal{S})$. As usual $\mathbb{R}, \mathbb{C}, \mathbb{Z}$ and $\mathbb{F}$ stands for the set of real numbers, complex numbers, integers, and a field respectively.

Consider $A \in \mathbb{R}^{n\times m}$. It can be interpreted as a digital picture seen on a screen. Then a_{ij} encodes the color and its strength in the location of (i, j). In many case m and n are very big, so it is very costly and time consuming to storage the information, or to transmit it. We know that there is a lot of redundancy in the picture. Is there a way to condense the information to have almost the same picture, when an average person looks at it? The answer is yes, and one way to achieve it is to use the *singular value decomposition* discussed later in this course.

An other possibility is that A represents DNA gene expression data, where a_{ij} is the expression level of the gene i in the experiment number j. The number of genes is huge, e.g. from 6,000 to 100,000 and the number of experiments can be from 4 to 30. This is done by lasers and computers, and certain percentage of entries is corrupted. To do some statistics on DNA we need the values of all entries of A. Is there a good way to *impute*, (complete), the values of A using matrix theory? The answer is yes, and one can use least squares and inverse eigenvalue techniques to do it.

In many applications one has a linear system given schematically by the *input-output* (black box) relation $\mathbf{x} \to \mathbf{y}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are column vectors with n coordinates, and $\mathbf{y} = A\mathbf{x}$, where $A \in \mathbb{R}^{n\times n}$. If one repeats this procedure m times, (closed loop, then $\mathbf{x}_m = A^m\mathbf{x}$, $m = 1, 2, \dots$. How does $\mathbf{x}_m$ loos like where m is very big? This question is very natural in stationary Markov chains, which are nowadays are very popular in many

simulations and algorithms for hard problems in combinatorics and computer science. The answer to this problem is given by using the Jordan canonical form.

Let $G = (V, E)$ be a digraph on the set of n vertices V, and the set of edges $E \subset V \times V$. Then G is represented by $A = (a_{ij}) \in \{0,1\}^{n \times n}$, where $a_{ij} = 0$ or $a_{ij} = 1$ if there is no edge or there is an edge from the vertex i to the vertex j respectively. Many properties of graph are reflected in the *spectrum*, (the set of eigenvalues), of A, and the corresponding eigenvectors, in particular to the eigenvector corresponding to the nonnegative eigenvalue of the maximal modulus. This topis is covered by the Perron-Frobenius theorem for nonnegative matrices.

2 Jordan Canonical Form

2.1 Statement of the Problem

Remark 2.1 *In this notes we sometimes are going to emphasize that certain results hold for a general field $\mathbb{F}$. The student unfamiliar with this notion can safely assume that $\mathbb{F}$ is either the field of complex numbers or the field of the real numbers.*

Let $\mathbf{V}$ be a *vector space* over the field $\mathbb{F}$ of dimension n, e.g. $\mathbf{V} = \mathbb{F}^n$. (Here $\mathbb{F}^n$ is the set of column vectors with n coordinates in the field $\mathbb{F}$. To save space we denote the column vector $\mathbf{x}$ with coordinates $x_1, \ldots, x_n$ as $\mathbf{x} = (x_1, \ldots, x_n)^\top$.) Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be a basis in $\mathbf{V}$, i.e. any vector $\mathbf{x} \in \mathbf{V}$, is uniquely expressed as a *linear combination* of $\mathbf{u}_1, \ldots, \mathbf{u}_n$: $\mathbf{x} = x_1\mathbf{u}_1 + x_2\mathbf{u}_2 + \ldots + x_n\mathbf{u}_n$. (Any set of n *linearly independent vectors* forms a basis in n-dimensional vector space.) The vector $(x_1, x_2, \ldots, x_n)^\top \in \mathbb{F}^n$ is called the *coordinate vector* of $\mathbf{x}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_n]$ of $\mathbf{V}$, and $x_1, \ldots, x_n$ are called the coordinates of $\mathbf{x}$ with respect to $[\mathbf{u}_1, \ldots, \mathbf{u}_n]$. It is convenient to use the *formalism*: $\mathbf{x} = [\mathbf{u}_1, \ldots, \mathbf{u}_n](x_1, \ldots, x_n)^\top$. In $\mathbb{F}^n$, ($\mathbb{R}^n$ or $\mathbb{C}^n$), we have the *standard basis*

$$\mathbf{e}_1 = (1, 0, \ldots, 0)^\top,\ \mathbf{e}_2 = (0, 1, 0, \ldots, 0)^\top, \ldots, \mathbf{e}_n = (0, \ldots, 0, 1)^\top,$$

where $\mathbf{e}_i$ has the i-th coordinate equal to 1, while all other coordinates are equal to 0. Let $[\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be an other basis in $\mathbf{V}$. Then there exists $U = (u_{ij}) \in \mathbb{F}^{n \times n}$ such that

$$[\mathbf{u}_1, \ldots, \mathbf{u}_n] = [\mathbf{v}_1, \ldots, \mathbf{v}_n]U \iff \mathbf{u}_i = \sum_{j=1}^{n} u_{ji}\mathbf{v}_j, \text{ for } i = 1, \ldots, n. \tag{2.1}$$

Furthermore, U is an invertible matrix, i.e. there exists $V \in \mathbb{F}^{n \times n}$ such that $UV = VU = I_n$, where I_n in $n \times n$, identity matrix whose i-th column is the vector $\mathbf{e}_i$, for $i = 1, \ldots, n$. (If no confusion arises we sometimes denote I_n by I.) V is a unique matrix which is denoted by U^{-1}, *the inverse* of U. U is called the the *transition matrix* from the base $[\mathbf{u}_1, \ldots, \mathbf{u}_n]$ to the base $[\mathbf{v}_1, \ldots, \mathbf{v}_n]$.

Let $[\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be a basis and define vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ as in (2.1). Then $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is a basis in $\mathbf{V}$ if and only U is an invertible matrix. Furthermore, if we multiply $[\mathbf{u}_1, \ldots, \mathbf{u}_n] = [\mathbf{v}_1, \ldots, \mathbf{v}_n]U$ by U^{-1} from the right we get that $[\mathbf{v}_1, \ldots, \mathbf{v}_n] = [\mathbf{u}_1, \ldots, \mathbf{u}_n]U^{-1}$, i.e. the transition matrix from "$\mathbf{v}$"- basis to "$\mathbf{u}$"-basis is given by the inverse of the transition matrix from "$\mathbf{u}$"- basis to "$\mathbf{v}$"-basis.

Let $T : \mathbf{V} \to \mathbf{V}$ be a *linear transformation*: $T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$ for all *scalars* $a, b \in \mathbb{F}$ and vectors $\mathbf{x}, \mathbf{y} \in \mathbf{V}$. For example for $A \in \mathbb{F}^{n \times n}$ $A : \mathbb{F}^n \to \mathbb{F}^n$ is given by $\mathbf{x} \mapsto A\mathbf{x}$, for any column vector $\mathbf{x} \in \mathbb{F}^n$, is linear transformation.

Any linear transformation T is determined uniquely by its representation matrix $A = (a_{ij}) \in \mathbb{F}^n$ in a given basis $[\mathbf{u}_1, \ldots, \mathbf{u}_n]$, defined as $T\mathbf{u}_i = a_{1i}\mathbf{u}_1 + a_{2i}\mathbf{u}_2 + \ldots + a_{ni}\mathbf{u}_n, i = 1, \ldots, n$. The formalism notation is

$$T[\mathbf{u}_1, \ldots, \mathbf{u}_n] := [T\mathbf{u}_1, \ldots, T\mathbf{u}_n] = [\mathbf{u}_1, \ldots, \mathbf{u}_n]A.$$

Note that if $\mathbf{x}$ and $\mathbf{y}$ are the coordinate vectors of $\mathbf{v}$ and $T\mathbf{v}$ respectively, then $\mathbf{y} = A\mathbf{x}$:

$$T\mathbf{v} = T(\sum_{i=1}^{n} x_i\mathbf{u}_i) = \sum_{i=1}^{n} x_i T\mathbf{u}_i = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ji}\mathbf{u}_j = \sum_{i=j=1}^{n} a_{ji}x_i\mathbf{u}_j = \sum_{j=1}^{n}(\sum_{i=1}^{n} a_{ji}x_i)\mathbf{u}_j = \sum_{j=1}^{n} y_j\mathbf{u}_j.$$

This easily follows from the formalism

$$T\mathbf{v} = T([\mathbf{u}_1,\dots,\mathbf{u}_n]\mathbf{x}) = (T[\mathbf{u}_1,\dots,\mathbf{u}_n])\mathbf{x} = ([\mathbf{u}_1,\dots,\mathbf{u}_n]A)\mathbf{x} = [\mathbf{u}_1,\dots,\mathbf{u}_n](A\mathbf{x}).$$

Let $[\mathbf{v}_1,\dots,\mathbf{v}_n]$ be another basis in $\mathbf{V}$. Assume (2.1). Then the representation matrix of T in "$\mathbf{v}$"-basis is given by $B = UAU^{-1}$:

$$T[\mathbf{u}_1,\dots,\mathbf{u}_n] = [\mathbf{u}_1,\dots,\mathbf{u}_n]A \Rightarrow T([\mathbf{v}_1,\dots,\mathbf{v}_n]U) = ([\mathbf{v}_1,\dots,\mathbf{v}_n]U)A \Rightarrow$$
$$(T[\mathbf{v}_1,\dots,\mathbf{v}_n])U = [\mathbf{v}_1,\dots,\mathbf{v}_n](UA) \Rightarrow T[\mathbf{v}_1,\dots,\mathbf{v}_n] = [\mathbf{v}_1,\dots,\mathbf{v}_n](UAU^{-1}).$$

Definition 2.2 *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* $\mathbb{F}$. $A, B \in \mathbb{F}^{n\times n}$ *are called* similar, *and this is denoted by* $A \sim B$, *if* $B = UAU^{-1}$ *for some* $U \in \mathrm{GL}(n,\mathbb{F})$. *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 2.3 *Let* $\mathbb{F}$ *be a field,* ($\mathbb{F} = \mathbb{R}, \mathbb{C}$). *Then the similarity relation on* $\mathbb{F}^{n\times n}$ *is an equivalence relation:*

$$A \sim A,\ A \sim B \iff B \sim A,\ A \sim B \text{ and } B \sim C \Rightarrow A \sim B.$$

Furthermore if $B = UAU^{-1}$ *then*

1. $\det(zI_n - B) = \det(zI_n - A)$, *i.e.* A *and* B *have the same characteristic polynomial.*
2. *For any integer* $m \geq 2$ $B^m = UA^mU^{-1}$.
3. *If in addition* A *is invertible, then* B *is invertible and* $B^m = UA^mU^{-1}$ *for any integer* m.

Corollary 2.4 *Let* $\mathbf{V}$ *be* 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 *is a similarity class. 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* $\mathrm{trace}\ T^m = \mathrm{trace}\ A^m$ *for any nonnegative integer. (*T^0 *is the identity operator, i.e* $T^0\mathbf{v} = \mathbf{v}$ *for all* $\mathbf{v} \in \mathbf{V}$, *and* $A^0 = I$. *Here by the trace of* $B \in \mathbb{F}^{n\times n}$, *denoted by* $\mathrm{trace}\ B$, *we mean the sum of all diagonal elements of* B.*)*

Problem 2.5 *(The representation problem.) Let* $\mathbf{V}$ *be* n*-dimensional vector space over* $\mathbb{F}$. *Assume that* $T : \mathbf{V} \to \mathbf{V}$ *is a linear transformation. Find a basis* $[\mathbf{v}_1,\dots,\mathbf{v}_n]$ *in which* T *has the simplest form. Equivalently, given* $A \in \mathbb{F}^{n\times n}$ *find* $B \sim A$ *of the simplest form.*

In the following case the answer is well known. Recall that $\mathbf{v} \in \mathbf{V}$ 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 $\det(\lambda I - A) = 0$. Hence 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 2.6 *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* 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.

Proposition 2.7 *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 dy a diagonal matrix*

$$\operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) := \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix},$$

if and only if the characteristic polynomial of T *is* $(z-\lambda_1)(z-\lambda_2)\dots(z-\lambda_n)$*, 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* $\operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ *if and only if* $\det(zI - A) = (z-\lambda_1)(z-\lambda_2)\dots(z-\lambda_n)$*, and* A *has* n*-linearly independent eigenvectors.*

Proof. Assume that there exists a basis $[\mathbf{u}_1, \dots, \mathbf{u}_n]$ in V such that T is represented in this basis dy a diagonal matrix $\Lambda := \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. 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, it follows that $T\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for $i = 1, \dots, n$. Since each $\mathbf{u}_i \neq \mathbf{0}$, we deduce that each $\mathbf{u}_i$ is an eigenvector of T. By our assumption $\mathbf{u}_1, \dots, \mathbf{u}_n$ for a basis in $\mathbf{V}$.

Assume now that $\mathbf{V}$ has a basis $[\mathbf{u}_1, \dots, \mathbf{u}_n]$ consisting eigenvectors of T. So $T\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for $i = 1, \dots, n$. Hence Λ is the representation matrix of T in the basis $[\mathbf{u}_1, \dots, \mathbf{u}_n]$.

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 $\mathbf{x} \in \mathbb{F}^n$. □

Theorem 2.8 *Let* $\mathbf{V}$ *be* 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* $\lambda_1, \dots, \lambda_n \in \mathbb{F}$*, 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 2.6). Hence to 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 Problem 1 of this section and Proposition 2.7. □

Given $A \in \mathbb{F}^{n\times n}$ it may happen that $\det(zI - A)$ does not have n roots in $\mathbb{F}$. (See for example Problem 2 of this section.) Hence we can not *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 diagform in general and Theorem diagthm in particular to see if A is diagonable.

Since $\mathbb{R}$ is not algebraically closed and $\mathbb{C}$ is, that is the reason that we sometimes we view a real valued matrix $A \in \mathbb{R}^{n\times n}$ as a complex valued matrix $A \in \mathbb{C}^{n\times n}$. (See Problem 2 of this section.)

Corollary 2.9 *Let* $A \in \mathbb{C}^{n\times n}$ *be nondiagonable. Then its characteristic polynomial must have a multiple root.*

See Problem 3 of this section.

Definition 2.10 *1. Let* k *be a positive integer and* $\lambda \in \mathbb{F}$*. Then* $J_k(\lambda) \in \mathbb{F}^{k\times k}$ *be a* $k \times k$ *be an upper diagonal matrix, with* λ *on the main diagonal,* 1 *on the next*

sub-diagonal and other entries are equal to 0 *for* $k > 1$*:*

$$J_k(\lambda) := \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix},$$

($J_1(\lambda) = [\lambda]$*.)*

2. *Let* $A_i \in \mathbb{F}^{n_i \times n_i}$ *for* $i = 1, \dots, l$*. Denote by*

$$\oplus_{i=1}^k A_i = A_1 \oplus A_2 \oplus \dots \oplus A_k = \operatorname{diag}(A_1, A_2, \dots, A_k) := \begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & A_k \end{bmatrix} \in \mathbb{F}^{n \times n}, \quad n = n_1 + n_2 + \dots + n_k,$$

the $n \times n$ *block diagonal matrix, whose blocks are* $A_1, A_2, \dots, A_k$*.*

Theorem 2.11 ***(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* $\lambda_1, \dots, \lambda_k \in \mathbb{C}$*, (*$\lambda_1, \dots, \lambda_k \in \mathbb{F}$*), and positive integers* $n_1, \dots, n_k$*. The Jordan canonical form is unique up to the permutations of the Jordan blocks* $J_{n_1}(\lambda_1), \dots, J_{n_k}(\lambda_k)$*.*

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* $J_{n_i}(\lambda_i), i = 1, \dots, k$ *are unique.*

Note that $A \in \mathbb{C}^{n \times n}$ is diagonable if and only in its Jordan canonical form $k = n$, i.e. $n_1 = \dots = n_n = 1$. For $k < n$, the *Jordan canonical form* is the simplest form of the similarity class of a nondiagonable $A \in \mathbb{C}^{n \times n}$.

We will prove Theorem 2.11 in the next several sections.

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 $i = 1, \dots, m$. Show by induction on m that if $\lambda_1, \dots, \lambda_m$ are m distinct scalars then $\mathbf{u}_1, \dots, \mathbf{u}_m$ 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 diagonable over the real numbers $\mathbb{R}$.

   (b) Show that A is diagonable over the complex numbers $\mathbb{C}$. Find $U \in C^{2 \times 2}$ 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 $A_i \in \mathbb{C}^{n_i \times n_i}, i = 1, \dots, k$. Show that $\det(zI_n - A) = \prod_{i=1}^k \det(zI_{n_i} - A_i)$. (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 diagonable 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 claims that the above identity holds for *any* $A \in \mathbb{C}^{n\times n}$.) **Hint:** Use the fact that A is diagonable.

2.2 Matrix polynomials

For a field $\mathbb{F}$, $\mathbb{F} = \mathbb{R}, \mathbb{C}$, denote by $\mathbb{F}[z]$, the *ring* of polynomials $p(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_n$ with coefficients in $\mathbb{F}$. The *degree* of p, 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. $a_j \neq 0$. So $\deg p = n$ if and only if $a_0 \neq 0$, the degree of a nonzero 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 n at most, 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 $\mathbb{F}[z]$ satisfies all the nice distribution identities as the addition and multiplication in $\mathbb{F}$. 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 variable over $\mathbb{F}$.)

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 $p(z)P(z) := (p(z)p_{ij}) \in \mathbb{F}[z]$. Again, this product satisfies nice distribution properties. Thus $\mathbb{F}[z]$ 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

$$p_{ij}(z) = \sum_{k=0}^{l} p_{ij,k} z^{l-k}, \; P_k := (p_{ij,k})_{i=j=1}^{m,n} \in \mathbb{F}^{m\times n} \text{ for } k = 0, \ldots, l.$$

Then

$$P(z) = P_0 z^l + P_1 z^{l-1} + \ldots + P_l, \quad P_i \in \mathbb{F}^{m\times n}, \; i = 0, \ldots, l, \tag{2.2}$$

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 (2.2) and any $A \in \mathbb{F}^{n\times n}$ we define

$$P(A) = \sum_{i=0}^{l} P_i A^{l-i} = P_0 A^l + P_1 A^{l-1} + \ldots + P_l, \text{ where } A^0 = I_n.$$

Recall that 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 2.12 *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 $A^iA^j = A^{i+j}$ for any nonnegative 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 2.13 *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*

$$P(z) = T_{left}(z)(zI - A) + P(A). \tag{2.3}$$

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 (2.2). Then $P(z) = P_0z^l + \tilde{P}(z)$, where $\tilde{P}(z) = \sum_{i=1}^{l} P_iz^{l-1}$. 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

$$P_0z^l = \hat{T}_{left}(z)(zI_n - A) + P_0A^l, \text{ where } \hat{T}_{left}(z) = \sum_{i=0}^{l-1} P_0A^iz^{l-i-1},$$

and $\hat{T}_{left}$ is unique. Hence $T_{left}(z) = \hat{T}_{left}(z) + \tilde{T}_{left}$ is unique, $P(A) = P_0A^l + \tilde{P}(A)$ and (2.3) 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 (2.3) 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 generalized to any $Q(z) = Q_0z^l + Q_1z^{l-1} + \ldots + Q_l \in \mathbb{F}[z]$, where $Q_0 \in \mathrm{GL}(n, \mathbb{F})$: There exists unique $T_{left}(z), R_{left}(z) \in \mathbb{F}[z]$ such that

$$P(z) = T_{left}(z)Q(z) + R_{left}(z),\ \deg R_{left} < \deg Q,\ Q(z) = \sum_{i=0}^{l} Q_iz^{l-i},\ Q_0 \in \mathrm{GL}(n, \mathbb{F}). \tag{2.4}$$

Here we agree that $(Az^i)(Bz^j) = (AB)z^{i+j}$ for any $A, B \in \mathbb{F}^{n\times n}$ and nonnegative integers i, j.

Theorem 2.14 *(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 $B(z) \in \mathbb{F}[z]^{n\times n}$. Recall 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 from the right $P(z)$. Lemma 2.13 yields that $P(A) = \mathbf{0}$. □

For $p, q \in \mathbb{F}[z]$ let (p, q) be 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. s is determined up to a multiple of a nonzero scalar. s can be chosen as a unique *monic* polynomial:

$$s(z) = z^l + s_1 z^{l-1} + \ldots + s_l \in \mathbb{F}[z]. \tag{2.5}$$

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

$$p_i(z) = t_i(z)p_{i+1}(z) + p_{i+2}(z),\ \deg p_{i+2} < \deg p_{i+1} \quad i = 1, \ldots \tag{2.6}$$

Start this algorithm with $p_1 = p, p_2 = q$. Continue it until $p_k = 0$ the first time. (Note that $k \geq 3$. Then $p_{k-1} = (p, q)$. It is straightforward to show from the above algorithm that

$$(p(z), q(z)) = u(z)p(z) + v(z)q(z),\ \text{for some } u(z), v(z) \in \mathbb{F}[z]. \tag{2.7}$$

(This formula holds for any $p, q \in \mathbb{F}[z]$.) $p, q \in \mathbb{F}[z]$ are called *coprime* if $(p, q) = 1$.

Corollary 2.15 *Let $p, q \in \mathbb{F}[z]$ be coprime. Then there exists $u, v \in \mathbb{F}[z]$ 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 $p, q \in \mathbb{F}[z]$ are both nonzero polynomials that split (to linear factors) over $\mathbb{F}$. So

$$p(z) = p_0(z - \alpha_1)\ldots(z - \alpha_i), p_0 \neq 0, \quad q(z) = q_0(z - \beta_1)\ldots(z - \beta_j), q_0 \neq 0.$$

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 is a nonzero 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 $P(z) = p(z)I_n$.

2.3 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 $A^0 \neq \mathbf{0}$.) So $\sum_{i=0}^m b_i A^{m-i} = 0$, and $(b_0, \ldots, b_m)^\top \neq \mathbf{0}$. If $b_0 = 0$ then $A^0, \ldots, A^{m-1}$ are linearly dependent, which contradicts the definition of m. Hence $b_0 \neq 0$. Divide the linear dependence by b_0 to obtain.

$$\psi(A) = 0, \quad \psi(z) = z^m + \sum_{i=1}^m a_i z^{m-i} \in \mathbb{F}[z], \quad a_i = \frac{b_i}{b_0} \text{ for } i = 1, \ldots, m. \tag{2.8}$$

ψ is called the *minimal polynomial* of A. In principle $m \leq n^2$, but in reality $m \leq n$:

Theorem 2.16 *Let $A \in \mathbb{F}^{n\times n}$ and $\psi(z)$ be its characteristic polynomial. Assume that $p(z) \in \mathbb{F}[z]$ is an annihilated polynomial of A, i.e. $p(A) = \mathbf{0}$. Then ψ divides p. In particular, the characteristic polynomial $p(z) = \det(zI_n - A)$ is divisible by $\psi(z)$. Hence $\deg \psi \leq \deg p = n$.*

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 2.12 yields that $p(A) = t(A)\psi(A) + r(A)$ which implies that $r(A) = 0$. Assume that $l = \deg r(z) \geq 0$, i.e. r is not identically the zero

polynomial. So $A^0, \dots, A^l$ 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 $\deg \psi \leq \deg p = n$. □

Definition 2.17 *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 2.18 *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$, and assume that $\mathbf{V}_1, \dots, \mathbf{V}_i$ nonzero subspaces of $\mathbf{V}$. Then $\mathbf{V}$ is a direct sum of $\mathbf{V}_1, \dots, \mathbf{V}_i$, 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 $j = 1, \dots, i$. Equivalently, let $[\mathbf{v}_{j1}, \dots, \mathbf{v}_{jl_j}]$ be a basis of $\mathbf{V}_j$ for $j = 1, \dots, i$. Then $\dim \mathbf{V} = \sum_{j=1}^{i} \dim \mathbf{V}_j = \sum_{j=1}^{i} l_j$ and the $\dim \mathbf{V}$ vectors $\mathbf{v}_{11}, \dots, \mathbf{v}_{1l_1}, \dots, \mathbf{v}_{i1}, \dots, \mathbf{v}_{il_i}$ 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 $\mathbf{V}$ and the zero subspace $\{\mathbf{0}\}$, (which consist only of the zero element), are invariant subspaces. Those are called *trivial* invariant subspaces. $\mathbf{U}$ is called a *nontrivial* invariant subspace if $\mathbf{U}$ 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. (See Problem 1 in the end of this section.) Equivalently $\psi(z)$ is the monic polynomial of the minimal degree which annihilates T: $\psi(T) = \mathbf{0}$.

Theorem 2.19 *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator on a finite dimensional space $\dim \mathbf{V} > 0$. 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 $k = 2$. So $\psi(z) = \psi_1(z)\psi_2(z)$. Let $\mathbf{V}_1 := \psi_2(T)\mathbf{V}, \mathbf{V}_2 = \psi_1(T)\mathbf{V}$ be the ranges 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 that $\psi_2(T) = 0$. By Theorem 2.16 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 nonzero 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 exists 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 $\mathbf{v} \in \mathbf{V}$. Then $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, 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 $i = 1, 2$. So $\bar{\mathbf{V}}_i$ 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 $\dim(\bar{V}_i \cap \mathbf{V}_j) \geq 0$. As before we conclude that $\mathbf{U} := \bar{V}_i \cap \mathbf{V}_j$ 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 $\hat{\psi}_2 := \psi_2 \dots \psi_k$. 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 to $T|_{\hat{V}_2}$ to deduce the theorem. □

Problems

1. Let $A, B \in \mathbb{F}^{n \times n}$ and $p(z) \in \mathbb{F}[z]$. Show
   (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 $\lambda_1, \dots, \lambda_k \in \mathbb{F}$ be k distinct elements. Let

$$A = \oplus_{i=j=1}^{k,l_i} J_{m_{ij}}(\lambda_i), \text{ where } m_i = m_{i1} \geq \dots \geq m_{il_i} \geq 1, \text{ for } i = 1, \dots, k. \quad (2.9)$$

   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 diagonable 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 diagonable is "much smaller" then the variety of matrices with multiple eigenvalue.

   This fact holds for any $n \times n$ matrices $\mathbb{R}^{n \times n}$ or $\mathbb{C}^{n \times n}$.

5. **Programming Problem**

   Spectrum and pseudo spectrum: Let $A = (a_{ij})_{i,j=1}^n \in \mathbb{C}^{n \times n}$. Then $\det(zI_n - A) = (z - \lambda_1) \dots (z - \lambda_n)$ and the *spectrum* of A is given as spec $A := \{\lambda_1, \dots, \lambda_n\}$. 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: $a_1.a_2 \dots a_8 \times 10^m$ for some integer m., 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). 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 student 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 $\cdot$ on the $X - Y$ axis, together with the plot of the spectrum of A. (Altogether you will have 4 graphs.)

   (b) Let $A := \text{diag}(0.1C, [-0.5])$, 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 3 above, and the second block is 1×1 matrix with the entry -0.5. Repeat 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 $1 \leq i \leq j$, 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?

2.4 Existence and uniqueness of the Jordan canonical form

Definition 2.20 *$A \in \mathbb{F}^{n\times n}$ or a linear transformation $T : \mathbf{V} \to \mathbf{V}$ is called nilpotent respectively, if $A^m = \mathbf{0}$ or $T^m = \mathbf{0}$. The minimal $m \geq 1$ for which $A^m = \mathbf{0}$ or $T^m = \mathbf{0}$ is called the index of nilpotency of A and T respectively, and denoted by* index A *or* index T *respectively.*

Assume that A or T are nilpotent, then the s-numbers are defined as

$$s_i(A) := \text{rank } A^{i-1} - 2\text{rank } A^i + \text{rank } A^{i+1},\ s_i(T) := \text{rank } T^{i-1} - 2\text{rank } T^i + \text{rank } T^{i+1},\ i = 1, \dots \tag{2.10}$$

Note that A or T are nilpotent with the index of nilpotency m if and only 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.

Proposition 2.21 *Let $T : \mathbf{V} \to \mathbf{V}$ be a nilpotent operator, with the index of nilpotency m, on the finite dimensional vector $\mathbf{V}$. Then*

$$\text{rank } T^i = \sum_{j=i+1}^{m} (j-i)s_j = (m-i)s_m + (m-i-1)s_{m-1} + \dots s_{i+1}, \quad i = 0, \dots, m-1. \tag{2.11}$$

Proof. Since $T^l = \mathbf{0}$ for $l \geq m$ it follows that $s_m(T) = \text{rank } T^{m-1}$ and $s_{m-1} = \text{rank } T^{m-2} - 2\text{rank } T^{m-1}$ if $m > 1$. This proves (2.11) for $i = m-1, m-2$. For other values of i (2.11) follows straightforward from (2.10) by induction on $m - i \geq 2$. □

Theorem 2.22 *Let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation on a finite dimensional space. Assume that T is nilpotent with the index of nilpotency m. Then $\mathbf{V}$ has a basis of the form*

$$\mathbf{x}_j, T\mathbf{x}_j, \dots T^{l_j-1}\mathbf{x}_j,\ j = 1, \dots, i,\ \text{where } l_1 = m \geq \dots \geq l_i \geq 1,\ \text{and } T^{l_j}\mathbf{x}_j = 0,\ j = 1, \dots, i. \tag{2.12}$$

More precisely, the number of l_j, which are equal to an integer $l \in [1, m]$, is equal to $s_l(T)$ given in (2.10).

Proof. Let $s_i := s_i(T), i = 1, \dots, m$ be given by (2.10). Since $T^l = \mathbf{0}$ for $l \geq m$ it follows that $s_m = \text{rank } T^{m-1} = \dim \text{range } T^{m-1}$. So $[\mathbf{y}_1, \dots, \mathbf{y}_{s_m}]$ is a basis for $T^{m-1}\mathbf{V}$. Clearly $\mathbf{y}_i = T^{m-1}\mathbf{x}_i$ for some $\mathbf{x}_1, \dots, \mathbf{x}_{s_m} \in \mathbf{V}$. We claim that the ms_m vectors

$$x_1, T\mathbf{x}_1, \dots, T^{m-1}\mathbf{x}_1, \dots, \mathbf{x}_{s_m}, T\mathbf{x}_{s_m}, \dots, T^{m-1}\mathbf{x}_{s_m} \tag{2.13}$$

are linearly independent. Suppose that there exists a linear combination of these vectors that is equal to $\mathbf{0}$:

$$\sum_{j=0}^{m-1} \sum_{k=1}^{s_m} \alpha_{jk} T^j \mathbf{x}_k = \mathbf{0}. \tag{2.14}$$

Multiply this equality by T^{m-1}. Thus we obtain $\sum_{j=0}^{m-1} \sum_{k=1}^{s_m} \alpha_{jk} T^{m-1+j}\mathbf{x}_k = \mathbf{0}$. Recall that $T^l = \mathbf{0}$ for any $l \geq m$. Hence this equality reduces to $\sum_{k=1}^{s_m} \alpha_{0k} T^{m-1}\mathbf{x}_k = \mathbf{0}$. Since $T^{m-1}\mathbf{x}_1, \dots, T^{m-1}\mathbf{x}_{s_m}$ form a basis in $T^{m-1}\mathbf{V}$ it follows that $\alpha_{0k} = 0$ for $k = 1, \dots, s_m$. If $m = 1$ we deduce that the vectors in (2.13) are linearly independent. Assume that $m > 1$. Suppose that we already proved that $\alpha_{jk} = 0$ for $k = 1, \dots, s_m$ and $j = 0, \dots, l-1$, where $1 \leq l \leq m-1$. Hence in (2.14) we can assume that the summation on j starts from $j = l$. Multiply (2.14) by T^{m-l+1} and use the above arguments to deduce that $\alpha_{lk} = 0$ for $k = 1, \dots, s_m$. Use this argument iteratively for $l = 1, \dots, m-1$ to deduce the linear independence of the vectors in (2.13).

Note that for $m = 1$ we proved the theorem. Assume that $m > 1$. Let $p \in [1, m]$ be an integer. We claim that the vectors

$$x_j, T\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j, \quad \text{for all } j \text{ such that } l_j \geq p \tag{2.15}$$

are linearly independent and satisfy the condition $T^{l_j}\mathbf{x}_j = \mathbf{0}$ for all $l_j \geq p$. Moreover, the vectors

$$T^{p-1}\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j, \quad \text{for all } j \text{ such that } l_j \geq p \tag{2.16}$$

is a basis for range T^{p-1}. Furthermore for each integer $l \in [p, m]$ the number of l_j, which are equal to l, is equal to $s_l(T)$.

We prove this claim by the induction on $m-p+1$. For $p = m$ our previous argument give this claim. Assume that the claim holds for $p = q \geq m$ and let $p = q - 1$. By the induction assumption the vectors in (2.15) are linearly independent for $l_j \geq q$. Hence that vectors $T^{q-2}\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j$ for all $l_j \geq q$ are linearly independent. Use the induction assumption that the number of $l_j = l \in [q, m]$ is equal to $s_l(T)$ to deduce that the number of this vectors is equal to $t_{q-2} := (m-q+2)s_m + (m-q+1)s_{m-1} + \ldots + 2s_q$. Also the number of $l_j \geq q$ is $L_q = s_m + s_{m-1} + \ldots s_q$. Use the formula for rank T^{q-2} in (2.11) to deduce that rank $T^{q-2} - t_{q-2} = s_{q-1}$.

Suppose first that $s_{q-1} = 0$. Hence the vectors $T^{q-2}\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j$ for all $l_j \geq q$ form a basis in range T^{q-2}. In this case we assume that there is no l_j that is equal to $q - 1$. This concludes the proof of the induction step and the proof of the theorem in this case.

Assume now that $s_{q-1} > 0$. Then there exist vectors $\mathbf{z}_1, \ldots, \mathbf{z}_{s_{q-1}}$ that together with the vectors $T^{q-2}\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j$ for all $l_j \geq q$ form a basis in $T^{q-2}\mathbf{V}$. Let $\mathbf{z}_k = T^{q-2}\mathbf{u}_k, k = 1, \ldots, s_{q-1}$. Observe next that by induction hypothesis the vectors given in (2.16) form a basis in range T^{p-1} for $p = q$. Hence $T^{q-1}\mathbf{u}_k = \sum_{j:l_j \geq q} \sum_{r=q-1}^{l_j-1} \beta_{k,r,j} T^r \mathbf{x}_j$. Let $\mathbf{v}_k := \mathbf{u}_k - \sum_{j:l_j \geq q} \sum_{r=q-1}^{l_j-1} \beta_{k,r,j} T^{r-q+1}\mathbf{x}_j$. Clearly $T^{q-1}\mathbf{v}_k = 0$ for $k = 1, \ldots, s_{q-1}$. Also $T^{q-2}\mathbf{v}_k = \mathbf{z}_k - \sum_{j:l_j \geq q} \sum_{r=q-1}^{l_j-1} \beta_{k,r,j} T^{r-1}\mathbf{x}_j$. Hence $T^{q-2}\mathbf{v}_1, \ldots, T^{q-2}\mathbf{v}_{s_{q-1}}$ and the vectors $T^{q-2}\mathbf{x}_j, \ldots, T^{l_j-1}\mathbf{x}_j$ for all $l_j \geq q$ form a basis in $T^{q-2}\mathbf{V}$. From the above definition of L_q $l_j \geq q$ if and only if $j = [1, L_q]$. Let $\mathbf{x}_j = \mathbf{v}_{j-L_q}$ and $l_j = s_{q-1}$ for $j = L_q + 1, \ldots, L_{q-1} := L_q + s_{q-1}$.

It is left to show that the vectors given in (2.15) are linearly independent for $p = q - 1$. This is done as in the beginning of the proof of the theorem. (Assume that a linear combination of these vectors is equal to $\mathbf{0}$. Then apply T^{q-2} and use the fact that $T^{l_j}\mathbf{x}_j = \mathbf{0}$ for $j = 1, \ldots, L_{q-1}$. Then continue as in the beginning of the proof of this theorem.) This concludes the proof of this theorem by induction. □

Corollary 2.23 *Let T satisfies the assumption of Theorem 2.22 hold. Denote $\mathbf{V}_j := \text{span}\,(T^{l_j-1}\mathbf{x}_j, \ldots, T\mathbf{x}_j, \mathbf{x}_j)$ 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 $2 \leq r, p_r \in [1, m-1], s_{p_r}(T) > 0$ and $k_{r-1} = \sum_{j=1}^{m-p_r} s_{m-j+1}(T)$.*

Definition 2.24 *$T : \mathbf{V} \to \mathbf{V}$ be a nilpotent operator. Then the sequence $(l_1, \ldots, l_i)$ defined in Theorem 2.22, 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 2.11 (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 $p(z) = \prod_{j=1}^{k}(z-\lambda_j)^{n_j}$. Here $\lambda_1, \dots, \lambda_k$ 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 2.16 $\psi(z)|p(z)$. Problem 1(c) of §2.3 we deduce that $\psi(\lambda_j) = 0$ for $j = 1, \dots, k$. Hence

$$\det(zI_n - A) = \prod_{j=1}^{k}(z-\lambda_j)^{n_j},\ \psi(z) = \prod_{j=1}^{k}(z-\lambda_j)^{m_j},\ 1 \leq m_j \leq n_j,\ \lambda_j \neq \lambda_i \text{ for } j \neq i,\ i,j = 1, \dots, k. \tag{2.17}$$

Let $\psi_j := (z - \lambda_j)^{m_j}$ for $j = 1, \dots, k$. 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 2.19 to obtain teh 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 $i = 1, \dots, k$. That is $T_i - \lambda_i I_i$, where I_i is the identity operator, i.e. $I_i\mathbf{v} = v$ 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 2.22 and Corollary 2.23 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 $j = 1, \dots, q_j$. According to Corollary 2.23

$$m_i = m_{i1} \geq \dots m_{iq_i} \geq 1, \quad i = 1, \dots, k. \tag{2.18}$$

Furthermore, the above sequence is completely determined by rank $Q_i^j, j = 0, 1, \dots$ for $i = 1, \dots, k$. 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 $j = 1, \dots, q_i$ and $i = 1, \dots, k$. 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 Problems 1,2 in §2.3.) Thus $m_{i1} = m_i$ for $i = 1, \dots, k$. Theorem 2.19 yields that the subspaces $\mathbf{V}_1, \dots, \mathbf{V}_k$ are uniquely determined by ψ. So each T_i and $Q_i = T - \lambda_i I_i$ are uniquely determined. Theorem 2.22 yields that rank $Q_i^j, j = 0, 1, \dots$ determines the sizes of the Jordan blocks of Q_i. Hence all the Jordan blocks corresponding to λ_i are uniquely determined for each $i \in [1, k]$. □

Problems

1. Let $T : \mathbf{V} \to \mathbf{V}$ be nilpotent with $m = \text{index } T$. Let $(\omega_1, \dots, \omega_m)$ be the Weyr characteristic. Show that rank $T^j = \sum_{p=1}^{j} \omega_j$ for $j = 1, \dots, m$.

2. Let $A \in \mathbb{F}^{n\times n}$. Denote by $p(z)$ and $\psi(z)$ its characteristic and minimal polynomials, by adj $(zI_n - A) \in \mathbb{F}[z]^{n\times n}$ the adjoint matrix if $zI_n - A$, $q(z)$ the g.c.d., (the greatest common divisor) of the entries of $zI_n - A$, which is the g.c.d of all $(n-1)\times(n-1)$ minors of $(zI_n - A)$. ($q(z)$ is a assumed to be a monic polynomial in $\mathbb{F}[z]$.) The aim of this problem to demonstrate the equality $\psi(z) = \frac{p(z)}{q(z)}$.

   (a) Show that $q(z)$ divides $p(z)$. (*Hint:* Expand $\det(zI_n - A)$ be the first row.) Let $\phi(z) := \frac{p(z)}{q(z)}$

   (b) Show that adj $(zI_n - A) = q(z)C(z)$ for some $C(z) \in \mathbb{F}[z]^{n\times n}$. Show that $\phi(z)I_n = C(z)(zI_n - A)$. (Recall the proof of Theorem 2.14 that $p(z)I_n =$ adj $(zI_n - A)(zI_n - A)$.) Show that $\phi(A) = \mathbf{0}$. Conclude that $\psi(z)|\phi(z)$.

   (c) Let $\theta(z) := \frac{p(z)}{\psi(z)}$. Show that $\theta(z) \in \mathbb{F}[z]$.

(d) Show that $\psi(z)I_n = D(z)(zI_n - A)$ for some $D(z) \in \mathbb{F}[z]^{n\times n}$. Conclude that $D(z) = \frac{1}{\theta(z)}\text{adj}\,(zI_n - A)$. Conclude that $\theta(z)|q(z)$. Show that $\phi(z) = \psi(z)$.

3. Let $A \in \mathbb{C}^{n\times n}$. Show that A is diagonable if and only if all the zeros of the minimal polynomial ψ of A are simple, i.e. ψ does not have multiple roots.

4. 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

$$s_i(A,\lambda_j) := \text{rank}\,(A-\lambda_j I_n)^{i-1} - 2\text{rank}\,(A-\lambda_j I_n)^{i} + \text{rank}\,(A-\lambda_j I_n)^{i+1}, \quad i = 1,\dots,n_j, j = 1,\dots,k. \tag{2.19}$$

(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 $i \in [1,n_j]$ such that $1s_1(A,\lambda_j) + 2s_2(A,\lambda_j)\dots + is_i(A,\lambda_j) = n_j$.

3 Applications of Jordan Canonical form

3.1 Functions of Matrices

Let $A \in \mathbb{C}^{n\times n}$. Consider the iterations

$$\mathbf{x}_l = A\mathbf{x}_{l-1},\ \mathbf{x}_{l-1} \in \mathbb{C}^n, \quad l = 1,\dots \tag{3.1}$$

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 to $10^8 n(2n-1)$ operations, if we simply program the iterations (3.1). 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. To do that we need to give the correct way to find directly A^{10^8}, or for that matter any $f(A)$, where $f(z)$ is either polynomial, or more complex functions as $e^z, \cos z, \sin z$, an entire function $f(z)$, or even more special functions.

Theorem 3.1 *Let $A \in \mathbb{C}^{n\times n}$ and*

$$\det(zI_n - A) = \prod_{i=1}^{k}(z-\lambda_i)^{n_i},\ \psi(z) = \prod_{i=1}^{k}(z-\lambda_i)^{m_i}, \tag{3.2}$$

$$1 \le m := \deg\psi = \sum_{i=1}^{k} m_i \le n = \sum_{i=1}^{k} n_i,\ 1 \le m_i \le n_i,\ \lambda_i \ne \lambda_j \text{ for } i \ne j,\ i,j = 1,\dots,k,$$

where $\psi(z)$ is the minimal polynomial of A. Then there exists unique m linearly independent matrices $Z_{ij} \in \mathbb{C}^{n\times n}$ for $i = 1,\dots,k$ and $j = 0,\dots,m_i - 1$, which depend on A, such that for any polynomial $f(z)$ the following identity holds

$$f(A) = \sum_{i=1}^{k}\sum_{j=0}^{m_i-1} \frac{f^{(j)}(\lambda_i)}{j!} Z_{ij}. \tag{3.3}$$

($Z_{ij}, i = 1,\dots,k, j = 1,\dots,m_i$ 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 $j \geq n$. 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} (3.3) holds for $K_l z^l$, where $K_l \in \mathbb{C}$ and $l = 0, 1, \ldots$. Hence (3.3) holds for any polynomial $f(z)$ for this choice of A.

Assume now that A is a direct sum of Jordan blocks as in (2.9): $A = \oplus_{i=j=1}^{k,l_i} J_{m_{ij}}(\lambda_i)$. Here $m_i = m_{i1} \geq \ldots \geq m_{il_i} \geq 1$ for $i = 1, \ldots, k$, and $\lambda_i \neq \lambda_j$ for $i \neq j$. Thus (3.2) holds with $n_i = \sum_{j=1}^{l_i} m_{ij}$ for $i = 1, \ldots, k$. 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 $l \in [1, k]$ 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 $j = 0, \ldots, m_i - 1$. Note that each Z_{ij} is a nonzero matrix with $0-1$ entries. Furthermore, two different Z_{ij} and $Z_{i'j'}$ do not have a common 1 entry. Hence $Z_{ij}, i = 1, \ldots, k, j = 0, \ldots, m_i - 1$ are linearly independent. It is straightforward to deduce (3.3) 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 (3.3) holds. Clearly

$$f(B) = Uf(A)U^{-1} = \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \frac{f^{(j)}(\lambda_i)}{j!} U Z_{ij} U^{-1}.$$

Hence (3.3) holds for B, where $UZ_{ij}U^{-1}, i = 1, \ldots, k, j = 0, \ldots, m_{ij-1}$ are the B-components.

The uniqueness of the A-components follows from the existence and uniqueness of the Lagrange-Sylvester interpolation polynomial as explained below.

□

Theorem 3.2 *(**The Lagrange-Sylvester interpolation polynomial**). Let $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$ be k-distinct numbers. Let $m_1, \ldots, m_k$ be k positive integers and let $m = m_1 + \ldots + m_k$. Let $s_{ij}, i = 1, \ldots, k, j = 0, \ldots, m_i - 1$ 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 $i = 1, \ldots, k, j = 0, \ldots, m_i - 1$ satisfying the conditions. (For $m_i = 1, i = 1, \ldots, k$ ϕ is the Lagrange interpolating polynomial.)*

Proof. The Lagrange interpolating polynomial is given by the formula

$$\phi(z) = \sum_{i=1}^{k} \frac{(z-\lambda_1)\ldots(z-\lambda_{i-1})(z-\lambda_{i+1})\ldots(z-\lambda_k)}{(\lambda_i-\lambda_1)\ldots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\ldots(\lambda_i-\lambda_k)} s_{i0}.$$

In the general case one determines $\phi(z)$ as follows. Let $\psi(z) := \prod_{i=1}^{k}(z-\lambda_i)^{m_i}$. Then

$$\phi(z) = \psi(z) \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \frac{t_{ij}}{(z-\lambda_i)^{m_i-j}}.$$

Now start to determine t_{ij} recursively starting with any i and $j = 0$. Then it is straightforward to show that $t_{i0} = \psi_i(\lambda_i)$, where $\psi_i(z) = \frac{\psi(z)}{(z-\lambda_i)^{m_i}}$. Now find t_{i1} by taking the derivative of the above formula for $\phi(z)$ and let $z = \lambda_i$. Continue this process until all $t_{ij}, i = 1, \dots, k, j = 1, \dots, m_i - 1$ are determined. Note that $\deg \phi \le m - 1$.

The uniqueness 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 $\deg \omega(z) \le m - 1$ it follows that $\omega(z)$ is the zero polynomial, i.e. $\phi(z) = \theta(z)$. □

Proof of the uniqueness of A-components. Let $\phi_{ij}(z)$ be the Lagrange-Sylvester polynomial given by the data $s_{i'j'}, i' = 1, \dots, k, j' = 1, \dots, m_{i'} - 1$. Assume $s_{ij} = j!$ and all other $s_{i'j'} = 0$. Then (3.3) yields that $Z_{ij} = \phi_{ij}(A)$. □

Proposition 3.3 *Let $A \in \mathbb{C}^{n\times n}$. Assume that the minimal polynomial $\psi(z)$ be given by (3.2) and denote $m = \deg \psi$. Then for each integers $u, v \in [1, n]$ 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*

$$\sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \binom{l}{j} \lambda_i^{\max(l-j,0)} (Z_{ij})_{uv} = a_{uv}^{(l)}, \quad l = 0, \dots, m - 1. \tag{3.4}$$

(Note that $\binom{l}{j} = 0$ for $j > l$.)

Proof. Consider the equality (3.3) 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 (3.4). Thus the systems (3.4) 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 (3.4) for each $u, v \in [1, n]$. 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 (3.4) has a unique solution. □

The algorithm for finding the A-components and its complexity.

1. (a) Set $i = 1$.
   (b) Compute and store A^i. Check if $I_n, A, \dots, A^i$ are linearly independent. If independent, set $i = i + 1$ and go to (b).
   (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 (3.4).
2. 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 $m = n$ C_{n^2+1} has las 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 (3.4), use Gauss-Jordan for the augmented matrix $[S\ T]$. Here $S \in \mathbb{C}^{n\times n}$ stands for the coefficient of the system (3.4), depending on $\lambda_1, \dots, \lambda_k$. $T \in \mathbb{C}^{n\times n^2}$ given the "left-hand side" of n^2 systems of (3.4). 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.)

Problems

1. Let $A \in \mathbb{C}^{4\times 4}$ be given as in Problem 3 of Section 2.3. Assume that the characteristic polynomial of A is $z^2(z-1)^2$.
   (a) Use Problem 4 of Section 2.4 to find the Jordan canonical form of A.
   (b) Assume that the minimal polynomial of A is $z(z-1)^2$. Find all the A-components.
   (c) Give the explicit formula for any A^l.
2. Let $A \in \mathbb{C}^{n\times n}$ and assume that $\det(zI_n - A) = \prod_{i=1}^k (z-\lambda_i)^{n_i}$, 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 $Z_{ij}Z_{ip}$?

3.2 Power stability, convergence and boundedness of matrices

Corollary 3.4 *Let $A \in \mathbb{C}^{n\times n}$ Assume that the minimal polynomial $\psi(z)$ be given by (3.2) and denote by $Z_{ij}, i = 1, \dots, k, j = 0, \dots, m_j - 1$ the A-components. Then for each positive integer l*

$$A^l = \sum_{i=1}^{k} \sum_{j=0}^{m_i-1} \binom{l}{j} \lambda_i^{\max(l-j,0)} Z_{ij}. \tag{3.5}$$

If we know the A-components then to compute A^l we need only around $2mn^2 \leq 2n^3$ 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. See Problem ? However roundoff error remains a problem for large l.

Definition 3.5 *Let $A \in \mathbb{C}^{n\times n}$. A is called power stable if $\lim_{l\to\infty} A^l = \mathbf{0}$. 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 3.6 *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.

Proof. Consider the formula (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 (3.4), 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 $\binom{l}{j}\lambda_i^{l-j}, l = j, j+1, \dots,$ 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 $\binom{l}{j}\lambda_i^{l-j}, l = j, j+1, \dots,$ is bounded if and only if either $|\lambda_i| < 1$ or $|\lambda_i| = 1$ and $j = 0$. Hence we deduce the condition *3* of the theorem.

□

Corollary 3.7 *Let* $A \in \mathbb{C}^{n\times n}$ *and consider the iterations* $\mathbf{x}_l = A\mathbf{x}_{l-1}$ *for* $l = 1, \dots$. *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. $\mathbf{x}_l, l = 0, 1, \dots$ *converges if and only if* A *is power convergent.*

3. $\mathbf{x}_l, l = 0, 1, \dots$ *is bounded if and only if* A *is power bounded.*

Proof. If A satisfies the conditions of an item i Theorem 3.6 then the corresponding condition i of the corollary clearly holds. Assume that the conditions of an item i of the corollary holds. Choose $\mathbf{x}_0 = \mathbf{e}_j = (\delta_{1j}, \dots, \delta_{nj})^\top$ for $j = 1, \dots, n$ to deduce the corresponding condition i of Theorem 3.6. □

Theorem 3.8 *Let* $A \in \mathbb{C}^{n\times n}$ *and consider the nonhomogeneous iterations*

$$\mathbf{x}_l = A\mathbf{x}_{l-1} + \mathbf{b}_l, \quad l = 0, \dots \tag{3.6}$$

Then

1. $\lim_{l\to\infty} \mathbf{x}_l = \mathbf{0}$ *for any* $\mathbf{x}_0 \in \mathbb{C}^n$ *and any sequence* $\mathbf{b}_0, \mathbf{b}_1, \dots$ *satisfying the condition* $\lim_{l\to\infty} \mathbf{b}_l = 0$ *if and only if* A *is power stable.*

2. *The sequence* $\mathbf{x}_l, l = 0, 1, \dots$ *converges for any* $\mathbf{x}_0$ *and any sequence* $\mathbf{b}_0, \mathbf{b}_1, \dots$ *satisfying the condition* $\sum_{l=0}^{l} \mathbf{b}_l$ *converges.*

3. *The sequence* $\mathbf{x}_l, l = 0, 1, \dots$ *is bounded for any* $\mathbf{x}_0$ *and any sequence* $\mathbf{b}_0, \mathbf{b}_1, \dots$ *satisfying the condition* $\sum_{l=0}^{l} ||\mathbf{b}_l||_\infty$ *converges. (Here* $||(x_1, \dots, x_n)||_\infty = \max_{i\in[1,n]} |x_i|$.*)*

Proof. Assume that $\mathbf{b}_l = 0$. Since $\mathbf{x}_0$ is arbitrary we deduce the necessity of all the conditions from Theorem 3.6. The sufficiency of the above conditions follow from the Jordan Canonical Form of A as follows.

Let $J = U^{-1}AU$ where U is an invertible matrix and J is the Jordan canonical form of A. By 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 system (3.6) 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 $n = 1$ (3.6) reduces to

$$x_l = \lambda x_{l-1} + b_l, \quad x_0, x_l, b_l \in \mathbb{C} \text{ for } l = 1, \ldots \tag{3.7}$$

It is straightforward to show, e.g. use induction that

$$x_l = \sum_{i=0}^{l} \lambda^i b_{l-i} = b_l + \lambda b_{l-1} + \ldots + \lambda^l b_0 \quad l = 1, \ldots, \text{ were } b_0 := x_0. \tag{3.8}$$

Let $\beta_m = \sup_{i\geq m} |b_i|$. 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 $l > m$

$$|x_l| \leq \sum_{i=0}^{l} |\lambda|^i |b_{l-i}| = \sum_{i=0}^{l-m} |\lambda|^i |b_{l-i}| + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \leq$$

$$\beta_m \sum_{i=0}^{l-m} |\lambda|^i + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \leq \beta_m \sum_{i=0}^{\infty} |\lambda|^i + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| =$$

$$\frac{\beta_m}{1-|\lambda|} + |\lambda|^{l-m} \sum_{j=1}^{m} |\lambda|^j ||b_{m-j}| \to \frac{\beta_m}{1-|\lambda|} \text{ as } l \to \infty.$$

That is $\limsup_{l\to\infty} |x_l| \leq \frac{\beta_m}{1-|\lambda|}$. 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 (3.6) 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 (3.6) 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 $l = 1, ldots$. 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 $\lim_{l\to\infty} x_{1,l} = 0$. Hence $\lim_{l\to\infty} \mathbf{x}_l = 0$. The proof of this case is concluded.

2. Assume that A satisfies the 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 (3.7) with $\lambda = 1$. (3.8) 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 (3.8) yields that $|x_l| \leq \sum_{i=0}^{l} |b_i|$. The assumption that $\sum_{i=1}^{\infty} |\mathbf{b}_i|$ converges imply that $|x_l| \leq \sum_{i=0}^{\infty} |b_i| < \infty$.

□

Remark 3.9 *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. (If time permits we revisit these problems later on in the course.)

Problems

1. Consider the nonhomogeneous system $\mathbf{x}_l = A_l\mathbf{x}_{l-1},\ A_l \in \mathbb{C}^{n\times n},\quad l = 1, \ldots$. 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 $B := A_pA_{p-1}\ldots A_1$ 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 satisfied. First, B is power convergent, i.e. $\lim_{l\to\infty} B^l = C$. Second, $A_iC = C$ 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.

3.3 e^{At} and stability of certain systems of ODE

Recall that 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

$$e^{At} := I_n + At + \frac{A^2t^2}{2} + \frac{A^3t^3}{6} + \ldots = \sum_{l=0}^{\infty} \frac{A^lt^l}{l!}.$$

Hence e^{At} satisfies the matrix differential equation

$$\frac{e^{At}}{dt} = Ae^{At} = e^{At}A. \tag{3.9}$$

Also one has the standard identity $e^{At}e^{Au} = e^{A(t+u)}$ for any complex numbers t, u.

Proposition 3.10 *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 $\mathbf{x}(t) \in \mathbb{C}^n$, 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 $\mathbf{b}(t) \in \mathbb{C}^n$ be any continuous vector function on $\mathbb{R}$ and consider the nonhomogeneous system of n ordinary differential equations with the initial condition:*

$$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + \mathbf{b}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{3.10}$$

Then this system has a unique solution of the form

$$\mathbf{x}(t) = e^{A(t-t_0)}\mathbf{x}_0 + \int_{t_0}^{t} e^{A(t-u)}\mathbf{b}(u)du. \tag{3.11}$$

Proof. The uniqueness of the solution of (3.10) follows from the uniqueness of solutions to system of ODE (Ordinary Differential Equations). The first part of the proposition follows from (3.9). 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 (3.10) 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 have 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 (3.11). □

Note: The second term in the formula (3.11) can be considered as a perturbation term to the solution $\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t), \mathbf{x}(t_0) = \mathbf{x}_0$, i.e. to the system (3.10) with $\mathbf{b}(t) \equiv \mathbf{0}$.

Use (3.3) for e^{zt} and the observation that $\frac{d^j e^{zt}}{dz^j} = t^j e^{zt}, j = 0, 1, \ldots$ to deduce:

$$e^{At} = \sum_{j=1}^{k} \sum_{j=0}^{m_i - 1} \frac{t^j e^{\lambda_i t}}{j!} Z_{ij}. \tag{3.12}$$

We can substitute this expression for e^{At} in (3.11) to get a simple expression of the solution $\mathbf{x}(t)$ of (3.10).

Definition 3.11 *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 3.12 *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) *$\lambda = 2\pi l\sqrt{-1}$ 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 (3.12). 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 (3.12), 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 3.13 *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 3.14 *Let $A \in \mathbb{C}^{n\times n}$ and consider the system of differential equations (3.10). Then for any $\mathbf{x}_0 \in \mathbb{C}^n$*

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. The necessity of the conditions of the theorem follow from Corollary 3.13 by choosing $\mathbf{b}(t) \equiv \mathbf{0}$.

1. Suppose that A is stable. Then Corollary 3.13 yields that $\lim_{t\to\infty} e^{At}\mathbf{x}_0 = \mathbf{0}$. Thus show that $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$, it is enough to show that the second term in (3.11) tends to $\mathbf{0}$. Use (3.12) to show that it is enough to demonstrate that

$$\lim_{t\to\infty} \int_{t_0}^{t} (t-u)^j e^{\lambda(t-u)} g(u)du = 0, \text{ where } \Re\lambda < 0,$$

for any continuous $g(t) \in [t_0, \infty)$, such that $\lim g(t) = 0$. For $\epsilon > 0$ there exists $T = T(\epsilon)$ such that $|g(t)| \le \epsilon$ for $t \ge T(\epsilon)$. Let $t > T(\epsilon)$. Then

$$\begin{aligned}
|\int_{t_0}^{t} (t-u)^j e^{\lambda(t-u)} g(u)du| &= |\int_{t_0}^{T(\epsilon)} (t-u)^j e^{\lambda(t-u)} g(u)du + \int_{T(\epsilon)}^{t} (t-u)^j e^{\lambda(t-u)} g(u)du| \\
&\le |\int_{t_0}^{T(\epsilon)} (t-u)^j e^{\lambda(t-u)} g(u)du| + |\int_{T(\epsilon)}^{t} (t-u)^j e^{\lambda(t-u)} g(u)du| \\
&\le |\int_{t_0}^{T(\epsilon)} (t-u)^j e^{\lambda(t-u)} g(u)du| + \epsilon \int_{T(\epsilon)}^{t} (t-u)^j e^{\Re\lambda(t-u)} du.
\end{aligned}$$

Consider the first term in the last inequality. Since $\lim_{t\to\infty} t^j e^{\lambda t} = 0$ it follows that the first term converges to zero. The second term bounded by ϵK for $K := \int_0^{\infty} t^j e^{\Re\lambda t} dt$. Hence as $\epsilon \to 0$ we deduce that $\lim_{t\to\infty} \int_{t_0}^{t} (t-u)^j e^{\lambda(t-u)} g(u)du = 0$.

2. Using part *1* we deduce the result for any eigenvalue λ with $\Re\lambda < 0$. It is left to discuss the case $\lambda = 0$. We assume that the Jordan blocks of A correspond to $\lambda = 0$ are of length one. So the A-component corresponding to $\lambda = 0$ is Z_{10}. The corresponding term is

$$\dots$$

4 Inner product spaces

4.1 Inner product

Definition 4.1 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and let $\mathbf{V}$ be a vector space over $\mathbb{F}$. Then $\langle \cdot, \cdot \rangle : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ is called an inner product if the following conditions hold:*

(a) $\langle a\mathbf{x}+b\mathbf{y},\mathbf{z}\rangle = a\langle \mathbf{x},\mathbf{z}\rangle + b\langle \mathbf{y},\mathbf{z}\rangle$, for all $a,b\in\mathbb{F},\ \mathbf{x},\mathbf{y},\mathbf{z}\in\mathbf{V}$,

(br) for $\mathbb{F}=\mathbb{R}$ $\langle \mathbf{y},\mathbf{x}\rangle = \langle \mathbf{x},\mathbf{y}\rangle$, for all $\mathbf{x},\mathbf{y}\in\mathbf{V}$;

(bc) for $\mathbb{F}=\mathbb{C}$ $\langle \mathbf{y},\mathbf{x}\rangle = \overline{\langle \mathbf{x},\mathbf{y}\rangle}$, for all $\mathbf{x},\mathbf{y}\in\mathbf{V}$;

(c) $\langle \mathbf{x},\mathbf{x}\rangle > 0$ for all $\mathbf{x}\in\mathbf{V}\backslash\{\mathbf{0}\}$.

$||\mathbf{x}|| := \sqrt{\langle \mathbf{x},\mathbf{x}\rangle}$ *is called the norm (length) of* $\mathbf{x}\in\mathbf{V}$.

Other standard properties of inner products are mentioned in Problems 4.2-4.3. We will use the abbreviation IPS for inner product space. In this chapter we assume that $\mathbb{F}=\mathbb{R},\mathbb{C}$ unless stated otherwise.

Proposition 4.2 *Let* $\mathbf{V}$ *be a vector space over* $\mathbb{R}$. *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*

$$(a+\sqrt{-1}b)(\mathbf{x},\mathbf{y}) := a(\mathbf{x},\mathbf{y}) + b(-\mathbf{y},\mathbf{x}), \quad \text{for all } a,b\in\mathbb{R},\ \mathbf{x},\mathbf{y}\in\mathbf{V}.$$

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* $\mathbf{V}_c$*:*

$$\langle(\mathbf{x},\mathbf{y}),(\mathbf{u},\mathbf{v})\rangle = \langle \mathbf{x},\mathbf{u}\rangle + \langle \mathbf{y},\mathbf{v}\rangle + \sqrt{-1}(\langle \mathbf{y},\mathbf{u}\rangle - \langle \mathbf{x},\mathbf{v}\rangle),\ \mathbf{x},\mathbf{y},\mathbf{u},\mathbf{v}\in\mathbf{V}.$$

We leave the proof of this proposition to the reader (Problem 4.4).

Definition 4.3 *Let* $\mathbf{V}$ *be an IPS. Then*
(a) $\mathbf{x},\mathbf{y}\in\mathbf{V}$ *are called orthogonal if* $\langle \mathbf{x},\mathbf{y}\rangle = 0$.
(b) $S,T\subset\mathbf{V}$ *are called orthogonal if* $\langle \mathbf{x},\mathbf{y}\rangle = 0$ *for any* $\mathbf{x}\in S,\ \mathbf{y}\in T$.
(d) For any $S\subset\mathbf{V}$, $S^{\perp}\subset\mathbf{V}$ *is the maximal orthogonal set to* S.
(e) $\mathbf{x}_1,...,\mathbf{x}_m$ *is called an orthonormal set if*

$$\langle \mathbf{x}_i,\mathbf{x}_j\rangle = \delta_{ij}, \quad i,j=1,...,m.$$

(f) $\mathbf{x}_1,...,\mathbf{x}_n$ *is called an orthonormal basis if it is an orthonormal set which is a basis in* $\mathbf{V}$.

Definition 4.4 *(***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* ($m\geq 0$). *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* $||\mathbf{y}_{k+1}||^{-1}\mathbf{y}_{k+1}$.

Corollary 4.5 *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:*

$$\begin{aligned}
&\mathbf{y}_1 := \mathbf{x}_1,\ r_{11} := ||\mathbf{y}_1||,\ \mathbf{e}_1 := \frac{\mathbf{y}_1}{r_{11}},\\
&r_{ji} := \langle \mathbf{x}_i,\mathbf{e}_j\rangle,\ j=1,...,i-1,\\
&\mathbf{y}_i := \mathbf{x}_i - \sum_{j=1}^{i-1} r_{ji}\mathbf{e}_j,\ r_{ii} := ||\mathbf{y}_i||,\ \mathbf{e}_i := \frac{\mathbf{y}_i}{r_{ii}},\ i=2,...,n.
\end{aligned} \tag{4.1}$$

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 4.5 for the definition of* $\operatorname{dist}(\mathbf{x}_i,S_{i-1})$*.)*

Corollary 4.6 *Any (ordered) basis in a finite dimensional IPS* $\mathbf{V}$ *induces an orthonormal basis by the Gram-Schmidt algorithm.*

See Problem 4.5 for some known properties related to the above notions.

Remark 4.7 *It is known, e.g. [8] that the Gram-Schmidt process as described in (4.1) is numerically unstable. That is, there is a severe loss of orthogonality of* $\mathbf{y}_1, \dots$ *as we proceed to compute* $\mathbf{y}_i$*. In computations one uses either a modified GSP or Householder orthogonalization [8].*

Definition 4.8 *(***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 (*$m \geq 0$*). 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:*

- *Initialize* $j = 1$ *and* $p = m$*.*
- *If* $\mathbf{x}_j \neq \mathbf{0}$ *let* $\mathbf{e}_j := \frac{1}{\|\mathbf{x}_j\|}\mathbf{x}_j$*. If* $\mathbf{x}_j = \mathbf{0}$ *replace* p *by* $p-1$ *and* $\mathbf{x}_i$ *by* $\mathbf{x}_{i+1}$ *for* $i = j, \dots, p$*.*
- $\mathbf{p}_i := \langle \mathbf{x}_i, \mathbf{e}_j \rangle \mathbf{e}_j$ *and replace* $\mathbf{x}_i$ *by* $\mathbf{x}_i := \mathbf{x}_i - \mathbf{p}_i$ *for* $i = j+1, \dots, p$*.*
- *Let* $j = j + 1$ *and repeat the process.*

MGS algorithm is stable, needs mn^2 flops, which is more time consuming then GS algorithm.

Problems

(4.2)

Let $\mathbf{V}$ be an IPS over $\mathbb{F}$. Show

$$\begin{aligned}
&\langle \mathbf{0}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{0} \rangle = 0, \\
&\text{for } \mathbb{F} = \mathbb{R} \ \langle \mathbf{z}, a\mathbf{x} + b\mathbf{y} \rangle = a\langle \mathbf{z}, \mathbf{x} \rangle + b\langle \mathbf{z}, \mathbf{y} \rangle, \text{ for all } a, b \in \mathbb{R}, \ \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}, \\
&\text{for } \mathbb{F} = \mathbb{C} \ \langle \mathbf{z}, a\mathbf{x} + b\mathbf{y} \rangle = \bar{a}\langle \mathbf{z}, \mathbf{x} \rangle + \bar{b}\langle \mathbf{z}, \mathbf{y} \rangle, \text{ for all } a, b \in \mathbb{C}, \ \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}.
\end{aligned}$$

(4.3)

Let $\mathbf{V}$ be an IPS. Show
(a) $||a\mathbf{x}|| = |a| \ ||\mathbf{x}||$ for $a \in \mathbb{F}$ and $\mathbf{x} \in \mathbf{V}$.
(b) The Cauchy-Schwarz inequality:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq ||\mathbf{x}|| \ ||\mathbf{y}||,$$

and equality holds if and only if $\mathbf{x}, \mathbf{y}$ are linearly dependent (collinear).
(c) The triangle inequality

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

and equality holds if either $\mathbf{x} = 0$ or $\mathbf{y} = a\mathbf{x}$ for $a \in \mathbb{R}_+$.

(4.4)

Prove Proposition 4.2.

(4.5)

Let $\mathbf{V}$ be a finite dimensional IPS of dimension n. Assume that $S \subset \mathbf{V}$. Show
(a) 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

$$\mathbf{x} = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i. \tag{4.6}$$

Furthermore for any $\mathbf{x}, \mathbf{y} \in \mathbf{V}$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{e}_i \rangle \overline{\langle \mathbf{y}, \mathbf{e}_i \rangle}. \tag{4.7}$$

(c) Assume that S is a finite set. Let $\tilde{S}$ be the set obtained by the Gram-Schmidt process. Show that $\tilde{S} = \emptyset \iff \text{span} S = \{\mathbf{0}\}$. Show that if $\tilde{S} \neq \emptyset$ then $\mathbf{e}_1, ..., \mathbf{e}_p$ is an orthonormal basis in span S.
(d) There exists an orthonormal basis $\mathbf{e}_1, ..., \mathbf{e}_n$ in $\mathbf{V}$ and $0 \leq m \leq n$ 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. 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

$$\begin{aligned} &\text{dist}(\mathbf{x}, S) := ||\mathbf{x} - P\mathbf{x}|| \leq ||\mathbf{x} - \mathbf{w}|| \text{ for any } \mathbf{w} \in S \\ &\text{and equality } \iff \mathbf{w} = P\mathbf{x}. \end{aligned} \tag{4.8}$$

(h) 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.
(i) 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**: $P\mathbf{x}$ is called *the least square approximation* to $\mathbf{x}$ in the subspace S.)

(4.9)

Let $X \in \mathbb{C}^{m \times n}$ and assume that $m \geq n$ 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 (4.5) 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}, $j \leq i$ given by (4.1). 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.

(4.10)

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 $\mathbb{R}^{2n}$. Let $C = A + \sqrt{-1}B$, $A, B \in \mathrm{M}_n(\mathbb{R})$.
a. Then $\hat{C} := \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \in \mathrm{M}_{2n}(\mathbb{R})$ 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 $\bar{\lambda} I + H$.

4.2 Geometric interpretation of the determinant

Definition 4.9 *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*

$\mathrm{P}(\mathbf{x}_1,\dots,\mathbf{x}_m)$ *are adjacent, i.e. connected by an edge in* $\mathrm{P}(\mathbf{x}_1,\dots,\mathbf{x}_m)$*, if* $||(a_1,\dots,a_m)^\top - (b_1,\dots,b_m)^\top|| = 1$*, i.e. the* $0-1$ *coordinates of* $(a_1,\dots,a_m)^\top$ *and* $(b_1,\dots,b_m)^\top$ *differ only at one coordinate* k*, for some* $k \in [1,m]$.

Note that if $\mathbf{e}_1,\dots,\mathbf{e}_n$ is the standard basis in $\mathbb{R}^n$, i.e. $\mathbf{e}_i = (\delta_{1i},\dots,\delta_{ni})^\top, i = 1,\dots,n$, then $\mathrm{P}(\mathbf{e}_1,\dots,\mathbf{e}_m)$ is the m-dimensional unit cube, whose edges are parallel to $\mathbf{e}_1,\dots,\mathbf{e}_m$ and its center (of gravity) is $\frac{1}{2}(\underbrace{1,\dots,1}_{},0,\dots,0)^\top$, where 1 appears m times for $1 \le m \le n$.

For $m > n$ $\mathrm{P}(\mathbf{x}_1,\dots,\mathbf{x}_m)$ is "flattened" parallelepiped, since $\mathbf{x}_1,\dots,\mathbf{x}_m$ are always linearly dependent in $\mathbb{R}^n$ for $m > n$.

Proposition 4.10 *Let* $A \in \mathbb{R}^{n\times n}$ *and view* $A = [\mathbf{c}_1\ \mathbf{c}_2\dots\mathbf{c}_n]$ *as an ordered set of* n *vectors, (columns),* $\mathbf{c}_1,\dots,\mathbf{c}_n$*. Then* $|\det A|$ *is the* n*-dimensional volume of the parallelepiped* $\mathrm{P}(\mathbf{c}_1,\dots,\mathbf{c}_n)$*. If* $\mathbf{c}_1,\dots,\mathbf{c}_n$ *are linearly independent then the orientation in* $\mathbb{R}^n$ *induced by* $\mathbf{c}_1,\dots,\mathbf{c}_n$ *is the same as the orientation induced by* $\mathbf{e}_1,\dots,\mathbf{e}_n$ *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 $\mathbf{c}_1,\dots,\mathbf{c}_n$ are linearly dependent, then $\mathrm{P}(\mathbf{c}_1,\dots,\mathbf{c}_n)$ lies in a subspace of $\mathbb{R}^n$, i.e. some $n-1$ dimensional subspace, and hence the n-dimensional volume of $\mathrm{P}(\mathbf{c}_1,\dots,\mathbf{c}_n)$ is zero.

Assume now that $\det A \neq 0$, i.e. $\mathbf{c}_1,\dots,\mathbf{c}_n$ are linearly independent. Perform that Gram-Schmidt process 4.4. Then $A = QR$, where $Q = [\mathbf{e}_1\ \mathbf{e}_2\dots\mathbf{e}_n]$ is an orthogonal matrix and $R = (r_{ji}) \in \mathbb{R}^{n\times n}$ is an upper diagonal matrix. (See Problem 4.9.) 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 $|\det A| = \det R = r_{11}r_{22}\dots r_{nn}$. 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 $\mathbf{e}_1,\dots,\mathbf{e}_{i-1}$ for $i = 2,\dots,n$. (See Problem 4.5 parts (f-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}. Hence 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 $\mathbf{e}_1,\dots,\mathbf{e}_n$ be the new basis $-\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_n$ and the rotate the new basis $-\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_n$ to the basis consisting of the columns of an orthogonal matrix Q, where $\det Q = -1$. □

Theorem 4.11 ***(The Hadamard determinantal 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* $i \neq j$*, 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 if and only if $\mathbf{c}_i = \mathbf{0}$ for some i.

Assume now that $\det A \neq 0$ and perform the Gram-Schmidt process. From (4.1) it follows that $A = QR$ where Q is a unitary matrix, i.e. $Q^*Q = I_n$ and $R = (r_{ji}) \in \mathbb{C}^{n\times n}$ 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 4.5 and the proof of Proposition 4.10 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 4.5 to deduce that $||\mathbf{c}_i|| = \operatorname{dist}(\mathbf{c}_i, \operatorname{span}(\mathbf{c}_1, \dots, \mathbf{c}_{i-1}))$ 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. □

Problems

1. Let $A = (a_{ij})_{i,j} \in \mathbb{C}^{n \times n}$. Assume that $|a_{ij}| \leq K$ for all $i, j = 1, \dots, n$. Show that $|\det A| \leq K^n n^{\frac{n}{2}}$.
2. Let $A = (a_{ij})_{i,j=1}^n \in \mathbb{C}^{n \times n}$ such that $|a_{ij}| \leq 1$ for $i, j = 1, \dots, n$. Show that $|\det A| = n^{\frac{n}{2}}$ if and only if $A^*A = AA^* = nI_n$. In particular, if $|\det A| = n^{\frac{n}{2}}$ then $|a_{ij}| = 1$ for $i, j = 1, \dots, n$.
3. 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 $i, j = 1, \dots, n$ and $|\det A| = n^{\frac{n}{2}}$.
4. Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ and assume that $a_{ij} = \pm 1, i, j = 1, \dots, n$. Show that if $n > 2$ then the assumption that $|\det A| = n^{\frac{n}{2}}$ yields that n is divisible by 4.
5. 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 $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ such that $a_{ij} = \pm 1$ for $i, j = 1, \dots, n$ and $|\det A| = n^{\frac{n}{2}}$ is called a *Hadamard* matrix. *It is conjectured* that for each n divisible by 4 there exists a Hadamard matrix.

4.3 Special transformations in IPS

Proposition 4.12 *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* $< T\mathbf{x}, \mathbf{y} >=< \mathbf{x}, T^*\mathbf{y} >$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$.

See Problems 4.3-4.4.

Definition 4.13 *Let* $\mathbf{V}$ *be an IPS and let* $T : \mathbf{V} \to \mathbf{V}$ *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* $T^*T = TT^*$.
Denote by $\mathbf{S}(\mathbf{V})$, $\mathbf{AS}(\mathbf{V})$, $\mathbf{U}(\mathbf{V})$, $\mathbf{N}(\mathbf{V})$ *the sets of self-adjoint, anti self-adjoint, unitary and normal operators on* $\mathbf{V}$ *respectively.*

Proposition 4.14 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$ *with an orthonormal basis* $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$. *Let* $T : \mathbf{V} \to \mathbf{V}$ *be a linear transformation. Let* $A = (a_{ij}) \in \mathbb{F}^{n \times n}$ *be the representation matrix of* T *in the basis* E*:*

$$a_{ij} =< T\mathbf{e}_j, \mathbf{e}_i >, \quad i, j = 1, ..., n. \tag{4.1}$$

Then for $\mathbb{F} = \mathbb{R}$*:*

(a) T^* is represented by $A^\top$,
(b) T is selfadjoint $\iff$ $A = A^\top$,
(c) T is anti selfadjoint $\iff$ $A = -A^\top$,
(d) T is unitary $\iff$ A is orthogonal $\iff$ $AA^\top = A^\top A = I$,
(e) T is normal $\iff$ A is normal $\iff$ $AA^\top = A^\top A$,

and for $\mathbb{F} = \mathbb{C}$*:*

(a) T^* is represented by A^* $(:= \bar{A}^\top)$,
(b) T is selfadjoint $\iff$ A is hermitian $\iff$ $A = A^*$,
(c) T is anti selfadjoint $\iff$ A is anti hermitian $\iff$ $A = -A^*$,
(d) T is unitary $\iff$ A is unitary $\iff$ $AA^* = A^*A = I$,
(e) T is normal $\iff$ A is normal $\iff$ $AA^* = A^*A$.

See Problem 4.5.

Proposition 4.15 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{R}$*, and let* $T \in \mathrm{Hom}(\mathbf{V})$*. Let* $\mathbf{V}_c$ *be the complexification of* $\mathbf{V}$*. Show that there exists a unique* $T_c \in \mathrm{Hom}(\mathbf{V}_c)$ *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.*

See Problem 4.6

Definition 4.16 *For a domain* $\mathcal{D}$ *with identity* 1 *let*

$$\begin{aligned}
&\mathbf{S}(n,\mathcal{D}) := \{A \in \mathcal{D}^{n\times n} : \quad A = A^\top\},\\
&\mathbf{AS}(n,\mathcal{D}) := \{A \in \mathrm{M}_n(\mathcal{D}) : \quad A = -A^\top\},\\
&\mathbf{O}(n,\mathcal{D}) := \{A \in \mathcal{D}^{n\times n} : \quad AA^\top = A^\top A = I\},\\
&\mathbf{SO}(n,\mathcal{D}) := \{A \in \mathbf{O}(n,\mathcal{D}) : \quad \det A = 1\},\\
&\mathbf{DO}(n,\mathcal{D}) := \mathbf{D}(n,\mathcal{D}) \cap \mathbf{O}(n,\mathcal{D}),\\
&\mathbf{N}(n,\mathbb{R}) := \{A \in \mathbb{R}^{n\times n} : \quad AA^\top = A^\top A\},\\
&\mathbf{N}(n,\mathbb{C}) := \{A \in \mathbb{C}^{n\times n} : \quad AA^* = A^*A\},\\
&\mathbf{H}_n := \{A \in \mathrm{M}_n(\mathbb{C}) : \quad A = A^*\},\\
&\mathbf{AH}_n := \{A \in \mathbb{C}^{n\times n} : \quad A = -A^*\},\\
&\mathbf{U}_n := \{A \in \mathbb{C}^{n\times n} : \quad AA^* = A^*A = I\},\\
&\mathbf{SU}_n := \{A \in \mathbf{U}_n : \quad \det A = 1\},\\
&\mathbf{DU}_n := \mathbf{D}(n,\mathbb{C}) \cap \mathbf{U}_n.
\end{aligned}$$

See Problem 4.7 for relations between these classes.

Theorem 4.17 *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 consiting 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

$$T(T^*\mathbf{x}) = (TT^*)\mathbf{x} = (T^*T)\mathbf{x} = T^*(T\mathbf{x}) = \lambda_1 T^*\mathbf{x} \Rightarrow T^*\mathbf{V}_1 \subset \mathbf{V}_1.$$

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 4.4). Hence for $T|\mathbf{V}_1$ the theorem trivially holds. For $T|\mathbf{V}_1^\perp$ the theorem follows by induction. □

The proof of Theorem 4.17 yields:

Corollary 4.18 *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 consiting of eigenvectors of* T*.*

Proposition 4.19 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{C}$*. Let* $T \in \mathbf{N}(\mathbf{V})$*. Then*

$$T \text{ is self} - \text{adjoint} \iff \text{spec}\,(T) \subset \mathbb{R},$$
$$T \text{ is unitary} \iff \text{spec}\,(T) \subset S^1 = \{z \in \mathbb{C} : \quad |z| = 1\}.$$

Proof. Since T is normal there exists an orthonormal basis $\mathbf{e}_1, ..., \mathbf{e}_n$ such that $T\mathbf{e}_i = \lambda_i \mathbf{e}_i,\ i = 1, ..., n$. Hence $T^*\mathbf{e}_i = \bar{\lambda}_i \mathbf{e}_i$. Then

$$T = T^* \iff \lambda_i = \bar{\lambda}_i,\ i = 1, ..., n,$$
$$TT^* = T^*T = I \iff |\lambda_i| = 1,\ i = 1, ..., n.$$

□

Combine Proposition 4.15 and Corollary 4.18 with the above proposition to deduce:

Corollary 4.20 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{R}$ *and let* $T \in \mathbf{S}(\mathbf{V})$*. Then* spec $(T) \subset \mathbb{R}$ *and* $\mathbf{V}$ *has an orthonormal basis consisting of the eigenvectors of* T*.*

Proposition 4.21 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{R}$ *and let* $T \in \mathbf{U}(\mathbf{V})$*. Then* $\mathbf{V} = \oplus_{i \in \{-1,1,2,...,k\}} \mathbf{V}_i$*, 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) $T|\mathbf{V}_i = \mathbf{V}_i,\ \dim \mathbf{V}_i = 2,\ \text{spec}\,(T|\mathbf{V}_i) \subset S^1 \backslash \{-1, 1\}$ *for* $i = 2, ..., k$*.*

See Problem 4.9.

Proposition 4.22 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{R}$ *and let* $T \in \mathbf{AS}(\mathbf{V})$*. Then* $\mathbf{V} = \oplus_{i \in \{1,2,...,k\}} \mathbf{V}_i$*, where* $k \geq 1$*,* $\mathbf{V}_i$ *and* $\mathbf{V}_j$ *are orthogonal for* $i \neq j$*, such that*
(a) $T|\mathbf{V}_1 = 0_{\mathbf{V}_1}\ \dim \mathbf{V}_1 \geq 0$,
(b) $T|\mathbf{V}_i = \mathbf{V}_i,\ \dim \mathbf{V}_i = 2,\ \text{spec}\,(T|\mathbf{V}_i) \subset \sqrt{-1}\mathbb{R} \backslash \{0\}$ *for* $i = 2, ..., k$*.*

See Problem 4.10.

Theorem 4.23 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{C}$ *of dimension* n*. Let* $T \in \text{Hom}(\mathbf{V})$*. (Here* $Hom(\mathbf{V})$ *stands for the algebra of all linear transformations from* $\mathbf{V}$ *to itself.) Let* $\lambda_1, ..., \lambda_n \in \mathbb{C}$ *be* n *eigenvalues of* T *counted with their multiplicities. Then there exists a unitary basis* $\mathbf{g}_1, ..., \mathbf{g}_n$ *of* $\mathbf{V}$ *with the following properties:*

$$T\text{span}(\mathbf{g}_1, ..., \mathbf{g}_i) \subset \text{span}(\mathbf{g}_1, ..., \mathbf{g}_i),\ \langle T\mathbf{g}_i, \mathbf{g}_i \rangle = \lambda_i,\ i = 1, ..., n. \tag{4.2}$$

Let $\mathbf{V}$ *be an IPS over* $\mathbb{R}$ *of dimension* n*. Let* $T \in \text{Hom}(\mathbf{V})$ *and assume that* spec $(T) \subset \mathbb{R}$*. Let* $\lambda_1, ..., \lambda_n \in \mathbb{R}$ *be* n *eigenvalues of* T *counted with their multiplicities. Then there exists an orthonormal basis* $\mathbf{g}_1, ..., \mathbf{g}_n$ *of* $\mathbf{V}$ *such that (4.2) 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 $\lambda_1 \in$ 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 $\mathbf{U} := \text{span}(\mathbf{g}_1)^\perp$. Let P be the orthogonal projection on $\mathbf{U}$. Let $T_1 := PT|_{\mathbf{U}}$. Then $T_1 \in \text{Hom}(\mathbf{U})$. 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

$$T_1\text{span}(\mathbf{g}_2, ..., \mathbf{g}_i) \subset \text{span}(\mathbf{g}_2, ..., \mathbf{g}_i),\ \langle T\mathbf{g}_i, \mathbf{g}_i \rangle = \tilde{\lambda}_i,\ i = 1, ..., n.$$

It is straightforward to show that $T\text{span}(\mathbf{g}_1, ..., \mathbf{g}_i) \subset \text{span}(\mathbf{g}_1, ..., \mathbf{g}_i)$ 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 $b_{ii} = \tilde{\lambda}_i,\ i = 2, ..., n$. Hence $\lambda_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_n$ 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 4.8 and 4.15 to deduce:

Corollary 4.24 *Let $A \in \mathbb{C}^{n\times n}$. Let $\lambda_1, ..., \lambda_n \in \mathbb{C}$ be n eigenvalues of A counted with their multiplicities. Then there exist an upper triangular matrix $B = (b_{ij})_1^n \in \mathrm{M}_n(\mathbb{C})$, 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 $A \in \mathrm{M}_n(\mathbb{R})$ and assume that 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* 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 4.14).

Definition 4.25 *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 $\mathbf{W} = \text{span}(\mathbf{v}, T\mathbf{v}, T^2\mathbf{v}, \ldots)$ 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 4.26 *Let $\mathbf{V}$ be a finite dimensional IPS. Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator. For $0 \neq \mathbf{v} \in \mathbf{V}$ let $\mathbf{W} = \text{span}(\mathbf{v}, T\mathbf{v}, ..., T^{r-1}\mathbf{v})$ 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 $[\mathbf{v}, T\mathbf{V}, ..., T^{r-1}\mathbf{v}]$ of $\mathbf{W}$. Then $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = 0$ for $1 \leq i \leq j - 2$, i.e. the representation matrix of $T|\mathbf{W}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ is upper Hessenberg. If T is self-adjoint then the representation matrix of $T|\mathbf{W}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ is a tridiagonal hermitian matrix.*

Proof. Let $\mathbf{W}_j = \text{span}(\mathbf{v}, \ldots, T^{j-1}\mathbf{v})$ 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 $[\mathbf{v}, \ldots, T^{r-1}\mathbf{v}]$ is a basis for $\mathbf{W}$. Recall that $\text{span}(\mathbf{u}_1, ..., \mathbf{u}_j) = \mathbf{W}_j$ for $j = 1, \ldots, r$. Let $r \geq j \geq i + 2$. 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 $r \geq i \geq j + 2$. 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 $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ is a tridiagonal hermitian matrix. □

Problems

(4.3)

Prove Proposition 4.12.

(4.4)

Let $P, Q \in \text{Hom}(\mathbf{V}), \mathbf{a}, b \in \mathbb{F}$. Show that $(aP + bQ)^* = \bar{a}P^* + \bar{b}Q^*$.

(4.5)

Prove Proposition 4.14.

(4.6)

Prove Proposition 4.15 for finite dimensional $\mathbf{V}$. (*Hint*: Choose an orthonormal basis in $\mathbf{V}$.)

(4.7)

Show the following

$$\begin{aligned}
&\mathbf{SO}(n,\mathcal{D}) \subset \mathbf{O}(n,\mathcal{D}) \subset \mathrm{GL}(n,\mathcal{D}),\\
&\mathbf{S}(n,\mathbb{R}) \subset \mathbf{H}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{AS}(n,\mathbb{R}) \subset \mathbf{AH}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{S}(n,\mathbb{R}), \mathbf{AS}(n,\mathbb{R}) \subset \mathbf{N}(n,\mathbb{R}) \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{O}(n,\mathbb{R}) \subset \mathbf{U}_n \subset \mathbf{N}(n,\mathbb{C}),\\
&\mathbf{SO}(n,\mathcal{D}),\ \mathbf{O}(n,\mathcal{D}),\ \mathbf{SU}_n,\ \mathbf{U}_n \ \text{ are groups}\\
&\mathbf{S}(n,\mathcal{D}) \ \text{ is a } \mathcal{D}-\text{module of dimension } \binom{n+1}{2},\\
&\mathbf{AS}(n,\mathcal{D}) \ \text{ is a } \mathcal{D}-\text{module of dimension } \binom{n}{2},\\
&\mathbf{H}_n \ \text{ is an } \mathbb{R}-\text{ vector space of dimension } n^2.\\
&\mathbf{AH}_n = \sqrt{-1}\,\mathbf{H}_n
\end{aligned} \tag{4.8}$$

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 tranfer matrix either from E to G or from G to E is a unitary matrix.

(4.9)

Prove Proposition 4.21

(4.10)

Prove Proposition 4.22

(4.11)

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 any $A \in \mathbf{SO}(2,\mathbb{R})$ is e^B for some $B \in \mathbf{AS}(2,\mathbb{R})$. (*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 4.21 and 4.22 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 $n(>1)$-dimensional IPS over $\mathbb{F} = \mathbb{R}$. Let $p \in \langle n-1\rangle$. Assume that $\mathbf{x}_1, ..., \mathbf{x}_p$ and $\mathbf{y}_1, ..., \mathbf{y}_p$ be two orthonormal systems in $\mathbf{V}$. Show that these two o.n.s. 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 o.n.s. and $\mathbf{z}_i(0) = \mathbf{x}_i, \mathbf{z}_i(1) = \mathbf{y}_i, i = 1, ..., p$.

(4.12)

a. Show that if Q is 3×3 orthogonal matrix with $\det Q = 1$ then 1 is an eigenvalue of Q.
b. Let Q be 3×3 orthogonal matrix with $\det Q = 1$. Show that there exists $\mathbf{e} \in \mathbb{R}^3, ||\mathbf{e}|| = 1$ and $\theta \in [0, 2\pi)$ such that for each $\mathbf{x} \in \mathbb{R}^3$ the vector $Q\mathbf{x}$ can be obtained as follows. Decompose $\mathbf{x} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} = \langle \mathbf{x}, \mathbf{e}\rangle \mathbf{e}$ and $\langle \mathbf{v}, \mathbf{e}\rangle = 0$. Let $S := \mathrm{span}(\mathbf{e})^\perp$ be the two dimensional subspace orthogonal to $\mathbf{e}$. Then $Q\mathbf{x} = \mathbf{u} + \mathbf{w}$, where $\mathbf{w} \in S$ is obtained by rotating $\mathbf{v} \in S$ by an angle θ. (This is result is called *Euler's* theorem, i.e. a rotation of a three dimensional body around its center of gravity can be obtained as a two dimensional rotation along some axis, (given by the direction of $\mathbf{e}$).)
c. For which values of n any $n \times n$ orthogonal matrix Q has an eigenvalue 1 or -1? Can you tell under what condition -1 is always an eigenvalue of Q?

(4.13)

a. Show that $\mathbf{U}_n = e^{\mathbf{AH}_n}$. (*Hint*: Use Proposition 4.19 and its proof.)
b. Show that $\mathbf{U}_n$ is path connected.
c. Prove Problem 4.11e for $\mathbb{F} = \mathbb{C}$.

(4.14)

Show
(a) $D_1 D D_1^* = D$ for any $D \in \mathbf{D}(n, \mathbb{C}),\ D_1 \in \mathbf{DU}_n$.
(b) $A \in \mathbf{N}(n, \mathbb{C}) \iff A = UDU^*,\ U \in \mathbf{SU}_n,\ D \in \mathbf{D}(n, \mathbb{C})$.
(c) $A \in \mathbf{N}(n, \mathbb{R}),\ \sigma(A) \subset \mathbb{R} \iff A = UDU^\top,\ U \in \mathbf{SO}_n,\ D \in \mathbf{D}(n, \mathbb{R})$.

(4.15)

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 $(BB^*)_{11} = (B^*B)_{11}$.)

(4.16)

Let the assumptions of Theorem 4.26 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 $\mathbf{w}_1 := \frac{\mathbf{v}}{||\mathbf{v}||}$. 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 $\mathbf{w}_{i+1} := \frac{\tilde{\mathbf{w}}_{i+1}}{||\tilde{\mathbf{w}}_{i+1}||}$ 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.)

4.4 Quadratic and hermitian forms

In this section you may assume that $\mathcal{D} = \mathbb{R}, \mathbb{C}$.

Definition 4.27 *Let $\mathbf{V}$ be a module over $\mathcal{D}$ and $Q : \mathbf{V} \times \mathbf{V} \to \mathcal{D}$. Q is called a quadratic form (on $\mathbf{V}$) if the following conditions are satisfied:*
(a) $Q(\mathbf{x}, \mathbf{y}) = Q(\mathbf{y}, \mathbf{x})$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ *(symmetricity);*
(b) $Q(a\mathbf{x} + b\mathbf{z}, y) = aQ(\mathbf{x}, \mathbf{y}) + bQ(\mathbf{z}, \mathbf{y})$ *for all* $a, b \in \mathcal{D}$ *and* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$ *(bilinearity).*

For $\mathcal{D} = \mathbb{C}$ Q is called hermitian form (on $\mathbf{V}$) if Q satisfies the conditions (a') and (b) where
(a') $Q(\mathbf{x}, \mathbf{y}) = Q(\bar{\mathbf{y}}, \mathbf{x})$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ *(barsymmetricity).*

The following results are elementary (see Problems 4.1-4.2):

Proposition 4.28 *Let $\mathbf{V}$ be a module over $\mathcal{D}$ with a basis $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$. Then there is $1-1$ correspondence between a quadratic form Q on $\mathbf{V}$ and $A \in \mathbf{S}(n, \mathcal{D})$:*

$$Q(\mathbf{x}, \mathbf{y}) = \eta^\top A \xi,$$
$$\mathbf{x} = \sum_{i=1}^{n} \xi_i \mathbf{e}_i,\ \mathbf{y} = \sum_{i=1}^{n} \eta_i \mathbf{e}_i,\ \xi = (\xi_1, ..., \xi_n)^\top, \eta = (\eta_1, ..., \eta_n)^\top \in \mathcal{D}^n.$$

Let $\mathbf{V}$ be a vector space over $\mathbb{C}$ with a basis $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$. Then there is $1-1$ correspondence between a hermitian form Q on $\mathbf{V}$ and $A \in \mathbf{H}_n$:

$$Q(\mathbf{x}, \mathbf{y}) = \eta^* A \xi,$$
$$\mathbf{x} = \sum_{i=1}^{n} \xi_i \mathbf{e}_i,\ \mathbf{y} = \sum_{i=1}^{n} \eta_i \mathbf{e}_i,\ \xi = (\xi_1, ..., \xi_n)^\top, \eta = (\eta_1, ..., \eta_n)^\top \in \mathbb{C}^n.$$

Definition 4.29 *Let the assumptions of Proposition 4.28 hold. Then A is called the representation matrix of Q in the basis E.*

Proposition 4.30 *Let the assumptions of Proposition 4.28 Let* $F = \{\mathbf{f_1}, ..., \mathbf{f_n}\}$ *be another basis of the* $\mathcal{D}$ *module* $\mathbf{V}$*. Then the quadratic form* Q *is represented by* $B \in \mathbf{S}(n, \mathcal{D})$ *in the basis* F*, where* B *is congruent* A*:*

$$B = U^\top AU, \quad U \in \mathrm{GL}(n, \mathcal{D})$$

and U *is the matrix corresponding to the basis change from* F *to* E*. For* $\mathcal{D} = \mathbb{C}$ *the hermitian form* Q *is presented by* $B \in \mathbf{H}_n$ *in the basis* F*, where* B *hermicongruent 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*.*

In what follows we assume that $\mathcal{D} = \mathbb{F} = \mathbb{R}, \mathbb{C}$.

Proposition 4.31 *Let* $\mathbf{V}$ *be an* n *dimensional vector space over* $\mathbb{R}$*. Let* $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ *be a quadratic form. Let* $A \in \mathbf{S}(n, \mathbb{R})$ *the representation matrix of* Q *with respect to a basis* E *in* $\mathbf{V}$*. Let* $\mathbf{V}_c$ *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* $\mathbf{V}_c$*.*

See Problem 4.3

Normalization 4.32 *Let* $\mathbf{V}$ *is a finite dimensional IPS over* $\mathbb{F}$*. Let* $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ *be either a quadratic 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 (see Problem 4.4).

Proposition 4.33 *Let* $\mathbf{V}$ *is an* n*-dimensional IPS over* $\mathbb{F}$*. Let* $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ *be either a quadratic 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* $Q(\mathbf{x}, \mathbf{y}) =< T\mathbf{x}, \mathbf{y} >$ *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*.* Q *has only real roots:*

$$\lambda_1(Q) \geq ... \geq \lambda_n(Q),$$

which are called the eigenvalues of Q*. Furthermore there exists an orthonormal basis* $F = \{\mathbf{f}_1, ..., \mathbf{f}_n\}$ *in* $\mathbf{V}$ *such that* $D = \mathrm{diag}(\lambda_1(Q), ..., \lambda_n(Q))$ *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*

$$Q(T, \mathbf{U})(\mathbf{x}, \mathbf{y}) :=< T\mathbf{x}, \mathbf{y} > \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathbf{U}$$

is either a quadratic 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 4.34 *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*

$$\lambda_1(T) \geq ... \geq \lambda_n(T).$$

Same normalization applies to real symmetric matrices and complex hermitian matrices.

Problems

(4.1)

Prove Proposition 4.28.

(4.2)

Prove Proposition 4.30.

(4.3)

Prove Proposition 4.31.

(4.4)

Prove Proposition 4.33.

4.5 Max-min characterizations

Theorem 4.35 *(**Convoy Principle** [12]) Let* $\mathbf{V}$ *be an* n*-dimensional IPS. Let* $\mathrm{Gr}(m,\mathbf{V})$ *be the space of all* m*-dimensional subspaces in* $\mathbf{U}$ *of dimension* $m \in [0,n]\cap\mathbb{Z}_+$*. Let* $T\in\mathbf{S}(\mathbf{V})$*. Then*

$$\lambda_k(T)=\max_{\mathbf{U}\in\mathrm{Gr}(k,\mathbf{V})}\min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{U}}\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}= \tag{4.1}$$
$$\max_{\mathrm{Gr}(k,\mathbf{V})}\lambda_k(Q(T,\mathbf{U})),\quad k=1,...,n,$$

where the quadratic form $Q(T,\mathbf{U})$ *is defined in Proposition 4.33. For* $k\in[1,n]\cap\mathbb{N}$ *let* $\mathbf{U}$ *be an invariant subspace of* T *spanned by eigenvectors* $\mathbf{e}_1,...,\mathbf{e}_k$ *corresponding to the eigenvalues* $\lambda_1(T),...,\lambda_k(T)$*. 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* $\mathbf{U}$ *contains and eigenvector of* T *corresponding to* $\lambda_k(T)$*.*

In particular

$$\lambda_1(T)=\max_{\mathbf{0}\neq\mathbf{x}\in\mathbf{V}}\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle},\quad \lambda_n(T)=\min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{V}}\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle} \tag{4.2}$$

Moreover for any $\mathbf{x}\neq\mathbf{0}$

$$\lambda_1(T)=\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}\iff T\mathbf{x}=\lambda_1(T)\mathbf{x},$$
$$\lambda_n(T)=\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}\iff T\mathbf{x}=\lambda_n(T)\mathbf{x},$$

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 (4.2) is called *convoy principle*.

Proof. Choose an orthonormal basis $E=\{\mathbf{e}_1,...,\mathbf{e}_n\}$ such that

$$T\mathbf{e}_i=\lambda_i(T)\mathbf{e}_i,\ \langle\mathbf{e}_i,\mathbf{e}_j\rangle=\delta_{ij}\quad i,j=1,...,n. \tag{4.3}$$

Then

$$\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}=\frac{\sum_{i=1}^n\lambda_i(T)|x_i|^2}{\sum_{i=1}^n|x_i|^2},\quad \mathbf{x}=\sum_{i=1}^n x_i\mathbf{e}_i\neq\mathbf{b}0. \tag{4.4}$$

The above equality yields straightforward (4.2) 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

$$\lambda_k(Q(T,\mathbf{U}))=\min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{U}}\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}\quad\text{for any }\mathbf{U}\in\mathrm{Gr}(k,\mathbf{U}). \tag{4.5}$$

Next there exists $\mathbf{b}0\neq\mathbf{x}\in\mathbf{U}$ such that $\langle\mathbf{x},\mathbf{e}_i\rangle=0$ for $i=1,...,k-1$. (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

$$\lambda_1(T)=...=\lambda_{n_1}(T)>\lambda(T)_{n_1+1}(T)=...=\lambda_{n_2}(T)>...>$$
$$\lambda_{n_{r-1}+1}(T)=...=\lambda_{n_r}(T)=\lambda_n(T),\quad n_0=0<n_1<...<n_r=n. \tag{4.6}$$

Assume that $n_{j-1}<k\leq n_j$. Suppose that $\lambda_k(Q(T,\mathbf{U}))=\lambda_k(T)$. Then for the $\mathbf{x}$ of the above form $\frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle\mathbf{x},\mathbf{x}\rangle}=\lambda_k(T)$. Hence $\mathbf{x}=\sum_{i=k}^{n_j}x_i\mathbf{e}_i$. Thus $T\mathbf{x}=\lambda_k(T)\mathbf{x}$.

Let $\mathbf{U}_k = \mathrm{span}(\mathbf{e}_1, ..., \mathbf{e}_k)$. Let $\mathbf{0} \neq \mathbf{x} = \sum_{i=1}^k x_i \mathbf{e}_i \in \mathbf{U}_k$. Then

$$\frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\sum_{i=1}^k \lambda_i(T)|x_i|^2}{\sum_{i=1}^k |x_i|^2} \geq \lambda_k(T) \Rightarrow \lambda_k(Q(T, \mathbf{U}_k)) \geq \lambda_k(T).$$

Hence $\lambda_k(Q(T, \mathbf{U}_k)) = \lambda_k(T)$. □

It can be shown that for $k > 1$ and $\lambda_1(T) > \lambda_k(T)$ there exist $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ 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 $\mathbf{e}_1, ..., \mathbf{e}_k$ satisfying (4.3). (See Problem 4.10.)

Corollary 4.36 *Let the assumptions of Theorem 4.35 hold. Let $1 \leq \ell \leq n$. Then*

$$\lambda_k(T) \geq \lambda_k(Q(T, \mathbf{W})) \text{ and } \lambda_k(T) = \max_{\mathbf{W} \in \mathrm{Gr}(\ell, \mathbf{V})} \lambda_k(Q(T, \mathbf{W})), \quad k = 1, ..., \ell. \tag{4.7}$$

Proof. For $k \leq \ell$ apply Theorem 4.35 to $\lambda_k(Q(T, \mathbf{W}))$ to deduce that $\lambda_k(Q(T, \mathbf{W})) \leq \lambda_k(T)$. Let $\mathbf{U}_\ell = \mathrm{span}(\mathbf{e}_1, ..., \mathbf{e}_\ell)$. Then

$$\lambda_k(Q(T, \mathbf{U}_\ell)) = \lambda_k(T), \quad k = 1, ..., \ell.$$

□

Corollary 4.37 *(Cauchy Interlacing Theorem) Let $A \in \mathbf{H}_n$ and let $B \in \mathbf{H}_{n-1}$ be the principal submatrix of A obtained by deleting the row and the column $i \in [1, n]$ of A. Denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_{n-1}$ the eigenvalues of A and B respectively. Then*

$$\lambda_1 \geq \nu_1 \geq \lambda_2 \geq \ldots \geq \nu_{n-1} \geq \lambda_n,$$

i.e. $\lambda_i \geq \nu_i \geq \lambda_{i+1}$ for $i = 1, \ldots, n-1$.

Proof. Let $\mathbf{U}_i := \mathrm{span}(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_n)$. Then the restriction of the quadratic form $\mathbf{x}^* A\mathbf{x}$ to $\mathbf{U}$ gives rise to the quadratic form induced by B. Corollary 4.36 yields the inequality $\lambda_i \geq \nu_i$ for $i = 1, \ldots, n-1$. Consider now $-A$ and its principal submatrix $-B$. Their eigenvalues arranged in a decreasing order are $-\lambda_n \geq -\lambda_{n-1} \geq \ldots \geq -\lambda_1$ and $-\nu_{n-1} \geq -\nu_{n-2} \geq \ldots \geq -\nu_1$ respectively. The above arguments yield $-\lambda_{n-i+1} \geq -\nu_{n-i}$ for $i = 1, \ldots, n-1$, which are equivalent to $\nu_j \geq \lambda_{j+1}$ for $j = 1, \ldots, n-1$. □

Definition 4.38 *For $T \in \mathbf{S}(\mathbf{V})$ denote by $\iota_+(T), \iota_0(T), \iota_-(T)$ the number of positive, negative and zero 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.*

Let $B > \mathbf{0}$, $B \geq \mathbf{0}$, $B \leq \mathbf{0}$, $B < \mathbf{0}$ if $\iota_0(T) + \iota_-(T) = 0$, $\iota_-(T) = 0$, $\iota_+(T) = 0$ and $\iota_+(T) + \iota_0(T) = 0$ respectively.

For $B \in \mathbf{H}_n$ $\iota(B) := (\iota_+(B), \iota_0(B), \iota_-(B))$ is the inertia of B, where $\iota_+(B), \iota_0(B), \iota_-(B)$ is the number of positive, negative and zero eigenvalues of B respectively.

Proposition 4.39 *Let $U \in \mathrm{Gr}(k, \mathbf{V})$.*

1. *Assume that $\lambda_k(Q(T, \mathbf{U})) > 0$, i.e. $Q(T, \mathbf{U}) > 0$. Then $k \leq \iota_+(T)$.*
2. *Assume that $\lambda_k(Q(T, \mathbf{U})) \geq 0$, i.e. $Q(T, \mathbf{U}) \geq 0$. Then $k \leq \iota_+(T) + \iota_0(T)$.*
3. *Assume that $\lambda_1(Q(T, \mathbf{U})) < 0$, i.e. $Q(T, \mathbf{U}) < 0$. Then $k \leq \iota_-(T)$.*
4. *Assume that $\lambda_1(Q(T, \mathbf{U})) \leq 0$, i.e. $Q(T, \mathbf{U}) \leq 0$. Then $k \leq \iota_-(T) + \iota_0(T)$.*

5. ***Sylvester Law of Inertia***: *Let $B \in \mathbf{H}_n$ or $B \in \mathrm{S}_n(\mathbb{R})$ and assume that $A = PBP^*$ or $A = PBP^\top$ for some $P \in \mathrm{GL}(n,\mathbb{C})$ or $P \in \mathrm{GL}(n,\mathbb{R})$ respectively. Then $\iota(A) = \iota(B)$. Furthermore, if $A, B \in \mathbf{H}_n$ or $A, B \in \mathrm{S}_n(\mathbb{R})$ have the same inertia then there exists $P \in \mathrm{GL}(n,\mathbb{C})$ or $P \in \mathrm{GL}(n,\mathbb{R})$ such that $A = PBP^*$ or $A = PBP^\top$ respectively.*

Proof. *1.* Corollary 4.36 yields that $\lambda_k(T) \geq \lambda_k(Q(T,\mathbf{U})) > 0$, hence $k \leq \iota_+(T)$. The proofs of *2,3,4* are similar.

5. Assume that $A = PBP^*$. Note that if $\mathbf{x}^*B\mathbf{x} > 0$ or $\mathbf{x}^*B\mathbf{x} \geq 0$ or for all $\mathbf{0} \neq \mathbf{x} \in \mathbf{U}$ then $\mathbf{y}^*A\mathbf{y} > 0$ $\mathbf{y}^*A\mathbf{y} \geq 0$ for all $\mathbf{y} \in P^*\mathbf{U}$ respectively. Hence $\iota_+(B) \leq \iota_+(A)$ and $\iota_+(B) + \iota_0(B) \leq \iota_+(A) + \iota_0(A)$. Since P is invertible $P^{-1} = Q$ and $B = QAQ^*$. Hence we deduce as above that $\iota_+(A) \leq \iota_+(B)$ and $\iota_+(A) + \iota_0(A) \leq \iota_+(B) + \iota_0(B)$. Thus $\iota(A) = \iota(B)$ and $\iota_0(A) = \iota_0(B)$. Since $\iota_+(A) + \iota_0(A) + \iota_-(A) = \iota_+(B) + \iota_0(B) + \iota_-(B) = n$ we deduce that $\iota_-(A) = \iota_-(B)$, i.e. $\iota(A) = \iota(B)$.

Assume now that $\iota(B) = \iota(A)$. Observe that $B = Q\Lambda Q^*$, where Q is unitary and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $f(x) = \frac{1}{\sqrt{|x|}}$ if $x \neq 0$ and $f(0) = 1$. Set $R = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n)) \in S_n(\mathbb{R})$. Then $C = R^*\Lambda\mathbb{R} = R\Lambda R$ is a diagonal matrix with $\iota_+(B)$ $1's$, $\iota_0(B)$ zeros and $\iota_-(B)$ $-1's$ on the main diagonal. So $(QR)^*B(QR) = C$. Similarly $C = (Q'R')^*A(Q'R')$. Hence $A = PBP^*$.

□

Theorem 4.40 *Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Then*

$$\lambda_k(T) = \min_{W \in \mathrm{Gr}(k-1,\mathbf{V})} \max_{\mathbf{0} \neq x \in \mathbf{W}^\perp} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}, \quad k = 1, ..., n.$$

See Problem 4.11 for the proof of the theorem and the following corollary.

Corollary 4.41 *Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Let $k, \ell \in [1, n-1]$ be integers satisfying $k \leq \ell$, $k + \ell > n$. Then*

$$\lambda_{k+\ell-n}(T) \leq \lambda_k(Q(T,\mathbf{W})) \leq \lambda_k(T), \quad \text{for any } \mathbf{W} \in \mathrm{Gr}(\ell, \mathbf{V}).$$

Definition 4.42 *Let $\mathbf{V}$ be an n-dimensional IPS. Fix an integer $k \in [1, n]$. Then $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\}$ is called an orthonormal k-frame if $\langle \mathbf{f}_i, \mathbf{f}_j\rangle = \delta_{ij}$ 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 orthonormal k-frames which span $\mathbf{U}$.

Theorem 4.43 *(**Ky Fan** [3]) Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Then for any integer $k \in [1, n]$*

$$\sum_{i=1}^k \lambda_i(T) = \max_{\{\mathbf{f}_1,...,\mathbf{f}_k\} \in \mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^k \langle T\mathbf{f}_i, \mathbf{f}_i\rangle.$$

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 spanF_k *is spanned by $\mathbf{e}_1, ..., \mathbf{e}_k$ satisifying (4.3).*

Proof. Define

$$\begin{aligned} \operatorname{tr} Q(T,\mathbf{U}) &:= \sum_{i=1}^{k} \lambda_i(Q(T,\mathbf{U})) \quad \text{for } \mathbf{U}\in \operatorname{Gr}(k,\mathbf{V}), \\ \operatorname{tr}_k T &:= \sum_{i=1}^{k} \lambda_i(T). \end{aligned} \tag{4.8}$$

Let $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \operatorname{Fr}(k,\mathbf{V})$. Set $\mathbf{U} = \operatorname{span} F_k$. Then in view of Corollary 4.36

$$\sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle = \operatorname{tr} Q(T,\mathbf{U}) \le \sum_{i=1}^{k} \lambda_i(T).$$

Let $E_k := \{\mathbf{e}_1, ..., \mathbf{e}_k\}$ where $\mathbf{e}_1, ..., \mathbf{e}_n$ are given by (4.3). Clearly $\operatorname{tr}_k T = \operatorname{tr} Q(T, \operatorname{span} E_k)$. This shows the maximal characterization of $\operatorname{tr}_k T$.

Let $\mathbf{U} \in \operatorname{Gr}(k,\mathbf{V})$ and assume that $\operatorname{tr}_k T = \operatorname{tr} Q(T,\mathbf{U})$. Hence $\lambda_i(T) = \lambda_i(Q(T,\mathbf{U}))$ for $i = 1, ..., k$. Then there exists $G_k = \{\mathbf{g}_1, ..., \mathbf{g}_k\} \in \operatorname{Fr}(k,\mathbf{U}))$ such that

$$\min_{\mathbf{0}\neq\mathbf{x}\in\operatorname{span}(\mathbf{g}_1,...,\mathbf{g}_i\}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} = \lambda_i(Q(T,\mathbf{U})) = \lambda_i(T),\ i = 1, ..., k.$$

Use Theorem 4.35 to deduce that $T\mathbf{g}_i = \lambda_i(T)\mathbf{g}_i$ for $i = 1, ..., k$. □

Theorem 4.44 *(J. Neumann) Let $A, B \in \mathbf{H}_n$. Denote by $\lambda_1(A) \ge \ldots \ge \lambda_n(A), \lambda_1(B) \ge \ldots \ge \lambda_n(B)$ the eigenvalues of A and B respectively. Then*

$$\lambda_1(A)\lambda_n(B) + \lambda_2(A)\lambda_{n-1}(B) + \ldots + \lambda_n(A)\lambda_1(B) \le \operatorname{tr}(AB) \le \sum_{i=1}^{n} \lambda_i(A)\lambda_i(B) \tag{4.9}$$

Equalities hold if and only if there is an orthonormal basis $\mathbf{g}_1, \ldots, \mathbf{g}_n$ of $\mathbb{C}^n$ such that

1. *For the equality case in the upper bound $A\mathbf{g}_i = \lambda_i(A)\mathbf{g}_i, B\mathbf{g}_i = \lambda_i(B)\mathbf{g}_i$ for $i = 1, \ldots, n$.*
2. *For the equality case in the lower bound $A\mathbf{g}_i = \lambda_i(A)\mathbf{g}_i, B\mathbf{g}_i = \lambda_{n-i+1}(B)\mathbf{g}_i$ for $i = 1, \ldots, n$.*

Proof. Note that for any invertible matrix U one has $\operatorname{tr}(AB) = \operatorname{tr}(UABU^{-1}) = \operatorname{tr}((UAU^{-1})(UBU^{-1}))$, since $tr(AB)$ is the sum of the eigenvalues of AB. Choose U unitary such that $UAU^{-1} = UAU^* = \Lambda := \operatorname{diag}(\lambda_1(A), \ldots, \lambda_n(A)$. Let $C = UBU^{-1} = UBU^*$. Then $C = (c_{ij})_{i,j=1}^n \in \mathbf{H}_n$ and $\lambda_i(C) = \lambda_i(B)$ for $i = 1, \ldots, n$. Clearly $\operatorname{tr}(\Lambda C) = \sum_{i=1}^n \lambda_i(A)c_{ii} = \sum_{i=1}^n \lambda_i(A)(\mathbf{e}_i^\top C\mathbf{e}_i)$, where $\mathbf{e}_i = (\delta_{i1}, \ldots, \delta_{in})^\top, i = 1, \ldots, n$ is the standard basis in $\mathbb{C}^n$. Observe next that

$$\sum_{i=1}^{n} \lambda_i(A)\mathbf{e}_i^\top C\mathbf{e}_i = \sum_{i=1}^{n-1}(\lambda_i(A) - \lambda_{i+1}(A)) \sum_{j=1}^{i} \mathbf{e}_j^\top C\mathbf{e}_j + \lambda_n \sum_{i=1}^{n} \mathbf{e}_i^\top C\mathbf{e}_i.$$

Since $\lambda_i(A) - \lambda_{i+1}(A) \ge 0$ Ky Fan inequality yields that $(\lambda_i(A) - \lambda_{i+1}(A)) \sum_{j=1}^i \mathbf{e}_j^\top C\mathbf{e}_j \le (\lambda_i(A) - \lambda_{i+1}(A)) \sum_{j=1}^i \lambda_j(C)$ for $i = 1, \ldots, n-1$. Combine all these inequalities with the equality $\sum_{i=1}^n \mathbf{e}_i^\top C\mathbf{e}_i = \operatorname{tr} C = \sum_{i=1}^n \lambda_i(C)$. to deduce the inequality $\operatorname{tr}(\Lambda C) \le \sum_{i=1}^n \lambda_i(A)\lambda_i(C)$. This gives the upper inequality in (4.9).

Equality case is slightly more delicate to analyze. If $\lambda_1(A) > \ldots > \lambda_n(A)$ then $C\mathbf{e}_i = \lambda_i(C)\mathbf{e}_i$ for $i = 1, \ldots, n$ if $\operatorname{tr}(\Lambda C) = \sum_{i=1}^n \lambda_i(A)\lambda_i(C)$.

To obtain the lower bound note that $\mathrm{tr}(A(-B)) \leq \sum_{i=1}^n \lambda_i(A)\lambda_i(-B)$. Use the identity $\lambda_i(-B) = -\lambda_{n-i+1}(B)$ for $i = 1, \ldots, n$ to deduce the lower bound. Equality case as for the upper bound. □

Definition 4.45 *Let*

$$\mathbb{R}^n_{\searrow} := \{\mathbf{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n : \quad x_1 \geq x_2 \geq ... \geq x_n\}.$$

For $\mathbf{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n$ let $\bar{\mathbf{x}} = (\bar{x}_1, ..., \bar{x}_n)^T \in \mathbb{R}^n_{\searrow}$ be the unique rearrangement of the coordinates of $\mathbf{x}$ in a decreasing order. That is there exists a permutation π on $\{1, ..., n\}$ such that $\bar{x}_i = x_{\pi(i)}$, $i = 1, ..., n$. Let $\mathbf{x} = (x_1, ..., x_n)^\top, \mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$. Then $\mathbf{x}$ is weakly majorized by $\mathbf{y}$ ($\mathbf{y}$ weakly majorizes $\mathbf{x}$), which is denoted by $\mathbf{x} \preceq \mathbf{y}$, if

$$\sum_{i=1}^k \bar{x}_i \leq \sum_{i=1}^k \bar{y}_i, \quad k = 1, ..., n.$$

$\mathbf{x}$ is majorized by $\mathbf{y}$ ($\mathbf{y}$ majorizes $\mathbf{x}$), which is denoted by $\mathbf{x} \prec \mathbf{y}$, if $\mathbf{x} \preceq \mathbf{y}$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

A remarkable inequality is attached to the notion of majorization [6], see also Problem **??** part (c).

Theorem 4.46 *Let $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top \prec \mathbf{y} = (y_1, ..., y_n)^T$. Let $\phi : [\bar{y}_n, \bar{y}_1] \to \mathbb{R}$ be a continuous convex function. Then*

$$\sum_{i=1}^n \phi(x_i) \leq \sum_{i=1}^n \phi(y_i).$$

Corollary 4.47 *Let $\mathbf{V}$ be an n-dimensional IPS. Let $T \in \mathbf{S}(\mathbf{V})$. Denote $\lambda(T) := (\lambda_1(T), ..., \lambda_n(T))^\top \in \mathbb{R}^n_{\searrow}$. Let $F_n = \{\mathbf{f}_1, ...\mathbf{f}_n\} \in \mathrm{Fr}(n, \mathbf{V})$. Then $(\langle T\mathbf{f}_1, \mathbf{f}_1\rangle, ..., \langle T\mathbf{f}_n, \mathbf{f}_n\rangle)^\top \prec \lambda(T)$. Let $\phi : [\lambda_n(T), \lambda_1(T)] \to \mathbb{R}$ be a continuous convex function. Then*

$$\sum_{i=1}^n \phi(\lambda_i(T)) = \max_{\{\mathbf{f}_1, ...\mathbf{f}_n\} \in \mathrm{Fr}(n, \mathbf{V})} \sum_{i=1}^n \phi(\langle T\mathbf{f}_i, \mathbf{f}_i\rangle).$$

See Problem 4.12

Definition 4.48 *A set $D \subseteq \mathbb{R}^n$ is called convex if for any $\mathbf{x}, \mathbf{y} \in D$ and any $t \in [0, 1]$ the linear combination $t\mathbf{x} + (1 - t)\mathbf{y}$ is in the set D.*

Let $D \subseteq \mathbb{R}^n$ be a convex set and $f : D \to \mathbb{R}$ be a function on D. Then f is called a convex function on D if for any $\mathbf{x}, \mathbf{y} \in D$ and any $t \in [0, 1]$ $f(t\mathbf{x} + (1 - t)\mathbf{y}) \leq tf(\mathbf{x}) + (1 - t)f(\mathbf{y})$.

Problems

1. Let k, n be positive integers and assume that $k \leq n$. Let $f_k : \mathrm{S}_n(\mathbb{R}) \to \mathbb{R}$ be the following function: $f_k(A) = \sum_{i=1}^k \lambda_i(A)$ for any $A \in \mathrm{S}_n(\mathbb{R})$. Show that f_k is convex on $\mathrm{S}_n(\mathbb{R})$. (See above for the definition of convexity.) What happens for $k = n$?

(4.10)

Let $\mathbf{V}$ be 3 dimensional IPS and $T \in \mathrm{Hom}(\mathbf{V})$ 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 $\mathbf{W} = \text{span}(\mathbf{e}_1, \mathbf{e}_3)$.
(a) Show that for each $t \in [\lambda_3(T), \lambda_1(T)]$ there exists a unique $\mathbf{W}(t) \in \text{Gr}(1, \mathbf{W})$ such that $\lambda_1(Q(T, \mathbf{W}(t))) = t$.
(b) Let $t \in [\lambda_2(T), \lambda_1(T)]$. Let $\mathbf{U}(t) = \text{span}(\mathbf{W}(t), \mathbf{e}_2) \in \text{Gr}(2, \mathbf{V})$. Show that $\lambda_2(T) = \lambda_2(Q(T, \mathbf{U}(t))$.

(4.11)

(a) Let the assumptions of Theorem 4.40 hold. Let $\mathbf{W} \in \text{Gr}(k-1, \mathbf{V})$. Show that there exists $\mathbf{b}0 \neq \mathbf{x} \in \mathbf{W}^\perp$ such that $\langle \mathbf{x}, \mathbf{e}_i \rangle = 0$ for $k+1, ..., n$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ satisfy (4.3). 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 $\mathbf{U}_\ell = \text{span}(\mathbf{e}_1, ..., \mathbf{e}_\ell)$. Show that $\lambda_1(Q(T, \mathbf{U}_\ell^\perp)) = \lambda_{\ell+1}(T)$ for $\ell = 1, ..., n-1$.
(c) Prove Theorem 4.40.
(d) Prove Corollary 4.41. (**Hint**: Choose $U \in \text{Gr}(k, \mathbf{W})$ such that $\mathbf{U} \subset \mathbf{W} \cap \text{span}(\mathbf{e}_{k+\ell-n+1}, ..., \mathbf{e}_n)^\perp$. Then $\lambda_{k+\ell-n}(T) \leq \lambda_k(Q(T, \mathbf{U})) \leq \lambda_k(Q(T, \mathbf{W}))$.)

(4.12)

Prove Corollary 4.47.

(4.13)

Let $B = (b_{ij})_1^n \in \mathbf{H}_n$. Show that $B > 0$ if and only if $\det(b_{ij})_1^k > 0$ for $k = 1, ..., n$.

4.6 Positive definite operators and matrices

Definition 4.49 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F} = \mathbb{C}, \mathbb{R}$. Let $S, T \in \mathbf{S}(\mathbf{V})$. Then $T > S$, $(T \geq S)$ if $\langle T\mathbf{x}, \mathbf{x} \rangle > \langle S\mathbf{x}, \mathbf{x} \rangle$, $(\langle T\mathbf{x}, \mathbf{x} \rangle \geq \langle S\mathbf{x}, \mathbf{x} \rangle)$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbf{V}$. T is called positive (nonnegative) definite if $T > \mathbf{0}$ $(T \geq \mathbf{0})$, where 0 is the zero operator in* $\text{Hom}(\mathbf{V})$.

Let P, Q be either quadratic forms if $\mathbb{F} = \mathbb{R}$ or hermitian forms if $\mathbb{F} = \mathbb{C}$. Then $Q > P$, $(Q \geq P)$ if $Q(\mathbf{x}, \mathbf{x}) > P(\mathbf{x}, \mathbf{x})$, $(Q(\mathbf{x}, \mathbf{x}) \geq P(\mathbf{x}, \mathbf{x}))$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbf{V}$. Q is called positive (nonnegative) definite if $Q > \mathbf{0}$ $(Q \geq \mathbf{0})$, where $\mathbf{0}$ is the zero operator in $\text{Hom}(\mathbf{V})$.

For $A, B \in \mathbf{H}_n$ $B > A$ $(B \geq A)$ if $\mathbf{x}^* B\mathbf{x} > \mathbf{x}^* A\mathbf{x}$ $(\mathbf{x}^* B\mathbf{x} \geq \mathbf{x}^* A\mathbf{x})$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n$. $B \in \mathbf{H}_n$ is called is called positive (nonnegative) definite if $B > \mathbf{0}$ $(B \geq \mathbf{0})$.

Use (4.1) to deduce.

Corollary 4.50 *Let $\mathbf{V}$ be n-dimensional IPS. Let $T \in \mathbf{S}(\mathbf{V})$. Then $T > \mathbf{0}$ $(T \geq \mathbf{0})$ 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 $i = 1, ..., n$.*

Proposition 4.51 *Let $\mathbf{V}$ be a finite dimensional IPS. Assume that $T \in \mathbf{S}(\mathbf{V})$. Then $T \geq \mathbf{0}$ if and only if there exists $S \in \mathbf{S}(\mathbf{V})$ such that $T = S^2$. Furthermore $T > \mathbf{0}$ if and only if S is invertible. For $\mathbf{0} \leq T \in \mathbf{S}(\mathbf{V})$ there exists a unique $\mathbf{0} \leq S \in \mathbf{S}(\mathbf{V})$ 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 $T \geq \mathbf{0}$. Let $\mathbf{e}_1, ..., \mathbf{e}_n$ be an orthonormal basis consisting of eigenvectors of T as in (4.3). Since $\lambda_i(T) \geq 0$, $i = 1, ..., n$ we can define $P \in \text{Hom}(\mathbf{V})$ as follows

$$P\mathbf{e}_i = \sqrt{\lambda_i(T)}\mathbf{e}_i, \quad i = 1, ..., n.$$

Clearly P is self-adjoint nonnegative 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 $T \geq \mathbf{0}$. Clearly $\langle T\mathbf{x}, \mathbf{x} \rangle = 0 \iff S\mathbf{x} = \mathbf{0}$. Hence $T > \mathbf{0} \iff S \in \text{GL}(\mathbf{V})$. Suppose that $S \geq \mathbf{0}$. Then $\lambda_i(S) = \sqrt{\lambda_i(T)}$, $i = 1, ..., n$. Furthermore each eigenvector of S is an eigenvector of T. It is straightforward to show that $S = P$, where P is defined above. □

Corollary 4.52 *Let $B \in \mathbf{H}_n$ $(\mathbf{S}(n,\mathbb{R}))$. Then $B \geq \mathbf{0}$ if and only there exists $A \in \mathbf{H}_n$ $(\mathbf{S}(n,\mathbb{R}))$ such that $B = A^2$. Furthermore $B > \mathbf{0}$ if and only if A is invertible. For $B \geq \mathbf{0}$ there exists a unique $A \geq \mathbf{0}$ such that $B = A^2$. This A is denoted by $B^{\frac{1}{2}}$.*

Theorem 4.53 *Let $\mathbf{V}$ be an IPS over $\mathbb{F} = \mathbb{C}, \mathbb{R}$. Let $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbf{V}$. Then the grammian matrix $G(\mathbf{x}_1, ..., \mathbf{x}_n) := (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_1^n$ is a hermitian nonnegative definite matrix. (If $\mathbb{F} = \mathbb{R}$ then $G(\mathbf{x}_1, ..., \mathbf{x}_n)$ is real symmetric nonnegative definite.) $G(\mathbf{x}_1, ..., \mathbf{x}_n) > \mathbf{0}$ if and only $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. Furthermore for any integer $k \in [1, n-1]$*

$$\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \leq \det G(\mathbf{x}_1, ..., \mathbf{x}_k) \ \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n). \tag{4.1}$$

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

$$\mathbf{a}^* G(\mathbf{x}_1, ..., \mathbf{x}_n)\mathbf{a} = \langle \sum_{i=1}^n a_i \mathbf{x}_i, \sum_{j=1}^n a_j \mathbf{x}_j \rangle \geq 0.$$

Equality holds if and only if $\sum_{i=1}^n a_i \mathbf{x}_i = 0$. Hence $G(\mathbf{x}_1, ..., \mathbf{x}_n) \geq \mathbf{0}$ and $G(\mathbf{x}_1, ..., \mathbf{x}_n) > \mathbf{0}$ if and only if $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. In particular $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \geq 0$ 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 (4.1). Assume first that the right-hand side of (4.1) 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 (4.1) 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 (4.1). It is left to show the inequality (4.1) 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 (4.1). Let $S_j = \text{span}(\mathbf{x}_1, ..., \mathbf{x}_j)$ for $j = 1, ..., n$. Corollary 4.1 yields that $\text{span}(\mathbf{e}_1, ..., \mathbf{e}_{n-1}) = S_{n-1}$. Hence $\mathbf{y}_n = \mathbf{x}_n - \sum_{j=1}^{n-1} b_j \mathbf{x}_j$ for some $b_2, ..., b_n \in \mathbb{F}$. 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

$$\begin{aligned} &\det G(\mathbf{x}_i, ..., \mathbf{x}_n) = \det G(\mathbf{x}_i, ..., \mathbf{x}_{n-1}) \ ||y_n||^2 = ... = \\ &\det G(\mathbf{x}_i, ..., \mathbf{x}_k) \prod_{i=k+1}^n ||y_i||^2 = \\ &\det G(\mathbf{x}_i, ..., \mathbf{x}_k) \prod_{i=k+1}^n \text{dist}(\mathbf{x}_i, S_{i-1})^2, \quad k = n-1, ..., 1. \end{aligned} \tag{4.2}$$

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

$$\hat{S}_j := \text{span}(\mathbf{x}_{k+1}, ..., \mathbf{x}_j) = \text{span}(\hat{\mathbf{y}}_{k+1}, ..., \hat{\mathbf{y}}_j), \ \text{dist}(\mathbf{x}_j, \hat{S}_{j-1}) = ||\hat{\mathbf{y}}_j||,$$

for $j = k+1, ..., n$, where $\hat{S}_k = \{\mathbf{b}0\}$. Use (4.2) 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

$$||\mathbf{y}_j|| = \text{dist}(\mathbf{x}_j, S_{j-1}) \leq \text{dist}(\mathbf{x}_j, \hat{S}_{j-1}) = ||\hat{\mathbf{y}}_j||, \ j = k+1, ..., n.$$

This shows (4.1). Assume now equality holds in (4.1). 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 $\text{dist}(\mathbf{x}_j, S_{j-1}) =$

$\operatorname{dist}(\hat{\mathbf{y}}_j, S_{j-1}) = ||\mathbf{y}_j||$. Hence $||\hat{\mathbf{y}}_j|| = \operatorname{dist}(\hat{\mathbf{y}}_j, S_{j-1})$. Part (h) of Problem 4.5 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. $< \mathbf{x}_j, \mathbf{x}_i >= 0$ for $j > k$ and $i \leq k$. Clearly, if the last condition holds then $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) = \det G(\mathbf{x}_1, ..., \mathbf{x}_k)\ \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n)$. □

$\det G(\mathbf{x}_1, ..., \mathbf{x}_n)$ has the following geometric meaning. Consider a parallelepiped Π in $\mathbf{V}$ spanned by $\mathbf{x}_1, ..., \mathbf{x}_n$ starting from the origin $\mathbf{b}0$. That is Π is a convex hull spanned by the vectors $\mathbf{b}0$ 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 (4.1) and equalities (4.2) are "obvious" from this geometrical point of view.

Corollary 4.54 *Let $\mathbf{0} \leq B = (b_{ij})_1^n \in \mathbf{H}_n$. Then*

$$\det B \leq \det(b_{ij})_1^k\ \det(b_{ij})_{k+1}^n, \text{ for } k = 1, ..., n-1.$$

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 4.52 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 $< \mathbf{x}, \mathbf{y} >= \mathbf{y}^* \mathbf{x}$. Since $X \in \mathbf{H}_n$ we deduce that $B = G(\mathbf{x}_1, ..., \mathbf{x}_n)$. □

Theorem 4.55 *Let $\mathbf{V}$ be an n-dimensional IPS. Let $T \in \mathbf{S}$. TFAE:*
(a) $T > \mathbf{0}$.
(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 4.51 $T = S^2$ 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 4.1 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 $\mathbf{U} := \operatorname{span}(\mathbf{g}_1, ..., \mathbf{g}_{n-1})$ and $Q := Q(T, \mathbf{U})$. Then there exists $P \in \mathbf{S}(\mathbf{U})$ such that $\langle P\mathbf{x}, \mathbf{y} \rangle = Q(\mathbf{x}, \mathbf{y}) = \langle T\mathbf{x}, \mathbf{y} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbf{U}$. By induction $P > 0$. Corollary 4.36 yields that $\lambda_{n-1}(T) \geq \lambda_{n-1}(P) > 0$. Hence T has at least $n-1$ positive eigenvalues. Let $\mathbf{e}_1, ..., \mathbf{e}_n$ be given by (4.3). 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 $\mathbf{g}_1, ..., \mathbf{g}_n$ to $\mathbf{e}_1, ..., \mathbf{e}_n$, i.e.

$$\mathbf{g}_i = \sum_{p=1}^n a_{pi} \mathbf{e}_p,\ i = 1, ..., n.$$

It is straightforward to show that

$$(\langle T\mathbf{g}_i, \mathbf{g}_j \rangle)_1^n = A^T (\langle T\mathbf{e}_p, \mathbf{e}_q \rangle) \bar{A} \Rightarrow \tag{4.3}$$

$$\det(\langle T\mathbf{g}_i, \mathbf{g}_j \rangle)_1^n = \det(\langle T\mathbf{e}_i, \mathbf{e}_j \rangle)_1^n |\det A|^2 = |\det A|^2 \prod_{i=1}^n \lambda_i(T).$$

Since $\det(\langle T\mathbf{g}_i, \mathbf{g}_j \rangle)_1^n > 0$ and $\lambda_1(T) \geq ... \geq \lambda_{n-1}(T) > 0$ it follows that $\lambda_n(T) > 0$. □

Corollary 4.56 *Let $B = (b_{ij})_1^n \in \mathbf{H}_n$. Then $B > \mathbf{0}$ if and only if $\det(b_{ij})_1^k > 0$ for $k = 1, ..., n$.*

The following result is straightforward (see Problem 4.5:

Proposition 4.57 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ with the inner product $\langle \cdot, \cdot \rangle$. Assume that $T \in \mathbf{S}(\mathbf{V})$. Then $T > \mathbf{0}$ 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 $(\mathbf{x}, \mathbf{y}) = \langle T\mathbf{x}, \mathbf{y} \rangle$ for a unique self-adjoint positive definite operator $T \in \mathrm{Hom}(\mathbf{V})$.*

Example 4.58 *Each $\mathbf{0} < B \in \mathbf{H}_n$ induces and inner product on $\mathbb{C}^n$: $(\mathbf{x}, \mathbf{y}) = \mathbf{y}^* B\mathbf{x}$. Each $\mathbf{0} < B \in \mathbf{S}(n, \mathbb{R})$ induces and inner product on $\mathbb{R}^n$: $(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T B\mathbf{x}$. 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.*

Definition 4.59 *Let $\mathbf{V}$ be a finite dimensional IPS with the inner product $\langle \cdot, \cdot \rangle$. Let $S \in \mathrm{Hom}(V)$. Then S is called symmetrizable if there exists an inner product $(\cdot, \cdot)$ on $\mathbf{V}$ such that S is self-adjoint with respect to $(\cdot, \cdot)$.*

Definition 4.60 *Let $A \in \mathbb{C}^{n \times n}$ and assume that $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A counted with their multiplicities, i.e. $\det(zI - A) = \prod_{i=1}^n (z - \lambda_i)$. Let $\iota_+(A), \iota_0(A), \iota_-(A)$ be the number of the eigenvalues of A satisfying $\Re\lambda_i > 0, \Re\lambda_i = 0, \Re\lambda_i < 0$ respectively. (Here $\Re z$ stands for the real part of the complex number $z \in \mathbb{C}$.) Then $\iota(A) := (\iota_+(A), \iota_0(A), \iota_-(A))$ is called the inertia of A.*

Note that for $A \in \mathbf{H}_n$ the inertia of A coincides with the inertia defined in Definition 4.38. Furthermore A is stable if and only if $\iota_-(A) = n$, i.e. $\iota_+(A) = \iota_0(A) = 0$.

Theorem 4.61 *Let $A \in \mathbb{C}^{n \times n}$ and $B \in \mathbf{H}_n$. Then $C := A^*B + BA \in \mathbf{H}_n$. If $C > \mathbf{0}$ then A, B are nonsingular and $\iota(A) = \iota(B)$. In particular, $\iota_0(A) = 0$, i.e. A does not have eigenvalues on the imaginary axis.*

Let $A \in \mathbb{C}^{n \times n}$. If A is stable then for any given $C \in \mathbf{H}_n$ the linear system $A^*B + BA = C$, in unknown matrix $B \in \mathbf{H}_n$, has a unique solution B.

*Moreover any $A \in \mathbb{C}^{n \times n}$ is stable if an only if the system $A^*B + BA = I$ has a unique solution $B \in \mathbf{H}_n$ which is negative definite, i.e. $B < \mathbf{0}$. (**Lyapunov criteria of stability**.)*

Proof. As B is hermitian $(A^*B + BA)^* = B^*A + A^*B^* = BA + A^*B$, i.e. C is hermitian. Suppose that $B\mathbf{x} = 0$. Then $\mathbf{x}^*B = \mathbf{0}^\top$ and $\mathbf{x}^*C\mathbf{x} = \mathbf{x}^*A\mathbf{0}^\top + \mathbf{0}^\top A\mathbf{x} = 0$. If $C > \mathbf{0}$ we deduce that $\mathbf{x} = 0$, i.e. B is nonsingular. Hence $B^2 > 0$. Suppose next that $A\mathbf{x} = \lambda\mathbf{x}$, and λ is purely imaginary, i.e. $\bar{\lambda} = -\lambda$. Then $\mathbf{x}^*A^* = (A\mathbf{x})^* = (\lambda\mathbf{x})^* = \bar{\lambda}\mathbf{x}^* = -\lambda\mathbf{x}^*$. Hence

$$\mathbf{x}^*C\mathbf{x} = \mathbf{x}^*A^*B\mathbf{x} + \mathbf{x}^*BA\mathbf{x} = -\lambda\mathbf{x}^*B\mathbf{x} + \lambda\mathbf{x}^*B\mathbf{x} = 0.$$

If $C > 0$ we deduce that $\mathbf{x} = 0$, i.e. $\iota_0(A) = 0$. Let $t \in [0, 1]$ and consider $A(t) = (1-t)A + tB$. Then $C(t) := A(t)^*B + BA(t) = (1-t)C + 2tB^2$. If $C > \mathbf{0}$ then $C(t) > \mathbf{0}$ for each $t \in [0, 1]$. Thus $i_0(A(t)) = 0$ for all $t \in [0, 1]$. The eigenvalues of $A(t)$ are continuous functions of $t \in [0, 1]$. Since $i_0(A(t) = 0$ it follows that $i_+(A(t))$ and $i_-(A(t))$ are constant integers, i.e. they do not depend on t. Hence $\iota(A(t)) = \iota(A(0)) = \iota(A) = \iota(A(1)) = \iota(B)$.

Assume now that $A \in \mathbb{C}^{n \times n}$ stable. Fix any $C \in \mathbf{H}_n$ and consider the equation

$$A^*B + BA = C, \quad C = X + \sqrt{-1}\, Y,\ X \in \mathrm{S}_n(\mathbb{R}), Y \in \mathbf{AS}(n, \mathbb{R}). \tag{4.4}$$

This is a system of n^2 real valued equations in n^2 real unknowns: the $\frac{n(n+1)}{2}$ entries of X and the $\frac{n(n-1)}{2}$ entries of Y. We claim that this system has a unique solution. To show that it is enough to show that for $C = \mathbf{0}$ the system (4.4) has a unique solution $B = \mathbf{0}$. Assume to the contrary that the system $A^*B + BA^* = \mathbf{0}$ has a nontrivial solution $\mathbf{0} \neq B \in \mathbf{H}_n$. Then $B = U\,\mathrm{diag}(D, \mathbf{0})U^*$ where $D \in \mathrm{S}_m(\mathbb{R})$ is a diagonal invertible matrix and $\mathbf{0} \in \mathrm{S}_{n-m}(\mathbb{R})$ and $m \in [1, n]$. Let $E = UAU^*$. Then $E^*\,\mathrm{diag}(D, \mathbf{0}) + \mathrm{diag}(D, \mathbf{0})E = \mathbf{0}$. Write E as a block matrix $\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$. Then the matrix equation for E implies that $E_{12} = E_{21} = \mathbf{0}$, i.e. $E = \mathrm{diag}(E_{11}, E_{22})$, and $E_{11}^*D = -DE_{11}$. So $E_{11}^* = -DE_{11}D^{-1}$. The

eigenvalues $-DE_{11}D^{-1}$ are the negative eigenvalues of E_{11}, while the eigenvalues of E_{11}^* are the conjugate of the eigenvalues of E_{11}. Hence each eigenvalue of E_{11} is minus a conjugate of another eigenvalue of E_{11}. This is impossible if A is stable, since all the eigenvalues of A, and hence of E_{11} have negative real parts. Hence $B = \mathbf{0}$, and the system (4.4) has a unique solution $B = \mathbf{H}_n$ for any $C \in \mathbf{H}_n$.

Let $A \in \mathbb{C}^{n\times n}$. Assume first that the system (4.4) for $C = I_n$ has a unique solution $B \in \mathbf{H}_n$. The first part of the theorem shows that A, B are nonsingular and $\iota(A) = \iota(B)$. If $B < \mathbf{0}$ we deduce that $\iota_-(A) = n$, i.e. A is stable.

Suppose now that A is stable, i.e. $\iota_-(A) = n$. The second part of the theorem implies that the system (4.4) has a unique solution $B \in \mathbf{H}_n$ for $C = I_n$. The first part of the theorem implies that $\iota(B) = \iota(A)$. Hence $B < \mathbf{0}$. □

Problems

(4.5)

Show Proposition 4.57.

4.7 Singular Value Decomposition

Let $\mathbf{U}, \mathbf{V}$, be finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, with the inner products $\langle\cdot,\cdot\rangle_{\mathbf{U}}, \langle\cdot,\cdot\rangle_{\mathbf{V}}$ 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 $A \in \mathbb{F}^{m\times n}$. 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 selfadjoint operators. As

$$\langle T^*T\mathbf{v}, \mathbf{v}\rangle_{\mathbf{V}} = \langle T\mathbf{v}, T\mathbf{v}\rangle_{\mathbf{V}} \geq 0, \quad \langle TT^*\mathbf{u}, \mathbf{u}\rangle_{\mathbf{U}} = \langle T^*\mathbf{u}, T^*\mathbf{u}\rangle_{\mathbf{U}} \geq 0$$

it follows that $T^*T \geq 0, TT^* \geq 0$. Let

$$T^*T\mathbf{c}_i = \lambda_i(T^*T)\mathbf{c}_i,\ \langle \mathbf{c}_i, \mathbf{c}_k\rangle_{\mathbf{V}} = \delta_{ik},\ i, k = 1, ..., n, \tag{4.1}$$
$$\lambda_1(T^*T) \geq ... \geq \lambda_n(T^*T) \geq 0,$$
$$TT^*\mathbf{d}_j = \lambda_j(TT^*)\mathbf{d}_j,\ \langle \mathbf{d}_j, \mathbf{d}_l\rangle_{\mathbf{U}} = \delta_{jl},\ j, l = 1, ..., m, \tag{4.2}$$
$$\lambda_1(TT^*) \geq ... \geq \lambda_m(TT^*) \geq 0,$$

Proposition 4.62 *Let $\mathbf{U}, \mathbf{V}$, be finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $T : \mathbf{V} \to \mathbf{U}$. Then* $\operatorname{rank} T = \operatorname{rank} T^* = \operatorname{rank} T^*T = \operatorname{rank} TT^* = r$. *Furthermore the selfadjoint nonnegative definite operators T^*T and TT^* have exactly r positive eigenvalues, and*

$$\lambda_i(T^*T) = \lambda_i(TT^*) > 0, \quad i = 1, ..., \operatorname{rank} T. \tag{4.3}$$

Moreover for $i \in [1, r]$ $T\mathbf{c}_i$ 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 (4.1) then $\tilde{\mathbf{d}}_i := \frac{T\mathbf{c}_i}{||T\mathbf{c}_i||}, i = 1, ..., r$ satisfy (4.2) for $i = 1, ..., r$. Similar result holds for $\mathbf{d}_1, ..., \mathbf{d}_r$.

Proof. Clearly $T\mathbf{x} = 0 \iff \langle T\mathbf{x}, T\mathbf{x}\rangle = 0 \iff T^*T\mathbf{x} = 0$. Hence

$$\operatorname{rank} T^*T = \operatorname{rank} T = \operatorname{rank} T^* = \operatorname{rank} TT^* = r.$$

Thus T^*T and TT^* have exactly r positive eigenvalues. Let $i \in [1, r]$. Then $T^*T\mathbf{c}_i \neq 0$. Hence $T\mathbf{c}_i \neq 0$. (4.1) 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 (4.3) holds. Assume that $\mathbf{c}_1, ..., \mathbf{c}_r$ satisfy (4.1). Let $\tilde{\mathbf{d}}_1, ..., \tilde{\mathbf{d}}_r$ be defined as above. By the definition $||\tilde{\mathbf{d}}_i|| = 1, i = 1, ..., r$. Let $1 \leq i < j \leq r$. 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. □

Let

$$\sigma_i(T) = \sqrt{\lambda_i(T^*T)} \text{ for } i = 1, ...r, \quad \sigma_i(T) = 0 \text{ for } i > r,$$
$$\sigma_{(p)}(T) := (\sigma_1(T), ..., \sigma_p(T))^\top \in \mathbb{R}^p_\searrow, \quad p \in \mathbb{N}. \tag{4.4}$$

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 (4.1). 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 $\text{span}(\mathbf{d}_1, ..., \mathbf{d}_r)$ is spanned by all eigenvectors of TT^* corresponding to nonzero eigenvalues of TT^* it follows that $\ker T^* = \text{span}(\mathbf{d}_{r+1}, ..., \mathbf{d}_m)$. Hence (4.2) holds. In these orthonormal bases of $\mathbf{U}$ and $\mathbf{V}$ the operators T and T^* represented quite simply:

$$T\mathbf{c}_i = \sigma_i\mathbf{d}_i,\ i = 1, ..., n, \quad \text{where } \mathbf{d}_i = 0 \text{ for } i > m,$$
$$T^*\mathbf{d}_j = \sigma_j\mathbf{c}_j,\ j = 1, ..., m, \quad \text{where } \mathbf{c}_j = 0 \text{ for } j > n.. \tag{4.5}$$

Let

$$\Sigma = (s_{ij})_{i,j=1}^{m,n},\ s_{ij} = 0 \text{ for } i \neq j,\ s_{ii} = \sigma_i \text{ for } i = 1, ..., \min(m.n). \tag{4.6}$$

In the case $m \neq n$ we call Σ a quasi 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^* represented by the matrices Σ and $\Sigma^\top$ respectively.

Corollary 4.63 *Let $[\mathbf{u}_1, ..., \mathbf{u}_m], [\mathbf{v}_1, ..., \mathbf{v}_n]$ be orthonormal bases in the vector spaces $\mathbf{U}, \mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ respectively. Then T and T^* are presented by the matrices $A \in \mathbb{F}^{m\times n}$ and $A^* \in \mathbb{F}^{n\times m}$ respectively. Let $U \in \mathbf{U}(m)$ and $V \in \mathbf{U}(n)$ be the unitary matrices representing the change of base $[\mathbf{d}_1, ..., \mathbf{d}_m]$ to $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\mathbf{c}_1, ..., \mathbf{c}_n]$ to $[\mathbf{v}_1, ..., \mathbf{v}_n]$ respectively. (If $\mathbb{F} = \mathbb{R}$ then U and V are orthogonal matrices.) Then*

$$A = U\Sigma V^* \in \mathbb{F}^{m\times n}, \quad U \in \mathbf{U}(m),\ V \in \mathbf{U}(n). \tag{4.7}$$

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 (4.5) to deduce that $U^*AV = \Sigma$. □

Definition 4.64 *(4.7) is called the* singular value decomposition *(SVD) of A.*

Proposition 4.65 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and denote by $\mathcal{R}_{mn,k}(\mathbb{F}) \subset \mathbb{F}^{m\times n}$ the set of all matrices of rank $k \in [1, \min(m, n)]$ at most. Then $A \in \mathcal{R}_{mn,k}(\mathbb{F})$ if and only if A can be expressed as a sum of at most k matrices of rank 1. Furthermore $\mathcal{R}_{mn,k}(\mathbb{F})$ is a variety in $\mathrm{M}_{mn}(\mathbb{F})$ given by the polynomial conditions: Each $(k + 1) \times (k + 1)$ minor of A is equal to zero.*

For the proof see Problem 4.20

Theorem 4.66 *For $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and $A = (a_{ij}) \in \mathbb{F}^{m\times n}$ the following conditions hold:*

$$||A||_F := \sqrt{\operatorname{tr} A^*A} = \sqrt{\operatorname{tr} AA^*} = \sqrt{\sum_{i=1}^{\operatorname{rank} A} \sigma_i(A)^2}. \tag{4.8}$$

$$||A||_2 := \max_{\mathbf{x}\in\mathbb{F}^n, ||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 = \sigma_1(A). \tag{4.9}$$

$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} ||A-B||_2 = \sigma_{k+1}(A), \quad k = 1, ..., \operatorname{rank} A - 1. \tag{4.10}$$

$$\begin{gathered}\sigma_i(A) \geq \sigma_i((a_{i_p j_q})_{p=1,q=1}^{m',n'}) \geq \sigma_{i+(m-m')+(n-n')}(A), \\ m' \in [1,m],\ n' \in [1,n],\ 1 \leq i_1 < ... < i_{m'} \leq m,\ 1 \leq j_1 < ... < j_{n'} \leq n.\end{gathered} \tag{4.11}$$

Proof. The proof of (4.21) is left a Problem 4.21. We now show the equality in (4.9). View A as an operator $A : \mathbb{R}^n \to \mathbb{R}^m$. From the definition of $||A||_2$ it follows

$$||A||_2^2 = \max_{0\neq\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{x}^*A^*A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \lambda_1(A^*A) = \sigma_1(A)^2,$$

which proves (4.9).

We now prove (4.10). In the SVD decomposition of A (4.7) assume that $U = (\mathbf{u}_1, ..., \mathbf{u}_m)$ and $V = (\mathbf{v}_1, ..., \mathbf{v}_n)$. Then (4.7) is equivalent to the following representation of A:

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,\ \mathbf{u}_1, ..., \mathbf{u}_r \in \mathbb{R}^m,\ \mathbf{v}_1, ..., \mathbf{v}_r \in \mathbb{R}^n,\ \mathbf{u}_i^*\mathbf{u}_j = \mathbf{v}_i^*\mathbf{v}_j = \delta_{ij},\ i,j = 1, ..., r, \tag{4.12}$$

where $r = \operatorname{rank} A$. Let $B = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* \in \mathcal{R}_{mn,k}$. Then in view of (4.9)

$$||A - B||_2 = ||\sum_{k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*||_2 = \sigma_{k+1}.$$

Let $B \in \mathcal{R}_{mn,k}$. To show (4.10) 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 $\operatorname{codim} \mathbf{W} \leq k$. 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 an a principal submatrix of A^*A of dimension n'. The interlacing inequalities between the eigenvalues of A^*A and C^*C yields (4.11) 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 (4.11). □

Corollary 4.67 *Let $\mathbf{U}$ and $\mathbf{V}$ be finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $T : \mathbf{V} \to \mathbf{U}$ be a linear operator. Then*

$$||T||_F := \sqrt{\operatorname{tr} T^*T} = \sqrt{\operatorname{tr} TT^*} = \sqrt{\sum_{i=1}^{\operatorname{rank} T} \sigma_i(T)^2}. \tag{4.13}$$

$$||T||_2 := \max_{\mathbf{x}\in\mathbf{V}, ||\mathbf{x}||_2=1} ||T\mathbf{x}||_2 = \sigma_1(A). \tag{4.14}$$

$$\min_{Q\in\mathrm{L}(\mathbf{V},\mathbf{U}), \operatorname{rank} Q\leq k} ||T - Q||_2 = \sigma_{k+1}(T), \quad k = 1, ..., \operatorname{rank} T - 1. \tag{4.15}$$

Theorem 4.68 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and assume that $A \in \mathrm{M}_{mn}(\mathbb{F})$. Define*

$$H(A) = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in \mathrm{H}_{m+n}(\mathbb{F}). \tag{4.16}$$

Then

$$\begin{aligned} &\lambda_i(H(A)) = \sigma_i(A),\ \lambda_{m+n+1-i}(H(A)) = -\sigma_i(A),\ i = 1, ..., \operatorname{rank} A, \\ &\lambda_j(H(A)) = 0,\ j = \operatorname{rank} A + 1, ..., n + m - \operatorname{rank} A. \end{aligned} \tag{4.17}$$

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 and in Proposition 4.62 respectively. Then

$$\begin{aligned} &\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix} = \sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix}, \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix} = -\sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix}, \\ &\quad i = 1, ..., \operatorname{rank} A, \\ &\ker H(A) = \operatorname{span}((\mathbf{d}_{r+1}^*, 0)^*, ..., (\mathbf{d}_m^*, 0)^*, (0, \mathbf{c}_{r+1}^*)*, ..., (0, \mathbf{c}_n^*)^*),\ r = \operatorname{rank} A. \end{aligned} \tag{4.18}$$

Proof. It is straightforward to show the equalities (4.18). Since all the eigenvectors appearing in (4.18) are linearly independent we deduce (4.17). □

Theorem 4.69 *Let $A, B \in \mathbb{C}^{m\times n}$, and assume that $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq 0, \sigma_1(B) \geq \sigma_2(B) \geq \ldots \geq 0$, where $\sigma_i(A) = 0$ and $\sigma_j(B) = 0$ for $i >$ rank A and $j >$ rank B respectively. Then*

$$-\sum_{i=1}^{m} \sigma_i(A)\sigma_i(B) \leq \Re \operatorname{tr} AB^* = \Re \operatorname{tr} A^*B \leq \sum_{i=1}^{m} \sigma_i(A)\sigma_i(B). \tag{4.19}$$

Equality holds if the A and B have common left and the right eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n$ corresponding to $\sigma_1(A), \ldots$ and $\sigma_1(B), \ldots$, respectively.

Proof. Note that $\operatorname{tr} AB^* = \operatorname{tr} B^*A$ and $\Re \operatorname{tr} AB^* = \Re \operatorname{tr}(AB^*)^* = \Re \operatorname{tr} BA^*$. Observe next that $\operatorname{tr} H(A)H(B) = 2\Re \operatorname{tr} AB^*$. Combine Theorems 4.68 and 4.44 to deduce the theorem. □

Corollary 4.70 *For $A \in \mathbb{C}^{m\times n}$ $\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} ||A - B||_F^2 = \sum_{i=k+1}^{m} \sigma_i(A)^2$.*

Proof. Let $B \in \mathbb{C}^{m\times n}$. Then

$$||A-B||_F^2 = \operatorname{tr}(A-B)(A^*-B^*) = ||A||_F^2+||B||_F^2-2\Re \operatorname{tr} AB^* = \sum_{i=1} \sigma_i(A)^2+\sigma_i(B)^2-2\Re \operatorname{tr} AB^*$$

Assume now that $B \in \mathcal{R}_{m,n,k}$. Then rank $B \leq k$ and denote by $x_1 \geq \ldots x_k \geq 0$ the nonzero singular values of B. Use Theorem 4.69 to deduce that

$$\begin{aligned} ||A - B||_F \geq ||A||_F^2 + \sum_{i=1}^{k} x_i^2 - 2\sum_{i=1}^{k} \sigma_i(A)x_i = \sum_{i=k+1}^{n} \sigma_i(A)^2 + \sum_{i=1}^{k} (x_i - \sigma_i(A))^2 \\ \geq \sum_{i=k+1}^{n} \sigma_i(A)^2. \end{aligned}$$

Let $A = \sum_{i=1}^{m} \sigma_i(A)\mathbf{u}_i\mathbf{v}_i^*$ is the singular decomposition of A. Choose $B = \sum_{i=1}^{k} \sigma_i(A)\mathbf{u}_i\mathbf{v}_i^* \in \mathcal{R}_{m.n,k}$ to see that $||A - B||_F^2 = \sum_{i=k+1}^{m} \sigma_i(A)^2$. □

Define by $\mathbb{R}^n_{+,\searrow} := \mathbb{R}^n_{\searrow} \cap \mathbb{R}^n_+$. Then $D \subset \mathbb{R}^n_{+,\searrow}$ is called a strong Schur set if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{+,\searrow}, \mathbf{x} \preceq \mathbf{y}$ we have the implication $\mathbf{y} \in D \Rightarrow \mathbf{x} \in D$.

Theorem 4.71 *Let $p \in \mathbb{N}$ and $D \subset \mathbb{R}^p_{\searrow} \cap \mathbb{R}^p_+$ be a regular convex strong Schur domain. Fix $m, n \in \mathbb{N}$ and let $\sigma_{(p)}(D) := \{A \in \mathbb{F}^{m\times n} : \sigma_{(p)}(A) \in D\}$. Let $h : D \to \mathbb{R}$ be a convex and strongly Schur's order preserving on D. Let $f : \sigma_{(p)} :\to \mathbb{R}$ be given as $h \circ \sigma_{(p)}$. Then f is a convex function.*

See Problem **??**.

Corollary 4.72 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$, $m, n, p \in \mathbb{N}$, $q \in [1, \infty)$ and $w_1 \geq w_2 \geq ... \geq w_p > 0$. Then the following function*

$$f : \mathbb{F}^{m\times n} \to \mathbb{R}, \quad f(A) := (\sum_{i=1}^{p} w_i \sigma_i(A)^q)^{\frac{1}{q}}, \ A \in \mathbb{F}^{m\times n}$$

is a convex function.

See Problem 4.22

Theorem 4.73 *Let $\mathbf{U}$ be IPS over $\mathbb{C}$. Let $T : \mathbf{U} \to \mathbf{U}$ be a linear operator. Then $\rho(T) \leq ||T||_2$. Furthermore equality holds if and only if the following conditions hold.*
(a) T and T^* have a common eigenvector $\mathbf{x}$ such that $T\mathbf{x} = \lambda\mathbf{x}$, $T^*\mathbf{x} = \bar{\lambda}\mathbf{x}$ and $|\lambda| = \rho(T)$.
(b) Let T_1 be the restriction of T to the invariant subspace $\mathbf{V} := \text{span}(\mathbf{x})^\perp$. Then $||T_1||_2 \leq \rho(T)$.

Proof. Let $T\mathbf{x} = \lambda\mathbf{x}$ where $||\mathbf{x}|| = 1$ and $\rho(T) = |\lambda|$. Recall $||T||_2 = \sigma_1(T)$, 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 $|\lambda|^2 = \langle T\mathbf{x}, T\mathbf{x}\rangle = \langle T^*T\mathbf{x}, \mathbf{x}\rangle \leq \lambda_1(T^*T) = ||T||_2^2$. Hence $\rho(T) \leq ||T||_2$.

Assume now that $\rho(T) = ||T||_2$. $\rho(T) = 0$ then $||T||_2 = 0 \Rightarrow T = \mathbf{0}$, and theorem holds trivially n 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 $\mathbf{U} = \text{span}(\mathbf{x})^\perp$ be the orthogonal complement of $\text{span}(\mathbf{x})$. Since $T\text{span}(\mathbf{x}) = \text{span}(\mathbf{x})$ it follows that $T^*\mathbf{U} \subseteq \mathbf{U}$. Similarly, since $T^*\text{span}(\mathbf{x}) = \text{span}(x)$ $T\mathbf{U} \subseteq \mathbf{U}$. Thus $\mathbf{V} = \text{span}(\mathbf{x}) \oplus \mathbf{U}$ and $\text{span}(\mathbf{x}), \mathbf{U}$ are invariant subspaces of T and T^*. Hence $\text{span}(\mathbf{x}), \mathbf{U}$ 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 $||T_1||_2^2 = \lambda_1(T_1 * T_1) \geq \lambda_1(T^*T) = ||T||_2^2$. This establishes the second part of theorem, labeled (a) and (b).

The above result imply that the conditions (a) and (b) of the theorem yield the equality $\rho(T) = ||T||_2$. □

Corollary 4.74 *Let $\mathbf{U}$ be an n-dimensional IPS over $\mathbb{C}$. Let $T : \mathbf{U} \to \mathbf{U}$ be a linear operator. Denote by $|\lambda(T)| = (|\lambda_1(T)|, ..., |\lambda_n(T)|)^\top$ 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.*

Problems

(4.20)

Prove Proposition 4.65. (Use SVD to prove the nontrivial part of the Proposition.)

(4.21)

Prove the equalities in (4.8).

(4.22)

a. Prove Corollary 4.72
b. Recall the definition of a norm on a vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Show that the function f defined in Corollary 4.72 is a norm. For $p = \min(m, n)$ and $w_1 = ... = w_p = 1$ this norm is called the $q - Schatten$ norm.

1. Let $A \in \mathrm{S}_n(\mathbb{R})$ and assume the $A = Q^T \Lambda Q$, where $Q \in \mathbf{O}(n, \mathbb{R})$ and $\Lambda = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ is a diagonal matrix, where $|\alpha_1| \geq \ldots \geq |\alpha_n| \geq 0$.

   (a) Find the SVD of A.

   (b) Show that $\sigma_1(A) = \max(\lambda_1(A), |\lambda_n(A)|)$, where $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$ are the n eigenvalues of A arranged in a decreasing order.

2. Let k, m, n be a positive integers such that $k \leq \min(m, n)$. Show that the function $f : \mathbb{R}^{m \times n} : [0, \infty)$ given by $f(A) = \sum_{i=1}^{k} \sigma_i(A)$ is a convex function on $\mathbb{R}^{m \times n}$.

4.8 Moore-Penrose generalized inverse

Let $A \in \mathbb{C}^{m \times n}$. Then (4.12) is called the *reduced* SVD of A. It can be written as

$$\begin{gathered} A = U_r \Sigma_r V_r^*, \quad r = \mathrm{rank}\, A, \quad \Sigma_r := \mathrm{diag}(\sigma_1(A), \ldots, \sigma_r(A)) \in S_r(\mathbb{R}), \\ U_r = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{C}^{m \times r}, V_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{C}^{n \times r}, U_r^* U_r = V_r^* V_r = I_r, . \end{gathered} \tag{4.23}$$

Recall that

$$AA^* \mathbf{u}_i = \sigma_i(A)^2 \mathbf{u}_i, A^* A \mathbf{v}_i = \sigma_i(A)^2 \mathbf{v}_i, \mathbf{v}_i = \frac{1}{\sigma_i(A)} A^* \mathbf{u}_i, \mathbf{u}_i = \frac{1}{\sigma_i(A)} A \mathbf{v}_i, i = 1, \ldots, r.$$

Then

$$A^\dagger := V_r \Sigma_r^{-1} U_r^* \in \mathbb{C}^{n \times m} \tag{4.24}$$

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, V are real values matrices over the real numbers $\mathbb{R}$.

Theorem 4.75 *Let $A \in \mathbb{C}^{m \times n}$ matrix. Then the Moore-Penrose generalized inverse $A^\dagger \in \mathbb{C}^{n \times m}$ satisfies the following properties.*

1. $\mathrm{rank}\, A = \mathrm{rank}\, A^\dagger$.

2. $A^\dagger A A^\dagger = A^\dagger$, $AA^\dagger A = A$, $A^* A A^\dagger = A^\dagger A A^* = A^*$.

3. *$A^\dagger A$ and $AA^\dagger$ are Hermitian nonnegative 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 $\mathrm{rank}\, A = n$ then $A^\dagger = (A^* A)^{-1} A^*$. In particular, if $A \in \mathbb{C}^{n \times n}$ is invertible then $A^\dagger = A^{-1}$.*

Proposition 4.76 *Let $E \in \mathbb{C}^{l \times m}, G \in \mathbb{C}^{m \times n}$. Then $\mathrm{rank}\, EG \leq \min(\mathrm{rank}\, E, \mathrm{rank}\, G)$. If $l = m$ and E is invertible then $\mathrm{rank}\, EG = \mathrm{rank}\, G$. If $m = n$ and G is invertible then $\mathrm{rank}\, EG = \mathrm{rank}\, G$.*

Proof. Let $\mathbf{e}_1, \ldots, \mathbf{e}_m \in \mathbb{C}^l, \mathbf{g}_1, \ldots, \mathbf{g}_n \in \mathbb{C}^m$ be the columns of E and G respectively. Then $\mathrm{rank}\, E = \dim \mathrm{span}(\mathbf{e}_1, \ldots, \mathbf{e}_l)$. Observe that $EG = [E\mathbf{g}_1, \ldots, E\mathbf{g}_n] \in \mathbb{C}^{l \times n}$. Clearly $E\mathbf{g}_i$ is a linear combination of the columns of E. Hence $E\mathbf{g}_i \in \mathrm{span}(\mathbf{e}_1, \ldots, \mathbf{e}_l)$. Therefore $\mathrm{span}(E\mathbf{g}_1, \ldots, E\mathbf{g}_n) \subseteq \mathrm{span}(\mathbf{e}_1, \ldots, \mathbf{e}_l)$, which implies that $\mathrm{rank}\, EG \leq \mathrm{rank}\, E$. Note that $(EG)^T = G^T E^T$. Hence $\mathrm{rank}\, EG = \mathrm{rank}\, (EG)^T \leq \mathrm{rank}\, G^T = \mathrm{rank}\, G$. Thus $\mathrm{rank}\, EG \leq \min(\mathrm{rank}\, E, \mathrm{rank}\, G)$. Suppose E is invertible. Then $\mathrm{rank}\, EG \leq \mathrm{rank}\, G = \mathrm{rank}\, E^{-1}(EG) \leq \mathrm{rank}\, EG$. Hence $\mathrm{rank}\, EG = \mathrm{rank}\, G$. Similarly $\mathrm{rank}\, EG = \mathrm{rank}\, E$ if G is invertible. □

Proof of Theorem 4.75.

1. Proposition 4.76 yields that rank $A^\dagger$ = rank $V_r\Sigma_r^{-1}U_r^* \leq$ rank $\Sigma_r^{-1}U_r^* \leq$ rank $\Sigma_r^{-1} = r =$ rank A. Since $\Sigma_r = V_r^*A^\dagger U_r$ Proposition 4.76 yields that rank $A^\dagger \geq$ 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. Hence $AA^\dagger$ is nonnegative 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 $||\mathbf{x}|| \geq ||A^\dagger\mathbf{b}||$ 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 (4.7). 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 $\mathbf{z} = (z_1, \ldots, z_n)^T := V^*\mathbf{x}$, $\mathbf{c} = (c_1, \ldots, c_m)^T := U^*\mathbf{b}$. 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 $z_1, \ldots, z_n$ variables is given in the form $\sigma_i(A)^2z_i = \sigma_i(A)c_i$ for $i = 1, \ldots, n$. Since $\sigma_i(A) = 0$ for $i > r$ we obtain that $z_i = \frac{1}{\sigma_i(A)}c_i$ for $i = 1, \ldots, r$ while $z_{r+1}, \ldots, z_n$ 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 $i = r+1, \ldots, n$. This solution corresponds 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 $\mathbf{b}$ $(A^*A)^{-1}A^*\mathbf{b} = A^\dagger\mathbf{b}$, hence $A^\dagger = (A^*A)^{-1}A^*$.

   If A is an $n \times n$ matrix and is invertible it follows that $(A^*A)^{-1}A^* = A^{-1}(A^*)^{-1}A^* = A^{-1}$. □

Problems

1. $P \in \mathbb{C}^{n\times n}$ is called a *projection* if $P^2 = P$. Show that P is a projection if and only if the following two conditions are satisfied:

   - Each eigenvalue of P is either 0 or 1.
   - P is a diagonable matrix.

2. $P \in \mathbb{R}^{n\times n}$ 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 $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbf{V}$ $||\mathbf{a} - \mathbf{b}|| \geq ||\mathbf{a} - P\mathbf{a}||$ and 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.

3. Let $A \in \mathbb{R}^{m\times n}$ and assume that the SVD of A is given by (4.7), 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!

4.9 Rank-constrained matrix approximations

Let $A \in \mathbb{C}^{m\times n}$ and assume that $A = U_A\Sigma_A V_A^*$ be the SVD of A given in (4.7). Let $U_A = [\mathbf{u}_1\ \mathbf{u}_2\ \dots \mathbf{u}_m], V_A = [\mathbf{v}_1\ \mathbf{v}_2\ \dots \mathbf{v}_n]$ be the representations of U, V in terms of their m, n columns respectively. Then

$$P_{A,L} := \sum_{i=1}^{\text{rank } A} \mathbf{u}_i\mathbf{u}_i^* \in \mathbb{C}^{m\times m}, \quad P_{A,R} := \sum_{i=1}^{\text{rank } A} \mathbf{v}_i\mathbf{v}_i^* \in \mathbb{C}^{n\times n}, \tag{4.25}$$

are the orthogonal projections on the range of A and A^* respectively. Denote by $A_k := \sum_{i=1}^k \sigma_i(A)\mathbf{u}_i\mathbf{v}_i^* \in \mathbb{C}^{m\times n}$ for $k = 1, \dots, \text{rank } A$. For $k > \text{rank } A$ we define $A_k := A$ $(= A_{\text{rank } A})$. For $1 \le k < \text{rank } A$, the matrix A_k is uniquely defined if and only if $\sigma_k(A) > \sigma_{k+1}(A)$.

Theorem 4.77 *Let $A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{m\times p}, C \in \mathbb{C}^{q\times n}$ be given. Then $X = B^\dagger(P_{B,L}AP_{C,R})_k C^\dagger$ is a solution to the minimal problem*

$$\min_{X\in\mathcal{R}(p,q,k)} ||A - BXC||_F, \tag{4.26}$$

having the minimal $||X||_F$. This solution is unique if and only if either $k \ge \text{rank } P_{B,L}AP_{C,R}$ or $1 \le k < \text{rank } P_{B,L}AP_{C,R}$ and $\sigma_k(P_{B,L}AP_{C,R}) > \sigma_{k+1}(P_{B,L}AP_{C,R})$.

Proof. Recall that the Frobenius norm is invariant under the multiplication from the left and the right by the corresponding unitary matrices. Hence $||A-BXC||_F = ||A_1-\Sigma_B Y\Sigma_C||$, where $\tilde{A} := U_B^*AV_C, \tilde{X} := V_B^*XU_C$. Clearly, X and $\tilde{X}$ have the same rank and the same Frobenius norm. Thus it is enough to consider the minimal problem $\min_{Y\tilde{X}\in\mathcal{R}(p,q,k)} ||\tilde{A} - \Sigma_B\tilde{X}\Sigma_C||_F$. Let $s = \text{rank } B, t = \text{rank } B$. Clearly if B or C is a zero matrix, then $X = \mathbf{0}$ is the solution to the minimal problem (4.26). In this case either $P_{B,L}$ or $P_{C,R}$ are zero matrices, and the theorem holds trivially in this case.

It is left to consider the case $1 \le s, 1 \le t$. Define $B_1 := \text{diag}(\sigma_1(B), \dots, \sigma_s(B)) \in \mathbb{C}^{s\times s}, C_1 := \text{diag}(\sigma_1(C), \dots, \sigma_t(C)) \in \mathbb{C}^{t\times t}$. Partition $\tilde{A}$ and $\tilde{X}$ to 2×2 block matrices $\tilde{A} = [A_{ij}]_{i,j=1}^2$ and $\tilde{X} = [X_{ij}]_{i,j=1}^2$, where $A_{11}, X_{11} \in \mathbb{C}^{s\times t}$. (For certain values of s and t, we may have to partition $\tilde{A}$ or $\tilde{X}$ to less than 2×2 block matrices.) Observe next that $Z := \Sigma_B Y\Sigma_C = [Z_{ij}]_{i,j=1}^2$, where $Z_{11} = B_1Y_{11}C_1$ and all other blocks Z_{ij} are zero matrices. Hence

$$||\tilde{A} - Z||_F^2 = ||A_{11} - Z_{11}||_F^2 + \sum_{2<i+j\le 4} ||A_{ij}||_F^2 \ge ||A_{11} - (A_{11})_k||_F^2 + \sum_{2<i+j\le 4} ||A_{ij}||_F^2.$$

Thus $\hat{X} = [X_{ij}]_{i,j=1}^2$, where $X_{11} = B_1^{-1}(A_{11})_kC_1^{-1}$ and $X_{ij} = \mathbf{0}$ for all $(i,j) \ne (1,1)$ is a solution $\min_{Y\tilde{X}\in\mathcal{R}(p,q,k)} ||\tilde{A} - \Sigma_B\tilde{X}\Sigma_C||_F$ with the minimal Frobenius form. This solution is unique if and only if the solution $Z_{11} = (A_{11})_k$ is the unique solution to $\min_{Z_{11}\in\mathcal{R}(s,t,k)} ||A_{11} - Z_{11}||_F$. This happens if either $k \ge \text{rank } A_{11}$ or $1 \le k < \text{rank } A_{11}$ and $\sigma_k(A_{11}) > \sigma_{k+1}(A_{11})$. A straightforward calculation shows that $\hat{X} = \Sigma_B^\dagger(P_{\Sigma_B,L}\tilde{A}P_{\Sigma_C,R})_k\Sigma_C^\dagger$. This shows that $X = B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger$ is a solution of (4.26) with the minimal Frobenius norm. This solution is unique if and only if either $k \ge \text{rank } P_{B,L}AP_{C,R}$ or $1 \le k < \text{rank } P_{B,L}AP_{C,R}$ and $\sigma_k(P_{B,L}AP_{C,R}) > \sigma_{k+1}(P_{B,L}AP_{C,R})$. □

4.10 Generalized Singular Value Decomposition

See [4] for more details on this section.

Proposition 4.78 *Let $\mathbf{0} < M \in \mathrm{S}_m(\mathbb{R}), \mathbf{0} < N \in \mathrm{S}_n(\mathbb{R})$ be positive definite symmetric matrices. Let $\langle \mathbf{x}, \mathbf{y}\rangle_M := \mathbf{y}^T M^2\mathbf{x}, \langle \mathbf{u}, \mathbf{v}\rangle_N := \mathbf{v}^T N^2\mathbf{u}$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m, \mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, be inner*

products in $\mathbb{R}^m, \mathbb{R}^n$ *respectively. Let* $A \in \mathbb{R}^{m\times n}$ *and view* A *as a linear operator* $A : \mathbb{R}^n \to \mathbb{R}^m, \mathbf{x} \mapsto A\mathbf{x}$. *Denote by* $A^c : \mathbb{R}^m \to \mathbb{R}^n$ *the adjoint operator with respect to the inner products* $\langle\cdot,\cdot\rangle_M, \langle\cdot,\cdot\rangle_N$. *That is* $\langle A\mathbf{u}, \mathbf{y}\rangle_M = \langle \mathbf{u}, A^c\mathbf{y}\rangle_N$. *Then* $A^c = N^{-2}A^T M^2$.

Proof. Clearly $\langle A\mathbf{u}, \mathbf{y}\rangle_M = \mathbf{y}^T M^2 A\mathbf{u}, \langle \mathbf{u}, A^c\mathbf{y}\rangle_N = (A^c\mathbf{y})^T N^2\mathbf{u}$. Hence $(A^c)^T N^2 = M^2 A$. Take the transpose of this identity and divide by N^{-2} from the left to deduce $A^c = N^{-2}A^T M^2$. □

Theorem 4.79 *Let the assumptions of Proposition 4.78 hold. Then the generalized singular value decomposition (GSVD) of* A *is*

$$A = U\Sigma V^T, \Sigma = \operatorname{diag}(\sigma_1, \dots,) \in \mathbb{R}^{m\times n}, \sigma_1 \geq \dots \sigma_r > 0,\ \sigma_i = 0 \text{ for } i > r := \operatorname{rank} A, \tag{4.27}$$
$$U \in \mathrm{GL}(m, \mathbb{R}), V \in \mathrm{GL}(n, \mathbb{R}), U^T M^2 U = I_m, V^T N^{-2} V = I_m.$$

Proof. Identify $\mathbb{R}^n$ and $\mathbb{R}^m$, with the inner products $\langle\cdot,\cdot\rangle_N, \langle\cdot,\cdot\rangle_M$, with IPS $\mathbf{V}, \mathbf{U}$ respectively. Identify A, A^c with the linear operators $T : \mathbf{V} \to V, T^* : \mathbf{U} \to \mathbf{V}$ respectively. Apply Proposition 4.62. Then $\mathbf{v}_1, \dots, \mathbf{v}_n$ and $\mathbf{u}_1, \dots, \mathbf{u}_m$ are orthonormal sets of eigenvectors of $A^c A$ and AA^c respectively, corresponding to the eigenvalues $\sigma_1^2, \dots, \sigma_n^2$ and $\sigma_1^2, \dots, \sigma_m^2$:

$$\begin{aligned}
&N^{-2}A^T M^2 A\mathbf{v}_i = \sigma_i^2\mathbf{v}_i,\ \mathbf{v}_j^T N^2\mathbf{v}_i = \delta_{ij},\ i, j = 1, \dots, n, V = N^2[\mathbf{v}_1\ \dots\ \mathbf{v}_n]\\
&AN^{-2}A^T M^2\mathbf{u}_i = \sigma_i^2\mathbf{u}_i,\ \mathbf{u}_j^T M^2\mathbf{u}_i = \delta_{ij},\ i, j = 1, \dots, m, U = [\mathbf{u}_1\ \dots\ \mathbf{u}_m],\\
&\mathbf{u}_i = \frac{1}{\sigma_i}A\mathbf{v}_i,\ i = 1, \dots, r = \operatorname{rank} A.
\end{aligned} \tag{4.28}$$

To justify the decomposition $A = U\Sigma V^T$, choose a vector $\mathbf{v} \in \mathbb{R}^n$ and write it up as $\mathbf{v} = \sum_{i=1}^n v_i\mathbf{v}_i$. Then $A\mathbf{v} = \sum_{i=1}^n Av_i$. Since $\sigma_i = 0$ for $i > r$ it follows that $A\mathbf{v}_i = \mathbf{0}$. Also $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$ for $i = 1, \dots, r$. Hence $A\mathbf{v} = \sum_{i=1}^r v_i\sigma_i\mathbf{u}_i$. Compare that with $U\Sigma V^T\mathbf{v}_i = U\Sigma[\mathbf{v}_1\ \dots\ \mathbf{v}_n]^T N^2\mathbf{v}_i$, which us equal to $\sigma_i\mathbf{u}_i$ if $i \leq r$ and $\mathbf{0}$ if $i > r$. Hence $A = U\Sigma V^T$. □

Corollary 4.80 *Let the assumptions and the notations of Theorem 4.79 hold. Then for* $k \in [1, r]$

$$A_k := U_k\Sigma_k V_k^T = \sum_{i=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^T N^2, \quad U_k \in \mathbb{R}^{m\times k}, V_k \in \mathbb{R}^{n\times k}, \tag{4.29}$$
$$\Sigma_k := \operatorname{diag}(\sigma_1, \dots, \sigma_k), U_k^T M^2 U_k = V_k^T N^{-2} V_k = I_k, U_k = [\mathbf{u}_1, \dots, \mathbf{u}_k], V_k = N^2[\mathbf{v}_1, \dots, \mathbf{v}_k]$$

is the best rank k*-approximation to* A *in the Frobenius and the operator norms with respect to the inner products* $\langle\cdot,\cdot\rangle_M, \langle\cdot,\cdot\rangle_N$ *on* $\mathbb{R}^m, \mathbb{R}^n$ *respectively.*

Theorem 4.81 *Let* $A \in \mathbb{R}^{m\times n}$ *and* $B \in \mathbb{R}^{l\times n}$. *Then there exists a generalized (common) singular value decomposition of* A *and* B*, named GSVD, of the form*

$$\begin{aligned}
&A := U_r\Sigma_r(A)V_r^T, \quad B = W_r\Sigma_r(B)V_r^T,\\
&\Sigma_r(A) = \operatorname{diag}(\sigma_1(A), \dots, \sigma_r(A)), \quad \Sigma_r(B) = \operatorname{diag}(\sigma_1(B), \dots, \sigma_r(B)),\\
&\sigma_i(A)^2 + \sigma_i(B)^2 = 1,\ \text{for } i = 1, \dots, r, \quad r := \operatorname{rank}[A^T\ B^T]\\
&U_r^T U_r = W_r^T W_r = V_r^T N^{-2} V_r = I_r.
\end{aligned} \tag{4.30}$$

Let $\mathbf{V} \subseteq \mathbb{R}^n$ *be the subspace spanned by the columns of* A^T *and* B^T. *Then* $\mathbf{0} \leq N \in \mathrm{S}_n(\mathbb{R})$ *is any positive definite matrix such that* $N\mathbf{V} = \mathbf{V}$ *and* $N^2|\mathbf{V} = A^T A + B^T B|\mathbf{V}$. *Furthermore, the GSVD of* A *and* B *is obtained as follows. Let* $P := A^T A + B^T B$. *Then* $\operatorname{rank} P = r$. *Let*

$$\begin{aligned}
&P := Q_r\Omega_r^2 Q_r^T,\ Q_r^T Q_r = I_r, Q_r := [\mathbf{q}_1\ \dots\ \mathbf{q}_r], \Omega_r := \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r}),\\
&P\mathbf{q}_i = \lambda_i\mathbf{q}_i, \mathbf{q}_j^T\mathbf{q}_i = \delta_{ij}, i, j = 1, \dots, n, \lambda_1 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_n,
\end{aligned} \tag{4.31}$$

be the spectral decomposition of P. Define

$$C_A := \Omega_r^{-1} Q_r^T A^T A Q_r \Omega_r^{-1}, \quad C_B := \Omega_r^{-1} Q_r^T B^T B Q_r \Omega_r^{-1} \in R^{r\times r}.$$

Then $C_A + C_B = I_r$. Let $C_A = R\Sigma_r(A)^2 R^T, R \in \mathbb{R}^{r\times r}, R^T R = I_r$ be the spectral decomposition of C_A. Then $C_B = R\Sigma_r(B)^2 R^T$ is the spectral decomposition of C_B. Furthermore $V = Q_r\Omega_r R$. The nonzero orthonormal columns of U_r and W_r corresponding to positive singular values $\sigma_i(A)$ and $\sigma_j(B)$ are uniquely determined by the equalities $U_r\Sigma_r(A) = AQ_r\Omega_r^{-1}R$ and $W_r\Sigma_r(B) = BQ_r\Omega_r^{-1}R$. Other columns of U_r and W_r is any set of orthonormal vectors in $\mathbb{R}^m$ and $\mathbb{R}^l$ respectively, which are orthogonal to the previously determined columns of U_r and W_r respectively.

Proof. We first prove the identities (4.30). Assume that $Q_r\Omega_r^{-1}R = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$. Clearly, range P = span$(\mathbf{q}_1, \ldots, \mathbf{q}_r)$ = span$(\mathbf{v}_1, \ldots, \mathbf{v}_r)$, and range (A^T), range $(B^T) \subseteq$ range P. Hence ker P = span$(\mathbf{u}_1, \ldots, \mathbf{u}_r)^\perp \subseteq \ker A, \ker P \subseteq \ker B$.

Let $U_r = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_r] \in \mathbb{R}^{m\times r}$. From the equality $U_r\Sigma_r(A) = AQ_r\Omega_r^{-1}R$ we deduce that $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$ for $i = 1, \ldots$. The equality $\Sigma_r(A)U_r^T U_r\Sigma_r(A) = (AQ_r\Omega_r^{-1}R)^T(AQ_r\Omega_r^{-1}R) = \Sigma_r(A)^2$ implies that the columns of U_r corresponding to positive $\sigma_i(A)$ form an orthonormal system. Since $V = Q_r\Omega_r R$ we obtain $V^T Q_r\Omega_r^{-1}R = I_r$. Hence $U_r\Sigma_r(A)V^T\mathbf{v}_i = \sigma_i(A)\mathbf{u}_i$ for $i = 1, \ldots, r$. Therefore $A = U_r\Sigma_r(A)V_r^T$. Similarly $B = W_r\Sigma_r(A)V_r^T$.

From the definitions of P, Q_r, C_A, C_B it follows that $C_A + C_B = I_r$. Hence $\Sigma_1(A)^2 + \Sigma_r(B)^2 = I_r$. Other claims of the theorem follow straightforward. □

We now discuss a numerical example in [4] which shows the sensitivity of GSVD of two matrices. We first generate at random two matrices $A_0 \in \mathbb{R}^{8\times 7}$ and $B_0 \in \mathbb{R}^{9\times 7}$, where rank A_0 = rank $B_0 = 2$ and rank $[A_0^T\ B_0^T] = 3$. These is done as follows. Choose at random $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^8, \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^9, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 \in \mathbb{R}^7$. Then $A_0 = \mathbf{x}_1\mathbf{z}_1^T + \mathbf{x}_2\mathbf{z}_2^T, B_0 = \mathbf{y}_1\mathbf{z}_1^T + \mathbf{y}_2\mathbf{z}_3^T$. The first three singular values of A_0, B_0 are given as follows.

$$27455.5092631633888,\ 17374.6830503566089,\ 3.14050409246786192 \times 10^{-12},$$

$$29977.5429571960522,\ 19134.3838220483449,\ 3.52429226420727071 \times 10^{-12},$$

i.e. the ranks of A_0 and B_0 are 2 within the double digit precision. The four first singular values of $P_0 = A_0^T A_0 + B_0^T B_0$ are

$$1.32179857269680762 \times 10^9,\ 6.04366385186753988 \times 10^8,$$
$$3.94297368116438210 \times 10^8,\ 1.34609524647135614 \times 10^{-7}.$$

Again rank $P_0 = 3$ within double digit precision. The 3 generalized singular values of A_0 and B_0 given by Theorem 4.81 are:

$$\phi_1 = 1,\ \phi_2 = 0.6814262563, \phi_3 = 3.777588180 \times 10^{-9},$$
$$\psi_1 = 0,\ \psi_2 = 0.7318867789,\ \psi_3 = 1.$$

So A_0 and B_0 have one "common" generalized right singular vector $\mathbf{v}_2$ with the corresponding singular value ϕ_2, ψ_2, which are relatively close numbers. The right singular vector $\mathbf{v}_1$ is present only in A_0 and the right singular vector $\mathbf{v}_3$ is present only in B_0.

We next perturb A_0, B_0 by letting $A = A_0 + X, B = B_0 + Y$, where $X \in \mathbb{R}^{8\times 7}, Y \in \mathbb{R}^{9\times 7}$. The entries of X and Y were chosen each at random. The 7 singular values of X and Y are given as follows:

$$(27490, 17450, 233, 130, 119, 70.0, 18.2),\ (29884, 19183, 250, 187, 137, 102, 19.7).$$

Note that $||X|| \sim 0.01||A_0||,\ ||Y|| \sim 0.01||B_0||$. Form the matrices $A := A_0 + X,\ B := B_0 + Y$. These matrices have the full ranks with corresponding singular values rounded off to three significant digits at least:

$$(27490, 17450, 233, 130, 119, 70.0, 18.2),\ (29884, 19183, 250, 187, 137, 102, 19.7).$$

We now replace A, B by A_1, B_1 of rank two using the first two singular values and the corresponding singular vectors in the SVD decompositions of A, B. Then two nonzero singular values of A_1, B_1 are $(27490, 17450), (29883, 19183)$, rounded to five significant digits. The singular values of the corresponding $P_1 = A_1^T A_1 + B_1^T B_1$ are up to 3 significant digits:

$$(1.32 \times 10^9, 6.07 \times 10^8, 3.96 \times 10^8, 1.31 \times 10^4, 0.068, 9.88 \times 10^{-3}, 6.76 \times 10^{-3}). \quad (4.32)$$

Assume that $\tilde{r} = 3$, i.e. P has three significant singular values. We now apply our Theorem 4.81. The three generalized singular values of A_1, B_1 are

$$(1.000000000, .6814704276, 0.7582758358 \times 10^{-8}), \quad (0., .7318456506, 1.0).$$

These result match the generalized singular values of A_0, B_0 at least up to four significant digits. Let $\hat{V}, \hat{U}_1, \hat{W}_1$ be the matrix V, the first two columns of U, the last two columns of W, which are computed for A_1, B_1. Then

$$\frac{||V - \hat{V}||}{||V||} \sim 0.0061, \ ||U_1 - \hat{U}_1|| \sim 0.0093, \ ||W_1 - \hat{W}_1|| \sim 0.0098.$$

Finally we discuss the critical issue of choosing correctly the number of significant singular values of noised matrices A, B and the corresponding matrix $P = A^T A + B^T B$. Assume that the numerical rank of P_1 is 4. That is in Theorem 4.81 assume that $r = 4$. Then the four generalized singular values of A_1, B_1 up to six significant digits are $(1, 1, 0, 0), (0, 0, 1, 1)$!

5 Tensors

5.1 Introduction

The common notion of *tensors* in mathematics is associated with differential geometry, covariant and contravariant derivatives, Christofel symbols and Einstein theory of general relativity. In engineering and other mundane applications as biology, psychology, the relevant notions in mathematics are related to *multilinear algebra*. Of course the notions in the two mentioned fields are related.

5.2 Tensor product of two vector spaces

Given two vector spaces $\mathbf{U}, \mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ one first defines one defines the $\mathbf{U} \otimes \mathbf{V}$ as a linear span of all vectors of the form $\mathbf{u} \otimes \mathbf{v}$, where $\mathbf{u} \in \mathbf{U}, \mathbf{v} \in \mathbf{V}$ satisfying the following *natural* properties:

- $a(\mathbf{u} \otimes \mathbf{v}) = (a\mathbf{u}) \otimes \mathbf{v} = \mathbf{u} \otimes (a\mathbf{v})$ for all $a \in \mathbb{F}$.
- $(a_1\mathbf{u}_1 + a_2\mathbf{u}_2) \otimes \mathbf{v} = a_1(\mathbf{u}_1 \otimes \mathbf{v}) + a_2(\mathbf{u}_2 \otimes \mathbf{v})$ for all $a_1, a_2 \in \mathbb{F}$. (Linearity in the first variable.)
- $\mathbf{u} \otimes (a_1\mathbf{v}_1 + a_2\mathbf{v}_2) = a_1(\mathbf{u} \otimes \mathbf{v}_1) + a_2(\mathbf{u} \otimes \mathbf{v}_2)$ for all $a_1, a_2 \in \mathbb{F}$. (Linearity in the second variable.)
- If $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \ldots \mathbf{v}_n$ are bases in $\mathbf{U}$ and $\mathbf{V}$ respectively, then $\mathbf{u}_i \otimes \mathbf{v}_j, i = 1, \ldots, m, j = 1, \ldots, n$ is a basis in $\mathbf{U} \otimes \mathbf{V}$.

The element $\mathbf{u} \otimes \mathbf{v}$ is called *decomposable* tensor, or decomposable element (vector), or *rank one tensor*.

It is not difficult to show that $\mathbf{U} \otimes \mathbf{V}$ always exists.

Example 1. Let $\mathbf{U}$ be the space of all polynomials in variable x of degree less than m: $p(x) = \sum_{i=0}^{m-1} a_i x^i$ with coefficients in $\mathbb{F}$. Let $\mathbf{V}$ be the space of all polynomials in

variable y of degree less than n: $q(y) = \sum_{j=0}^{n-1} b_j x^j$ with coefficients in $\mathbb{F}$. Then $\mathbf{U} \otimes \mathbf{V}$ is identified with the vector space of all polynomials in two variables x, y of the form $f(x,y) = \sum_{i=j=0}^{m-1,n-1} c_{ij} x^i y^j$ with the coefficients in $\mathbb{F}$. The decomposable elements are $p(x)q(y), p \in \mathbf{U}, q \in \mathbf{V}$.

The tensor products of this kind is the basic tool for solving PDE (partial differential equations), using *separation of variables*, i.e. Fourier series.

Example 2. Let $\mathbf{U} = \mathbb{F}^m, \mathbf{V} = \mathbb{F}^n$ then $\mathbf{U} \otimes \mathbf{V}$ is identified with the space of $m \times n$ matrices $\mathbb{F}^{m \times n}$. The decomposable tensor $\mathbf{u} \otimes \mathbf{v}$ is identified with $\mathbf{u}\mathbf{v}^T$. Note $\mathbf{u}\mathbf{v}^T$ is indeed rank one matrix.

Assume now that in addition $\mathbf{U}, \mathbf{V}$ are IPS with the inner products $\langle \cdot, \cdot \rangle_{\mathbf{U}}, \langle \cdot, \cdot \rangle_{\mathbf{V}}$. Then there exists a unique inner product $\langle \cdot, \cdot \rangle_{\mathbf{U} \otimes \mathbf{V}}$ which satisfies the property

$$\langle \mathbf{u} \otimes \mathbf{v}, \mathbf{x} \otimes \mathbf{y} \rangle_{\mathbf{U} \otimes \mathbf{V}} = \langle \mathbf{u}, \mathbf{x} \rangle_{\mathbf{U}} \langle \mathbf{v}, \mathbf{y} \rangle_{\mathbf{V}} \text{ for all } \mathbf{u}, \mathbf{x} \in \mathbf{U} \text{ and } \mathbf{v}, \mathbf{y} \in \mathbf{V}.$$

This follows from the fact that if $\mathbf{u}_1, \dots, \mathbf{u}_m$ and $\mathbf{v}_1, \dots, \mathbf{v}_m$ are orthonormal bases in $\mathbf{U}$ and $\mathbf{V}$ respectively, then $\mathbf{u}_i \otimes \mathbf{v}_j, i = 1, \dots, m, j = 1, \dots, n$ is an orthonormal basis in $\mathbf{U} \otimes \mathbf{V}$.

In Example 1, if one has the standard inner product in $\mathbb{F}^m$ and $\mathbb{F}^n$ then these inner product induce the following inner product in $\mathbb{F}^{m \times n}$: $\langle A, B \rangle = \operatorname{tr} AB^*$. If $A = \mathbf{u}\mathbf{v}^T, B = \mathbf{x}\mathbf{y}^T$ then $trAB^* = (\mathbf{x}^*\mathbf{u})(\mathbf{y}^*\mathbf{v})$.

Any $\tau \in \mathbf{U} \otimes \mathbf{V}$ can be viewed as a linear transformation $\tau_{\mathbf{U},\mathbf{V}} : \mathbf{U} \to \mathbf{V}$ and $\tau_{\mathbf{V},\mathbf{U}}$ as follows. Assume for simplicity that $\mathbf{U}, \mathbf{V}$ are IPS over $\mathbb{R}$. Then

$$(\mathbf{u} \otimes \mathbf{v})_{\mathbf{U},\mathbf{V}} : \mathbf{U} \to \mathbf{V} \text{ is given by } \mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{u} \rangle_{\mathbf{U}} \mathbf{v},$$
$$(\mathbf{u} \otimes \mathbf{v})_{\mathbf{V},\mathbf{U}} : \mathbf{V} \to \mathbf{U} \text{ is given by } \mathbf{y} \mapsto \langle \mathbf{y}, \mathbf{v} \rangle_{\mathbf{V}} \mathbf{u}.$$

Since any $\tau \in \mathbf{U} \otimes \mathbf{V}$ is a linear combination of rank one tensors, equivalently is linear combination of rank one matrices, the above definitions extend to any $\tau \in \mathbf{U} \otimes \mathbf{V}$. Thus if $A \in \mathbb{F}^{m \times n} = \mathbb{F}^m \otimes \mathbb{F}^n$ then $A_{\mathbb{F}^m,\mathbb{F}^n}\mathbf{u} = A^T\mathbf{u}, A_{\mathbb{F}^n,\mathbb{F}^m}\mathbf{v} = A\mathbf{v}$.

For $\tau \in \mathbf{U} \otimes \mathbf{V}$ rank $_{\mathbf{U}}\tau :=$ rank $\tau_{\mathbf{V},\mathbf{U}}$ and rank $_{\mathbf{V}}\tau :=$ rank $\tau_{\mathbf{U},\mathbf{V}}$. The rank of τ, denoted by rank τ, is the minimal decomposition of τ to a sum of rank one nonzero tensors: $\tau = \sum_{i=1}^k \mathbf{u}_i \otimes \mathbf{v}_i$, where $\mathbf{u}_i, \mathbf{v}_i \neq \mathbf{0}$ for $i = 1, \dots, k$.

Proposition 5.1 *Let $\tau \in \mathbf{U} \otimes \mathbf{V}$. Then* rank $\tau =$ rank $_{\mathbf{U},\mathbf{V}}\tau =$ rank $_{\mathbf{V},\mathbf{U}}\tau$.

Proof. Let $\tau = \sum_{i=1}^k \mathbf{u}_i \otimes \mathbf{v}_i$. Then $\tau(\mathbf{v}) = \sum_{i=1}^k \langle \mathbf{v}, \mathbf{v}_i \rangle_{\mathbf{V}} \mathbf{u}_i \in \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Hence Range $\tau_{\mathbf{V},\mathbf{U}} \subseteq \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Therefore

$$\text{rank } \tau_{\mathbf{V},\mathbf{U}} = \dim \text{Range } \tau_{\mathbf{V},\mathbf{U}} \leq \dim \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k) \leq k.$$

(It is possible that $\mathbf{u}_1, \dots, \mathbf{u}_k$ are linearly dependent.) Since τ is represented by a matrix A we know that rank $\tau_{\mathbf{U},\mathbf{V}} =$ rank $A^T =$ rank $A =$ rank $\tau_{\mathbf{V},\mathbf{U}}$. Also rank τ is the minimal number k that A is represented as rank one matrices. The singular value decomposition of A yields that one can represent A as sum of rank A of rank one matrices. □

Let $T_i : \mathbf{U}_i \to \mathbf{V}_i$ be linear operators. Then they induce a linear operator on $T_1 \otimes T_2 : \mathbf{U}_1 \otimes \mathbf{U}_2 \to \mathbf{V}_1 \otimes \mathbf{V}_2$ such that $(T_1 \otimes T_2)(\mathbf{u}_1 \otimes \mathbf{u}_2) = (T_1\mathbf{u}_1) \otimes (T_2\mathbf{u}_2)$ for all $\mathbf{u}_1 \in \mathbf{U}_2, \mathbf{u}_2 \in \mathbf{U}_2$. We will see in the next section that $T_1 \otimes T_2$ is a special *4* tensor, i.e. $T_1 \otimes T_2 \in \mathbf{U}_1 \otimes \mathbf{V}_1 \otimes \mathbf{U}_2 \otimes \mathbf{V}_2$.

If furthermore $P_i : \mathbf{V}_i \to \mathbf{W}_i, i = 1, 2$ then we have the following composition $(P_1 \otimes P_2)(T_1 \otimes T_2) = (P_1T_1) \otimes (P_2T_2)$. This equality follows from

$$(P_1 \otimes P_2)\big((T_1 \otimes T_2)(\mathbf{u}_1 \otimes \mathbf{u}_2)\big) = (P_1 \otimes P_2)(T_1\mathbf{u}_1 \otimes T_2\mathbf{u}_2) = (P_1T_1\mathbf{u}_1) \otimes (P_2T_2\mathbf{u}_2).$$

Since each linear operator $T_i : \mathbf{U}_i \to \mathbf{V}_i, i = 1, 2$ is represented by a matrix, one can reduce the definition of $T_1 \otimes T_2$ to the notion of *tensor* product of two matrices $A \in \mathbb{F}^{m_1 \times n_1}, B \in \mathbb{F}^{m_2 \times n_2}$. This tensor product is called the *Kronecker* product.

Let $A = [a_{ij}]_{i,j=1}^{m_1,n_1} \in \mathbb{F}^{m_1 \times n_1}$. Then $A \otimes B \in \mathbb{F}^{m_1 m_2 \times n_1 n_2}$ is the following block matrix:

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & ... & a_{1n_1}B \\ a_{21}B & a_{22}B & ... & a_{2n_1}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m_1 1}B & a_{m_1 2}B & ... & a_{m_1 n_1}B \end{bmatrix} \tag{5.1}$$

Let us try to understand the logic of this notation. Let $\mathbf{x} = [x_1, \dots, x_{n_1}]^T \in \mathbb{F}^{n_1} = \mathbb{F}^{n_1 \times 1}, = [y_1, \dots, y_{n_2}]^T \in \mathbb{F}^{n_2} = \mathbb{F}^{n_2 \times 1}$. Then we view $\mathbf{x} \otimes \mathbf{y}$ as a column vector of dimension $n_1 n_2$ given by the above formula. So the coordinates of $\mathbf{x} \otimes \mathbf{y}$ are $x_j y_l$ where the double indices are arranged in the lexicographic order, (the order of a dictionary):

$$(1,1), (1,2), \dots, (1,n_2), (2,1), \dots, (2,n_1), \dots, (n_1,1), \dots, (n_1,n_2).$$

The entries of $A \otimes B$ are $a_{ij}b_{kl} = c_{(i,k)(j,l)}$. So we view (i,k) as the row index of $A \otimes B$ and (j,l) as the column index of $A \otimes B$. Then

$$[(A \otimes B)(\mathbf{x} \otimes \mathbf{y})]_{(i,k)} = \sum_{j,l=1}^{n_1,n_2} c_{(i,k)(j,l)} x_j y_k = \big(\sum_{j=1}^{n_1} a_{ij}x_j\big)\big(\sum_{l=1}^{n_2} b_{kl}y_l\big) = (A\mathbf{x})_i (B\mathbf{y})_k.$$

As should be according to our notation. Thus as in operator case

$$(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1 B_1) \otimes (A_2 B_2) \text{ if } A_i \in \mathbb{F}^{m_i \times n_i}, B_i \in \mathbb{F}^{n_i \times l_i}, i = 1, 2.$$

Note that $I_m \otimes I_n = I_{mn}$. Moreover if A and B are diagonal matrices then $A \otimes B$ is a diagonal matrix. If A and B are upper or lower triangular then $A \otimes B$ is upper or lower triangular respectively. So if $A \in \mathrm{GL}(n, \mathbb{F}), B \in \mathrm{GL}(n, \mathbb{F})$ then $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. Also $(A \otimes B)^T = A^T \otimes B^T$. So if A, B are symmetric then $A \otimes B$ is symmetric. If A and B are orthogonal matrices then $A \otimes B$ is orthogonal. The following results follows straightforward from the above properties.

Proposition 5.2 . *The following facts hold*

1. *Let $A_i \in \mathbb{R}^{m_i \times n_i}$ for $i = 1, 2$. Assume that $A_i = U_i \Sigma_i V_i^T, U_i \in \mathbf{O}(m_i, \mathbb{R}), V_i \in \mathbf{O}(n_i, \mathbb{R})$ is the standard SVD decomposition for $i = 1, 2$. Then $A_1 \otimes A_2 = (U_1 \otimes U_2)(\Sigma_1 \otimes \Sigma_2)(V_1^T \otimes V_2^T)$ is a singular decomposition of $A_1 \otimes A_2$, except that the diagonal entries of $\Sigma_1 \otimes \Sigma_2$ are not arranged in a decreasing order. In particular all the nonzero singular values of $A_1 \otimes A_2$ are of the form $\sigma_i(A_1)\sigma_j(A_2)$, where $i = 1, \dots, \operatorname{rank} A_1$ and $j = 1, \dots, \operatorname{rank} A_2$. Hence $\operatorname{rank} A_1 \otimes A_2 = \operatorname{rank} A_1 \operatorname{rank} A_2$.*

2. *Let $A_k \in \mathbb{F}^{n_k \times n_k}, k = 1, 2$. Assume that $\det(zI_{n_k} - A_k) = \prod_{i=1}^{n_k}(z - \lambda_{i,k})$ for $k = 1, 2$. Then $\det(zI_{n_1 n_2} - A_1 \otimes A_2) = \prod_{i,j=1}^{n_1 n_2}(z - \lambda_{i,1}\lambda_{j,2})$.*

3. *Assume that $A_i \in \mathrm{S}_{n_i}(\mathbb{R})$ and $A_i = Q_i \Lambda_i Q_i^T$ is the spectral decomposition of A_i, i.e. Q_i is orthogonal and Λ_i is diagonal, for $i = 1, 2$. Then $A_1 \otimes A_2 = (Q_1 \otimes Q_2)(\Lambda_1 \otimes \Lambda_2)(Q_1^T \otimes Q_2^T)$ is the spectral decomposition of $A_1 \otimes A_2 \in S_{n_1 n_2}(\mathbb{R})$.*

In the next section we will show that the singular decomposition of $A_1 \otimes A_2$ is a minimal decomposition of the *4* tensor $A_1 \otimes A_2$. In the rest of the section we discuss the symmetric and skew symmetric tensor products of $\mathbf{U} \otimes \mathbf{U}$.

Definition: Let $\mathbf{U}$ be a vector space of dimension m over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Denote $\mathbf{U}^{\otimes 2} := \mathbf{U} \otimes \mathbf{U}$. The subspace $\mathrm{Sym}^2\mathbf{U} \subset \mathbf{U}^{\otimes 2}$, called a 2-*symmetric power* of $\mathbf{U}$, is spanned by tensors

of the form $\mathrm{sym}^2(\mathbf{u},\mathbf{v}) := \mathbf{u}\otimes\mathbf{v}+\mathbf{v}\otimes\mathbf{u}$ for all $\mathbf{u},\mathbf{v}\in\mathbf{U}$. $\mathrm{sym}^2(\mathbf{u},\mathbf{v}) = \mathrm{sym}^2(\mathbf{v},\mathbf{u})$ is called a 2-*symmetric product* of $\mathbf{u}$ and $\mathbf{v}$, or simply a *symmetric product* Any vector $\tau\in\mathrm{Sym}^2\mathbf{U}$ is a called a 2-*symmetric tensor*, or simply a *symmetric tensor*. The subspace $\bigwedge^2\mathbf{U}\subset\mathbf{U}^{\otimes 2}$, called 2-*exterior power* of $\mathbf{U}$, is spanned by all tensors of the form $\mathbf{u}\wedge\mathbf{v} := \mathbf{u}\otimes\mathbf{v}-\mathbf{v}\otimes\mathbf{u}$, for all $\mathbf{u},\mathbf{v}\in\mathbf{U}$. $\mathbf{u}\wedge\mathbf{v} = -\mathbf{v}\wedge\mathbf{u}$ is called the *wedge product* of $\mathbf{u}$ and $\mathbf{v}$. Any vector $\tau\in\bigwedge^2\mathbf{U}$ is a called a 2-*skew symmetric tensor*, or simply a *skew symmetric tensor*.

Since $2\mathbf{u}\otimes\mathbf{v} = \mathrm{sym}^2(\mathbf{u},\mathbf{v})+\mathbf{u}\wedge\mathbf{v}$ it follows that $\mathbf{U}^{\otimes 2} = \mathrm{Sym}^2(\mathbf{U})\oplus\bigwedge^2\mathbf{U}$. That is, any tensor $\tau\in\mathbf{U}^{\otimes 2}$ can be decomposed uniquely to a sum $\tau=\tau_s+\tau_a$ where $\tau_s,\tau_a\in\mathbf{U}^{\otimes 2}$ are symmetric and skew symmetric tensors respectively.

In terms of matrices, we identify $\mathbf{U}$ with $\mathbb{F}^m$, $\mathbf{U}^{\otimes 2}$ with $\mathbb{F}^{m\times m}$, i.e. the algebra of $m\times m$ matrices. Then $\mathrm{Sym}^2\mathbf{U}$ is identified with $\mathrm{S}_m(\mathbb{F})\subset\mathbb{F}^{m\times m}$, the space of $m\times m$ symmetric matrices: $A^T=A$, and $\bigwedge^2\mathbf{U}$ is identified with $\mathbf{AS}(m,\mathbb{F})$, the space of $m\times m$ skew symmetric matrices: $A^T=-A$. Note that any matrix $A\in\mathbb{F}^{m\times m}$ is of the form $A=\frac{1}{2}(A+A^T)+\frac{1}{2}(A-A^T)$, which is the unique decomposition to a sum of symmetric and skew symmetric matrices.

Proposition 5.3 *Let $\mathbf{U}$ be a finite dimensional space over $\mathbb{F}=\mathbb{R},\mathbb{C}$. Let $T:\mathbf{U}\to\mathbf{U}$ be any linear operator. Then $\mathrm{Sym}^2\mathbf{U}$ and $\bigwedge^2\mathbf{U}$ are invariant subspaces of $T^{\otimes 2} := T\otimes T : \mathbf{U}^{\otimes 2}\to\mathbf{U}^{\otimes 2}$.*

Proof. Observe that $T^{\otimes 2}\mathrm{sym}^2(\mathbf{u},\mathbf{v}) = \mathrm{sym}^2(T\mathbf{u},T\mathbf{v})$ and $T^{\otimes 2}\mathbf{u}\wedge\mathbf{v} = (T\mathbf{u})\wedge(T\mathbf{v})$. □

It can be shown that $\mathrm{Sym}^2\mathbf{U}$ and $\bigwedge^2\mathbf{U}$ are the only invariant subspaces of $T^{\otimes 2}$ for all choices of linear transformations $T:\mathbf{U}\to\mathbf{U}$.

5.3 Tensor product of many vector spaces

Let $\mathbf{U}_i$ be vector spaces of dimension m_i for $i=1,\dots,k$ over $\mathbb{F}=\mathbb{R},\mathbb{C}$. Then $\mathbf{U} := \otimes_{i=1}^k\mathbf{U}_i = U_1\otimes\mathbf{U}_2\otimes\ldots\otimes\mathbf{U}_k$ is the *tensor product space* of $\mathbf{U}_1,\dots,\mathbf{U}_k$ of dimension $m_1m_2\dots m_k$. $\mathbf{U}$ is spanned by the *decomposable tensors* $\otimes_{i=1}^k\mathbf{u}_i = \mathbf{u}_1\otimes\mathbf{u}_2\otimes\ldots\otimes\mathbf{u}_k$, called also *rank one tensors*, where $\mathbf{u}_i\in\mathbf{U}_i$ for $i=1,\dots,k$. As in the case of $k=2$ we have the basic identity:

$$a(\mathbf{u}_1\otimes\mathbf{u}_2\otimes\ldots\otimes\mathbf{u}_k) = (a\mathbf{u}_1)\otimes\mathbf{u}_2\otimes\ldots\otimes\mathbf{u}_k = \mathbf{u}_1\otimes(a\mathbf{u}_2)\otimes\ldots\otimes\mathbf{u}_k = \ldots = \mathbf{u}_1\otimes\mathbf{u}_2\otimes\ldots\otimes(a\mathbf{u}_k).$$

Also the above decomposable tensor is multilinear in each variable. The definition and the existence of k-tensor products can be done *recursively* as follows. For $k=2$ $\mathbf{U}_1\otimes\_2$ is defined in the previous section. Then define recursively $\otimes_{i=1}^k\mathbf{U}_i$ as $(\otimes_{i=1}^{k-1}\mathbf{U}_i)\otimes U_k$ for $k=3,\ldots$.

$$\otimes_{j=1}^k\mathbf{u}_{i_j,j},\ i_j=1,\dots,m_j, j=1,\dots,k \text{ is a basis of } \otimes_{i=1}^k\mathbf{U}_i \quad \text{if } \mathbf{u}_{1,i},\dots,\mathbf{u}_{m_i,i} \text{ is a basis of } \mathbf{U}_i \text{ for } i=1,\dots,k. \tag{5.2}$$

Assume now that in addition $\mathbf{U}_i$ is IPS with the inner product $\langle\cdot,\cdot\rangle_{\mathbf{U}_i}$ for $i=1,\dots,k$. Then there exists a unique inner product $\langle\cdot,\cdot\rangle_{\otimes_{i=1}^k\mathbf{U}}$ which satisfies the property

$$\langle\otimes_{i=1}^k\mathbf{u}_i,\otimes_{i=1}^k\mathbf{v}_i\rangle_{\otimes_{i=1}^k\mathbf{U}_i} = \prod_{i=1}^k\langle\mathbf{u}_i,\mathbf{v}_i\rangle_{\mathbf{U}_i} \text{ for all } \mathbf{u}_i,\mathbf{v}_i\in\mathbf{U}_i, i=1,\dots,k.$$

In particular, if $\mathbf{u}_{1,i},\dots,\mathbf{u}_{m_i,i}$ is an orthonormal basis in $\mathbf{U}_i$, for $i=1,\dots,k$, then $\otimes_{j=1}^k\mathbf{u}_{i_j,j}$, where $i_j=1,\dots,m_j$ and $j=1,\dots,k$ is an orthonormal basis in $\otimes_{i=1}^k\mathbf{U}_i$.

Then any $\tau \in \otimes_{i=1}^k \mathbf{U}$ can be represented as

$$\tau = \sum_{i_j \in [1,m_j], j=1,\ldots,k} t_{i_1 \ldots i_k} \otimes_{j=1}^k \mathbf{u}_{i_j,j}. \tag{5.3}$$

The above decomposition is called the *TUCKER model*. Then $\mathcal{T} = (t_{i_1 \ldots i_k})_{i_1=\ldots=i_k=1}^{m_1,\ldots,m_k}$ is called *the core tensor*. If $\mathbf{u}_{1,i}, \ldots, \mathbf{u}_{m_i,i}$ is an orthonormal basis in $\mathbf{U}_i$, for $i = 1, \ldots, k$, then the TUCKER model is referred is as *Higher-Order Singular Value Decomposition*, or HOSVD. The core tensor $\mathcal{T}$ is called is called diagonal if $t_{i_1 i_2 \ldots i_k} = 0$ whenever the equality $i_1 = \ldots = i_k$ *is not satisfied*.

We now discuss the change in the core tensor when we replace the base $[\mathbf{u}_{1,i}, \ldots, \mathbf{u}_{m_i,i}]$ in $\mathbf{U}_i$ to the base $[\mathbf{v}_{1,i}, \ldots, \mathbf{v}_{m_i,i}]$ in $\mathbf{U}_i$ for $i = 1, \ldots, k$. We fist recall the change in the coordinates of the vector $\mathbf{u} \in \mathbf{U}$ when we change the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_m]$ to the basis $[\mathbf{v}_1, \ldots, \mathbf{v}_m]$ in $\mathbf{U}$. Let $\mathbf{u} = \sum_{i=1}^m x_i \mathbf{u}_i$. Then $\mathbf{x} := [x_1, \ldots, x_m]^T$ is the coordinate vector of $\mathbf{u}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_m]$. It is convenient to represent $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]\mathbf{x}$. Suppose that $[\mathbf{v}_1, \ldots, \mathbf{v}_m]$ is another basis in $\mathbf{U}$. Then $Q = [q_{ij}]_{i,j=1}^m \in \mathbb{F}^{m \times m}$ is called the *transition matrix* from the base $[\mathbf{u}_1, \ldots, \mathbf{u}_m]$ to the base $[\mathbf{v}_1, \ldots, \mathbf{v}_m]$ if

$$[\mathbf{u}_1, \ldots, \mathbf{u}_m] = [\mathbf{v}_1, \ldots, \mathbf{v}_m]Q, \iff \mathbf{u}_i = \sum_{j=1}^m q_{ji} \mathbf{v}_j, \ j = 1, \ldots, m.$$

So Q^{-1} is the transition matrix from the base $[\mathbf{v}_1, \ldots, \mathbf{v}_m]$ to $[\mathbf{u}_1, \ldots, \mathbf{u}_m]$, i.e. $[\mathbf{v}_1, \ldots, \mathbf{v}_m] = [\mathbf{u}_1, \ldots, \mathbf{u}_m]Q^{-1}$. Hence

$\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]\mathbf{x} = [\mathbf{v}_1, \ldots, \mathbf{v}_m](Q\mathbf{x})$, i.e. $\mathbf{y} = Q\mathbf{x}$ coordinate vector of $\mathbf{u}$ in $[\mathbf{v}_1, \ldots \mathbf{v}_n]$ basis

Thus if $Q_l = [q_{ij,l}]_{i,j=1}^m \in \mathbb{F}^{m \times m}$ is the transition matrix from the base $[\mathbf{u}_{1,l}, \ldots, \mathbf{u}_{m_l,l}]$ to the base $[\mathbf{v}_{1,l}, \ldots, \mathbf{v}_{m_l,l}]$ for $l = 1, \ldots, k$ then the core tensor $\mathcal{T}' = (t'_{j_1 \ldots j_k})$ corresponding to the new basis $\otimes_{j=1}^k \mathbf{v}_{i_j,j}$ is given by the formula:

$$t'_{j_1,\ldots,j_k} = \sum_{i_1,\ldots,i_k=1}^{m_1,\ldots,m_k} \big(\prod_{l=1}^k q_{j_l i_l,l}\big) t_{i_1 \ldots i_k}, \text{ denoted as } \mathcal{T}' = \mathcal{T} \times Q_1 \times Q_2 \times \ldots \times Q_k. \tag{5.4}$$

Any tensor can be decomposed to a sum of rank one tensors:

$$\tau = \sum_{i=1}^R \otimes_{l=1}^k \mathbf{u}_{l,i}, \text{ where } \mathbf{u}_{l,i} \in \mathbf{U}_l \text{ for } i = 1, \ldots, j, l = 1, \ldots, k. \tag{5.5}$$

This decomposition is called *CANDECOMP-PARAFAC* decomposition. This decomposition is not unique. For example, we can obtain CANDECOMP-PARAFAC decomposition from Tucker decomposition by replacing the nonzero term $t_{i_1 \ldots i_k} \otimes_{j=1}^k \mathbf{u}_{i_j,j}, t_{i_1 \ldots i_k} \neq 0$ with $(t_{i_1 \ldots i_k} \mathbf{u}_{i_1}) \otimes \mathbf{u}_{i_2} \otimes \ldots \otimes \mathbf{u}_{i_k}$.

The value of the minimal R is called the *tensor rank* of τ, and is denoted by rank τ. That is, rank τ is the *minimal number* of rank one tensors in the decomposition of τ to a sum of rank one tensors. In general, it is a *difficult* problem to determine the exact value of rank τ for $\tau \in \otimes_{i=1}^k \mathbf{U}_i$ and $k \geq 3$.

Any $\tau \in \otimes_{i=1}^k \mathbf{U}$ can be viewed as a linear transformation

$$\tau_{\otimes_{l=1}^p \mathbf{U}_{i_l}, \otimes_{l=1}^{p'} \mathbf{U}_{i'_l}} : \otimes_{l=1}^p \mathbf{U}_{i_l} \to \otimes_{l=1}^{p'} \mathbf{U}_{i'_l}, 1 \leq i_1 < \ldots < i_p \leq k, 1 \leq i'_1 < \ldots < i_{p'} \leq k, \tag{5.6}$$

where the two sets of nonempty indices $\{i_1, \ldots, i_p\}, \{i'_1, \ldots, i_{p'}\}$ are complementary, i.e. $1 \leq p, p' < k, p+p' = 1$ and $\{i_1, \ldots, \mathbf{i}_p\} \cap \{i'_1, \ldots, i_{p'}\} = \emptyset, \{i_1, \ldots, \mathbf{i}_p\} \cup \{i'_1, \ldots, i_{p'}\} = \{1, \ldots, k\}$. The above transformation is obtained by *contracting the indices* $i_1, \ldots, i_p$. Assume for simplicity that $\mathbf{U}_i$ is IPS over $\mathbb{R}$ for $i = 1, \ldots, k$. Then for decomposable tensor transformation (5.6) is given as

$$(\otimes_{i=1}^{k}\mathbf{u}_i)(\otimes_{l=1}^{p}\otimes\mathbf{v}_{i_l}) = \big(\prod_{l=1}^{p}\langle\mathbf{v}_{i_l},\mathbf{u}_{i_l}\rangle_{\mathbf{U}_{i_l}}\big)\otimes_{l=1}^{p'}\otimes\mathbf{u}_{i'_l}. \tag{5.7}$$

For example for $k=3$ $(\mathbf{u}_1\otimes\mathbf{u}_2\otimes\mathbf{u}_3)(\mathbf{v}_2) = \big(\langle\mathbf{u}_2,\mathbf{v}_2\rangle_{\mathbf{U}_2}\big)\mathbf{u}_1\otimes\mathbf{u}_3$, where $\mathbf{u}_1\in\mathbf{U}_1,\mathbf{u}_2,\mathbf{v}_2\in\mathbf{U}_2,\mathbf{u}_3\in\mathbf{U}_3$ and $p=1,i_1=2,p'=2,i'_1=1,i'_3=3$.

Then

$$\text{rank }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}} := \dim\text{Range }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}}. \tag{5.8}$$

It is easy to compute the above ranks, as this rank is the rank of the corresponding matrix representing this linear transformation. As in the case of matrices it is straightforward to show that

$$\text{rank }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}} := \text{rank }\tau_{\otimes_{l=1}^{p'}\mathbf{U}_{i'_l},\otimes_{l=1}^{p}\mathbf{U}_{i_l}}. \tag{5.9}$$

In view of the above equalities, for $k=3$, i.e. $\mathbf{U}_1\otimes\mathbf{U}_2\otimes\mathbf{U}_3$ we have three different ranks:

$$\text{rank }_{\mathbf{U}_1}\tau := \text{rank }\tau_{\mathbf{U}_2\otimes\mathbf{U}_3,\mathbf{U}_1}, \text{rank }_{\mathbf{U}_2}\tau := \text{rank }\tau_{\mathbf{U}_1\otimes\mathbf{U}_3,\mathbf{U}_2}, \text{rank }_{\mathbf{U}_3}\tau := \text{rank }\tau_{\mathbf{U}_1\otimes\mathbf{U}_2,\mathbf{U}_3}.$$

Thus rank $_{\mathbf{U}_1}\tau$ can be viewed as the dimension of the subspace of $\mathbf{U}_1$ obtained by all possible contractions of τ with respect to $\mathbf{U}_2,\mathbf{U}_3$.

In general, let

$$\text{rank }_i\tau = \text{rank }\tau_{\mathbf{U}_1\otimes\ldots\otimes\mathbf{U}_{i-1}\otimes\mathbf{U}_{i+1}\otimes\ldots\otimes\mathbf{U}_k,\mathbf{U}_i},\ i=1,\ldots,k. \tag{5.10}$$

Let $\tau\in\otimes_{i=1}^{k}\mathbf{U}_i$ be fixed. Choose a basis of U_i such that $\mathbf{u}_{1,i},\ldots,\mathbf{u}_{\text{rank }_i\tau,i}$ is a basis in Range $\tau_{\mathbf{U}_1\otimes\ldots\otimes\mathbf{U}_{i-1}\otimes\mathbf{U}_{i+1}\otimes\ldots\otimes\mathbf{U}_k,\mathbf{U}_i}$. Then we obtain a more precise version of the TUCKER decomposition:

$$\tau = \sum_{i_j\in[1,\text{rank }_j\tau],j=1,\ldots,k} t_{i_1\ldots i_k}\otimes_{j=1}^{k}\mathbf{u}_{i_j,j}. \tag{5.11}$$

The following lower bound on rank τ can be computed easily:

Proposition 5.4 *Let $\mathbf{U}_i$ be a vector space of dimension m_i for $i=1,\ldots,k\geq 2$. Let $\tau\in\otimes_{i=1}^{k}\mathbf{U}_i$. Then for any set of complementary indices $\{i_1,\ldots,i_p\},\{i'_1,\ldots,i'_{p'}\}$*

$$\text{rank }\tau\geq\text{rank }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}}.$$

Proof. Let τ be of the form (5.5). Then

$$\text{rank }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}} = \dim\text{Range }\tau_{\otimes_{l=1}^{p}\mathbf{U}_{i_l},\otimes_{l=1}^{p'}\mathbf{U}_{i'_l}} = \dim\text{span}(\otimes_{l=1}^{p'}\mathbf{u}_{i'_l,1},\ldots,\otimes_{l=1}^{p'}\mathbf{u}_{i'_l,j})\leq j.$$

□

Proposition 5.5 *Let $\mathbf{U}_i$ be a vector space of dimension m_i for $i=1,\ldots,k\geq 2$. Let $\tau\in\otimes_{i=1}^{k}\mathbf{U}_i$. Then*

$$\text{rank }\tau\leq\frac{m_1\ldots m_k}{\max_{i\in[1,k]}m_i}.$$

Proof. The proof is by induction on k. For $k=2$, recall that any $\tau\in\mathbf{U}_1\otimes\mathbf{U}_2$ can be represented by $A\in\mathbb{F}^{m_1\times m_2}$. Hence rank $\tau=$ rank $A\leq\min(m_1,m_2)=\frac{m_1m_2}{\max(m_1,m_2)}$. Assume that the proposition holds of $k=n\geq 2$ and let $k=n+1$. By permuting the factor $\mathbf{U}_1,\ldots,\mathbf{U}_k$, one can assume that $m_n\leq m_i$ for $i=1,\ldots,n-1$. Let $\mathbf{u}_{1,n},\ldots,\mathbf{u}_{m_n,n}$ be a basis of $\mathbf{U}_n$. It is straightforward to show that $\tau\in\otimes_{i=1}^{n}\mathbf{U}_i$ is of the form $\tau=\sum_{p=1}^{m_n}\tau_p\otimes\mathbf{u}_{p,n}$ for unique $\tau_p\in\otimes_{i=1}^{n-1}\mathbf{U}_i$. Decompose each τ_i to a minimal sum of rank one tensors in $\otimes_{i=1}^{n-1}\mathbf{U}_i$. Now use the induction hypothesis for each τ_i to obtain a decomposition of τ as a sum of at most $\frac{m_1\ldots m_k}{\max_{i\in[1,k]}m_i}$ rank one tensors. □

5.4 Examples and results for 3-tensors

Let us start with the simplest case $\mathbf{U}_1 = \mathbf{U}_2 = \mathbf{U}_3 = \mathbb{R}^2, \mathbb{C}^2$. Let $\mathbf{e}_1 = [1,0]^T, \mathbf{e}_2 = [0,1]^T$ be the standard basis in $\mathbb{R}^2$ or $\mathbb{C}^2$. Then $\tau = \sum_{i=j=k=1}^{2} t_{ijk}\mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k$, and $\mathcal{T} = (t_{ijk})_{i=j=k=1}^{i=j=k=2}$. Let $T_k := [t_{ijk}]_{i,j=1}^{2}$ for $k = 1, 2$. Any $B = [b_{ij}]_{i,j=1}^{2} \in \mathbb{F}^2$ we identify with the tensor $\sum_{i,j=1}^{2} b_{ij}\mathbf{e}_i \otimes \mathbf{e}_j$. Using this identification we view $\tau = T_1 \otimes \mathbf{e}_1 + T_2 \otimes \mathbf{e}_2$.

Example 1: Assume that

$$t_{111} = t_{112} = 1,\ t_{221} = t_{222} = 2,\ t_{211} = t_{121} = t_{212} = t_{122} = 0. \tag{5.12}$$

To find $R_1 = \operatorname{rank}_{\mathbf{U}_2 \otimes \mathbf{U}_3, \mathbf{U}_1}$, we construct a matrix $A_1 := [a_{pq,1}]_{p,q=1}^{2,4} \in \mathbb{R}^{2\times 4}$, where $p = i = 1, 2$ and $q = (j,k), j, k = 1, 2$ and $a_{pq,1} = t_{ijk}$. Then $A_1 := \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \end{bmatrix}$, and $R_1 = \operatorname{rank} A_1 = 2$. Next $A_2 := [a_{pq,2}]_{p,q=1}^{2,4} \in \mathbb{R}^{2\times 4}$, where $p = i = 1, 2$ and $q = (j,k), j, k = 1, 2$ and $a_{pq,2} = t_{jik}$. Then $A_2 = A_1$ and $R_2 = \operatorname{rank} A_2 = 2$. Next $A_3 := [a_{pq,3}]_{p,q=1}^{2,4} \in \mathbb{R}^{2\times 4}$, where $p = i = 1, 2$ and $q = (j,k), j, k = 1, 2$ and $a_{pq,3} = t_{jki}$. Then $A_3 := \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 \end{bmatrix}$, and $R_3 = \operatorname{rank} A_3 = 1$. Hence Proposition 5.4 yields that $\operatorname{rank} \tau \geq 2$. We claim that $\operatorname{rank} \tau = 2$ as a tensor over real or complex numbers. Observe that

$$T_1 = T_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \equiv \mathbf{e}_1 \otimes \mathbf{e}_1 + 2\mathbf{e}_2 \otimes \mathbf{e}_2 \Rightarrow \tau = (\mathbf{e}_1 \otimes \mathbf{e}_1 + 2\mathbf{e}_2 \otimes \mathbf{e}_2) \otimes \mathbf{e}_1 + (\mathbf{e}_1 \otimes \mathbf{e}_1 + 2\mathbf{e}_2 \otimes \mathbf{e}_2) \otimes \mathbf{e}_2.$$

Hence

$$\tau = (\mathbf{e}_1 \otimes \mathbf{e}_1 + 2\mathbf{e}_2 \otimes \mathbf{e}_2) \otimes (\mathbf{e}_1 + \mathbf{e}_2) = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes (\mathbf{e}_1 + \mathbf{e}_2) + (2\mathbf{e}_2) \otimes \mathbf{e}_2 \otimes (\mathbf{e}_1 + \mathbf{e}_2).$$

Thus $\operatorname{rank} \tau \leq 2$ and we finally deduce that $\operatorname{rank} \tau = 2$.

Example 2: Assume that

$$t_{211} = t_{121} = t_{112} = 1, \quad t_{111} = t_{222} = t_{122} = t_{212} = t_{221}. \tag{5.13}$$

Let A_1, A_2, A_3 be the matrices defined as in Example 1. Then $A_1 = A_2 = A_3 := \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$. Hence $R_1 = R_2 = R_3 = 2$ and $\operatorname{rank} \tau \geq 2$. Observe next that

$$T_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \equiv \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1,\ T_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \equiv \mathbf{e}_1 \otimes \mathbf{e}_1 \Rightarrow \tau = T_1 \otimes \mathbf{e}_1 + T_2 \otimes \mathbf{e}_2.$$

Hence $\tau = \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2$. Thus $2 \leq \operatorname{rank} \tau \leq 3$. We will show that $\operatorname{rank} \tau = 3$, using the results that follow.

Proposition 5.6 *Let $\mathbf{u}_{1,i}, \ldots, \mathbf{u}_{m_i,i}$ be a basis of $\mathbf{U}_i$ for $i = 1, 2, 3$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $\tau \in \mathbf{U}_1 \otimes \mathbf{U}_2 \otimes \mathbf{U}_3$, and assume that $\tau = \sum_{i=j=k=1}^{m_1,m_2,m_3} t_{ijk}\mathbf{u}_{i,1} \otimes \mathbf{u}_{j,2} \otimes \mathbf{u}_{k,3}, i = 1, \ldots, m_1, j = 1, \ldots, m_2, k = 1, \ldots, m_3$. Let $T_k := [t_{ijk}]_{i,j=1}^{m_1,m_2} \in \mathbb{F}^{m_1 \times m_2}$ for $k = 1, \ldots, m_3$, i.e. $T_i = \sum_{i,j=1}^{m_1,m_2} t_{ijk}\mathbf{u}_{i,1} \otimes \mathbf{u}_{j,2}$ and $\tau = sum_{k=1}^{m_3} T_k \otimes \mathbf{u}_{k,3}$. Assume that a basis $[\mathbf{u}_{1,3}, \ldots, \mathbf{u}_{m_3,3}]$ of $\mathbf{U}_3$ is changed to a basis $[\mathbf{v}_{1,3}, \ldots, \mathbf{v}_{m_3,3}]$, where $[\mathbf{u}_{1,3}, \ldots, \mathbf{u}_{m_3,3}] = [\mathbf{v}_{1,3}, \ldots, \mathbf{v}_{m_3,3}]Q_3, Q_3 = [q_{pq,3}]_{p,q=1}^{m_3} \in \mathrm{GL}(m_3, \mathbb{F})$. Then $\tau = \sum_{k=1}^{m_3} T'_k \mathbf{v}_{k,3}$, where $T'_k = [t'_{ijk}]_{i,j=1}^{m_1,m_2} = \sum_{l=1}^{k} q_{kl,3}T_l$. In particular, $\mathcal{T}' := (t'_{ijk})_{i,j,k=1}^{m_1,m_2,m_2}$ is the core tensor corresponding to τ in the basis $\mathbf{u}_{i,1} \otimes \mathbf{u}_{j,2} \otimes \mathbf{v}_{k,3}$ for $i = 1, \ldots, m_1, j = 2, \ldots, m_2, k = 3, \ldots, m_3$. Furthermore, $R_3 = \operatorname{rank}_{\mathbf{U}_1 \otimes \mathbf{U}_2, \mathbf{U}_3} \tau = \operatorname{rank}_{\mathbf{U}_3, \mathbf{U}_1 \otimes \mathbf{U}_2} \tau = \dim \operatorname{span}(T_1, \ldots, T_{m_3})$, where each T_k is viewed as a vector in $\mathbb{F}^{m_1 \times m_2}$. In particular, one choose a basis $[\mathbf{v}_{1,3}, \ldots, \mathbf{v}_{m_3,3}]$ in $\mathbf{U}_3$ such that the matrices $T'_1, \ldots, T'_{R_3}$ are linearly independent. Furthermore, if $m_3 > R_3$ then we can assume that $T'_k = \mathbf{0}$ for $k > R_3$.*

Assume that $[\mathbf{v}_{1,3},\dots,\mathbf{v}_{m_3,3}]$ *in* $\mathbf{U}_3$ *was chosen as above. Let* $[\mathbf{v}_{1,1},\dots,\mathbf{v}_{m_1,1}],[\mathbf{v}_{1,2},\dots,\mathbf{v}_{m_2,2}]$ *be two bases in* $\mathbf{U}_1,\mathbf{U}_2$ *respectively, where* $[\mathbf{u}_{1,1},\dots,\mathbf{u}_{m_1,1}]=[\mathbf{v}_{1,1},\dots,\mathbf{v}_{m_1,1}]Q_1,[\mathbf{u}_{1,2},\dots,\mathbf{u}_{m_2,2}]=[\mathbf{v}_{1,2},\dots,\mathbf{v}_{m_2,2}]Q_2$ *and* $Q_1=[q_{pq,1}]_{p,q=1}^{m_1}\in\mathrm{GL}(m_1,\mathbb{F}),Q_2=[q_{pq,2}]_{p,q=1}^{m_2}\in\mathrm{GL}(m_2,\mathbb{F})$. *Let* $\tau=\sum_{i,j,k=1}^{m_1,m_2,m_3}\tilde{t}_{ijk}\mathbf{v}_{i,1}\otimes\mathbf{v}_{j,2}\otimes\mathbf{v}_{j,3}$, $\tilde{T}_k:=[\tilde{t}_{ijk}]_{i,j=1}^{m_1,m_2}\in\mathbb{F}^{m_1\otimes m_2}$ *for* $k=1,\dots,m_3$. *Then* $\tilde{T}_k=Q_1T_k'Q_2^T$ *for* $k=1,\dots,m_3$. *Furthermore, if one chooses the bases in* $\mathbf{U}_2,\mathbf{U}_2$ *such that* $\mathrm{Range}\,\tau_{\mathbf{U}_2\otimes\mathbf{U}_3,\mathbf{U}_1}=\mathrm{span}(\mathbf{v}_{1,1},\dots,\mathbf{v}_{R_1,1})$ *and* $\mathrm{Range}\,\tau_{\mathbf{U}_1\otimes\mathbf{U}_3,\mathbf{U}_2}=\mathrm{span}(\mathbf{v}_{1,2},\dots,\mathbf{v}_{R_2,2})$, *then each* $\tilde{T}_k$ *is a block diagonal matrix* $\tilde{T}_k=\mathrm{diag}(\hat{T}_k,\mathbf{0})$, *where* $\hat{T}_k=[\hat{t}_{ijk}]_{i,j=1}^{R_1,R_2}\in\mathbb{F}^{R_1\times R_2}$ *for* $k=1,\dots,m_3$. *Recalling that* $T_k'=0$ *for* $k>R_3$ *we get the representation* $\tau=\sum_{i,j,k=1}^{R_1,R_2,R_3}\hat{t}_{ijk}\mathbf{v}_{i,1}\otimes\mathbf{v}_{j,2}\otimes\mathbf{v}_{j,3}$.

The proof of this proposition is straightforward and is left to the reader.

By interchanging the factors $\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3$ in the tensor product $\mathbf{U}_1\otimes\mathbf{U}_2\otimes\mathbf{U}_3$ it will be convenient to assume that $m_1\geq m_2\geq m_3\geq 1$. Also the above proposition implies that for $\tau\neq\mathbf{0}$ we may assume that $m_1=R_1\geq m_2=R_2\geq m_3=R_3$.

Proposition 5.7 *Let* $\tau\in\mathbf{U}_1\otimes\mathbf{U}_2\otimes\mathbf{U}_3$. *If* $R_3=1$ *then* $\mathrm{rank}\,\tau=R_1=R_2$.

Proof. We may assume that $m_3=1$. In that case $\tau=T_1\otimes\mathbf{v}_{1,3}$. Hence $\mathrm{rank}\,\tau=\mathrm{rank}\,T_1=R_1=R_2$. □

Thus we need consider the case $m_1=R_1\geq m_2=R_2\geq m_3=R_3\geq 2$. We now consider the *generic case*, i.e. where $T_1,\dots,T_k\in\mathbb{F}^{m_1\times m_2}$ are *generic*. That is each $T_i\neq\mathbf{0}$ is chosen at random, where $\frac{T_i}{||T_i||_F}$ has a uniform distribution on the matrices of norm one. It is well known that a generic matrix $T\in\mathbb{F}^{m_1\times m_2}$ has rank m_2, since $m_1\geq m_2$.

Theorem 5.8 *Let* $\tau=\sum_{i,j,k=1}^{n,n,2}t_{ijk}\in\mathbb{C}^n\otimes\mathbb{C}^n\otimes\mathbb{C}^2$ *where* $n\geq 2$. *Denote* $T_1=[t_{ij1}]_{i,j=1}^n,T_2=[t_{ij2}]_{i,j=1}^n\in\mathbb{C}^{n\times n}$. *Suppose that there exists a linear combination* $T=aT_1+bT_2\in\mathrm{GL}(n,\mathbb{C})$, *i.e.* T *is invertible. Then* $n\leq\mathrm{rank}\,\tau\leq 2n-1$. *In particular, for a generic tensor* $\mathrm{rank}\,\tau=n$.

Proof. Recall that by changing a basis in $\mathbb{C}^2$, we may assume that $T_1'=aT_1+bT_2=T$. Suppose first that $T_2'=0$. Then $\mathrm{rank}\,\tau=\mathrm{rank}\,T=n$. Hence we have always the inequality $\mathrm{rank}\,\tau\geq n$.

Assume now that $R_3=2$. Choose first $Q_1=T^{-1}$ and $Q_2=I_n$. Then $\tilde{T}_2=I_n$. Thus, for simplicity of notation we may assume to start with that $T_1=I_n$. Let λ be an eigenvalue T_2. Then change the basis in $\mathbb{C}^2$ such that $T_1'=I_n$ and $T_2'=T_2-\lambda T_1=T_2-\lambda I_n$. So $\det T_2'=0$. Hence $r=\mathrm{rank}\,T_2'\leq n-1$. So SVD of $T_2=\sum_{i=1}^r\mathbf{u}_i\otimes(\sigma_i(T_2)\bar{\mathbf{u}}_i)$. Now $I_n=\sum_{i=1}^n\mathbf{e}_i\otimes\mathbf{e}_i$. Hence $\tau=\sum_{i=1}^n\mathbf{e}_i\otimes\mathbf{e}_1\otimes\mathbf{e}_1+\sum_{i=1}^r\mathbf{u}_i\otimes(\sigma_i(T_2)\bar{\mathbf{u}}_i)\otimes\mathbf{e}_2$. Hence $\mathrm{rank}\,\tau\leq 2n-1$.

We now consider the generic case. In that case T_1 is generic, so $\mathrm{rank}\,T_1=n$ is invertible. Choose $Q_1=T^{-1},Q_2=I_n,Q_3=I_2$. Then $\tilde{T}_1=I_n,\tilde{T}_2=T_1^{-1}T_2$. $\tilde{T}_2$ is generic. Hence it is diagonable. So $\hat{T}_2=X\,\mathrm{diag}(\lambda_1,\dots,\lambda_n)X^{-1}$ for some invertible X. Now we again change a basis in $\mathbf{U}_1=\mathbf{U}_2=\mathbb{R}^n$ by letting $Q_1=X^{-1},Q_2=X^T$. The new matrices $\hat{T}_1=X^{-1}I_nX=I_n,\hat{T}_2=X^{-1}\tilde{T}_2X=\mathrm{diag}(\lambda_1,\dots,\lambda_n)$. In this basis

$$\tau=\hat{T}_1\otimes\mathbf{e}_1+\hat{T}_2\otimes\mathbf{e}_2=\sum_{i=1}^n\mathbf{e}_i\otimes\mathbf{e}_i\otimes\mathbf{e}_1+\sum_{i=1}^n\lambda_i\mathbf{e}_i\otimes\mathbf{e}_i\otimes\mathbf{e}_2=\sum_{i=1}^n\mathbf{e}_i\otimes\mathbf{e}_i\otimes(\mathbf{e}_1+\lambda_i\mathbf{e}_2).$$

Hence the rank of generic tensor $\tau\in\mathbb{C}^n\otimes\mathbb{C}^n\otimes C^2$ is n. □

I believe that for any $\tau\in\mathbb{C}^n\otimes\mathbb{C}^n\otimes\mathbb{C}^2$ $\mathrm{rank}\,\tau\leq 2n-1$. The case $\mathrm{rank}\,\tau=2n-1$ would correspond to $T_1=I_n$ and T_2 a nilpotent matrix with one Jordan block.

The analysis of the proof of the above theorem yields.

Corollary 5.9 *Let* $\tau = \sum_{i,j,k=1}^{n,n,2} t_{ijk} \in \mathbb{F}^n \otimes \mathbb{F}^n \otimes \mathbb{F}^2$ *where* $n \geq 2$ *and* $\mathbb{F} = \mathbb{R}, \mathbb{C}$. *Denote* $T_1 = [t_{ij1}]_{i,j=1}^n, T_2 = [t_{ij2}]_{i,j=1}^n \in \mathbb{F}^{n \times n}$. *Suppose that there exists a linear combination* $T = aT_1 + bT_2 \in \mathrm{GL}(n, \mathbb{F})$, *i.e.* T *is invertible. Then* $\operatorname{rank} \tau = n$, *if and only if the matrix* $T^{-1}(-bT_1 + aT_2)$ *is diagonable over* $\mathbb{F}$.

This result shows that it is possible that for $\tau \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^2$ $\operatorname{rank}_{\mathbb{R}} \tau > n$, while $\operatorname{rank}_{\mathbb{C}} \tau = n$. For simplicity choose $n = 2$, $\tau = T_1 \otimes \mathbf{e}_1 + T_2 \otimes \mathbf{e}_2$, $T_1 = I_2, T_2 = -T_2^T \neq \mathbf{0}$. Then T_2 has two complex conjugate eigenvalues. Hence T_2 is not diagonable over $\mathbb{R}$. The above Corollary yields that $\operatorname{rank}_{\mathbb{R}} > 2$. However, since T_2 is normal T_2 is diagonable over $\mathbb{C}$. Hence the above Corollary yields that $\operatorname{rank}_{\mathbb{C}} \tau = 2$.

Proof of the claim in Example 2 Observe that T_1 is invertible $T_1^{-1} = T_1$. Consider $T_1^{-1} T_2 = S = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. Note that the Jordan canonical form of S is S^T. Hence S is not diagonable. The above Corollary shows that $\operatorname{rank}_{\mathbb{R}} \tau, \operatorname{rank}_{\mathbb{C}} \tau > 2$. Hence $\operatorname{rank}_{\mathbb{R}} \tau = \operatorname{rank}_{\mathbb{C}} \tau = 3$. This shows that Theorem 5.8 is sharp for $n = 2$.

References

[1] R.B. Bapat and T.E.S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, Cambridge, UK, 1997.

[2] A. Berman and R.J. Plemmons, *Nonnegative Matrices in Mathematical Sciences*, Academic Press, New York 1979.

[3] K. Fan, On a theorem of Weyl concerning eigenvalues of linear transformations. I., *Proc. Nat. Acad. Sci. U. S. A.* 35 (1949), 652–655.

[4] S. Friedland, A New Approach to Generalized Singular Value Decomposition, to appear in SIMAX, 2006.

[5] F.R. Gantmacher, *The Theory of Matrices*, Vol. I and II, Chelsea Publ. Co., New York 1959, Reprinted by Amer. Math. Soc..

[6] G.H. Hardy, J.E. Littlewood and G. Pólya, *Inequalities*, Cambridge Univ. Press, Second edition, 1952.

[7] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, New York 1988.

[8] G.H. Golub and C.F. Van Loan, *Matrix Computations*, John Hopkins Univ. Press, 1985.

[9] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, 2nd ed., New York 1982.

[10] S.J. Leon, *Linear Algebra with Applications*, MacMillan, 6th edition, 2002.

[11] A.J. Laub, *Matrix Analysis for Scientists & Engineers*, SIAM, 2005.

[12] G. Pólya and M. Schiffer, Convexity of functionals by transplantation, *J. Analyse Math.* 3 (1954), 245–346.