MATRICES

Shmuel Friedland[1]
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607-7045, USA
E-mail: friedlan@uic.edu

August 16, 2010

[1] *All rights reserved*

Contents

1 Domains, Modules and Matrices 1

1.1 Rings, Domains and Fields 1
1.2 Bezout Domains 3
1.3 $\mathbb{D}_U, \mathbb{D}_P$ and $\mathbb{D}_E$ domains 7
1.4 Factorizations in $\mathbb{D}[x]$ 9
1.5 Elementary Divisor Domain 12
1.6 Modules 13
1.7 Algebraically closed fields 17
1.8 The resultant and the discriminant 19
1.9 The ring $\mathbb{F}[x_1, ..., x_n]$ 22
1.10 Matrices and homomorphisms 23
1.11 Hermite normal form 26
1.12 Systems of linear equations over Bezout domains 34
1.13 Smith normal form 38
1.14 The ring of local analytic functions in one variable 42
1.15 The local-global domains in $\mathbb{C}^p$ 47
1.16 Historical remarks 52

2 Canonical Forms for Similarity 55

2.1 Strict equivalence of pencils 55
2.2 Similarity of matrices 61
2.3 The companion matrix 63
2.4 Splitting to invariant subspaces 65
2.5 An upper triangular form 69
2.6 Jordan canonical form 72
2.7 Some applications of Jordan canonical form 75
2.8 The matrix equation $AX - XB = 0$ 78
2.9 A criterion for similarity of two matrices 84
2.10 The matrix equation $AX - XB = C$ 87
2.11 A case of two nilpotent matrices 92

2.12 Historical remarks 94

3 Functions of Matrices and Analytic Similarity 95

3.1 Components of a matrix and functions of matrices 95
3.2 Cesaro convergence of matrices 99
3.3 An iteration scheme 101
3.4 Cauchy integral formula for functions of matrices 103
3.5 A canonical form over H_A 111
3.6 Analytic, pointwise and rational similarity 117
3.7 A Global Splitting 120
3.8 First variation of a geometrically simple eigenvalue 121
3.9 Analytic similarity over H_0 123
3.10 Similarity to diagonal matrices 130
3.11 Strict similarity of matrix polynomials 131
3.12 Similarity to diagonal matrices 134
3.13 Property L 140
3.14 Strict similarity of pencils and analytic similarity 146
3.15 Historical remarks 152

4 Inner product spaces 153

4.1 Inner product 153
4.2 Special transformations in IPS 157
4.3 Symmetric bilinear and hermitian forms 163
4.4 Max-min characterizations of eigenvalues 166
4.5 Positive definite operators and matrices 172
4.6 Majorization and applications 177
4.7 Spectral functions 182
4.8 Inequalities for traces 191
4.9 Singular Value Decomposition 193
4.10 Characterizations of singular values 198
4.11 Moore-Penrose generalized inverse 206
4.12 Approximation by low rank matrices 209
4.13 CUR-approximations 211
4.14 Some special maximal spectral problems 215
4.15 Multiplicity index of a subspace of $\mathbf{S}(\mathbf{V})$ 223
4.16 Analytic functions of hermitian matrices 223
4.17 Eigenvalues of sum of hermitian matrices 227

5 Elements of Multilinear Algebra 229

5.1 Tensor product of two free modules 229
5.2 Tensor product of several free modules 236
5.3 Sparse bases of subspaces 244

5.4 Tensor products of inner product spaces 257
5.5 Tensor products of exponents 268

6 Nonnegative matrices 281
6.1 Graphs 281
6.1.1 Undirected graphs 281
6.1.2 Directed graphs 282
6.1.3 Multi graphs 282
6.1.4 Matrices and graphs 286
6.1.5 Problems 288
6.2 Perron-Frobenius theorem 289
6.3 Index of primitivity 300
6.4 Reducible matrices 301
6.5 Stochastic matrices and Markov Chains 308
6.6 Friedland-Karlin results 312
6.7 Convexity and log-convexity 320
6.8 Min-max characterizations of $\rho(A)$ 324
6.9 Application to cellular communication 329
6.9.1 Introduction 329
6.9.2 Statement of problems 329
6.9.3 Relaxations of optimal problems 331
6.9.4 Preliminary results 332
6.9.5 Reformulation of optimal problems 335
6.9.6 Algorithms for sum rate maximization 337

7 Convexity 343
7.1 Convex sets 343
7.2 Doubly stochastic matrices 347
7.3 Convex functions 352
7.4 Norms over vector spaces 356
7.5 Numerical ranges and radii 363
7.6 Superstable norms 370
7.7 Operator norms 373
7.8 Tensor products of convex sets 377
7.9 Variation of tensor powers and spectra 384
7.10 Variation of permanents 392
7.11 The complexity of $\operatorname{conv} \Omega_n \odot \Omega_m$ 396
7.12 Vivanti-Pringsheim theorem and applications 398
7.13 Inverse eigenvalue problem for nonnegative matrices 406
7.14 Cones 413

Chapter 1

Domains, Modules and Matrices

1.1 Rings, Domains and Fields

Definition 1.1.1 *A non empty set R is called a ring if R has two binary operations, called addition and multiplication, such that for all $a, b, c \in R$ the following holds:*

(1.1.1) $a + b \in R$;

(1.1.2) $a + b = b + a$ (the commutative law);

(1.1.3) $(a + b) + c = a + (b + c)$ (the associative law) :

(1.1.4) $\exists\ 0 \in R$ such that $a + 0 = 0 + a = a,\ \forall\ a \in R$;

(1.1.5) $\forall\ a \in R,\ \exists - a \in R$ such that $a + (-a) = 0$;

(1.1.6) $ab \in R$;

(1.1.7) $a(bc) = (ab)c$ (the associative law);

(1.1.8) $a(b + c) = ab + ac, \quad (b + c)a = ba + ca,$ (the distribution laws).

R has an *identity* element 1 if $a1 = 1a$ for all $a \in R$. R is called *commutative* if

$$ab = ba, \quad \text{for all } a, b \in R. \tag{1.1.9}$$

Note that the properties $(1.1.2) - (1.1.8)$ imply that $a0 = 0a = 0$. If a and b are two nonzero elements such that

$$ab = 0 \tag{1.1.10}$$

then a and b are called zero divisors.

Definition 1.1.2 *$\mathbb{D}$ is called an integral domain if $\mathbb{D}$ is a commutative ring without zero divisors and containing identity* 1.

The classical example of an integral domain is the ring of integers $\mathbb{Z}$. In this book we shall use the following example of an integral domain.

Example 1.1.3 *Let $\Omega \subset \mathbb{C}^n$ be a nonempty set. Then $\mathrm{H}(\Omega)$ denotes the ring of analytic functions $f(z_1, ..., z_n)$ such that for each $\zeta \in \Omega$ there exists an open neighborhood $O(f, \zeta)$ of ζ such that f is analytic on $O(f, \zeta)$. If Ω is open we assume that f is defined only on Ω. If Ω consists of one point ζ then H_ζ stands for $\mathrm{H}(\{\zeta\})$.*

Note that zero element is the zero function and the identity element is the constant function which is equal to 1. The properties of analytic functions imply that $\mathrm{H}(\Omega)$ is an integral domain if and only if Ω is a *connected* set. (Ω is connected if for any open set $O \supset \Omega$ there exists an open connected set O' such that $O \supset O' \supset \Omega$.) In this book we shall assume that Ω is connected unless otherwise stated. See [Rud74] and [GuR65] for properties of analytic functions in one and several complex variables.

For $a, b \in \mathbb{D}$, *a divides b*, (or a is a *divisor* of b), denoted by $a|b$, if $b = ab_1$ for some $b_1 \in \mathbb{D}$. An element a is called *invertible*, (unit, unimodular), if $a|1$. $a, b \in \mathbb{D}$ are *associates*, denoted by $a \equiv b$, if $a|b$ and $b|a$. Denote $\{\{b\}\} = \{a \in \mathbb{D} : \ a \equiv b\}$. The associates of a and units are called *improper* divisors of a. For an invertible a denote by a^{-1} the unique element such that

$$aa^{-1} = a^{-1}a = 1. \tag{1.1.11}$$

$f \in \mathrm{H}(\Omega)$ is invertible if and only if f does not vanish at any point of Ω.

Definition 1.1.4 *A field $\mathbb{F}$ is an integral domain $\mathbb{D}$ such that any non zero element is invertible.*

The familiar examples of fields are the set of rational numbers $\mathbb{Q}$, the set of real numbers $\mathbb{R}$, and the set of complex numbers $\mathbb{C}$. Given an integral domain $\mathbb{D}$ there is a standard way to construct the *field* $\mathbb{F}$ of its *quotients.* $\mathbb{F}$ is formed by the set of equivalence classes of all quotients $\frac{a}{b}, b \neq 0$ such that

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \quad \frac{a}{b}\frac{c}{d} = \frac{ac}{bd}, \quad b, d \neq 0. \tag{1.1.12}$$

Definition 1.1.5 *For $\Omega \subset \mathbb{C}^n, \zeta \in \mathbb{C}^n$ let $\mathcal{M}(\Omega), \mathcal{M}_\zeta$ denote the quotient fields of $\mathrm{H}(\Omega), \mathrm{H}_\zeta$ respectively.*

Definition 1.1.6 *Let $\mathbb{D}[x_1, ..., x_n]$ be the ring of all polynomials in n variables with coefficients in $\mathbb{D}$:*

$$p(x_1, ..., x_n) = \sum_{|\alpha|\le m} a_\alpha x^\alpha, \tag{1.1.13}$$

$$\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{Z}_+^n, \quad |\alpha| := \sum_{i=1}^n \alpha_i, \quad x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}.$$

The *degree* of $p(x_1, ..., x_n) \ne 0$ (deg p) is m if there exists $a_\alpha \ne 0$ such that $|\alpha| = m$. (deg $0 = 0$.) A polynomial p is called homogeneous if $a_\alpha = 0$ for all $|\alpha| <$ deg p. It is a standard fact that $\mathbb{D}[x_1, ..., x_n]$ is an integral domain. (See Problems 2-3.) As usual $\mathbb{F}(x_1, ..., x_n)$ denotes the quotient field of $\mathbb{F}[x_1, ..., x_n]$.

Problems

1. Let $C[a, b]$ be the set of real valued continuous functions on the interval $[a, b], a < b$. Show that $C[a, b]$ is a commutative ring with identity and zero divisors.

2. Prove that $\mathbb{D}[x]$ is an integral domain.

3. Prove that $\mathbb{D}[x_1, ..., x_n]$ is an integral domain. (Use the previous problem and the identity $\mathbb{D}[x_1, ..., x_n] = \mathbb{D}[x_1, ..., x_{n-1}][x_n]$.)

4. Let $p(x_1, ..., x_n) \in \mathbb{D}[x_1, ..., x_n]$. Show that $p = \sum_{i\le \deg p} p_i$, where each p_i is either a zero polynomial or a homogeneous polynomial of degree i for $i \ge 1$. If p is not a constant polynomial then $m = \deg p \ge 1$ and $p_m \ne 0$. The polynomial p_m is called the *principle part* of p and is denoted by p_π. (If p is a constant polynomial then $p_\pi = p$.)

5. Let $p, q \in \mathbb{D}[x_1, ..., x_n]$. Show $(pq)_\pi = p_\pi q_\pi$.

1.2 Bezout Domains

Let $a_1, \ldots, a_n \in \mathbb{D}$. Assume first that not all of $a_1, \ldots, a_n$ are equal to zero. An element $d \in \mathbb{D}$ is a *greatest common divisor* (g.c.d) of $a_1, ..., a_n$ if $d|a_i$ for $i = 1, ..., n$, and for any d' such that $d'|a_i, i = 1, ..., n$, $d'|d$. Denote by $(a_1, ..., a_n)$ any g.c.d. of $a_1, ..., a_n$. Then $\{\{(a_1, \ldots, a_n)\}\}$ is the equivalence class of all g.c.d. of $a_1, \ldots, a_n$. For $a_1 = \ldots = a_n = 0$, we define 0 to be the g.c.d. of $a_1, \ldots, a_n$, i.e. $(a_1, ..., a_n) = 0$. The elements $a_1, ..., a_n$ are called *coprime* if $\{\{(a_1, ..., a_n)\}\} = \{\{1\}\}$.

Definition 1.2.1 $\mathbb{D}$ *is called a greatest common divisor domain, or simply GCD domain and denoted by* $\mathbb{D}_G$*, if any two elements in* $\mathbb{D}$ *have a g.c.d..*

A simple example of $\mathbb{D}_G$ is $\mathbb{Z}$. See Problem 5 for a non GCD domain.

Definition 1.2.2 *A subset* $I \subset \mathbb{D}$ *is called an ideal if for any* $a, b \in I$ *and* $p, q \in \mathbb{D}$ *the element* $pa + qb$ *belongs to* I.

In $\mathbb{Z}$ any nontrivial ideal is the set of all numbers divisible by an integer $k \neq 0$. In H(Ω), the set of functions which vanishes on a prescribed set $U \subset \Omega$, i.e.

$$I(U) := \{f \in H(\Omega) : \quad f(\zeta) = 0, \ \zeta \in U\}, \tag{1.2.1}$$

is an ideal. Ideal in I is called *prime* if $ab \in I$ implies that either a or b is in I. $I \subset \mathbb{Z}$ is a prime ideal if and only if I is the set of integers divisible by some prime number p. An ideal I is called *maximal* if the only ideals which contain I are I and $\mathbb{D}$. I is called *finitely generated* if there exists k elements (*generators*) $p_1, ..., p_k \in I$ such that any $i \in I$ is of the form

$$i = a_1 p_1 + \cdots + a_k p_k \tag{1.2.2}$$

for some $a_1, ..., a_k \in \mathbb{D}$. For example, in $\mathbb{D}[x, y]$ the set of all polynomials $p(x, y)$ such that

$$p(0, 0) = 0, \tag{1.2.3}$$

is an ideal generated by x and y. An ideal is called *principal* ideal if it is generated by one element p.

Definition 1.2.3 $\mathbb{D}$ *is called a Bezout domain, or simply BD and denoted by* $\mathbb{D}_B$*, if any two elements* $a, b \in \mathbb{D}$ *have g.c.d.* (a, b) *such that*

$$(a, b) = pa + qb, \tag{1.2.4}$$

for some $p, q \in \mathbb{D}$.

It is easy to show by induction that for $a_1, ..., a_n \in \mathbb{D}_B$

$$(a_1, ..., a_n) = \sum_{i=1}^{n} p_i a_i, \text{ for some } p_1, \ldots, p_n \in \mathbb{D}_B. \tag{1.2.5}$$

Lemma 1.2.4 *An integral domain is a Bezout domain if and only if any finitely generated ideal is principal.*

Proof. Assume that an ideal of $\mathbb{D}_B$ is generated by $a_1, ..., a_n$. Then (1.2.5) implies that $(a_1, ..., a_n) \in I$. Clearly $(a_1, ..., a_n)$ is a generator of I. Assume now that any finitely generated ideal of $\mathbb{D}$ is principal. For given $a, b \in \mathbb{D}$ let I be the ideal generated by a and b. Let d be a generator of I. So

$$d = pa + qb. \tag{1.2.6}$$

Since d generates I d divides a and b. (1.2.6) implies that if d' divides a and b then $d'|d$. Hence $d = (a, b)$ and $\mathbb{D}$ is $\mathbb{D}_B$. □

Let $I \subset \mathbb{D}[x, y]$ be the ideal given by (1.2.3). Clearly $(x, y) = 1$. As $1 \notin I$, I is not principal. As x, y generate I we obtain that $\mathbb{D}[x, y]$ is not $\mathbb{D}_B$. In particular $\mathbb{F}[x_1, ..., x_n]$ is not $\mathbb{D}_B$ for $n \geq 2$. The same argument shows that $H(\Omega)$ is not $\mathbb{D}_B$ for $\Omega \subset \mathbb{C}^n$ and $n \geq 2$. It is a standard fact that $\mathbb{F}[x]$ is a Bezout domain [Lan67]. (See 1.3.) For a connected set $\Omega \subset \mathbb{C}$ $\mathrm{H}(\Omega)$ is $\mathbb{D}_B$. This result is implied by the following interpolation theorem [Rud74, Thms 15.11, 15.15]:

Theorem 1.2.5 *Let $\Omega \subset \mathbb{C}$ be an open set, $A \subset \Omega$ be a countable set with no accumulation point in Ω. Assume that for each $\zeta \in A$, $m(\zeta)$ and $w_{0,\zeta}, ..., w_{m(\zeta),\zeta}$ are a nonnegative integer and $m(\zeta) + 1$ complex numbers, respectively. Then there exists $f \in \mathrm{H}(\Omega)$ such that*

$$f^{(n)}(\zeta) = n! w_{n,\zeta}, \quad n = 0, ..., m(\zeta), \quad \text{for all } \zeta \in A.$$

Furthermore, if all $w_{n,\zeta} = 0$ then there exists $g \in \mathrm{H}(\Omega)$ such that all zeros of g are in A and g has a zero of order $m(\zeta) + 1$ at each $\zeta \in A$.

Theorem 1.2.6 *Let $\Omega \subset \mathbb{C}$ be an open connected set. Then for $a, b \in \mathrm{H}(\Omega)$ there exists $p \in H(\Omega)$ such that $(a, b) = pa + b$.*

Proof. If $a = 0$ or $b = 0$ then $(a, b) = 1a + 1b$. Assume that $ab \neq 0$. Let A be the set of common zeros of $a(z)$ and $b(z)$. For each $\zeta \in A$ let $m(\zeta) + 1$ be the minimum multiplicity of the zero $z = \zeta$ of $a(z)$ and $b(z)$. Theorem 1.2.5 implies the existence of $f \in \mathrm{H}(\Omega)$ which has its zeros at A, such that at each $\zeta \in A$ $f(z)$ has a zero of order $m(\zeta) + 1$. Hence

$$a = \hat{a} f, \ b = \hat{b} f, \ \hat{a}, \hat{b} \in \mathrm{H}(\Omega).$$

Thus $\hat{a}$ and $\hat{b}$ do not have common zeros. If A is empty then $\hat{a} = a$, $\hat{b} = b$. Let $\hat{A}$ be the set of zeros of $\hat{a}$. Assume that for each $\zeta \in \hat{A}$ $\hat{a}$ has a zero of multiplicity $n(\zeta) + 1$. Since $b(\zeta) \neq 0$ for any $\zeta \in \hat{A}$, Theorem 1.2.5

implies the existence of a function $g \in \mathrm{H}(\Omega)$ which satisfies the interpolation conditions:

$$\frac{d^k}{dz^k}(e^{g(z)})|_{z=\zeta} = \frac{d^k}{dz^k}\hat{b}(z)|_{z=\zeta}, \quad k = 0, ..., n(\zeta), \quad \zeta \in \hat{A}.$$

Then

$$p = \frac{e^g - \hat{b}}{\hat{a}}, \quad (a, b) = fe^g = pa + b$$

and the theorem follows. □

Corollary 1.2.7 *Let $\Omega \subset \mathbb{C}$ be a connected set. Then $\mathrm{H}(\Omega)$ is a Bezout domain.*

Problems

1. Let $a, b, c \in \mathbb{D}_B$. Assume that $(a, b) = 1$, $(a, c) = 1$. Show that $(a, bc) = 1$.

2. Let I be a prime ideal in $\mathbb{D}$. Show that $\mathbb{D}/I$ (the set of all cosets of the form $I + a$) is an integral domain.

3. Let I an ideal in $\mathbb{D}$. For $p \in \mathbb{D}$ denote by $I(p)$ the set:

$$I(p) := \{a \in \mathbb{D} : \quad a = bp + q, \text{ for all } b \in \mathbb{D}, \ q \in I\}.$$

Show that $I(p)$ is an ideal. Prove that I is a maximal ideal if and only if for any $p \notin I$ $I(p) = \mathbb{D}$.

4. Show that an ideal I is maximal if and only if $\mathbb{D}/I$ is a field.

5. Let $\mathbb{Z}[\sqrt{-3}] = \{a \in \mathbb{C}, \ a = p + q\sqrt{-3}, \ p, q \in \mathbb{Z}\}$. Show

   (a) $\mathbb{Z}[\sqrt{-3}]$, viewed as a subset of $\mathbb{C}$, is a domain with respect to the addition and multiplication in $\mathbb{C}$.

   (b) Let $z = a + b\sqrt{-3} \in \mathbb{Z}[\sqrt{-3}]$. Then

   $$|z| = 1 \iff z = \pm 1, \quad |z| = 2 \iff z = \pm 2 \text{ or } z = \pm 1 \pm \sqrt{-3}.$$

   $|z| \geq \sqrt{7}$ for all other values of $z \neq 0$. In particular if $|z| = 2$ then z is a prime.

   (c) Let

   $$a = 4 = 2{\cdot}2 = (1+\sqrt{-3})(1-\sqrt{-3}), \ b = (1+\sqrt{-3}){\cdot}2 = -(1-\sqrt{-3})^2.$$

   Then any d that divides a and b divides one of the following primes $d_1 := 1 + \sqrt{-3}$, $\bar{d}_1 = 1 - \sqrt{-3}$, $d_2 := 2$.

   (d) $\mathbb{Z}[\sqrt{-3}]$ is not GCD domain.

1.3 $\mathbb{D}_U, \mathbb{D}_P$ and $\mathbb{D}_E$ domains

$p \in \mathbb{D}$ is *irreducible* (*prime*) if it is not a unit and every divisor of p is improper. A positive integer $p \in \mathbb{Z}$ is irreducible if and only if p is prime. A linear polynomial in $\mathbb{D}[x_1, ..., x_n]$ is irreducible.

Lemma 1.3.1 *Let $\Omega \subset \mathbb{C}$ be a connected set. Then all irreducible elements of $H(\Omega)$ (up to multiplication by invertible element) are of the form $z - \zeta$ for each $\zeta \in \Omega$.*

Proof. Let $f \in H(\Omega)$ be noninvertible. Then there exists $\zeta \in \Omega$ such that $f(\zeta) = 0$. Hence $z - \zeta | f(z)$. Therefore the only irreducible elements are $z - \zeta$. Clearly $\frac{z-\eta}{z-\zeta}$ is analytic in Ω if and only if $\eta = \zeta$. □

For $\zeta \in \mathbb{C}$ H_ζ has one irreducible element $z - \zeta$.

Definition 1.3.2 $\mathbb{D}$ *is unique factorization domain, or simply UFD and denoted by $\mathbb{D}_U$, if any nonzero, noninvertible element a can be factored as a product of irreducible elements*

$$a = p_1 \cdots p_r, \tag{1.3.1}$$

and these primes are uniquely determined within order and invertible factors.

$\mathbb{Z}$ and H_ζ, $\zeta \in \mathbb{C}$ are $\mathbb{D}_U$. $\mathbb{F}[x_1, ..., x_n]$ is $\mathbb{D}_U$ [Lan67].

Lemma 1.3.3 *Let $\Omega \subset \mathbb{C}$ be a connected open set. Then $H(\Omega)$ is not unique factorization domain.*

Proof. Theorem 1.2.6 yields the existence of a nonzero function $a(z) \in H(\Omega)$ which has a countable infinite number of zeros Ω (which do not accumulate in Ω). Use Lemma 1.3.1 to deduce that a can not be a product of a finite number of irreducible elements. □.

A straightforward consequence of this lemma that for any open set $\Omega \subset \mathbb{C}^n$, $H(\Omega)$ is not $\mathbb{D}_U$. See Problem 2.

Definition 1.3.4 $\mathbb{D}$ *is principal ideal domain, or simply PID and denoted by $\mathbb{D}_P$, if every ideal of $\mathbb{D}$ is principal.*

$\mathbb{Z}$ and $\mathbb{F}[z]$ are $\mathbb{D}_P$. It is known that any $\mathbb{D}_P$ is $\mathbb{D}_U$ [Lan67] or [vdW59]. Thus $H(\Omega)$ is not $\mathbb{D}_P$ for any open connected set $\Omega \subset \mathbb{C}^n$.

Definition 1.3.5 $\mathbb{D}$ *is a Euclidean domain, or simply ED and denoted by $\mathbb{D}_E$, if there exists a function $d : \mathbb{D}\backslash\{0\} \to \mathbb{Z}_+$ such that:*

$$\text{for all } a, b \in \mathbb{D},\ ab \neq 0 \quad d(a) \leq d(ab); \tag{1.3.2}$$

$$\text{for any } a, b \in \mathbb{D},\ ab \neq 0,\ \text{there exists } t, r \in \mathbb{D} \text{ such that}$$

$$a = tb + r,\ \text{where either } r = 0 \text{ or } d(r) < d(b). \tag{1.3.3}$$

We define $d(0) = -\infty$.

Standard examples of Euclidean domains are $\mathbb{Z}$ and $\mathbb{F}[x]$, see Problem 1.

Lemma 1.3.6 *Any ideal* $\{0\} \neq I \subset \mathbb{D}_E$ *is principal.*

Proof. Let $\min_{x\in I\setminus\{0\}} d(x) = d(a)$. Then I is generated by a. □

Lemma 1.3.7 *Let* $\Omega \subset \mathbb{C}$ *be a compact connected set. Then* $H(\Omega)$ *is* $\mathbb{D}_E$. *Here* $d(a)$ *is the number of zeros of a nonzero function* $a \in H(\Omega)$ *counted with their multiplicities.*

Proof. Let a be a nonzero analytic functions on an open connected set $O \supset \Omega$. Since each zero of a is an isolated zero of finite multiplicity, the assumption that Ω is compact yields that a has a finite number of zeros in Ω. Hence $d(a) < \infty$. Let p_a be a nonzero polynomial of degree $d(a)$ such that $a_0 := \frac{a}{p_a}$ does not vanish on Ω. By the definition $d(a) = d(p_a) = \deg p$. Let $a, b \in H(\Omega)$, $ab \neq 0$. Since $\mathbb{C}[z]$ is $\mathbb{D}_E$ we deduce that

$$p_a(z) = t(z)p_b(z) + r(z), \quad r = 0 \text{ or } d(r) < d(p_b).$$

Hence

$$a = \frac{a_0 t}{b_0} b + a_0 r, \quad a_0 r = 0 \text{ or } d(a_0 r) = d(r) < d(p_b) = d(b).$$

□

The Weierstrass preparation theorem [GuR65] can be used to prove the following extension of the above lemma to several complex variables:

Lemma 1.3.8 *Let* $\Omega \subset \mathbb{C}^n$ *be a compact connected set. Then* $H(\Omega)$ *is* $\mathbb{D}_U$.

Let $a_1, a_2 \in D_E\setminus\{0\}$. Assume that $d(a_1) \geq d(a_2)$. The *Euclidean* algorithm consists of a sequence $a_1, ..., a_{k+1}$ which is defined recursively as follows:

$$a_i = t_i a_{i+1} + a_{i+2}, \quad a_{i+2} = 0 \text{ or } d(a_{i+2}) < d(a_{i+1}). \tag{1.3.4}$$

Since $d(a) \geq 0$ the Euclidean algorithm terminates $a_1 \neq 0, \ldots, a_k \neq 0$ and $a_{k+1} = 0$. Hence

$$(a_1, a_2) = a_k. \tag{1.3.5}$$

See Problem 3.

Problems

1. Show that the following domains are Euclidean.

   (a) $\mathbb{Z}$, where $d(a) = |a|$ for any $a \in \mathbb{Z}$.

   (b) $\mathbb{F}[x]$, where $d(p(x)) = \deg p(x)$ for each nonzero polynomial $p(x) \in \mathbb{F}[x]$.

2. Let $\Omega \subset \mathbb{C}^n$ be an open set. Construct a nonzero function f depending on one variable in Ω, which has an infinite number of zeros in Ω. Prove that f can not be decomposed to a finite product of irreducible elements. Hence $H(\Omega)$ is not $\mathbb{D}_U$.

3. Consider the equation (1.3.3) for $r \neq 0$. Show that $(a, b) = (a, r)$. Using this result prove (1.3.5).

1.4 Factorizations in $\mathbb{D}[x]$

Let $\mathbb{F}$ be the field of quotients of $\mathbb{D}$. Assume that $p(x) \in \mathbb{D}[x]$. Suppose that

$$p(x) = p_1(x)p_2(x), \quad p_1(x), p_2(x) \in \mathbb{F}[x].$$

We discuss the problem when $p_1(x), p_2(x) \in \mathbb{D}[x]$. One has to take in account that for any $q(x) \in \mathbb{F}[x]$

$$q(x) = \frac{p(x)}{a}, \quad p(x) \in \mathbb{D}[x],\ a \in \mathbb{D}. \tag{1.4.1}$$

Definition 1.4.1 *Let*

$$p(x) = a_0x^m + \cdots + a_m \in \mathbb{D}[x]. \tag{1.4.2}$$

$p(x)$ is called normalized if $a_0 = 1$. Let $\mathbb{D}$ be GCD domain and denote $c(p) = (a_0, ..., a_m)$. $p(x)$ is called primitive if $c(p) = 1$.

The following result follows from Problem 2.

Lemma 1.4.2 *Let $\mathbb{F}$ be the quotient field of $\mathbb{D}_G$. Then for any $q(x) \in \mathbb{F}[x]$ there exists a decomposition (1.4.1) where $(c(p), a) = 1$. The polynomial $p(x)$ is uniquely determined up to an invertible factor in $\mathbb{D}_G$. Furthermore,*

$$q(x) = \frac{b}{a}r(x), \quad r(x) \in \mathbb{D}_G[x],\ a, b \in \mathbb{D}_G, \tag{1.4.3}$$

where $(a, b) = 1$ and $r(x)$ is primitive.

Lemma 1.4.3 (**Gauss's lemma**) *Let $p(x), q(x) \in \mathbb{D}_U[x]$ be primitive. Then $p(x)q(x)$ is primitive.*

The proof of Gauss lemma follows from the following proposition.

Proposition 1.4.4 *Let $p, q \in \mathbb{D}_G[x]$. Assume that $\pi \in \mathbb{D}$ is a prime element which divides $c(pq)$. Then π divides either $c(p)$ or $c(q)$.*

Proof. Clearly, it is enough to assume that $p, q \neq 0$. We prove the Proposition by induction on $k = \deg p + \deg q$. For $k = 0$ $p(x) = p_0, q(x) = q_0$. Hence $c(pq) = p_0 q_0$. Since $\pi | p_0 q_0$ we deduce that π divides either $p_0 = c(p)$ or $q_0 = c(q)$.

Assume that the proposition holds for $k \leq l$ and assume that $k = l + 1$. Let $p = \sum_{i=0}^{m} a_i x^i, q = \sum_{j=0}^{n} b_j x^j$, where $a_m b_n \neq 0$ and $l + 1 = m + n$. So $\pi | p_m q_m$. Without loss of generality we may assume that nontrivial case $\pi | p_m$ and $m > 0$. Let $r(x) = \sum_{i=0}^{m-1} a_i x^i$. Since $\pi | c(pq)$ it is straightforward that $\pi c(rq)$. As $\deg r + \deg q \leq l$ we deduce that $\pi | c(r)c(q)$. If $\pi | c(q)$ the proposition follows. If $\pi | c(r)$ then $\pi | c(p)$ and the proposition follows in this case too. □

Corollary 1.4.5 *Let $p(x) \in \mathbb{D}_U[x]$ be primitive. Assume that $p(x)$ is irreducible in $\mathbb{F}[x]$, where $\mathbb{F}$ is the quotient field of $\mathbb{D}_U$. Then $p(x)$ is irreducible in $\mathbb{D}_U[x]$.*

Theorem 1.4.6 *Let $\mathbb{F}$ be the quotient field of $\mathbb{D}_U$. Then any $p(x) \in \mathbb{D}_U[x]$ has unique decomposition (up to invertible elements in $\mathbb{D}_U$):*

$$p(x) = aq_1(x) \cdots q_s(x), \quad q_1, ..., q_s \in \mathbb{D}_U[x], \ a \in \mathbb{D}_U, \tag{1.4.4}$$

where $q_1(x), ..., q_s(x)$ are primitive and irreducible in $\mathbb{F}[x]$ and a has decomposition (1.3.1). Hence $\mathbb{D}_U[x]$ is UFD.

See [Lan67] and Problems 3-5.

Normalization 1.4.7 *Let $\mathbb{F}$ be a field an assume that $p(x) \in \mathbb{F}[x]$ is a nonconstant normalized polynomial in $\mathbb{F}[x]$. Let (1.4.4) be a decomposition to irreducible factors. Normalize the decomposition (1.4.4) by letting $q_1(x), ..., q_s(x)$ to be normalized irreducible polynomial in $\mathbb{F}[x]$. (Then $a = 1$.)*

Lemmas 1.4.3 and 1.4.5 yield (see Problem 5):

Theorem 1.4.8 *Let $p(x)$ be a normalized nonconstant polynomial in $\mathbb{D}_U[x]$. Let (1.4.4) be a normalized decomposition in $\mathbb{F}[x]$, where $\mathbb{F}$ is the quotient field of $\mathbb{D}_U$. Then $q_1(x),...,q_s(x)$ are irreducible polynomials in $\mathbb{D}_U[x]$.*

Theorem 1.4.9 *Let $\Omega \subset \mathbb{C}^n$ be a connected set. Assume that $p(x)$ is a normalized nonconstant polynomial in $H(\Omega)[x]$. Let (1.4.4) be a normalized decomposition in $\mathcal{M}[x]$, where $\mathcal{M}$ is the field of meromorphic functions in Ω. Then each $q_j(x)$ is an irreducible polynomial in $H(\Omega)[x]$.*

Proof. By the definition of $H(\Omega)$ we may assume that $p(x) \in H(\Omega_0)[x],\ q_j(x) \in \mathcal{M}(\Omega_0)[x],\ j = 1,...,s$ for some open connected $\Omega_0 \supset \Omega$. Let

$$q(x,z) = x^t + \sum_{r=1}^{t} \frac{\alpha_r(z)}{\beta_r(z)} x^{t-r}, \tag{1.4.5}$$
$$x \in \mathbb{C},\ z \in \Omega_0,\ \alpha_r(z), \beta_r(z) \in H(\Omega_0),\ r = 1,...,t.$$

Then $q(x,z)$ is analytic on $\Omega_0 \backslash \Gamma$, where Γ is an analytic variety given by

$$\Gamma = \{z \in \Omega_0 : \quad \prod_{r=1}^{t} \beta_r(z) = 0\}.$$

Let $x_1(z),...,x_t(z)$ be the roots of $q(x,z) = 0$, which is well defined as unordered set of functions $\{x_1(z),...,x_t(z)\}$ on $\Omega \backslash \Gamma$. Suppose that each $x_k(z)$ is bounded on some neighborhood O of a point $\zeta \in \Gamma$. Then each $\frac{\alpha_j(z)}{\beta_j(z)}$, which is the j symmetric function of $\{x_1(z),...,x_t(z)$, is bounded on O. The Riemann extension theorem [GrH78] implies that $\frac{\alpha_j(z)}{\beta_j(z)}$ is analytic in O. If each $x_k(z)$ is bounded in the neighborhood of each $\zeta \in \Gamma$ it follows that $\frac{\alpha_k(z)}{\beta_k(z)} \in H(\Omega_0),\ k = 1,...,t$.

The assumption that $p(x,z)$ is a normalized polynomial in $H(\Omega_0)$ yields that all the roots of $p(x,z) = 0$ are bounded on any compact set $S \subset \Omega$. The above arguments show that each $q_j(x,z)$ in the decomposition (1.4.4) of $p(x,z)$ is an irreducible polynomial in $H(\Omega)[x]$. □

Problems

1. $a_1,\ldots,a_k \in \mathbb{D}\backslash\{0\}$ are said to have the least common multiple, denoted by $\mathrm{lcm}(a_1,\ldots,a_k)$ and abbreviated as the lcm, if the following conditions hold. Assume that $b \in \mathbb{D}$ is divisible by each $a_i, i = 1,\ldots,k$. Then $\mathrm{lcm}(a_1,\ldots,a_k)|b$. (Note that the lcm is defined up to an invertible element.) Let $\mathbb{D}$ be GCD domain. Show

(a) $\mathrm{lcm}(a_1, a_2) = \frac{a_1 a_2}{(a_1,a_2)}$.

(b) For $k > 2$ $\mathrm{lcm}(a_1, \ldots, a_k) = \frac{\mathrm{lcm}(a_1,\ldots,a_{k-1})a_k}{(\mathrm{lcm}(a_1,\ldots,a_{k-1}),a_k)}$.

2. Let $\mathbb{F}$ be the division field of $\mathbb{D}_G$. Assume that $0 \neq q(x) \in \mathbb{F}[x]$. Write $q(x) = \sum_{i\in I} \frac{b_i}{a_i} x^i$ where $a_i, b_i \in \mathbb{D}_G \backslash \{0\}$ for each $i \in I$, and $I = \{0 \le i_1 < \ldots < i_k\}$ is a finite subset of $\mathbb{Z}_+$. Let $a'_i = \frac{a_i}{(a_i,b_i)}, b'_i = \frac{b_i}{(a_i,b_i)}$ for $i \in I$. Then (1.4.1) holds, where $a = \mathrm{lcm}(a'_{i_1}, \ldots, a'_{i_k})$ and $p(x) = \sum_{i\in I} \frac{b'_i a}{a'_i} x^i$. Show that $(c(p), a) = 1$. Furthermore, if $q(x) = \frac{r(x)}{c}$ for some $r(x) \in \mathbb{D}_G[x], c \in \mathbb{D}_G$ then $c = ea, r(x) = ep(x)$ for some $e \in \mathbb{D}_G \backslash \{0\}$.

3. Let $p(x)$ be given by (1.4.2) and put

$$q(x) = b_0 x^n + \cdots + b_n, \quad r(x) = p(x)q(x) = c_0 x^{m+n} + \cdots + c_{m+n}.$$

Assume that $p(x), q(x) \in \mathbb{D}_U[x]$. Let π be an irreducible element in $\mathbb{D}_U$ such that

$$\pi | a_i,\ i = 0, ..., \alpha, \quad \pi | b_j,\ j = 0, ..., \beta, \quad \pi | c_{\alpha+\beta+2}.$$

Then either $\pi | a_{\alpha+1}$ or $\pi | b_{\beta+1}$.

4. Prove that if $p(x), q(x) \in \mathbb{D}_U[x]$ then $c(pq) = c(p)c(q)$.

Deduce from the above equality Lemma 1.4.3. Also if $p(x)$ and $q(x)$ normalized polynomials then $p(x)q(x)$ is primitive.

5. Prove Theorem 1.4.8.

6. Using the equality $\mathbb{D}[x_1, ..., x_{n-1}][x_n] = \mathbb{D}[x_1, ..., x_n]$ prove that $\mathbb{D}_U[x_1, ..., x_n]$ is UFD. Deduce that $\mathbb{F}[x_1, ..., x_n]$ is UFD.

1.5 Elementary Divisor Domain

Definition 1.5.1 $\mathbb{D}_G$ *is elementary divisor domain, or simply EDD and denoted by* $(\mathbb{D}_{ED})$*, if for any three elements* $a, b, c \in \mathbb{D}$ *there exists* $p, q, x, y \in \mathbb{D}$ *such that*

$$(a, b, c) = (px)a + (py)b + (qy)c. \tag{1.5.1}$$

By letting $c = 0$ we obtain that (a, b) is a linear combination of a and b. Hence an elementary divisor domain is a Bezout domain.

Theorem 1.5.2 *Let $\mathbb{D}$ be a principal ideal domain. Then $\mathbb{D}$ is an elementary divisor domain.*

Proof. Without loss of generality we may assume that $abc \neq 0,\ (a, b, c) = 1$. Let $(a, c) = d$. Since $\mathbb{D}$ is $\mathbb{D}_U$ ([Lan67]), we decompose $a = a'a''$, where in the prime decomposition (1.3.1) of a, a' contains all the irreducible factors of a, which appear in the decomposition of d to irreducible factors. Thus

$$a = a'a'',\ (a', a'') = 1,\ (a', c) = (a, c),\ (a'', c) = 1, \tag{1.5.2}$$
$$\text{and if } a', f \text{ are not coprime then } c, f \text{ are not coprime.}$$

Hence there exists q and α such that

$$b - 1 = -qc + \alpha a''. \tag{1.5.3}$$

Let $d' = (a, b + qc)$. The above equality implies that $(d', a'') = 1$. Suppose that d' is not coprime with a'. Then there exists a noninvertible f such that f divides d' and a'. According to (1.5.2) $(f, c) = f'$ and f' is not invertible. Thus $f'|b$ which implies that f' divides a, c and b. Contradictory to our assumption that $(a, b, c) = 1$. So $(d', a') = 1$ which implies $(d', a) = 1$. Therefore there exists $x, y \in \mathbb{D}$ such that $xa + y(b + qc) = 1$. This shows (1.5.1) with $p = 1$. □

Theorem 1.5.3 *Let $\Omega \subset \mathbb{C}$ be a connected set. Then $H(\Omega)$ is an elementary divisor domain.*

Proof. Given $a, b, c \in H(\Omega)$ we may assume that $a, b, c \in H(\Omega_0)$ for some open connected set $\Omega_0 \supset \Omega$. Theorem 1.2.6 yields

$$(a, b, c) = (a, (b, c)) = a + y(b, c) = a + y(b + qc). \tag{1.5.4}$$

□

Problems

1. $\mathbb{D}$ is called *adequate* if for any $0 \neq a, c \in \mathbb{D}$ (1.5.2) holds. Use the proof of Theorem 1.5.2 to show that any adequate $\mathbb{D}_B$ is $\mathbb{D}_{ED}$.

2. Prove that for any connected set $\Omega \subset \mathbb{C}$, $H(\Omega)$ is an adequate domain ([Hel43]).

1.6 Modules

Definition 1.6.1 M *is an abelian group if it has a binary operation, denoted by +, which satisfies the conditions* (1.1.1 − 1.1.5).

Definition 1.6.2 *Let S be a ring with identity. An abelian group $\mathbf{M}$, which has an operation $+$, is called a (left) S-module if for each $r \in S$, $\mathbf{v} \in \mathbf{M}$ the product $r\mathbf{v}$ is an element of $\mathbf{M}$ such that the following properties hold:*

$$\begin{aligned} &r(\mathbf{v}_1+\mathbf{v}_2)=r\mathbf{v}_1+r\mathbf{v}_2, \quad (r_1+r_2)\mathbf{v}=r_1\mathbf{v}+r_2\mathbf{v}, \\ &(rs)\mathbf{v}=r(s\mathbf{v}), \quad 1\mathbf{v}=\mathbf{v}. \end{aligned} \tag{1.6.1}$$

$\mathbf{N} \subset \mathbf{M}$ is called a submodule if $\mathbf{N}$ is an S-module.

Assume that S does not have zero divisors. (I.e. if $r, s \in S$ and $rs = 0$ then either $r = 0$ or $s = 0$.) Then $\mathbf{M}$ does not have zero divisors if

$$r\mathbf{v}=\mathbf{0} \textit{ if and only if } \mathbf{v}=\mathbf{0} \textit{ for any } r \neq 0. \tag{1.6.2}$$

Assume that $\mathbb{D}$ is a domain. Then $\mathbf{M}$ is a called a $\mathbb{D}$-module if in addition to the above property $\mathbf{M}$ does not have zero divisors.

A standard example of S-module is

$$S^m := \{\mathbf{v}=(v_1,...,v_m)^\top : \quad v_i \in S,\ i=1,...,m\}, \tag{1.6.3}$$

where

$$\begin{aligned} &\mathbf{u}+\mathbf{v}=(u_1+v_1,...,u_m+v_m)^\top, \\ &r\mathbf{u}=(ru_1,...,ru_m)^\top, \quad r \in S. \end{aligned} \tag{1.6.4}$$

Note that if S does not have zero divisors then S^n is an S-module with no zero divisors.

Definition 1.6.3 *A $\mathbb{D}$-module $\mathbf{M}$ is finitely generated if there exist n-elements (generators) $\mathbf{v}_1,...,\mathbf{v}_n \in \mathbf{M}$ such that any $\mathbf{v} \in \mathbf{M}$ is of the form*

$$\mathbf{v}=\sum_{i=1}^{n} a_i\mathbf{v}_i, \quad a_i \in \mathbb{D},\ i=1,...,n. \tag{1.6.5}$$

If each $\mathbf{v}$ can be expressed uniquely in the above form then $\mathbf{v}_1,...,\mathbf{v}_n$ is called a basis in $\mathbf{M}$, and $\mathbf{M}$ is said to have a finite basis. We denote by $[\mathbf{v}_1,...,\mathbf{v}_n]$ a basis in $\mathbf{M}$.

Note that $\mathbb{D}^m$ has a standard basis $\mathbf{v}_i=(\delta_{i1},\ldots,\delta_{in})^\top, i=1,\ldots,n$.

Let $\mathbb{F}$ be a field. Then an $\mathbb{F}$-module is called a vector space $\mathbf{V}$ over $\mathbb{F}$. It is a standard fact in linear algebra [HJ88] that a finitely generated $\mathbf{V}$ has a finite basis. A finitely generated vector space is called *finite dimensional.* The number of vectors of a basis for a finite dimensional vector space $\mathbf{V}$ is constant. It is called the *dimension* of $\mathbf{V}$, and is denoted by dim $\mathbf{V}$. A submodule of $\mathbf{V}$ is called a *subspace* of $\mathbf{V}$.

Let $\mathbf{M}$ be a $\mathbb{D}$-module with a finite basis. Let $\mathbb{F}$ be a quotient ring of $\mathbb{D}$. It is possible to imbed $\mathbf{M}$ in a vector space $\mathbf{V}$ by considering all vectors $\mathbf{v}$ of the form (1.6.5), where $a_i \in \mathbb{F},\ i = 1, ..., n$. (For more general statement see Problem 1.) Thus dim $\mathbf{V} = \mathbf{n}$. Using this fact we obtain:

Lemma 1.6.4 *Any two finite bases of a* $\mathbb{D}$ *module contain the same number of elements* dim $\mathbf{V}$.

One of the standard examples of submodules in $\mathbb{D}^n$ is as follows. Consider the linear homogeneous system

$$\sum_{j=1}^{n} a_{ij}x_j = 0, \quad a_{ij}, x_j \in \mathbb{D},\ i = 1, ..., m,\ j = 1, ..., n. \tag{1.6.6}$$

Then the set of solutions $\mathbf{x} = (x_1, ..., x_n)^\top$ is a submodule of $\mathbb{D}^n$. In §1.12 we show that the above module has a basis if $\mathbb{D}$ is a Bezout domain.

Definition 1.6.5 *Let* $\mathbf{M}$ *be a module over* $\mathbb{D}$. *Assume that* $\mathbf{M}_i$ *is a submodule of* $\mathbf{M}$ *for* $i = 1, \ldots, k$. *Then* $\mathbf{M}$ *is called a direct sum of* $\mathbf{M}_1, \ldots, \mathbf{M}_k$*, and denoted as* $\mathbf{M} = \oplus_{i=1}^k \mathbf{M}_i$*, if every element* $\mathbf{m} \in \mathbf{M}$ *can be expressed in unique way as a sum* $\mathbf{m} = \sum_{i=1}^k \mathbf{m}_i$*, where* $\mathbf{m}_i \in \mathbf{M}_i$ *for* $i = 1, \ldots, k$.

Definition 1.6.6 *Then ring of quaternions* $\mathbb{H}$ *is a four dimensional vector space over* $\mathbb{R}$ *with the basis* $1, \mathbf{i}, \mathbf{j}, \mathbf{k}$*, i.e. vectors of the form*

$$\mathbf{q} = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \quad a, b, c, d \in \mathbb{R}, \tag{1.6.7}$$

where

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1,\ \mathbf{ij} = -\mathbf{ji} = \mathbf{k},\ \mathbf{jk} = -\mathbf{kj} = \mathbf{i},\ \mathbf{ki} = -\mathbf{ik} = \mathbf{j}. \tag{1.6.8}$$

It is known that $\mathbb{H}$ is a noncommutative division algebra over $\mathbb{R}$. See Problem 5.

Problems

1. Let $\mathbf{M}$ be a finite generated module over $\mathbb{D}$. Let $\mathbb{F}$ be the quotient field of $\mathbb{D}$. Show

(a) Assume that $\mathbf{M}$ is generated by $\mathbf{a}_1, \ldots, \mathbf{a}_m$. Let $\mathbf{N} = \{\mathbf{x} = (x_1, \ldots, x_m)^\top \in \mathbb{D}^m, \sum_{i=1}^m x_i \mathbf{a}_i = \mathbf{0}\}$. Then $\mathbf{N}$ is a $\mathbb{D}$-module.

(b) Let $\mathbf{U} \subseteq \mathbb{F}^m$ be the subspace generated by all vectors in $\mathbf{N}$. (Any vector in $\mathbf{U}$ is a finite linear combination of vectors in $\mathbf{N}$.) Then any vector in $\mathbf{u} \in \mathbf{U}$ is of the form $\frac{1}{c}\mathbf{b}$, where $\mathbf{b} \in \mathbf{N}$.

(c) Let $\mathbf{V} = \mathbb{F}^m/\mathbf{U}$. ($\mathbf{V}$ can be constructed as follows. Assume that $\dim \mathbf{U} = l$. Pick a basis $[\mathbf{u}_1, \ldots, \mathbf{u}_l]$ in $\mathbf{U}$ and complete this basis to a basis in $\mathbb{F}^m$. So $[\mathbf{u}_1, \ldots, \mathbf{u}_l, \mathbf{w}_1, \ldots, \mathbf{w}_{m-l}]$ is a basis in $\mathbb{F}^m$. Let $\mathbf{W} = \text{span}\,(\mathbf{w}_1, \ldots, \mathbf{w}_{m-l})$. Then any vector in $\mathbf{V}$ is of the form of a coset $\mathbf{w} + \mathbf{U}$ for a unique vector $\mathbf{w} \in \mathbf{W}$.)

(d) Define $\phi : \mathbf{M} \to \mathbf{V}$ as follows. Let $\mathbf{a} \in \mathbf{M}$ and write $\mathbf{a} = \sum_{i=1}^m a_i \mathbf{a}_i$. Set $\phi(\mathbf{a}) = (a_1, \ldots, a_m)^\top + \mathbf{U}$. Then

i. ϕ is well defined, i.e. does not depend on a particular representation of $\mathbf{a}$ as a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_m$.

ii. $\phi(\mathbf{a}) = \phi(\mathbf{b}) \iff \mathbf{a} = \mathbf{b}$.

iii. $\phi(a\mathbf{a} + b\mathbf{b}) = a\phi(\mathbf{a}) + b\phi(\mathbf{b})$ for any $a, b \in \mathbb{D}$ and $\mathbf{a}, \mathbf{b} \in \mathbf{M}$.

iv. For any $\mathbf{v} \in \mathbf{V}$ there exists $a \in \mathbb{D}$ and $\mathbf{a} \in \mathbf{M}$ such that $\phi(\mathbf{a}) = a\mathbf{v}$.

(e) Let $\mathbf{Y}$ be a finite dimensional vector space over $\mathbb{F}$ with the following properties.

i. There is an injection $\phi : \mathbf{M} \to \mathbf{Y}$, i.e. ϕ is one to one, such that $\phi(a\mathbf{m} + b\mathbf{n}) = a\phi(\mathbf{m}) + b\phi(\mathbf{n})$ for any $a, b \in \mathbb{D}$ and $\mathbf{m}, \mathbf{n} \in \mathbf{M}$.

ii. For any $\mathbf{y} \in \mathbf{Y}$ there exists $a \in \mathbb{D}$ and $\mathbf{m} \in \mathbf{M}$ such that $\phi(\mathbf{m}) = a\mathbf{y}$.

Then $\dim \mathbf{X} = \dim \mathbf{V}$, where $\mathbf{V}$ is defined in 1c.

Definition 1.6.7 *$\mathbb{D}$-module* $\mathbf{M}$ *is called k-dimensional, if* $\mathbf{M}$ *is finitely generated and* $\dim \mathbf{V} = k$.

2. Let $\mathbf{M}$ be a $\mathbb{D}$-module with a finite basis. Let $\mathbf{N}$ be a submodule of $\mathbf{M}$. Show that if $\mathbb{D}$ is $\mathbb{D}_P$ then $\mathbf{N}$ has a finite basis.

3. Let $\mathbf{M}$ be a $\mathbb{D}$-module with a finite basis. Assume that $\mathbf{N}$ is a finitely generated submodule of $\mathbf{M}$. Show that if $\mathbb{D}$ is $\mathbb{D}_B$ then $\mathbf{N}$ has a finite basis.

4. Let $\mathbf{M}$ be a module over $\mathbb{D}$. Assume that $\mathbf{M}_i$ is a submodule of $\mathbf{M}$ for $i = 1, \ldots, k$. Then $\mathbf{N} := M_1 + \ldots + M_k$ is the set of all $\mathbf{m}$ of the form $\mathbf{m}_1 + \ldots + \mathbf{m}_k$, where $\mathbf{m}_i \in \mathbf{M}_i$ for $i = 1, \ldots, k$. Show

(a) $\mathbf{N}$ is a submodule of $\mathbf{M}$.

(b) $\cap_{i=1}^{k}\mathbf{N}_i$ is a submodule of $\mathbf{M}$.

(c) Assume that $\mathbf{M}_1, \ldots, \mathbf{M}_k$ are finitely generated. Then $\mathbf{N}$ is finitely generated and $\dim \mathbf{N} \le \sum_{i=1}^{k} \dim \mathbf{M}_i$.

(d) Assume that $\mathbf{M}_1, \ldots, \mathbf{M}_k$ are have bases and $\mathbf{N} = \oplus_{i=1}^{k}\mathbf{M}_i$. Then $\mathbf{N}$ has a basis and $\dim \mathbf{N}l = \sum_{i=1}^{k} \dim \mathbf{M}_i$.

5. Show

(a) $\mathbb{H}$ can be viewed as $\mathbb{C}^2$, where each $\mathbf{q} \in \mathbb{H}$ of the form (1.6.7) can be written as $\mathbf{q} = z + w\mathbf{j}$, where $z = a + b\mathbf{i}, w = c + d\mathbf{i} \in \mathbb{C}$. Furthermore, for any $z \in \mathbb{C}$, $\mathbf{j}z = \bar{z}\mathbf{j}$.

(b) $\mathbf{H}$ is a ring with the identity $1 = 1 + 0\mathbf{i} + 0\mathbf{j} + 0\mathbf{k}$.

(c) $(r\mathbf{q})\mathbf{s} = \mathbf{q}(r\mathbf{s})$ for any $\mathbf{q}, \mathbf{s} \in \mathbf{H}$ and $r \in \mathbb{R}$. Hence $\mathbb{H}$ is an algebra over $\mathbb{R}$.

(d) Denote $|\mathbf{q}| = \sqrt{a^2 + b^2 + c^2 + d^2}$, $\bar{\mathbf{q}} = a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$ for any $\mathbf{q}$ of the form (1.6.7). Then $\bar{\mathbf{q}}\mathbf{q} = \mathbf{q}\bar{\mathbf{q}} = |\mathbf{q}|^2$. Hence $|\mathbf{q}|^{-2}\bar{\mathbf{q}}$ is the right and the left inverse of $\mathbf{q} \neq 0$.

1.7 Algebraically closed fields

Definition 1.7.1 *A field $\mathbb{F}$ is algebraically closed if any polynomial $p(x) \in \mathbb{F}[x]$ of the form* (1.4.2) *splits to linear factors in $\mathbb{F}$:*

$$p(x) = a_0 \prod_{i=1}^{m} (x - \xi_i), \quad \xi_i \in \mathbb{F},\ i = 1, ..., m,\ a_0 \neq 0. \tag{1.7.1}$$

The classical example of an algebraically closed field is the field of complex numbers $\mathbb{C}$. The field of real numbers $\mathbb{R}$ is not algebraically closed.

Definition 1.7.2 *Let $\mathbb{K} \supset \mathbb{F}$ be fields. Then $\mathbb{K}$ is an extension field of $\mathbb{F}$. $\mathbb{K}$ is called a finite extension of $\mathbb{F}$ if $\mathbb{K}$ is a finite dimensional vector space over $\mathbb{F}$. The dimension of the vector space $\mathbb{K}$ over $\mathbb{F}$ is called the degree of $\mathbb{K}$ and is denoted by* $[\mathbb{K} : \mathbb{F}]$.

Thus $\mathbb{C}$ is a finite extension of $\mathbb{R}$ of degree 2. It is known [Lan67], see Problems 1-2:

Theorem 1.7.3 *Let $p(x) \in \mathbb{F}[x]$. Then there exists a finite extension $\mathbb{K}$ of $\mathbb{F}$ such that $p(x)$ splits into linear factors in $\mathbb{K}[x]$.*

The classical Weierstrass preparation theorem in two complex variables is an explicit example of the above theorem. We state the Weierstrass preparation theorem in a form needed later [GuR65].

Theorem 1.7.4 *Let* H_0 *be the ring of analytic functions in one variable in the neighborhood of the origin* $0 \in \mathbb{C}$. *Let* $p(\lambda) \in \mathrm{H}_0[\lambda]$ *be a normalized polynomial of degree* n

$$p(\lambda, z) = \lambda^n + \sum_{j=1}^{n} a_j(z)\lambda^{n-j}, \quad a_j(z) \in \mathrm{H}_0,\ j = 1, ..., n. \tag{1.7.2}$$

Then there exists a positive integer $s|n!$ *such that*

$$p(\lambda, w^s) = \prod_{j=1}^{n}(\lambda - \lambda_j(w)), \quad \lambda_j(w) \in \mathrm{H}_0,\ j = 1, ..., n. \tag{1.7.3}$$

In this particular case the extension field $\mathbb{K}$ of $\mathbb{F} = \mathcal{M}_0$ is the set of multivalued functions in z, which are analytic in $z^{\frac{1}{s}}$ in the neighborhood of the origin. Thus $\mathbb{K} = \mathcal{M}_0(w)$, where

$$w^s = z. \tag{1.7.4}$$

The degree of $\mathbb{K}$ over $\mathbb{F}$ is s.

Problems

1. Let $\mathbb{F}$ be a field and assume that $p(x) = x^d + a_d x^{d-1} + \ldots + a_1 \in \mathbb{F}[x]$, where $d > 1$. On the vector field $\mathbb{F}^d$ define a product as follows. $(b_1, \ldots, b_d)(c_1, \ldots, c_d) = (r_1, \ldots, r_d)$, where $(r_1, \ldots, r_d)$ defined as follows. Let $b(x) = \sum_{i=1}^{d} b_i x^{i-1}, c(x) = \sum_{i=1}^{d} c_i x^{i-1}, r(x) = \sum_{i=1}^{d} r_i x^{i-1}$. Then $r(x)$ is the remainder of $b(x)c(x)$ be the division by $p(x)$. I.e. $b(x)c(x) = g(x)p(x) + r(x)$ where $\deg r(x) < d$. Let $\mathcal{P}_d$ be $\mathbb{F}^d$ with the above product.

   Show

   (a) $\mathcal{P}_d$ is a commutative ring with identity $\mathbf{e}_1 = (1, 0, \ldots, 0)$.

   (b) $\mathbb{F}$ is isomorphic to span $(\mathbf{e}_1)$, where $f \mapsto f\mathbf{e}_1$.

   (c) Let $\mathbf{e}_i = (\delta_{1i}, \ldots, \delta_{di}), i = 2, \ldots, d$. Then

   $$\mathbf{e}_2^i = \mathbf{e}_{1+i}, i = 0, \ldots, d-1, \quad p(\mathbf{e}_2) = 0.$$

   (d) $\mathcal{P}_d$ is a domain if and only if $p(x)$ is an irreducible polynomial over $\mathbb{F}[x]$.

(e) $\mathcal{P}_d$ is a field if and only if $p(x)$ is an irreducible polynomial over $\mathbb{F}[x]$.

(f) Assume that $p(x) \in \mathbb{F}[x]$ is irreducible. Then $\mathbb{K} := \mathcal{P}_d$ is an extension field of $\mathbb{F}$ with $[\mathbb{K} : \mathbb{F}] = d$. Furthermore $p(x)$ viewed as $p(x) \in \mathbb{K}[x]$ decompose to $p(x) = (x - \mathbf{e}_2)q(x)$, where $q(x) = x^{d-1} + \sum_{i=1}^{d} g_i x^{i-1} \in \mathbb{K}[x]$.

2. Let $\mathbb{F}$ be a field and $p(x) \in \mathbb{F}[x]$. Show that there exists a finite extension field $\mathbb{K}$ such that $p(x)$ splits in $\mathbb{K}$. Furthermore $[\mathbb{K} : \mathbb{F}] \le (\deg p)!$

1.8 The resultant and the discriminant

Let $\mathbb{D}$ be an integral domain. Suppose that

$$p(x) = a_0 x^m + \cdots + a_m,\ q(x) = b_0 x^n + \cdots + b_n \in \mathbb{D}[x]. \tag{1.8.1}$$

Assume furthermore that $m, n \ge 1$ and $a_0 b_0 \ne 0$. Let $\mathbb{F}$ be the quotient field $\mathbb{D}$ and assume that $\mathbb{K}$ is a finite extension of $\mathbb{F}$ such that $p(x)$ and $q(x)$ split to linear factors in $\mathbb{K}$. That is

$$\begin{aligned} p(x) &= a_0 \prod_{i=1}^{m} (x - \xi_i), \qquad \xi_i \in \mathbb{K},\ i = 1, ..., m,\ a_0 \ne 0. \\ q(x) &= b_0 \prod_{j=1}^{n} (x - \eta_j), \quad \eta_j \in \mathbb{K},\ j = 1, ..., n,\ b_0 \ne 0. \end{aligned} \tag{1.8.2}$$

Then the resultant $R(p, q)$ of p, q and the discriminant $D(p)$ of p are defined as follows.

$$\begin{aligned} R(p, q) &= a_0^n b_0^m \prod_{i,j=1}^{m,n} (\xi_i - \eta_j), \\ D(p) &= a_0^{m(m-1)} \prod_{1 \le i < j \le m} (\xi_i - \xi_j)^2. \end{aligned} \tag{1.8.3}$$

It is a classical result that $R(p, q) \in \mathbb{D}[a_0, ..., a_m, b_0, ..., b_n]$ and $D(p) \in \mathbb{D}[a_0, ..., a_m]$, e.g. [vdW59]. More precisely, we have.

Theorem 1.8.1 *Let*

$$\mathbf{a} = (a_0, \ldots, a_m) \in \mathbb{D}^{m+1}, \mathbf{b} = (b_0, \ldots, b_n) \in \mathbb{D}^{n+1},$$
$$p(x) = \sum_{i=0}^{m} a_i x^{m-i}, q(x) = \sum_{j=0}^{n} b_j x^{n-j}.$$

Then $R(p,q) = \det C(\mathbf{a}, \mathbf{b})$, *where*

$$C(\mathbf{a},\mathbf{b}) = \begin{bmatrix} a_0 & a_1 & a_2 & \ldots & a_m & 0 & 0 & \ldots & 0 \\ 0 & a_0 & a_1 & \ldots & a_{m-1} & a_m & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_0 & a_1 & a_2 & \ldots & a_m \\ b_0 & b_1 & \ldots & b_{n-2} & b_{n-1} & b_n & 0 & \ldots & 0 \\ 0 & b_0 & b_1 & \ldots & b_{n-2} & b_{n-1} & b_n & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \ldots & b_0 & b_1 & b_2 & \ldots & b_n \end{bmatrix}$$

is an $(m+n) \times (m+n)$ *matrix.*

Proof. Let $\mathbb{F}$ be the quotient field of $\mathbb{D}$, and assume that $p, q \in \mathbb{D}[x]$ split in a finite extension field $\mathbb{K}$ of $\mathbb{F}$. Let $c(x) = \sum_{i=0}^{n-1} c_i x^{n-1-i}, d(x) = \sum_{j=0}^{m-1} d_j x^{m-1-j} \in \mathbb{F}[x]$. Then $c(x)p(x) + d(x)q(x) = \sum_{l=0}^{m+n-1} g_l x^{m+n-1-l}$. Denote

$\mathbf{f} = (c_0, \ldots, c_{n-1}, d_0, \ldots, d_{m-1}), \mathbf{g} = (g_0, \ldots, g_{m+n-1}) \in \mathbb{D}^{m+n}$. A straightforward calculation show that $\mathbf{f}C(\mathbf{a},\mathbf{b}) = \mathbf{g}$.

Assume that $\det C(\mathbf{a},\mathbf{b}) \neq 0$. Let $\mathbf{f} = (0, \ldots, 0, 1)C(\mathbf{a},\mathbf{b})^{-1}$. Hence there exists $c(x), d(x) \in \mathbb{F}[x]$ of the above form such that $c(x)p(x) + d(x)q(x) = 1$. Hence p, q do not have common zeros in $\mathbb{K}$.

We now show that if $a_0 b_0 \neq 0$ then $R(p,q) = \det C(\mathbf{a},\mathbf{b})$. Divide the first n rows of $C(\mathbf{a},\mathbf{b})$ by a_0 and the last m rows of $C(\mathbf{a},\mathbf{b})$ by b_0, to deduce that it is enough to show the equality $R(p,q) = \det C(\mathbf{a},\mathbf{b})$ in the case $a_0 = b_0 = 1$. Then $p(x) = \prod_{i=1}^{m}(x - u_i), q(x) = \prod_{j=1}^{n}(x - v_j) \in \mathbb{K}[x]$. Recall that $(-1)^i a_i$ and $(-1)^j b_j$ the $i-th$ and $j-th$ elementary symmetric polynomials in $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$, respectively:

$$(1.8.4) \qquad a_i = (-1)^i \sum_{1 \le l_1 < \ldots < l_i \le m} u_{l_1} \ldots u_{l_i}, \quad i = 1, \ldots, m,$$
$$b_j = (-1)^j \sum_{1 \le l_1 < \ldots < l_j \le n} v_{l_1} \ldots v_{l_j}, \quad j = 1, \ldots, n.$$

Then $C(\mathbf{a},\mathbf{b})$ is a matrix with polynomial entries in $\mathbf{u} = (u_1,\ldots,u_m), \mathbf{v} = (v_1,\ldots,v_n)$. Hence $s(\mathbf{u},\mathbf{v}) := \det C(\mathbf{a},\mathbf{b})$ is a polynomial in $m+n$ variables.

Assume that $u_i = v_j$ for some $i \in [1,m], j \in [1,n]$. Then $p(x)$ and $q(x)$ have a common factor $x - u_i = x - v_j$. The above arguments shows that $s(\mathbf{u},\mathbf{v}) = 0$. Hence $s(\mathbf{u},\mathbf{v})$ is divisible by $t(\mathbf{u},\mathbf{v}) = \prod_{i=1,j=1}^{m,n}(u_i - v_j)$. So $s(\mathbf{u},\mathbf{v}) = h(\mathbf{u},\mathbf{v})t(\mathbf{u},\mathbf{v})$, for some polynomial $h(\mathbf{u},\mathbf{v})$.

Consider $s(\mathbf{u},\mathbf{v}), t(\mathbf{u},\mathbf{v}), h(\mathbf{u},\mathbf{v})$ as polynomials in $\mathbf{v}$ with coefficients in $\mathbb{D}[\mathbf{u}]$. Then $\deg_{\mathbf{v}} t(\mathbf{u},\mathbf{v}) = nm$ and the term of the highest degree is $(-1)^{mn}v_1^m \ldots v_n^m$. Observe next that the contribution of the variables $\mathbf{v}$ in $\det C(\mathbf{a},\mathbf{b})$ comes from it last m rows. The term of the maximal degree in each such row is n which comes only from $b_n = (-1)^n v_1 \ldots v_n$. Hence the coefficient of the product b_n^m comes from the minor of $C(\mathbf{a},\mathbf{b})$ based on the first n rows and columns. Clearly, this minor is equal to $a_0^n = 1$. So $h(\mathbf{u},\mathbf{v})$ is only polynomial in $\mathbf{u}$. Furthermore $h(\mathbf{u}) = 1$.

□

If $\mathbb{F}$ is a field of characteristic 0 then

$$D(p) = \pm a_0^{-1} R(p,p'). \tag{1.8.5}$$

Note that if a_i, b_i are given the weight i for $i = 0, ...$, then $R(p,q)$ and $D(p)$ are polynomials with total degrees mn and $m(m-1)$ respectively. See Problem 4.

Problems

1. Let $\mathbb{D}$ be a domain and assume that $p(x) = x^m, q(x) = (x+1)^n$. Show

   (a) $R(p,q) = 1$.

   (b) Let $\mathbf{a} = (1,0,\ldots,0) \in \mathbb{D}^{m+1}, \mathbf{b} = (\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}) \in \mathbb{D}^{n+1}$. Let $C(\mathbf{a},\mathbf{b})$ be defined as in Theorem 1.8.1. Then $\det C(\mathbf{a},\mathbf{b}) = 1$.

2. Let $\mathbf{u} = (u_1,\ldots,u_m), \mathbf{v} = (v_1,\ldots,v_n)$. Assume that each $a_i \in \mathbb{D}[\mathbf{u}], b_j \in \mathbb{D}[\mathbf{v}]$, is a multilinear polynomial for $i = 0,\ldots,m, j = 0,\ldots,n$. (The degree of a_i, b_j with respect to any variable is at most 1.) Let $C(\mathbf{a},\mathbf{b})$ be defined as in Theorem 1.8.1. Show that $\det C(\mathbf{a},\mathbf{b})$ is a polynomial of degree at most n and m with respect to u_i and v_j respectively, for any $i = 1,\ldots,m$ and $j = 1,\ldots,n$.

3. Let the assumptions of Theorem 1.8.1 hold. Show

   (a) If $a_0 = b_0$ then $\det C(\mathbf{a},\mathbf{b}) = 0$.

(b) Assume that $p(x)$ is not a zero polynomial and $a_0 = 0, b_0 \neq 0$. Then det $C(\mathbf{a}, \mathbf{b}) = 0$ if and only if p and q have a common root in an extension field $\mathbb{K}$ of $\mathbb{F}$, where p and q split.

4. Let $C(\mathbf{a}, \mathbf{b})$ be defined as Theorem 1.8.1. View det $C(\mathbf{a}, \mathbf{b})$ as a polynomial $F(\mathbf{a}, \mathbf{b})$. Assume that the weight $\omega(a_i) = i, \omega(b_j) = j$. Then the weight of a monomial in the variables $\mathbf{a}, \mathbf{b}$ is the sum of the weights of each variable times the number of times in appears in this monomial. Show

   (a) Each nontrivial monomial in $F(\mathbf{a}, \mathbf{b})$ is of weight mn.

   (b) Assume as in the proof of Theorem 1.8.1 that $a_0 = b_0 = 1$ and a_i and b_j are the $i-th$ and $j-th$ elementary symmetric polynomials in $\mathbf{u}$ and $\mathbf{v}$ respectively. Then each nontrivial monomial in $\mathbf{u}, \mathbf{v}$ appearing in $F(\mathbf{a}(\mathbf{u}), \mathbf{b}(\mathbf{v}))$ is of total degree mn.

1.9 The ring $\mathbb{F}[x_1, ..., x_n]$

In 1.2 we pointed out that $\mathbb{F}[x_1, ..., x_n]$ is not $\mathbb{D}_B$ for $n \geq 2$. It is known [Lan67] that $\mathbb{F}[x_1, ..., x_n]$ is Noetherian:

Definition 1.9.1 *$\mathbb{D}$ is Noetherian, denoted by $\mathbb{D}_N$, if any ideal of $\mathbb{D}$ is finitely generated.*

In what follows we assume that $\mathbb{F}$ is algebraically closed. Let $p_1, ..., p_k \in \mathbb{F}[x_1, ..., x_n]$. Denote by $U(p_1, ..., p_k)$ the common set of zeros of $p_1, ..., p_k$:

$$U(p_1, ..., p_k) = \{\mathbf{x} = (x_1, ..., x_n)^T : \quad p_j(\mathbf{x}) = 0,\ j = 1, ..., k\}. \tag{1.9.1}$$

$U(p_1, ..., p_k)$ may be an empty set. $U(p_1, ..., p_k)$ is called an algebraic variety (in $\mathbb{F}^n$). It is known [Lan67] that any nonempty variety in $\mathbb{F}^n$ splits as

$$U = \cup_{i=1}^k V_i, \tag{1.9.2}$$

where each V_i is an irreducible algebraic variety, which is not contained in any other V_j. Over $\mathbb{C}$ each irreducible variety $V \subset \mathbb{C}^n$ is a closed connected set. Furthermore, there exists a strict subvariety $W \subset V$ (of singular points of V) such that $V \backslash W$ is a connected analytic manifold of complex dimension d in $\mathbb{C}^n$. dim $V := d$ is called the dimension of V. If $d = 0$ then V consists of one point. For any set $U \subset \mathbb{F}^n$ let $I(U)$ be the ideal of polynomials vanishing on U:

$$I(U) = \{p \in \mathbb{F}[x_1, ..., x_n] : \quad p(\mathbf{x}) = 0,\ \forall \mathbf{x} \in U\}. \tag{1.9.3}$$

Theorem 1.9.2 (Hilbert Nullstellensatz) *Let $\mathbb{F}$ be an algebraically closed field. Let $I \subseteq \mathbb{F}[x_1, ..., x_n]$ be an ideal generated by $p_1, ..., p_k$. Assume that $g \in \mathbb{F}[x_1, ..., x_n]$. Then $g^j \in I$ for some positive integer j if and only if $g \in I(U(p_1, ..., p_k))$.*

Corollary 1.9.3 *Let $p_1, ..., p_k \in \mathbb{F}[x_1, ..., x_n]$, where $\mathbb{F}$ is algebraically closed field. Then $p_1, ..., p_k$ generate $\mathbb{F}[x_1, ..., x_n]$ if and only if $U(p_1, ..., p_k) = \emptyset$.*

1.10 Matrices and homomorphisms

Notation 1.10.1 *For a set S denote by $S^{m\times n}$ the set of all $m \times n$ matrices $A = [a_{ij}]_{i=j=1}^{i=m,j=n}$, where each $a_{ij} \in S$.*

Definition 1.10.2 *Let $\mathbf{M}$, $\mathbf{N}$ be $\mathbb{D}$-modules. Let $T : \mathbf{N} \to \mathbf{M}$. T is a homomorphism if*

$$T(a\mathbf{u} + b\mathbf{v}) = aT\mathbf{u} + bT\mathbf{v}, \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbf{N},\ a, b \in \mathbb{D}. \tag{1.10.1}$$

Let

$$\begin{aligned} \text{Range } T &= \{\mathbf{u} \in \mathbf{M} : \quad \mathbf{u} = T\mathbf{v},\ \mathbf{v} \in \mathbf{N}\}, \\ \text{Ker } T &= \{\mathbf{v} \in \mathbf{N} : \quad T\mathbf{v} = \mathbf{0}\}, \end{aligned}$$

be the range and the kernel of T. Denote by $\mathrm{Hom}(\mathbf{N}, \mathbf{M})$ *the set of all homomorphisms of* $\mathbf{N}$ *to* $\mathbf{M}$.

$T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M})$ *is an isomorphism if there exists $Q \in \mathrm{Hom}(\mathbf{M}, \mathbf{N})$ such that QT and TQ are the identity maps on* $\mathbf{M}$ *and* $\mathbf{N}$ *respectively.* $\mathbf{M}$ *and* $\mathbf{N}$ *are isomorphic if there exists an isomorphism $T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M})$.*

$\mathrm{Hom}(\mathbf{N}, \mathbf{M})$ is a $\mathbb{D}$-module with

$$(aS + bT)\mathbf{v} = aS\mathbf{v} + bT\mathbf{v}, \quad a, b \in \mathbb{D},\ S, T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M}),\ \mathbf{v} \in \mathbf{N}.$$

Assume that $\mathbf{M}$ and $\mathbf{N}$ have finite bases. Let $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\mathbf{v}_1, ..., \mathbf{v}_n]$ be bases in $\mathbf{M}$ and $\mathbf{N}$ respectively. Then there exists a natural isomorphism between $\mathrm{Hom}(\mathbf{N}, \mathbf{M})$ and $\mathbb{D}^{m\times n}$. For each $T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M})$ let $A = [a_{ij}] \in \mathbb{D}^{m\times n}$ be defined as follows:

$$T\mathbf{v}_j = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i, \quad j = 1, ..., n. \tag{1.10.2}$$

Conversely, for each $A = [a_{ij}] \in \mathbb{D}^{m\times n}$ there exists a unique $T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M})$ which satisfies (1.10.2). The matrix A is called the representation matrix of T in the bases $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\mathbf{v}_1, ..., \mathbf{v}_n]$.

Notation 1.10.3 *For a positive integer n denote $[n] = \{1, \ldots, n\}$. For $k \in [n]$ denote by $[n]_k$ the set of all subsets of $[n]$ of cardinality k. Each $\alpha \in [n]_k$ is represented by $(\alpha_1, \ldots, \alpha_k)$, where $\alpha_1, \ldots, \alpha_k$ are integers satisfying $1 \le \alpha_1 < \ldots < \alpha_k \le n$.*

Definition 1.10.4 *Let $\mathbb{D}$ be a domain and $A \in \mathbb{D}^{m\times n}$. Assume that $\alpha = (\alpha_1, \ldots, \alpha_k) \in [m]_k, \beta = (\beta_1, \ldots, \beta_l) \in \ [n]_l$. Denote by $A[\alpha, \beta] = [a_{\alpha_i\beta_j}]_{i,j=1}^{k,l}$ the $k \times l$ submatrix of A. For $k = l$, $\det A[\alpha, \beta]$ is called an (α, β) minor, k-minor, or simply a minor of A. The rank of A, denoted by $\mathrm{rank}\ A$, is the maximal size of a nonvanishing minor of A. (The rank of the zero matrix is 0.) The nullity of A, denoted by $\mathrm{nul}\ A$, is $n - \mathrm{rank\ A}$.*

Any $A \in \mathbb{D}^{m\times n}$ can be viewed as $T \in \mathrm{Hom}(\mathbb{D}^n, \mathbb{D}^m)$, where $T\mathbf{x} := A\mathbf{x},\ \mathbf{x} = (x_1, ..., x_n)^\top$. We will sometime denote T by A. If $\mathbb{D}$ is $\mathbb{D}_B$ then Range A has a finite basis of dimension rank A (Problem 1).

We now study the relations between the representation matrices of a fixed $T \in \mathrm{Hom}(\mathbf{N}, \mathbf{M})$ with respect to different bases in $\mathbf{M}$ and $\mathbf{N}$.

Definition 1.10.5 *$U \in \mathbb{D}^{n\times n}$ is called invertible (unimodular) if $\det U$ is an invertible element in $\mathbb{D}$.*

Proposition 1.10.6 *$U \in \mathbb{D}^{n\times n}$ is invertible if and only if there exists $V \in \mathbb{D}^{n\times n}$ such that either UV or VU is equal to the identity matrix I.*

Proof. Let $\mathbb{F}$ be the divison field of $\mathbb{D}$. Assume first that $\det U$ is an invertible element in $\mathbb{D}$. Then U is an invertible matrix in $\mathbb{F}^{n\times n}$, where

$$U^{-1} = \frac{1}{\det U}\ \mathrm{adj}\ U, \tag{1.10.3}$$

$$\mathrm{adj}\ A = [(-1)^{i+j}\det A[[n]\setminus\{j\}, [n]\setminus\{i\}]]_{i=j=1}^n. \tag{1.10.4}$$

Clearly $V := U^{-1} \in \mathbb{D}^{n\times n}$. Assume now that there exists $V \in \mathbb{D}^{n\times n}$ such that $VU = I$. Then $1 = \det VU = \det V \det U$ and $\det \mathrm{U}^{-1} = \det \mathrm{V} \in \mathbb{D}$. Similarly $\det U$ is invertible in $\mathbb{D}$ if $UV = I$. □

Notation 1.10.7 *Denote by $\mathbf{GL}(n, \mathbb{D})$ the group of invertible matrices in $\mathbb{D}^{n\times n}$.*

Lemma 1.10.8 *Let $\mathbf{M}$ be a $\mathbb{D}$-module with a finite basis $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$. Then $[\mathbf{u}_1, ..., \mathbf{u}_m]$ is a basis in $\mathbf{M}$ if and only if the matrix $Q = [q_{ki}] \in \mathbb{D}^{m\times m}$ given by the equalities*

$$\mathbf{u}_i = \sum_{k=1}^m q_{ki}\tilde{\mathbf{u}}_k, \quad i = 1, ..., m, \tag{1.10.5}$$

is an invertible matrix.

Proof. Suppose first that $[\mathbf{u}_1, ..., \mathbf{u}_m]$ is a basis in $\mathbf{M}$. Then

$$\tilde{\mathbf{u}}_k = \sum_{l=1}^{m} r_{lk}\mathbf{u}_l, \quad k = 1, ..., m. \tag{1.10.6}$$

Let $R = [r_{lk}]_1^m$. Insert (1.10.6) to (1.10.5) and use the assumption that $[\mathbf{u}_1, ..., \mathbf{u}_m]$ is a basis to obtain that $RQ = I$. Proposition 1.10.6 yields that $Q \in \mathbf{GL}(m, \mathbb{D})$. Assume now that Q is invertible. Let $R = Q^{-1}$. Hence (1.10.6) holds. It is straightforward to deduce that $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ is a basis in $\mathbf{M}$. □

Definition 1.10.9 *Let $A, B \in \mathbb{D}^{m\times n}$. Then A and B are right equivalent, left equivalent and equivalent if the following conditions hold respectively:*

(1.10.7) $B = AP \quad$ for some $P \in \mathbf{GL}(n, \mathbb{D}) \quad (A \sim_r B)$,

(1.10.8) $B = QA \quad$ for some $Q \in \mathbf{GL}(m, \mathbb{D}) \quad (A \sim_l B)$,

(1.10.9) $B = QAP \quad$ for some $P \in \mathbf{GL}(n, \mathbb{D}),\ Q \in \mathbf{GL}(m, \mathbb{D}) \quad (A \sim B)$.

Clearly, all the above relations are equivalence relations.

Theorem 1.10.10 *Let $\mathbf{M}$ and $\mathbf{N}$ be $\mathbb{D}$-modules with finite bases having m and n elements respectively. Then $A, B \in \mathbb{D}^{m\times n}$ represent some $T \in$ $\mathrm{Hom}(\mathbf{N}, \mathbf{M})$ in certain bases as follows:*
(l) *$A \sim_l B$ if and only if A and B represent T in the corresponding bases of $\mathbf{U}$ and $\mathbf{V}$ respectively*

$$[\mathbf{u}_1, ..., \mathbf{u}_m],\ [\mathbf{v}_1, ..., \mathbf{v}_n] \quad \text{and} \quad [\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m],\ [\mathbf{v}_1, ..., \mathbf{v}_n]. \tag{1.10.10}$$

(r) *$A \sim_r B$ if and only if A and B represent T in the corresponding bases of $\mathbf{U}$ and $\mathbf{V}$ respectively*

$$[\mathbf{u}_1, ..., \mathbf{u}_m],\ [\mathbf{v}_1, ..., \mathbf{v}_n] \quad \text{and} \quad [\mathbf{u}_1, ..., \mathbf{u}_m],\ [\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_n]. \tag{1.10.11}$$

(e) *$A \sim B$ if and only if A and B represent T in the corresponding bases of $\mathbf{U}$ and $\mathbf{V}$ respectively*

$$[\mathbf{u}_1, ..., \mathbf{u}_m],\ [\mathbf{v}_1, ..., \mathbf{v}_n] \quad \text{and} \quad [\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m],\ [\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_n]. \tag{1.10.12}$$

Sketch of a proof. Let A be the representation matrix of T in the bases $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\mathbf{v}_1, ..., \mathbf{v}_n]$ given in (1.10.2). Assume that the relation between the bases $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ is given by (1.10.5). Then

$$T\mathbf{v}_j = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i = \sum_{i=k=1}^{m} q_{ki}a_{ij}\tilde{\mathbf{u}}_k, \quad j = 1, ..., n.$$

Hence the representation matrix B in bases $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ and $[\mathbf{v}_1, ..., \mathbf{v}_n]$ is given by (1.10.8).

Change the basis $[\mathbf{v}_1, ..., \mathbf{v}_n]$ to $[\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_n]$ according to

$$\tilde{\mathbf{v}}_j = \sum_{l=1}^{n} p_{lj}\mathbf{v}_l, \ j = 1, ..., n, \quad P = [p_{lj}] \in \mathbf{GL}(n, \mathbb{D}).$$

Then a similar computation shows that T is presented in the bases $[\mathbf{u}_1, ..., \mathbf{u}_m]$ and $[\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_n]$ by AP. Combine the above results to deduce that the representation matrix B of T in bases $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ and $[\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_n]$ is given by (1.10.9). □

Problems

1. Let $A \in \mathbb{D}_B^{m\times n}$. View A a as linear transformation from $A : \mathbb{D}_B^n \to \mathbb{D}_B^m$. to show that Range A is a module with basis of dimension rank A. (**Hint**: Use Problem 1.6.3.)

2. For $A, B \in \mathbb{D}^{m\times n}$ show.

   (a) If $A \sim_l B$ then Ker $A =$ Ker B and Range A and Range B are isomorphic.

   (b) $A \sim_r B$ then Range $A =$ Range B and Ker A and Ker B are isomorphic.

1.11 Hermite normal form

We start this section with two motivating problems.

Problem 1.11.1 *Given $A, B \in \mathbb{D}^{m\times n}$. When are A and B*
(l) *left equivalent;*
(r) *right equivalent;*
(e) *equivalent.*

Problem 1.11.2 *For a given $A \in \mathbb{D}^{m\times n}$ characterize the equivalence classes corresponding to the left equivalence, to the right equivalence and to the equivalence relation as defined in Problem* 1.11.1.

For $\mathbb{D}_G$ the equivalence relation has the following natural invariants:

Lemma 1.11.3 *For $A \in \mathbb{D}_G^{m\times n}$ let*

$$
\begin{aligned}
&\mu(\alpha, A) := \text{g.c.d. } (\{\det A[\alpha,\theta],\ \theta \in [n]_k\}, \quad \alpha \in [m]_k),\\
&\nu(\beta, A) := \text{g.c.d. } (\{\det A[\phi,\beta],\ \phi \in [m]_k\}), \quad \beta \in [n]_k,\\
&\delta_k(A) := \text{g.c.d. } (\{\det A[\phi,\theta],\ \phi \in [m]_k,\ \theta \in [n]_k\}),
\end{aligned}
\tag{1.11.1}
$$

($\delta_k(A)$ is called the k-th determinant invariant of A.) Then

$$
\begin{aligned}
&\mu(\alpha, A) \equiv \mu(\alpha, B) \quad \text{for all } \alpha \in [m]_k \quad \text{if } A \sim_r B,\\
&\nu(\beta, A) \equiv \nu(\beta, B) \quad \text{for all } \beta \in [n]_k \quad \text{if } A \sim_l B,\\
&\delta_k(A) \equiv \delta_k(B) \quad \text{if } A \sim B,
\end{aligned}
\tag{1.11.2}
$$

for $k = 1, ..., \min(m, n)$. (Recall that for $a, b \in \mathbb{D}$ $a \equiv b$ if $a = bc$ for some invertible $c \in \mathbb{D}$.)

Proof. Suppose that (1.10.7) holds. Then the Cauchy-Binet formula (e.g. [Gan59]) implies

$$\det B[\alpha,\gamma] = \sum_{\theta\in[n]_k} \det A[\alpha,\theta] \det P[\theta,\gamma].$$

Hence $\mu(\alpha, A)$ divides $\mu(\alpha, B)$. As $A = BP^{-1}$ we get $\mu(\alpha, B)|\mu(\alpha, A)$. Thus $\mu(\alpha, A) \equiv \mu(\alpha, B)$. The other equalities in (1.11.2) are established in a similar way. □

Clearly

$$A \sim_l B \iff A^T \sim_r B^T, \quad A, B \in \mathbb{D}^{m\times n}. \tag{1.11.3}$$

Hence it is enough to consider the left equivalence relation. We characterize the left equivalence classes for Bezout domains $\mathbb{D}_B$. To do that we need a few notations.

Recall that $P \in \mathbb{D}^{n\times n}$ is called a permutation matrix if P is a matrix having at each row and each column one nonzero element which is equal to the identity element 1. A permutation matrix is invertible since $P^{-1} = P^T$.

Definition 1.11.4 *Let $\Pi_n \subset \mathbf{GL}(n, \mathbb{D})$ be the group of $n \times n$ permutation matrices.*

Definition 1.11.5 *An invertible matrix $U \in \mathbf{GL}(n, \mathbb{D})$ is called simple if there exists $P, Q \in \Pi_n$ such that*

$$U = P \begin{bmatrix} V & 0 \\ 0 & I_{n-2} \end{bmatrix} Q, \tag{1.11.4}$$

where

$$V = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \in \mathbf{GL}(2, \mathbb{D}), \quad (\alpha\delta - \beta\gamma \text{ is invertible}). \tag{1.11.5}$$

U is called elementary if U is of the form (1.11.4) *and*

$$V = \begin{bmatrix} \alpha & \beta \\ 0 & \delta \end{bmatrix} \in \mathbf{GL}(2, \mathbb{D}), \ \textit{and } \alpha,\ \delta \textit{ are invertible.} \tag{1.11.6}$$

Definition 1.11.6 *Let $A \in \mathbb{D}^{m\times n}$. The following row* (*column*) *operations are called elementary:*
(a) *interchange any two rows* (*columns*) *of A;*
(b) *multiply row* (*column*) *i by an invertible element a;*
(c) *add to row* (*column*) *j b times row* (*column*) *i* ($i \neq j$)*.*
The following row (*column*) *operation is called simple:*
(d) *replace row* (*column*) *i by a times row* (*column*) *i plus b times row* (*column*) *j,*
and row (*column*) *j by c times row* (*column*) *i plus d times row* (*column*) *j, where $i \neq j$ and $ad - bc$ is invertible in $\mathbb{D}$.*

It is straightforward to see that the elementary row (column) operations can be carried out by multiplication of A by a suitable elementary matrix from left (right), and the simple row (column) operations are carried out by multiplication of A by a simple matrix U from (left) right.

Theorem 1.11.7 *Let $\mathbb{D}_B$ be a Bezout domain. Let $A \in \mathbb{D}_B^{m\times n}$. Assume that* rank A = r*. Then there exists $B = [b_{ij}] \in \mathbb{D}_B^{m\times n}$ which is equivalent to A and satisfies the following conditions:*

$$i - \text{th row of } B \text{ is a nonzero row if and only if } i \leq r. \tag{1.11.7}$$

Let b_{in_i} be the first nonzero entry in i-th row for $i = 1, ..., r$. Then

$$1 \leq n_1 < n_2 < \cdots < n_r \leq n. \tag{1.11.8}$$

The numbers $n_1, ..., n_r$ are uniquely determined and the elements b_{in_i}, $i = 1, ..., r$, which are called pivots, are uniquely determined, up to invertible factors, by the conditions

$$\begin{aligned} &\nu((n_1, ..., n_i), A) = b_{1n_1} \cdots b_{in_i}, \quad i = 1, ..., r, \\ &\nu(\alpha, A) = 0,\ \alpha \in Q_{i,n_i-1}, \quad i = 1, ..., r. \end{aligned} \tag{1.11.9}$$

For $1 \le j < i \le r$, adding to the row j a multiple of the row i does not change the above form of B. Assume that $B = [b_{ij}], C = [c_{ij}] \in \mathbb{D}^{m\times n}$ are left equivalent to A and satisfy the above conditions. If $b_{jn_i} = c_{jn_i}$, $j = 1, ..., i, i = 1, \ldots, r$ then $B = C$. The invertible matrix Q which satisfies (1.10.8) *can be given by a finite product of simple matrices.*

Proof. Clearly, it is enough to consider the case $A \ne 0$, i.e. $r \ge 1$. Our proof is by induction on n and m. For $n = m = 1$ the theorem is obvious. Let $n = 1$ and assume that for a given $m \ge 1$ there exists a matrix Q, which is a finite product of simple matrices, such that the entries $(i, 1)$ of Q are zero for $i = 2, ..., m$ if $m \ge 2$. Let $A_1 = [a_{i1}] \in \mathbb{D}_B^{(m+1)\times 1}$ and denote by A the submatrix $[a_{i1}]_{i=1}^m$. Set

$$Q_1 := \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix}.$$

Then the $(i, 1)$ entries of $A_2 = [a_{i1}^{(2)}] = Q_1 A_1$ are equal to zero for $i = 2, ..., m$. Interchange the second and the last row of A_2 to obtain A_3. Clearly $A_3 = [a_{i1}^{(3)}] = Q_2 A_2$ for some permutation matrix Q_2. Let $A_4 = (a_{11}^{(3)}, a_{21}^{(3)})^\top$. As $\mathbb{D}_B$ is Bezout domain there exists $\alpha, \beta \in \mathbb{D}_B$ such that

$$\alpha a_{11}^{(3)} + \beta a_{21}^{(3)} = (a_{11}^{(3)}, a_{21}^{(3)}) = d. \tag{1.11.10}$$

As $(\alpha, \beta) = 1$ there exists $\gamma, \delta \in \mathbb{D}_B$ such that

$$\alpha\delta - \beta\gamma = 1. \tag{1.11.11}$$

Let V be a 2×2 invertible matrix given by (1.11.5). Then

$$A_5 = V A_4 = \begin{bmatrix} d \\ d' \end{bmatrix}.$$

Lemma 1.11.3 implies $\nu((1), A_5) = \nu((1), A_4) = d$. Hence $d' = pd$ for some $p \in \mathbb{D}_B$. Thus

$$\begin{bmatrix} d \\ 0 \end{bmatrix} = W A_5, \quad W = \begin{bmatrix} 1 & 0 \\ -p & 1 \end{bmatrix} \in \mathbf{GL}(2, \mathbb{D}_B).$$

Let

$$Q_3 = \begin{bmatrix} W & 0 \\ 0 & I_{m-1} \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_{m-1} \end{bmatrix}.$$

Then the last m rows of $A_6 = [a_{i1}^{(6)}] = Q_3 A_3$ are zero rows. So $a_{11}^{(6)} = \nu((1), A_6) = \nu((1), A_1)$ and the theorem is proved in this case.

Assume now that we proved the theorem for all $A_1 \in \mathbb{D}_B^{m\times n}$ where $n \le p$. Let $n = p+1$ and $A \in \mathbb{D}_B^{m\times(p+1)}$. Let $A_1 = [a_{ij}]_{i=j=1}^{mp}$. The induction hypothesis implies the existence of $Q_1 \in \mathbf{GL}(m, \mathbb{D}_B)$, which is a finite product of simple matrices, such that $B_1' = [b_{ij}^{(1)}]_{i=j=1}^{mp} = Q_1 A_1$ satisfies the assumptions of our theorem. Let $n_1', ..., n_s'$ be the integers defined by A_1. Let $B_1 = [b_{ij}^{(1)}]_{i=j=1}^{mn} = Q_1 A$. If $b_{in}^{(1)} = 0$ for $i > s$ then $n_i = n_i'$, $i = 1, ..., s$ and B_1 is in the right form. Suppose now that $b_{in}^{(1)} \ne 0$ for some $s < i \le m$. Let $B_2 = [b_{in}^{(1)}]_{i=s+1}^{m} \in \mathbb{D}_B^{(m-s)\times 1}$. We proved above that there exists $Q_2 \in \mathbf{GL}(m-s, \mathbb{D}_B)$ such that $Q_2 B_2 = (c, 0, ..., 0)^T$, $c \ne 0$. Then

$$B_3 = \begin{bmatrix} I_s & 0 \\ 0 & Q_2 \end{bmatrix} B_1$$

is in the right form with

$$s = r-1,\ n_1 = n_1', ..., n_{r-1} = n_{r-1}',\ n_r = n.$$

We now show (1.11.9). First if $\alpha \in [n_i - 1]_i$ then any matrix $B[\beta|\alpha]$, $\beta \in [m]_i$ has at least one zero row. Hence $\det B[\beta|\alpha] = 0$. Therefore $\nu(\alpha, B) = 0$. Lemma 1.11.3 yields that $\nu(\alpha, A) = 0$. Let $\alpha = (n_1, ..., n_i)$. Then $B[\beta|\alpha]$, $\beta \in [m]_i$. Then $B[\beta|\alpha]$ has at least one zero row unless β is equal to $\gamma = (1, 2, ..., i)$. Therefore

$$\nu(\alpha, A) = \nu(\alpha, B) = \det B[\gamma|\alpha] = b_{1n_1} \cdots b_{in_i} \ne 0.$$

This establishes (1.11.9).

It is obvious that $b_{1n_1}, ..., b_{rn_r}$ are determined up to invertible elements. For $1 \le j < i \le r$ we can perform the following elementary row operation on B: add to row j a multiple of row i. The new matrix C will satisfy the assumption of the theorem. It is left to show that if $B = [b_{ij}], C = [c_{ij}] \in \mathbb{D}_B^{m\times n}$ are left equivalent to A, have the same form given by the theorem and satisfying

$$b_{jn_i} = c_{jn_i}, \quad j = 1, ..., i,\ i = 1, ..., r, \tag{1.11.12}$$

then $B = C$. See Problem 1. □

A matrix $B \in \mathbb{D}_B^{m\times n}$ is said to to be in a *Hermite normal form*, abbreviated as *HNF*, if it satisfies conditions (1.11.7-1.11.8).

Normalization 1.11.8 *Let $B = [b_{ij}] \in \mathbb{D}_B^{m\times n}$ in a Hermite normal form. If b_{in_i} is invertible we set $b_{in_i} = 1$ and $b_{jn_i} = 0$ for $i < j$.*

Theorem 1.11.9 *Let U be an invertible matrix over a Bezout domain. Then U is a finite product of simple matrices.*

Proof. Since $\det U$ is invertible, Theorem 1.11.7 yields that b_{ii} is invertible. Normalization 1.11.8 implies that the Hermite normal form of U is I. Hence the inverse of U is a finite product of simple matrices. Therefore U itself is a finite product of simple matrices. □

Normalization 1.11.10 *For Euclidean domains assume*

$$\textit{either } b_{jn_i} = 0 \quad \textit{or } d(b_{jn_i}) < d(b_{in_i}) \textit{ for } j < i. \tag{1.11.13}$$

For $\mathbb{Z}$ we assume that $b_{in_i} \ge 1$ and $0 \le b_{jn_i} < b_{in_i}$ for $j < i$. For $\mathbb{F}[x]$ we assume that b_{in_i} is a normalized polynomial.

Corollary 1.11.11 *Let $\mathbb{D}_E = \mathbb{Z}, \mathbb{F}[x]$. Under Normalization* 1.11.10 *any $A \in \mathbb{D}_E^{m\times n}$ has a unique Hermite normal form.*

It is a well known fact that over Euclidean domains Hermite normal form can be achieved by performing elementary row operations.

Theorem 1.11.12 *Let $A \in \mathbb{D}_E^{m\times n}$. Then $B = QA$, $Q \in \mathbf{GL}(m, \mathbb{D}_E)$ where B is in a Hermite normal form satisfying Normalization* 1.11.10 *and Q is a product of finite elementary matrices.*

Proof. From the proof of Theorem 1.11.7 it follows that it is enough to show that any $A \in \mathbf{GL}(2, \mathbb{D}_E)$ is a finite product of elementary invertible matrices in $\mathbf{GL}(2, \mathbb{D}_E)$. As I_2 is the Hermite normal form of any 2×2 invertible matrix, it suffices to show that any $A \in \mathbb{D}_E^{2\times 2}$ can be brought to its Hermite form by a finite number of elementary row operations. Let

$$A_i = \begin{bmatrix} a_i & b_i \\ a_{i+1} & b_{i+1} \end{bmatrix}, \quad A_1 = PA,$$

where P is a permutation matrix such that $d(a_1) \ge d(a_2)$. Suppose first that $a_2 \ne 0$. Compute a_{i+2} by (1.3.4). Then

$$A_{i+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -t_i \\ 0 & 1 \end{bmatrix} A_i, \quad i = 1, ...$$

As the Euclidean algorithm terminates after a finite number of steps we obtain that $a_{k+1} = 0$. Then A_k is the Hermite normal form of A. If

$b_{k+1} = 0$ we are done. If $b_{k+1} \neq 0$ substract from the first row of A_k a corresponding multiple of the second row of A_k to obtain the matrix

$$B' = \begin{bmatrix} a_k & b'_k \\ 0 & b_{k+1} \end{bmatrix}, \quad d(b_{k+1}) > d(b'_k).$$

Multiply each row of B' by invertible element if necessary to obtain Hermite's form of B according to Normalization 1.11.10. We obtained B by a finite number of elementary row operations. If $a_1 = a_2 = 0$ perform the Euclid algorithm on the second column of A. □

Corollary 1.11.13 *Let $U \in \mathbf{GL}(n, \mathbb{D}_E)$. Then U is a finite product of elementary invertible matrices.*

Corollary 1.11.14 *Let $\mathbb{F}$ be a field. Then $A \in \mathbb{F}^{m\times n}$ can be brought to its unique reduced row echelon form given by Theorem* 1.11.7 *with*

$$b_{in_i} = 1,\ b_{jn_i} = 0, \quad j = 1, ..., i-1,\ i = 1, ..., r,$$

by a finite number of elementary row operations.

Problems

1. Show

   (a) Let $A, B \in \mathbb{D}^{m\times m}$ be two upper triangular matrices with the same nonzero diagonal entries. Assume that $QA = B$ for some $Q \in \mathbb{D}^{m\times m}$. Then Q is un upper triangular matrix with one on the main diagonal. (**Hint**: First prove this claim for the quotient field $\mathbb{F}$ of $\mathbb{D}$.)

   (b) Let $Q \in \mathbb{D}^{m\times m}$ be an upper triangular matrix with 1 on the main diagonal. Show that $Q = R_2 \dots R_m = T_m \dots T_2$ where $R_i - I_m, Q_i - I_m$ may have nonzero entries only in the places (j, i) for $j = 1, \dots, i-1$ and $i = 2, \dots, m$.

   (c) Let $A, B \in \mathbb{D}^{m\times n}$. Assume that $A \sim_l B$, and A and B are in *HNF* and have the same pivots. Then B can be obtained from A, by adding multiples of the row b_{in_i} to the rows $j = 1, \dots, i-1$ for $i = 2, \dots, r$.

   (d) Let $B = [b_{ij}], C = [c_{ij}] \in \mathbb{D}^{m\times n}$. Assume that B, C are in Hermite's normal form with the same r numbers $1 \le n_1 < \cdots <$

$n_r \le n$.Suppose furthermore (1.11.12) holds and $B = QC$ for some $Q \in \mathbb{D}^{m\times m}$. Then

$$Q = \begin{bmatrix} I_r & * \\ 0 & * \end{bmatrix} \quad \Rightarrow \quad B = C.$$

(Here $*$ denotes a matrix of a corresponding size.)

(e) Let $\mathbf{M}$ be a $\mathbb{D}_B$ module, $\mathbf{N} = \mathbb{D}_B^n$ and $T \in \mathrm{Hom}\,(\mathbf{N}, \mathbf{M})$. Let Range (T) be the range of T in $\mathbf{M}$. Then the module Range (T) has a basis $T\mathbf{u}_1, ..., T\mathbf{u}_k$ such that

$$\mathbf{u}_i = \sum_{j=1}^{i} c_{ij}\mathbf{v}_j, \quad c_{ii} \ne 0, \quad i = 1, ..., k, \tag{1.11.14}$$

where $\mathbf{v}_1, ..., \mathbf{v}_n$ is a permutation of the standard basis

$$\mathbf{e}_i = (\delta_{i1}, ..., \delta_{in})^T, \quad i = 1, ..., n. \tag{1.11.15}$$

2. Let $A \in \mathbb{D}_B^{m\times n}$ and assume that B is it's Hermite's normal form. Assume that $n_i < j < n_{i+1}$. Prove that

$$\nu(\alpha, A) = b_{1n_1} \cdots b_{(i-1)n_{i-1}} b_{ij}, \quad \text{for } \alpha = (n_1, ..., n_{i-1}, j).$$

3. **Definition 1.11.15** *Let $\mathbb{F}$ be a field and $\mathbf{V}$ a vector space over $\mathbb{F}$ of dimension n. A flag $\mathbf{F}_*$ on $\mathbf{V}$ is a strictly increasing sequence of subspaces*

$$\begin{gathered} 0 = \mathbf{F}_0 \subset \mathbf{F}_1 \subset \cdots \subset \mathbf{F}_n = \mathbf{V}, \\ \dim \mathbf{F}_i = i, \quad i = 1, ..., n = \dim V. \end{gathered} \tag{1.11.16}$$

Show

(a) Let $\mathbf{L}$ be a subspace of $\mathbf{V}$ of dimension ℓ. Then

$$\dim \mathbf{L}\cap\mathbf{F}_{i-1} \le \dim \mathbf{L}\cap\mathbf{F}_i \le \dim \mathbf{L}\cap\mathbf{F}_{i-1} + 1, \; i = 1, ..., n. \tag{1.11.17}$$

(b) Let $\mathrm{Gr}(\ell, \mathbf{V})$ be the space of all ℓ-dimensional subspaces of $\mathbf{V}$. Let $J = \{1 \le j_1 < \cdots < j_\ell \le n\}$ be a subset of $< n >$ of cardinality $\ell = |J|$. Then

$$\begin{gathered} \Omega^o(J, \mathbf{F}_*) := \{\mathbf{L} \in \mathrm{Gr}(\ell, \mathbf{V}) : \quad \dim \mathbf{L} \cap \mathbf{F}_{j_i} = i, \quad i = 1, ..., \ell\}, \\ \Omega(J, \mathbf{F}_*) := \{\mathbf{L} \in \mathrm{Gr}(\ell, \mathbf{V}) : \quad \dim \mathbf{L} \cap \mathbf{F}_{j_i} \ge i, \quad i = 1, ..., \ell\}, \end{gathered} \tag{1.11.18}$$

which are called the open and the closed Schubert cell in $\mathrm{Gr}(\ell, \mathbf{V})$ respectively. Show that a given $\mathbf{L} \in \mathrm{Gr}(\ell, \mathbf{V})$ belongs to the smallest open Grassmanian cell Ω^o_J, where $J = J(\mathbf{L}, \mathbf{F}_*)$ given by the condition

$$\dim \mathbf{L} \cap \mathbf{F}_{j_i} = i, \quad \dim \mathbf{L} \cap \mathbf{F}_{j_i - 1} = i - 1, \quad i = 1, ..., \ell. \tag{1.11.19}$$

(c) Let $\mathbf{V} = \mathbb{F}^n$ and assume that $\mathbf{e}_1, .., \mathbf{e}_n$ is the standard basis of $\mathbb{F}^n$. Let

$$\mathbf{F}_i = \text{span} \ (\mathbf{e}_n, \mathbf{e}_{n-1}, ..., \mathbf{e}_{n-i+1}), \quad i = 1, ..., n \tag{1.11.20}$$

be the *reversed* standard flag in $\mathbb{F}^n$. Let $A \in \mathbb{F}^{m \times n}$. Assume that $\ell = \mathrm{rank\ A} \geq 1$. Let $\mathbf{L} \in \mathrm{Gr}(\ell, \mathbb{F}^n)$ be the vector space spanned by the columns of A^T. Let $N = \{1 \leq n_1 < \cdots < n_\ell \leq n\}$ be the integers given by the row echelon form of A. Then $J(\mathbf{L}, \mathbf{F}_*) = N$.

1.12 Systems of linear equations over Bezout domains

Consider a system of m linear equations in n unknowns:

$$\begin{gathered} \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, ..., m, \\ a_{ij}, b_i \in \mathbb{D}, \ i = 1, ..., m, \ j = 1, ..., n. \end{gathered} \tag{1.12.1}$$

In matrix notation (1.12.1) is equivalent to

$$A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{D}^{m \times n}, \ \mathbf{x} \in \mathbb{D}^n, \ \mathbf{b} \in \mathbb{D}^m. \tag{1.12.2}$$

Let

$$\hat{A} = [A, \mathbf{b}] \in \mathbb{D}^{m \times (n+1)}. \tag{1.12.3}$$

The matrix A is called the *coefficient* matrix and the matrix $\hat{A}$ is called the *augmented coefficient* matrix. If $\mathbb{D}$ is a field, the classical Kronecker-Capelli theorem states [Gan59] that (1.12.1) is solvable if and only if

$$\mathrm{rank\ A} = \mathrm{rank\ \hat{A}}. \tag{1.12.4}$$

Let $\mathbb{F}$ be the quotient field of $\mathbb{D}$. If (1.12.1) is solvable over $\mathbb{D}$ it is also solvable over $\mathbb{F}$. Therefore (1.12.4) is a necessary condition for the solvability of (1.12.1) over $\mathbb{D}$. Clearly, even in the case $m = n = 1$ this condition is not sufficient. In this section we give necessary and sufficient conditions on $\hat{A}$ for the solvability of (1.12.1) over a Bezout domain. First we need the following lemma:

Lemma 1.12.1 *Let* $0 \neq A \in \mathbb{D}_B^{m\times n}$. *Then there exist* $P \in \Pi_m$, $U \in$ $\mathbf{GL}(n, \mathbb{D}_B)$ *such that*

$$\begin{aligned} &C = [c_{ij}] = PAU, \\ &c_{ii} \neq 0, \quad i = 1, ..., \mathrm{rank\ A}, \\ &c_{ij} = 0 \quad \text{if either } j > i \textit{ or } j > \mathrm{rank\ A}. \end{aligned} \tag{1.12.5}$$

Proof. Consider the matrix $A^\top$. By interchanging the columns of A^T, i.e. multiplying $A^\top$ from the right by some permutation matrix P^T, we can assume that the Hermite normal form of $A^\top P^\top$ satisfies $n_i = i$, $i = 1, ..., \mathrm{rank\ A}$. □

Theorem 1.12.2 *Let* $\mathbb{D}$ *be a Bezout domain. Then the system* (1.12.1) *is solvable if and only if*

$$r = \mathrm{rank\ A} = \mathrm{rank\ \hat{A}}, \quad \mathrm{\delta_r(A) \equiv \delta_r(\hat{A})}. \tag{1.12.6}$$

Proof. Assume first the existence of $\mathbf{x} \in \mathbb{D}^n$ which satisfies (1.12.2). Hence (1.12.4) holds, i.e. the first part of (1.12.6) holds. As any minor $r \times r$ of A is a minor of $\hat{A}$ we deduce that $\delta_r(\hat{A})|\delta_r(A)$. (1.12.2) implies that $\mathbf{b}$ is a linear combination of the columns of A. Consider any $r \times r$ minor of $\hat{A}$ which contains the $n+1$-st column $\mathbf{b}$. Since $\mathbf{b}$ is a linear combination of columns of A it follows that $\delta_r(A)$ divides this minor. Hence $\delta_r(A)|\delta_r(\hat{A})$, which establishes the second part of (1.12.6). (Actually we showed that if (1.12.1) is solvable over $\mathbb{D}_G$ then (1.12.6) holds.)

Assume now that (1.12.6) holds. Let

$$V\hat{A} = \hat{B} = [B, \hat{\mathbf{b}}] \in \mathbb{D}^{m\times(n+1)}, \ V \in \mathbf{GL}(m, \mathbb{D})$$

be Hermite's normal form of $\hat{A}$. Hence B is Hermite's normal form of A. Furthermore

$$\begin{aligned} &VA = B, \quad \mathrm{rank\ B = rank\ A = rank\ \hat{A} = rank\ \hat{B} = r}, \\ &\delta_r(B) \equiv \delta_r(A) \equiv \delta_r(\hat{A}) \equiv \delta_r(\hat{B}). \end{aligned}$$

Hence n_r in Hermite's normal form of $\hat{A}$ is at most n. Note that the last $m - r$ equations of $B\mathbf{x} = \hat{\mathbf{b}}$ are the trivial equations $0 = 0$. That is, it is enough to show the solvability of the system (1.12.2) under the assumptions (1.12.6) with $r = m$. By changing the order of equations in (1.12.1) and introducing a new set of variables

$$\mathbf{y} = U^{-1}\mathbf{x}, \quad U \in \mathbf{GL}(n, \mathbb{D}), \tag{1.12.7}$$

we may assume that the system (1.12.2) is

$$C\mathbf{y} = \mathbf{d}, \quad C = PAU, \ \mathbf{d} = (d_1, ..., d_m)^T = P\mathbf{b}, \tag{1.12.8}$$

where C is given as in Lemma 1.12.5 with $r = m$. Let $\hat{C} = [C, \mathbf{d}]$. It is straightforward to see that $A \sim C, \ \hat{A} \sim \hat{C}$. Hence

$$\mathrm{rank\ C} = \mathrm{rank\ A} = \mathrm{rank\ \hat{A}} = \mathrm{rank\ \hat{C}} = \mathrm{m}, \quad \delta_{\mathrm{m}}(\mathrm{C}) \equiv \delta_{\mathrm{m}}(\mathrm{A}) \equiv \delta_{\mathrm{m}}(\hat{\mathrm{A}}) \equiv \delta_{\mathrm{m}}(\hat{\mathrm{C}}).$$

Thus it is enough to show that the system (1.12.8) is solvable. In view of the form of C the solvability of the system (1.12.8) over $\mathbb{D}$ is equivalent the solvability of the system

$$\tilde{C}\tilde{\mathbf{y}} = \mathbf{d}, \quad \tilde{C} = [c_{ij}]_{i=j=1}^{m} \in \mathbb{D}^{m \times m}, \ \tilde{\mathbf{y}} = (y_1, ..., y_m)^T. \tag{1.12.9}$$

Note that $\delta_m(C) = \delta_m(\tilde{C}) = \det \tilde{C}$. Cramer's rule for the above system in the quotient field $\mathbb{F}$ of $\mathbb{D}$ yields

$$y_i = \frac{\det \tilde{C}_i}{\det \tilde{C}}, \quad i = 1, ..., m.$$

Here $\tilde{C}_i$ is obtained by replacing column i of $\tilde{C}$ by $\mathbf{d}$. Clearly $\det \tilde{C}_i$ is an $m \times m$ minor of $\hat{C}$ up to the factor ±1. Hence it is divisible by $\delta_m(\hat{C}) \equiv \delta_m(C) = \det\,(\tilde{C})$. Therefore $y_i \in \mathbb{D}, \ i = 1, ..., m$. □

Theorem 1.12.3 *Let $A \in \mathbb{D}_B^{m \times n}$. Then* Range A *and* Ker A *are modules in $\mathbb{D}_B^m$ and $\mathbb{D}_B^n$ having finite bases with* rank A *and* nul A *elements respectively. Moreover, the basis of* Ker A *can be completed to a basis of $\mathbb{D}_B^n$.*

Proof. As in the proof of Theorem 1.12.2 we may assume that rank A = m and $A = C$, where C is given by (1.12.5) with $r = m$. Let $\mathbf{e}_1, ..., \mathbf{e}_n$ be the standard basis of $\mathbb{D}_B^n$. Then $C\mathbf{e}_1, ..., C\mathbf{e}_m$ is a basis in Range C and $\mathbf{e}_{m+1}, ..., \mathbf{e}_n$ is a basis for Ker A. □

Let $A \in \mathbb{D}_G^{m \times n}$. Expand any $q \times q$ minor of A by any $q - p$ rows, where $1 \le p < q$. We then deduce

$$\delta_p(A) | \delta_q(A) \quad \text{for any } 1 \le p \le q \le \min(m, n). \tag{1.12.10}$$

Definition 1.12.4 *For $A \in \mathbb{D}_G^{m\times n}$ let*

$$i_j(A) := \frac{\delta_j(A)}{\delta_{j-1}(A)}, \quad j = 1, ..., \mathrm{rank\ A}, \quad (\delta_0(\mathrm{A}) = 1),$$
$$i_j(A) = 0 \quad \mathrm{for\ rank\ A < j \le min(m, n)},$$

be the invariant factors of A. $i_j(A)$ is called a trivial factor if $i_j(A)$ is invertible in $\mathbb{D}_G$.

Suppose that (1.12.1) is solvable over $\mathbb{D}_B$. Using the fact that $\mathbf{b}$ is a linear combination of the columns of A and Theorem 1.12.2 we get an equivalent version of Theorem 1.12.2. (See Problem 2.)

Corollary 1.12.5 *Let $A \in \mathbb{D}_B^{m\times n}$, $\mathbf{b} \in \mathbb{D}_B^m$. Then the system* (1.12.1) *is solvable over $\mathbb{D}_B$ if and only if*

(1.12.11) $$r = \mathrm{rank\ A = rank\ \hat{A}, \quad i_k(A) \equiv i_k(\hat{A}),\ k = 1, ..., r.}$$

Problems

1. Let $A \in \mathbb{D}_G^{n\times n}$. Assume that $r = \mathrm{rank\ A}$. Show
   (a)
   (1.12.12)
   $\delta_j(A) = \omega_j i_1(A) \cdots i_j(A)$, where ω_j is invertible in $\mathbb{D}_G$ for $j = 1, \dots, n$.
   (b) $i_1(A)|i_j(A)$ for $j = 2, \dots, r$. (**Hint**: Expand any minor of order j by any row.)
   (c) Let $2 \le k, 2k-1 < j \le r$. Then $i_1(A) \dots i_k(A)|i_{j-k+1}(A) \dots i_j(A)$.
2. Give a complete proof of Corollary 1.12.5.
3. Let $A \in \mathbb{D}_B^{m\times n}$. Assume that all the pivots in HNF of $A^\top$ are invertible elements. Show
   (a) Any basis of Range A can be completed to a basis in $\mathbb{D}_B^m$.
   (b) $i_1(A) = \dots = i_{\mathrm{rank\ A}}(A) = 1$.
4. Assume that $\mathbb{D} = \mathbb{D}_B$, $\mathbf{M}$ is a $\mathbb{D}$-module with a basis, $\mathbf{M}_1, \mathbf{M}_2$ are finitely generated modules of $\mathbf{M}$. Show
   (a) $\mathbf{M}_1 \cap \mathbf{M}_2$ has a basis which can be completed to bases in $\mathbf{M}_1$ and $\mathbf{M}_2$.

(b) $\mathbf{M}_i = \mathbf{M}_1 \cap \mathbf{M}_2 \oplus N_i$ for $i = 1, 2$, where each $\mathbf{N}_i$ has a basis.

$$\dim \mathbf{M}_i = \dim (\mathbf{M}_1 \cap \mathbf{M}_2) + \dim \mathbf{N}_i, i = 1, 2,$$
$$\mathbf{M}_1 + \mathbf{M}_2 = (\mathbf{M}_1 \cap \mathbf{M}_2) \oplus \mathbf{N}_1 \oplus \mathbf{N}_2.$$

In particular, $\dim (\mathbf{M}_1 + \mathbf{M}_2) = \dim \mathbf{M}_1 + \dim \mathbf{M}_2 - \dim (\mathbf{M}_1 \cap \mathbf{M}_2)$.

1.13 Smith normal form

A matrix $D = [d_{ij}] \in \mathbb{D}^{m\times n}$ is called a diagonal matrix if $d_{ij} = 0$ for all $i \neq j$. The entries $d_{11}, ..., d_{\ell\ell}$, $\ell = \min(m, n)$ are called the diagonal entries of D. D is denoted as $D = \mathrm{diag}(d_{11}, ..., d_{\ell\ell})$.

Theorem 1.13.1 *Let $0 \neq A \in \mathbb{D}^{m\times n}$. Assume that $\mathbb{D}$ is an elementary divisor domain. Then A is equivalent to a diagonal matrix*

$$B = \mathrm{diag}(i_1(A), ..., i_r(A), 0, ..., 0), \quad r = \text{rank A}. \tag{1.13.1}$$

Furthermore

$$i_{j-1}(A)|i_j(A), \quad \text{for } j = 2, ..., \text{rank A}. \tag{1.13.2}$$

Proof. Recall that an elementary divisor domain is a Bezout domain. For $n = 1$ Hermite's normal form of A is a diagonal matrix with $i_1(A) = \delta_1(A)$. Next we consider the case $m = n = 2$. Let

$$A_1 = WA = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}, \quad W \in \mathbf{GL}(2, \mathbb{D}),$$

be Hermite's normal form of A. As $\mathbb{D} = \mathbb{D}_{ED}$ there exists $p, q, x, y \in \mathbb{D}$ such that

$$(px)a + (py)b + (qy)c = (a, b, c) = \delta_1(A).$$

Clearly $(p, q) = (x, y) = 1$. Hence there exist $\bar{p}, \bar{q}, \bar{x}, \bar{y}$ such that

$$p\bar{p} - q\bar{q} = x\bar{x} - y\bar{y} = 1.$$

Let

$$V = \begin{bmatrix} p & q \\ \bar{q} & \bar{p} \end{bmatrix}, \quad U = \begin{bmatrix} x & \bar{y} \\ y & \bar{x} \end{bmatrix}.$$

Thus

$$G = VAU = \begin{bmatrix} \delta_1(A) & g_{12} \\ g_{21} & g_{22} \end{bmatrix}.$$

Since $\delta_1(G) \equiv \delta_1(A)$ we deduce that $\delta_1(A)$ divides g_{12} and g_{21}. Apply appropriate elementary row and column operations to deduce that A is equivalent to a diagonal matrix $C = \mathrm{diag}(i_1(A), d_2)$. As $\delta_2(C) = i_1(A)d_2 \equiv \delta_2(A)$ we see that C is equivalent to the matrix of the form (1.13.1), where we can assume that $d_2 = i_2(A)$. Since $i_1(A)|d_2$ we have that $i_1(A)|i_2(A)$. We now prove the theorem in the case $m \ge 3,\ n = 2$ by induction starting from $m = 2$. Let $A = [a_{ij}] \in \mathbb{D}^{m\times 2}$ and denote by $\bar{A} = [a_{ij}]_{i=j=1}^{m-1,2}$. Use the induction hypothesis to assume that $\bar{A}$ is in the form (1.13.1). Interchange the second row of A with the last one to obtain $A_1 \in \mathbb{D}^{m\times 2}$. Apply simple row and column operations on the first two rows and columns of A_1 to obtain $A_2 = [a_{ij}^{(2)}] \in \mathbb{D}^{m\times 2}$, where $a_{11}^{(2)} = i_1(A)$. Use the elementary row and column operations to obtain A_3 of the form

$$A_3 = \begin{bmatrix} i_1(A) & 0 \\ 0 & A_4 \end{bmatrix}, \quad A_4 \in \mathbb{D}^{(m-1)\times 1}. \tag{1.13.3}$$

Recall that $i_1(A)$ divides all the entries of A_4. Hence $A_4 = i_1(A)B_4$ and $i_1(A_4) = i_1(A)i_1(B_4)$. Use simple row operations on the rows $2, ..., m$ of A_3 to bring B_4 to a diagonal form. Thus A is equivalent to the diagonal matrix
$C = \mathrm{diag}(i_1(A), i_1(A)i_1(B_4)) \in \mathbb{D}^{m\times 2}$. Recall that

$$i_1(A) = \delta_1(A) \equiv \delta_1(C), \quad \delta_2(A) \equiv \delta_2(C) = i_1(A)i_1(A_4) = i_1(A)i_1(A)i_1(B_4).$$

Thus $i_1(A)|i_1(A_4)$ so $i_1(A) = i_1(C)$ and $i_2(A) \equiv i_1(A_4)$. Hence C is equivalent to B of the form (1.13.1) and $i_1(A)|i_2(A)$.

By considering $A^\top$ we deduce that we proved the theorem in the case $\min(m,n) \le 2$. We now prove the remaining cases by a double induction on $m \ge 3$ and $n \ge 3$. Assume that the theorem holds for all matrices in $\mathbb{D}^{(m-1)\times n}$ for $n = 2, 3, ...$ Assume that $m \ge 3$ and is fixed, and theorem holds for any $E \in \mathbb{D}^{m\times(n-1)}$ for $n \ge 3$. Let $A = [a_{ij}] \in \mathbb{D}^{m\times n}$, $\bar{A} = [a_{ij}]_{i=j=1}^{m,n-1}$. Use the induction hypothesis to assume that $\bar{A} = \mathrm{diag}(d_1, ..., d_l)$, $l = \min(m, n-1)$. Here $d_1|d_i,\ i = 2, ..., l$. Interchange the second and the last column of A to obtain $A_1 = [a_{ij}^{(1)}] \in \mathbb{D}^{m\times n}$. Perform simple row operations on the rows of A_1 and simple column operations on the the first $m-1$ columns of A_1 to obtain the matrix $A_2 = [a_{ij}^{(2)}] \in \mathbb{D}^{m\times n}$ such that $\bar{A}_2 = [a_{ij}^{(2)}]_{i=j=1}^{m,n-1} = \mathrm{diag}(a_{11}^{(2)}, ..., a_{ll}^{(2)})$ is Smith's normal form of $\bar{A}_1 = [a_{ij}^{(1)}]_{i=j=1}^{m,n-1}$. The definition of A_2 yields that $i_1(A) \equiv a_{11}^{(2)}$. Use elementary row operations to obtain an equivalent matrix to A_2:

$$A_3 = \begin{bmatrix} i_1(A) & 0 \\ 0 & A_4 \end{bmatrix}, \quad A_4 \in \mathbb{D}^{(m-1)\times(n-1)}.$$

As $i_1(A) = \delta_1(A) = \delta_1(A_3)$ it follows that $i_1(A)$ divides all the entries of A_4. So $A_4 = i_1(A)B_4$. Hence $i_j(A_4) = i_1(A)i_j(B_4)$ Use simple row and column operations on the last $m-1$ rows and the last $n-1$ columns of A_3 to bring B_4 to Smith's normal form using the induction hypothesis:

$$A \sim A_5 = \begin{bmatrix} i_1(A) & 0 \\ 0 & i_1(A)\,\mathrm{diag}(i_1(B_4), ..., i_l(B_4)) \end{bmatrix}.$$

By induction hypothesis

$$i_j(B_4)|i_{j+1}(B_4),\ j = 1, ..., \mathrm{rank\ A} - 1, \quad \mathrm{i_j(B_4) = 0,\ j > rank\ A - 1.}$$

Similar claim holds for A_4. Hence

$$\delta_k(A) \equiv \delta_k(A_5) = i_1(A)i_1(A_4)\cdots i_{k-1}(A_4), \quad k = 2, ..., \mathrm{rank\ A}.$$

Thus

$$i_j(A_4) \equiv i_{j+1}(A), \quad j = 1, ..., \mathrm{rank\ A} - 1$$

and A_5 is equivalent to B given by (1.13.1). Furthermore, we showed (1.13.2). □

The matrix (1.13.1) is called the *Smith normal* form of A.

Corollary 1.13.2 *Let $A, B \in \mathbb{D}_{ED}^{m\times n}$. Then A and B are equivalent if and only if A and B have the same rank and the same invariant factors.*

Over an elementary divisor domain, the system (1.12.2) is equivalent to a simple system

$$i_k(A)y_k = c_k, \quad k = 1, ..., \mathrm{rank\ A},$$

(1.13.4)

$$0 = c_k, \quad k = \mathrm{rank\ A} + 1, ..., \mathrm{m},$$

(1.13.5) $$\mathbf{y} = P^{-1}\mathbf{x}, \quad \mathbf{c} = Q\mathbf{b}.$$

Here P and Q are the invertible matrices appearing in (1.10.9) and B is of the form (1.13.1). For the system (1.13.4) Theorems 1.12.2 and 1.12.3 are straightforward. Clearly

Theorem 1.13.3 *Let $A \in \mathbb{D}_{ED}^{m\times n}$. Assume that all the invariant factors of A are invertible elements in $\mathbb{D}_{ED}$. Then the basis of* Range A *can be completed to a basis of $\mathbb{D}_{ED}^{n}$.*

In what follows we adopt

Normalization 1.13.4 *Let $A \in \mathbb{F}[x]^{m\times n}$. Then the invariant polynomials (the invariant factors) of A are assumed to be normalized polynomials.*

Notation 1.13.5 *Let $A_i \in \mathbb{D}^{m_i\times n_i}$ for $i = 1,\ldots,k$. Then $\oplus_{i=1}^k A_i = \mathrm{diag}(A_1,\ldots,A_k)$ denote the block diagonal matrix $B = [B_{ij}]_{i,j=1}^k \in \mathbb{D}^{m\times n}$, where $B_{ij} \in \mathbb{D}^{m_i\times n_j}$ for $i,j = 1,\ldots,k$, $m = \sum_{i=1}^k m_i$, $n = \sum_{j=1}^k n_j$, such that $B_{ii} = A_i$ and $B_{ij} = 0$ for $i \neq j$.*

Problems

1. Let $A = \begin{bmatrix} p & 0 \\ 0 & q \end{bmatrix} \in \mathbb{D}_B^{2\times 2}$. Then A is equivalent to $\mathrm{diag}((p,q), \frac{pq}{(p,q)})$.

2. Let $A \in \mathbb{D}_G^{m\times n}$, $B \in \mathbb{D}_G^{p\times q}$. Suppose that either $i_s(A)|i_t(B)$ or $i_t(B)|i_s(A)$ for $s = 1,...,\text{rank A} = \alpha$, $\text{t} = 1,...,\text{rank B} = \beta$. Show that the set of the invariant factors $A\oplus B$ is $\{i_1(A),...,i_\alpha(A), i_1(B),...,i_\beta(B)\}$.

3. Let $\mathbf{M} \subset \mathbf{N}$ be $\mathbb{D}_{ED}$ modules with finite bases. Prove that there exists a basis $\mathbf{u}_1,...,\mathbf{u}_n$ in $\mathbf{N}$ such that $i_1\mathbf{u}_1,...,i_r\mathbf{u}_r$ is a basis in $\mathbf{M}$, where $i_1,...,i_r \in \mathbb{D}_{ED}$ and $i_j|i_{j+1}$ for $j = 1,...,r-1$.

4. Let $\mathbf{M}$ be a $\mathbb{D}$-module and $\mathbf{N}_1, \mathbf{N}_2 \subset \mathbf{M}$ be submodules. $\mathbf{N}_1$ and $\mathbf{N}_2$ are called equivalent if there exists an isomorphism $T \in \mathrm{Hom}\ (\mathbf{M},\mathbf{M})$ ($T^{-1} \in \mathrm{Hom}\ (\mathbf{M},\mathbf{M})$) such that $T\mathbf{N}_1 = \mathbf{N}_2$. Suppose that $\mathbf{M},\ \mathbf{N}_1,\ \mathbf{N}_2$ have bases $[\mathbf{u}_1,...,\mathbf{u}_m]$, $[\mathbf{v}_1,...,\mathbf{v}_n]$ and $[\mathbf{w}_1,...,\mathbf{w}_n]$ respectively. Let

$$\mathbf{v}_j = \sum_{i=1}^m a_{ij}\mathbf{u}_i, \quad \mathbf{w}_j = \sum_{i=1}^m b_{ij}\mathbf{u}_i, \quad j = 1,...,n,$$
$$A = [a_{ij}]_{i=j=1}^{m,n}, \quad B = [b_{ij}]_{i=j=1}^{m,n}. \tag{1.13.6}$$

   Show that $\mathbf{N}_1$ and $\mathbf{N}_2$ are equivalent if and only if $A \sim B$.

5. Let $\mathbf{N} \subset \mathbf{M}$ be $\mathbb{D}$ modules with bases. Assume that $\mathbf{N}$ has the division property: if $a\mathbf{x} \in \mathbf{N}$ for $0 \neq a \in \mathbb{D}$ and $\mathbf{x} \in \mathbf{M}$ then $\mathbf{x} \in \mathbf{N}$. Show that if $\mathbb{D}$ is an elementary divisor domain and $\mathbf{N}$ has the division property then any basis in $\mathbf{N}$ can be completed to a basis in $\mathbf{M}$.

6. Let $\mathbb{D}$ be elementary divisor domain. Assume that $\mathbf{N} \subset \mathbb{D}^m$ is a submodule with basis of dimension $k \in [1,m]$. Let $\mathbf{N}' \subset \mathbb{D}^m$ be the following set. $\mathbf{n} \in \mathbb{N}'$ if there exists $0 \neq a \in \mathbb{D}$ such that $a\mathbf{n} \in \mathbb{N}$. Show that $\mathbf{N}'$ is a submodule of $\mathbb{D}^m$, which has the division property.

Furthermore, $\mathbf{N}'$ has a basis of dimension k which can be obtained from a basis of $\mathbf{N}$ as follows. Let $\mathbf{w}_1, \ldots, \mathbf{w}_k$ be a basis of $\mathbf{N}$. Let $W \in \mathbb{D}^{m\times k}$ be the matrix whose columns are $\mathbf{w}_1, \ldots, \mathbf{w}_k$. Assume that $D = \mathrm{diag}(n_1, \ldots, n_k)$ is the Smith normal form of W. So $W = UDV, U \in \mathbf{GL}(m, \mathbb{D}), V \in \mathbf{GL}(k, \mathbb{D})$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are the first k columns of U. Then $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is a basis of $\mathbf{N}'$.

1.14 The ring of local analytic functions in one variable

In this section we consider applications of the Smith normal to the system of linear equations over H_0, the ring of local analytic functions in one variable at the origin. In 1.3 we showed that the only noninvertible irreducible element in H_0 is z. Let $A \in \mathrm{H}_0^{m\times n}$. Then $A = A(z) = [a_{ij}(z)]_{i=j=1}^{m,n}$ and $A(z)$ has the McLaurin expansion

$$A(z) = \sum_{k=0}^{\infty} A_k z^k, \quad A_k \in \mathbb{C}^{m\times n}, \ k = 0, ..., \tag{1.14.1}$$

which converges in some disk $|z| < R(A)$. Here $R(A)$ is a positive number which depends on A. That is, each entry $a_{ij}(z)$ has convergent McLaurin series for $|z| < R(A)$.

Notations and Definitions 1.14.1 *Let $A \in \mathrm{H}_0^{m\times n}$. Then local invariant polynomials (the invariant factors) of A are normalized to be*

$$i_k(A) = z^{\iota_k(A)}, \quad 0 \le \iota_1(A) \le \iota_2(A) \le ... \le \iota_r(A), \quad r = \mathrm{rank\ A}. \tag{1.14.2}$$

The number $\iota_r(A)$ is called the index of A and is denoted by $\eta = \eta(A)$. For a nonnegative integer p denote by $\kappa_p = \kappa_p(A)$-the number of local invariant polynomials of A whose degree is equal to p.

We start with the following perturbation result.

Lemma 1.14.2 *Let $A, B \in \mathrm{H}_0^{m\times n}$. Let*

$$C(z) = A(z) + z^{k+1}B(z), \tag{1.14.3}$$

where k is a nonnegative integer. Then A and C have the same local invariant polynomials up to degree k. Moreover, if k is equal to the index of A, and A and C have the same ranks then A is equivalent to C.

Proof. Since H_0 is an Euclidean domain we may already assume that A is in Smith's normal form

$$A = \mathrm{diag}(z^{\iota_1}, ..., z^{\iota_r}, 0, ..., 0). \tag{1.14.4}$$

Let $s = \sum_{j=0}^{k} \kappa_j(A)$. Assume first that $s \ge t \in \mathbb{N}$. Consider any any $t \times t$ submatrix $D(z)$ of $C(z) = [c_{ij}(z)]$. View $\det D(z)$ as a sum of $t!$ products. As $k+1 > \iota_t$ it follows each such product is divisible by $z^{\iota_1+\cdots+\iota_t}$. Let $D(z) = [c_{ij}(z)]_{i=j=1}^{t}$. Then the product of the diagonal entries is of the form $z^{\iota_1+\cdots+\iota_t}(1+zO(z))$. All other $t!-1$ products appearing in $\det D(z)$ are divisible by $z^{\iota_1+\cdots\iota_{t-2}+2(k+1)}$. Hence

$$\delta_t(C) = z^{\iota_1+\cdots+\iota_t} = \delta_t(A), \quad t = 1, ..., s, \tag{1.14.5}$$

which implies that

$$\iota_t(C) = \iota_t(A), \quad t = 1, ..., s. \tag{1.14.6}$$

As $s = \sum_{j=0}^{k} \kappa_j(A)$ it follows that

$$\kappa_j(C) = \kappa_j(A), \quad j = 0, ..., k-1, \quad \kappa_k(A) \le \kappa_k(C).$$

Write $A = C - z^{k+1}B$ and deduce from the above arguments that

$$\kappa_j(C) = \kappa_j(A), \quad j = 0, ..., k. \tag{1.14.7}$$

Hence A and C have the same local invariant polynomials up to degree k. Suppose that rank A = rank C. Then (1.14.6) implies that A and C have the same local invariant polynomials. Hence $A \sim B$. $\square$

Consider a system of linear equations over H_0

$$A(z)\mathbf{u} = \mathbf{b}(z), \quad A(z) \in \mathrm{H}_0^{m\times n}, \ \mathbf{b}(z) \in \mathrm{H}_0^{m}, \tag{1.14.8}$$

where we look for a solution $\mathbf{u}(z) \in \mathrm{H}_0^n$. Theorem 1.12.2 claims that the above system is solvable if and only if rank A = rank Â = r and the g.c.d. of all $r \times r$ minors of A and $\hat{A}$ are equal. In the area of analytic functions it is common to try to solve (1.3.6) by the method of power series. Assume that $A(z)$ has the expansion (1.14.1) and $\mathbf{b}(z)$ has the expansion

$$\mathbf{b}(z) = \sum_{k=0}^{\infty} \mathbf{b}_k z^k, \quad \mathbf{b}_k \in \mathbb{C}^m, \ k = 0, ... \tag{1.14.9}$$

Then one looks for a formal solution

$$\mathbf{u}(z) = \sum_{k=0}^{\infty} \mathbf{u}_k z^k, \quad \mathbf{u}_k \in \mathbb{C}^n, \ k = 0, ..., \tag{1.14.10}$$

which satisfies

$$\sum_{j=0}^{k} A_{k-j}\mathbf{u}_j = \mathbf{b}_k, \tag{1.14.11}$$

for $k = 0, \ldots$. A vector $\mathbf{u}(z)$ is called a formal solution of (1.14.8) if (1.14.11) holds for any $k \in \mathbb{Z}_+$. A vector $\mathbf{u}(z)$ is called an analytic solution if $\mathbf{u}(z)$ is a formal solution and the series (1.14.10) converges in some neighborhood of the origin, i.e. $\mathbf{u}(z) \in \mathrm{H}_{\mathrm{o}}^n$. We now give the exact conditions for which (1.14.11) is solvable for $k = 0, ..., q$.

Theorem 1.14.3 *Consider the system* (1.14.11) *for* $k = 0, ..., q \in \mathbb{Z}_+$. *Then this system is solvable if and only if* $A(z)$ *and* $\hat{A}(z)$ *have the same local invariant polynomials up to degree q:*

$$\kappa_j(A) = \kappa_j(\hat{A}), \quad j = 0, ..., q. \tag{1.14.12}$$

Assume that the system (1.14.8) *is solvable over* H_0. *Let* $q = \eta(A)$ *and suppose that* $\mathbf{u}_0, ..., \mathbf{u}_q$ *satisfies* (1.14.11) *for* $k = 0, ..., q$. *Then there exists* $\mathbf{u}(z) \in \mathrm{H}_0^n$ *satisfying* (1.14.8) *and* $\mathbf{u}(0) = \mathbf{u}_0$.

Let $q \in \mathbb{Z}_+$ *and* $\mathbf{W}_q \subset \mathbb{C}^n$ *be the subspace of all vectors* $\mathbf{w}_0$ *such that* $\mathbf{w}_0, ..., \mathbf{w}_q$ *is a solution to the homogenous system*

$$\sum_{j=0}^{k} A_{k-j}\mathbf{w}_j = 0, \quad k = 0, ..., q. \tag{1.14.13}$$

Then

$$\dim \mathbf{W}_q = n - \sum_{j=0}^{q} \kappa_j(A) \tag{1.14.14}$$

In particular, for $\eta = \eta(A)$ *and any* $\mathbf{w}_0 \in \mathbf{W}_\eta$ *there exists* $\mathbf{w}(z) \in \mathrm{H}_0^n$ *such that*

$$A(z)\mathbf{w}(z) = \mathbf{0}, \quad \mathbf{w}(0) = \mathbf{w}_0. \tag{1.14.15}$$

Proof. Let

$$\mathbf{u}_k = (u_{k,1}, ..., u_{k,n})^\top, \quad k = 0, ..., q.$$

We first establish the theorem when $A(z)$ is Smith's normal form (1.14.4). In that case the system (1.14.11) reduces to

$$\begin{aligned} u_{k-i_s,s} &= b_{k,s} \quad \text{if } \iota_s \le k, \\ 0 &= b_{k,s} \quad \text{if either } \iota_s > k \text{ or } s > \text{rank A}. \end{aligned} \tag{1.14.16}$$

The above equations are solvable for $k = 0, ..., q$ if and only if z^{ι_s} divides $b_s(z)$ for all $\iota_s \le q$, and for $\iota_s > q$, z^{q+1} divides $b_s(z)$. If $\iota_s \le q$ then subtract from the last column of $\hat{A}$ the s-column times $\frac{b_s(z)}{z^{\iota_s}}$. So $\hat{A}$ is equivalent to the matrix

$$A_1(z) = \mathrm{diag}(z^{\iota_1}, ..., z^{\iota_l}) \oplus z^{q+1} A_2(z),$$
$$l = \sum_{j=0}^{q} \kappa_j(A), \quad A_2 \in \mathrm{H}_0^{(m-l)\times(n+1-l)}.$$

According to Problem 2 the local invariant polynomials of $A_1(z)$ whose degree does not exceed q are $z^{\iota_1}, ..., z^{\iota_l}$. So $A(z)$ and $A_1(z)$ have the same local invariant polynomials up the degree q. Assume now that A and $\hat{A}$ have the same local invariant polynomial up to degree q. Hence

$$z^{\iota_1 + ... + \iota_k} = \delta_k(A) = \delta_k(\hat{A}), \quad k = 1, ..., l,$$
$$z^{\iota_1 + ... + \iota_l + q + 1} | \delta_{l+1}(\hat{A}).$$

The first set of the equalities implies that $z^{\iota_s} | b_s(z),\ s = 1, ..., l$. The last equality yields that for $s > l,\ z^{q+1} | b_s(z)$. Hence (1.14.11) is solvable for $k = 0, ..., q$ if and only if A and $\hat{A}$ have the same local invariant polynomials up to the degree q.

Assume next that (1.14.8) is solvable. Since $A(z)$ is of the form (1.14.4) the general solution of (1.14.8) in that case is

$$u_j(z) = \frac{b_i(z)}{z^{\iota_j}}, \quad j = 1, ..., \text{rank A},$$
$$u_j(z) \text{ is an arbitrary function in } \mathrm{H}_0, \quad j = \text{rank A} + 1, ..., \text{n}.$$

Hence

$$u_j(0) = b_{i_j}, \quad j = 1, ..., \text{rank A},$$
(1.14.17)
$$u_j(0) \text{ is an arbitrary complex number}, \quad j = \text{rank A} + 1, ..., \text{n}.$$

Clearly (1.14.16) implies that $u_{0,s} = u_s(0)$ for $k = i_s$. The solvability of (1.14.8) implies that $b_s(z) = 0$ for $s >$ rank A. So $u_{0,s}$ is not determined by (1.14.16) for $s >$ rank A. This proves the existence of $\mathbf{u}(z)$ satisfying (1.14.8) such that $\mathbf{u}(0) = \mathbf{u}_0$. Consider the homogeneous system (1.14.13) for $k = 0, ..., q$. Then $w_{0,s} = 0$ for $i_s \le$ and otherwise $u_{0,s}$ is a free variable. Hence (1.14.14) holds. Let $q = \eta = \eta(A)$. Then the system (1.14.13) implies that the coordinates of $\mathbf{w}_0$ satisfy the conditions (1.14.17). Hence the system (1.14.15) is solvable.

Assume now that $A(z) \in \mathrm{H}_0^{m\times n}$ is an arbitrary matrix. Theorem 1.13.1 implies the existence of

$P(z) \in \mathbf{GL}(n, \mathrm{H}_0),\ Q(z) \in \mathbf{GL}(m, \mathrm{H}_0)$ such that

$$Q(z)A(z)P(z) = B(z) = \mathrm{diag}(z^{\iota_1}, ..., z^{\iota_r}, 0, ..., 0), \quad 0 \le \iota_1 \le ... \le \iota_r,\ r = \mathrm{rank\ A}.$$

It is straightforward to show that $P(z) \in \mathbf{GL}(n, \mathrm{H}_0)$ if and only if $P(z) \in \mathrm{H}_0^{n\times n}$ and $P(0)$ is invertible. To this end let

$$P(z) = \sum_{k=0}^{\infty} P_k z^k, \quad P_k \in \mathbb{C}^{n\times n},\ k = 0, ..., \quad \det P_0 \ne 0,$$
$$Q(z) = \sum_{k=0}^{\infty} Q_k z^k, \quad Q_k \in \mathbb{C}^{n\times n},\ k = 0, ..., \quad \det Q_0 \ne 0.$$

Introduce a new set of variables $\mathbf{v}(z)$ and $\mathbf{v}_0, \mathbf{v}_1, ...$ such that

$$\mathbf{u}(z) = P(z)\mathbf{v}(z),$$
$$\mathbf{u}_k = \sum_{j=0}^{k} P_{k-j}\mathbf{v}_j, \quad k = 0, ...$$

Since $\det P_0 \ne 0$ $\mathbf{v}(z) = P(z)^{-1}\mathbf{u}(z)$ and we can express each $\mathbf{v}_k$ in terms of $\mathbf{u}_k, ..., \mathbf{u}_0$ for $k = 0, 1, ...$ Then (1.14.8) and (1.14.11) are respectively equivalent to

$$B(z)\mathbf{v}(z) = \mathbf{c}(z), \quad \mathbf{c}(z) = Q(z)\mathbf{b}(z),$$
$$\sum_{j=0}^{k} B_{k-j}\mathbf{v}_j = \mathbf{c}_k, \quad k = o, ..., q.$$

As $B \sim A$ and $\hat{B} = Q\hat{A}(P \oplus I_1) \sim \hat{A}$ we deduce the theorem. □

Problems

1. The system (1.14.8) is called solvable in the punctured disc if the system

   (1.14.18) $$A(z_0)\mathbf{u}(z_0) = \mathbf{b}(z_0),$$

   is solvable for any point $0 < |z_0| < R$ as a linear system over $\mathbb{C}$ for some $R > 0$, i.e.

   (1.14.19) $$\mathrm{rank\ A(z_0) = rank\ \hat{A}(z_0), \quad for\ all\ 0 < |z_0| < R.}$$

   Show that (1.14.8) is solvable in the punctured disk if and only if (1.14.8) is solvable over $\mathcal{M}_0$–the quotient field of H_0.

2. The system (1.14.8) is called pointwise solvable if (1.14.18) is solvable for all z_0 in some open disk $|z_0| < R$. Show that (1.14.8) is pointwise solvable if and only if (1.14.8) is solvable over $\mathcal{M}_0$ and

$$\text{rank A}(0) = \text{rank } \hat{\text{A}}(0). \tag{1.14.20}$$

3. Let $A(z) \in \mathrm{H}_0^{m\times n}$. $A(z)$ is called *generic* if whenever the system (1.3.6) is pointwise solvable then it is analytically solvable, i.e. solvable over H_0. Prove that $A(z)$ is generic if and only if $\eta(A) \le 1$.

4. Let $\Omega \subset \mathbb{C}$ be a domain and consider the system

$$A(z)\mathbf{u} = \mathbf{b}(z), \quad A(z) \in \mathrm{H}(\Omega)^{m\times n},\ \mathbf{b}(z) \in \mathrm{H}(\Omega)^m. \tag{1.14.21}$$

Show that the above system is solvable over $\mathrm{H}(\Omega)$ if and only if for each $\zeta \in \Omega$ this system is solvable in H_ζ. (*Hint*: As $\mathrm{H}(\Omega)$ is $\mathbb{D}_{ED}$ it suffices to analyze the case where $A(z)$ is in its Smith's normal form.)

5. Let $A(z)$ and $\mathbf{b}(z)$ satisfy the assumptions of Problem 4. $A(z)$ is called *generic* if whenever (1.14.21) is pointwise solvable it is solvable over $\mathrm{H}(\Omega)$. Show that $A(z)$ is generic if and only the invariant functions (factors) of $A(z)$ have only simple zeros. (ζ is a simple zero of $f \in \mathrm{H}(\Omega)$ if $f(\zeta) = 0$ and $f'(\zeta) \neq 0$.)

6. Let $A(z) \in \mathrm{H}(\Omega)^{m\times n}$, where Ω is a domain in $\mathbb{C}$. Prove that the invariant factors of $A(z)$ are invertible in $\mathrm{H}(\Omega)$ if and only if

$$\text{rank A}(\zeta) = \text{rank A}, \quad \text{for all } \zeta \in \Omega. \tag{1.14.22}$$

7. Let $A(z) \in \mathrm{H}(\Omega)^{m\times n}$, where Ω is a domain in $\mathbb{C}$. Assume that (1.14.22) holds. View $A(z) \in \mathrm{Hom}\,(\mathrm{H}(\Omega)^n, \mathrm{H}(\Omega)^m)$. Show that Range $A(z)$ has a basis which can be completed to a basis in $\mathrm{H}(\Omega)^m$. (*Hint*: Use Theorem 1.14.5.)

1.15 The local-global domains in $\mathbb{C}^p$

Let p be a positive integer and assume that $\Omega \subseteq \mathbb{C}^p$ is a domain. Consider the system of m nonhomogeneous equations in n unknowns:

$$A(\mathbf{z})\mathbf{u} = \mathbf{b}(\mathbf{z}), \quad A(\mathbf{z}) \in \mathrm{H}(\mathit{\Omega})^{m\times n},\ \mathbf{b}(\mathbf{z}) \in \mathrm{H}(\mathit{\Omega})^m. \tag{1.15.1}$$

In this section we are concerned with the problem of existence of a solution $\mathbf{u}(\mathbf{z}) \in \mathrm{H}(\mathit{\Omega})^m$ to the above system. Clearly a necessary condition for the solvability is the local condition:

Condition 1.15.1 *Let $\Omega \subseteq \mathbb{C}^p$ be a domain. For each $\zeta \in \Omega$ the system $A(\mathbf{z})\mathbf{u} = \mathbf{b}(\mathbf{z})$ has a solution $\mathbf{u}_\zeta(\mathbf{z}) \in \mathrm{H}_\zeta^m$.*

Definition 1.15.2 *A domain $\Omega \subseteq \mathbb{C}^p$ is called a local-global domain, if any system of the form (1.15.1), satisfying the condition 1.15.1, has a solution $\mathbf{u}(\mathbf{z}) \in \mathrm{H}(\mathit{\Omega})^m$.*

Problem 1.14.4 implies that any domain $\Omega \subset \mathbb{C}$ is a local-global domain. In this section we assume that $p > 1$. Problem 1 shows that not every domain in $\mathbb{C}^p$ is a local-global domain. We give a sufficient condition on domain Ω to be a local-global domain.

Definition 1.15.3 *A domain $\Omega \subset \mathbb{C}^p$ is called a domain of holomorphy, if there exist $f \in \mathrm{H}(\Omega)$ such that for any larger domain $\Omega_1 \subset \ C^p$, strictly containing Ω, there is no $f_1 \in \mathrm{H}(\Omega_1)$ which coincides with f on Ω.*

The following theorem is a very special case of Hartog's theorem [GuR65].

Theorem 1.15.4 *Let $\Omega \subseteq \mathbb{C}^p, p > 1$ be a domain. Assume that $\zeta \in \Omega$ and $f \in \mathrm{H}(\Omega\backslash\{\zeta\})$. Then $f \in \mathrm{H}(\Omega)$.*

Thus $\mathrm{H}(\Omega\backslash\{\zeta\})$ is not domain of holomorphy. A simple example of domain of holomorphy is [GuR65]:

Example 1.15.5 *Let $\Omega \subseteq \mathbb{C}^p$ be an open convex set. Then Ω is domain of holomorphy.*

(See 7.1 for the definition of convexity.) The main result of this section is:

Theorem 1.15.6 *Let $\Omega \subseteq \mathbb{C}^p, p > 1$ be domain of holomorphy. Then Ω is a local-global domain.*

The proof needs basic knowledge of sheaves and is brought for the reader who was exposed to the basic concepts in this field. See for example [GuR65]. We discuss only very special type of sheaves which are needed for the proof of Theorem 1.15.6.

Definition 1.15.7 *Let $\Omega \subseteq \mathbb{C}^p$ be an open set. Then*

1. *$\mathcal{F}(\Omega)$, called the sheaf of rings of holomorphic functions on Ω, is the union all $\mathrm{H}(U)$, where U ranges over all open subsets of Ω. Then for each $\zeta \in \Omega$ the local ring H_ζ is viewed as a subset of $\mathcal{F}(\Omega)$ and is called the stalk of $\mathcal{F}(\Omega)$ over ζ. A function $f \in \mathrm{H}(U)$ is called a section of $\mathcal{F}(\Omega)$ on U.*

2. *For an integer $n \geq 1$, $\mathcal{F}_n(\Omega)$, called an $\mathcal{F}(\Omega)$–sheaf of modules, is the union all $\mathrm{H}(U)^n$, where U ranges over all open subsets of Ω. Then for each $\zeta \in \Omega$ the local module H_ζ^n is viewed as a subset of $\mathcal{F}_n(\Omega)$ and is called the stalk of $\mathcal{F}_n(\Omega)$ over ζ. (Note H_ζ^n is an H_ζ module.) A vector $\mathbf{u} \in \mathrm{H}(U)^n$ is called a section on U. (If $U = \emptyset$ then $\mathrm{H}(U)^n$ consists of the zero element 0.)*

3. *$\mathcal{F} \subseteq \mathcal{F}_n(\Omega)$ is called a subsheaf if the following conditions holds. First $\mathcal{F} \cap \mathrm{H}^n(U)$ contains the trivial section 0 for each open set $U \subseteq \Omega$. Second, assume that $\mathbf{u} \in \mathrm{H}(U)^n \cap \mathcal{F}, \mathbf{v} \in \mathrm{H}(V)^n \cap \mathcal{F}$ and $W \subseteq U \cap V$ is an open nonempty set. Then for any $f, g \in \mathrm{H}(W)$ the vector $f\mathbf{u}+g\mathbf{v} \in \mathcal{F} \cap \mathrm{H}(W)^n$. (**Restriction property.**) Third, if $\mathbf{u} = \mathbf{v}$ on W then the section $\mathbf{w} \in \mathbf{H}^n(U \cup V)$, which coincides with $\mathbf{u}, \mathbf{v}$ on U, V respectively, belongs to $\mathcal{F} \cap \mathrm{H}^n(U \cup V)$. (**Extension property.**) $\mathcal{F}_\zeta := \mathcal{F} \cap H_\zeta^n$ is the stalk of $\mathcal{F}$ over $\zeta \in \Omega$.*

   (a) *Let U be an open subset of Ω. Then $\mathcal{F}(U) := \mathcal{F} \cap \mathcal{F}_n(U)$ is called the restriction of the sheaf $\mathcal{F}$ to U.*

   (b) *Let U be an open subset of Ω. The sections $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathcal{F} \cap \mathrm{H}(U)^n$ are said to generate $\mathcal{F}$ on U, if for any $\zeta \in U$ $\mathcal{F}_\zeta$ is generated by $\mathbf{u}_1, \ldots, \mathbf{u}_k$ over H_ζ. $\mathcal{F}$ is called finitely generated over U if such $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathcal{F} \cap \mathrm{H}(U)^n$ exists. $\mathcal{F}$ is called finitely generated if it is finitely generated over Ω. $\mathcal{F}$ is called of finite type if for each for each $\zeta \in \Omega$ there exists an open set $U_\zeta \subset \Omega$, containing ζ, such that $\mathcal{F}$ is finitely generated over U. (I.e. each $\mathcal{F}_\zeta$ is finitely generated.)*

   (c) *$\mathcal{F}$ is called a coherent sheaf if the following two conditions hold. First $\mathcal{F}$ is finite type. Second, for each open set $U \subseteq \Omega$ and for any $q \geq 1$ sections $\mathbf{u}_1, \ldots, \mathbf{u}_q \in \mathcal{F} \cap \mathrm{H}(\Omega)^n$ let $\mathcal{G} \subseteq \mathcal{F}_q(U)$ be a subsheaf generated by the condition $\sum_{i=1}^q f_i \mathbf{u}_i = 0$. That is, $\mathcal{G}$ is a union of all $(f_1, \ldots, f_q)^\top \in \mathrm{H}(V)^q$ satisfying the condition $\sum_{i=1}^q f_i \mathbf{u}_i = 0$ for all open $V \subseteq U$. Then $\mathcal{G}$ is of finite type.*

The following result is a straight consequence of Oka's coherence theorem [GuR65].

Theorem 1.15.8 *Let $\Omega \subseteq \mathbb{C}^P$ be an open set. Then*

- *The sheaf $\mathcal{F}_n(\Omega)$ is coherent.*
- *Let $A \in \mathrm{H}(\Omega)^{m\times n}$ be given. Let $\mathcal{F} \subseteq \mathcal{F}(\Omega)$ be the subsheaf consisting of all $\mathbf{u} \in \mathrm{H}(U)^n$ satisfying $A\mathbf{u} = 0$ for all open sets $U \subseteq \Omega$. Then $\mathcal{F}$ is coherent.*

Note that $\mathcal{F}_n(\Omega)$ is generated by n constant sections $\mathbf{u}_i := (\delta_{i1}, \ldots, \delta_{in})^\top \in \mathrm{H}(\Omega)^n, i = 1, \ldots, n$. The following theorem is a special case of Cartan's Theorem A.

Theorem 1.15.9 *Let $\Omega \subseteq \mathbb{C}^p$ be a domain of holomorphy. Let $\mathcal{F} \subset \mathcal{F}_n(\Omega)$ be a subsheaf defined in Definition 1.15.7. If $\mathcal{F}$ is coherent then $\mathcal{F}$ is finitely generated.*

Corollary 1.15.10 *Let $\Omega \subseteq \mathbb{C}^p$ be a domain of holomorphy and $A \in \mathrm{H}(\Omega)^{m\times n}$. Then there exists $\mathbf{u}_1, \ldots \mathbf{u}_l \in \mathrm{H}(\Omega)^n$, such that for any $\zeta \in \Omega$, every solution of the system $A\mathbf{u} = \mathbf{0}$ over H_ζ^n is of the form $\sum_{i=1}^{l} f_i\mathbf{u}_i$ for some $f_1, \ldots, f_l \in \mathrm{H}_\zeta$.*

We now introduce the notion of *sheaf cohomology* of $\mathcal{F} \subseteq \mathcal{F}_n(\Omega)$.

Definition 1.15.11 *Let $\Omega \subseteq \mathbb{C}^p$ be an open set. Let $\mathcal{U} := \{U_i \subseteq \Omega,\ i \in \mathcal{I}\}$ be an open cover of Ω. (I.e. each U_i is open, and $\cup_{i\in\mathcal{I}} U_i = \Omega$.) For each integer $p \geq 0$ and $p + 1$ tuples of indices $(i_0, \ldots, i_p) \in \mathcal{I}^{p+1}$ denote $U_{i_0\ldots i_p} := \cap_{j=0}^{p} U_{i_j}$.*

Assume that $\mathcal{F} \subseteq \mathcal{F}_n(\Omega)$ is a subsheaf. A p-cochain c is a map carrying each $p+1$–tuples of indices $(i_0, \ldots, i_p)$ to a section $\mathcal{F} \cap \mathrm{H}^n(U_{i_0\ldots i_p})$ satisfying the following properties.

1. *$c(i_0, \ldots, i_p) = 0$ if $U_{i_0\ldots i_p} = \emptyset$.*
2. *$c(\pi(i_0), \ldots, \pi(i_p)) = \mathrm{sgn}(\pi)c(i_0, \ldots, i_p)$ for any permutation $\pi : \{0, \ldots, p\} \to \{0, \ldots, p\}$. (Note that $c(i_0, \ldots, i_p)$ is the trivial section if $i_j = i_k$ for $j \neq k$.)*

Zero cochain is the cochain which assigns a zero section to any $(i_0 \ldots i_p)$. Two cochains c, d are added and subtracted by the identity $(c\pm d)(i_0 \ldots i_p) = c(i_0, \ldots, i_p)\pm d(i_0, \ldots, i_p)$. Denote by $\mathrm{C}^{\mathrm{p}}(\Omega, \mathcal{F}, \mathcal{U})$ the group of $p+1$ cochains.

The $p - th$ coboundary operator $\delta_p : \mathrm{C}^{\mathrm{p}}(\Omega, \mathcal{F}, \mathcal{U}) \to \mathrm{C}^{\mathrm{p+1}}(\Omega, \mathcal{F}, \mathcal{U})$ is defined as follows:

$$(\delta_p c)(i_0, \ldots, i_{p+1}) = \sum_{j=0}^{p+1} (-1)^j c(i_0, \ldots, \hat{i}_j, \ldots, i_{p+1}),$$

where $\hat{i}_j$ is a deleted index. Then $p - th$ cohomology group is given by

1. $\mathrm{H}^0(\Omega,\mathcal{F},\mathcal{U}) := \mathrm{Ker}\ \delta_0$.

2. For $p \geq 1$ $\mathrm{H}^p(\Omega,\mathcal{F},\mathcal{U}) := \mathrm{Ker}\ \delta_p/\mathrm{Range}\ \delta_{p-1}$. *(See Problem 2.)*

Lemma 1.15.12 *Let the assumptions of Definition 1.15.11 hold. Let* $c \in \mathrm{C}^0(\Omega,\mathcal{F},\mathcal{U})$. *Then* $c \in \mathrm{H}^0(\Omega,\mathcal{F},\mathcal{U})$ *if and only if* c *represents a global section* $\mathbf{u} \in \mathcal{F} \cap \mathrm{H}(\Omega)^n$.

Proof. Let $c \in \mathrm{C}^0(\Omega,\mathcal{F},\mathcal{U})$. Assume that $c \in \mathrm{H}^0(\Omega,\mathcal{F},\mathcal{U})$. Let U_0, U_1 be two open sets in $\mathcal{U}$. Then $c(i_0)-c(i_1)$ is the zero section on $U_0 \cap U_1$. Thus for each $\zeta \in U_0 \cap U_1$ $c(i_0)(\zeta) = c(i_1)(\zeta)$. Let $\mathbf{u}(\mathbf{z}) := c(i_0)(\mathbf{z}) \in \mathbb{C}^n$. It follows that $\mathbf{u} \in \mathrm{H}(\Omega)^n$. The extension property of subsheaf $\mathcal{F}$ yields that $\mathbf{u} \in \mathcal{F} \cap \mathrm{H}^n(\Omega)$. Vice versa, assume that $\mathbf{u} \in \mathcal{F} \cap \mathrm{H}^n(\Omega)$. Define $c(i_0) = \mathbf{u}|U_0$. Then $c \in \mathrm{H}^0(\Omega,\mathcal{F},\mathcal{U})$. □

We identify $\mathrm{H}^0(\Omega,\mathcal{F},\mathcal{U})$ with the set of global sections $\mathcal{F} \cap \mathrm{H}(\Omega)^n$. The cohomology groups $\mathrm{H}^p(\Omega,\mathcal{F},\mathcal{U}), p \geq 1$ depend on the open cover $\mathcal{U}$ of Ω. By refining the covers of Ω one can define the cohomology groups $\mathrm{H}^p(\Omega,\mathcal{F}), p \geq 0$. See Problem 3. Cartan's Theorem B claims [GuR65].

Theorem 1.15.13 *Let* $\Omega \subseteq \mathbb{C}^p$ *be domain of holomorphy. Assume that the sheaf* $\mathcal{F}$ *given in Definition 1.15.7 is coherent. Then* $\mathrm{H}^p(\Omega,\mathcal{F})$ *is trivial for any* $p \geq 1$.

Proof of Theorem 1.15.6. Consider the system (1.15.1). Let $\mathcal{F}$ be the coherent sheaf defined in Theorem 1.15.8. Assume that the system (1.15.1) is locally solvable over Ω. Let $\zeta \in \Omega$. Then there exists an open set $U_\zeta \subseteq \Omega$ such that there exists $\mathbf{u}_\zeta \in \mathrm{H}^n(U_\zeta)$ satisfying (1.15.1) over $\mathrm{H}(U_\zeta)$. Let $\mathcal{U} := \{U_\zeta, \zeta \in \Omega\}$ be an open cover of Ω. Define $c \in \mathrm{C}^1(\Omega,\mathcal{F},\mathcal{U})$ by $c(\zeta,\eta) = \mathbf{u}_\zeta - \mathbf{u}_\eta$. Note that

$$(\delta_1 c)(\zeta,\eta,\theta) = c(\eta,\theta) - c(\zeta,\theta) + c(\zeta,\eta) = 0.$$

Hence $c \in \mathrm{Ker}\ \delta_1$. Since $\mathcal{F}$ is coherent Cartan's Theorem B yields that $\mathrm{H}^1(\Omega,\mathcal{F})$ is trivial. Hence $\mathrm{H}^1(\Omega,\mathcal{F},\mathcal{U})$ is trivial. (See Problem 3c.) Thus, there exists an element $d \in \mathrm{C}^0(\Omega,\mathcal{F},\mathcal{U})$ such that $\delta_0 d = c$. Thus for each $\zeta,\eta \in \Omega$ such that $U_\zeta \cap U_\eta$ there exist sections $\mathbf{d}(\zeta) \in \mathcal{F} \cap \mathrm{H}^n(U_\zeta), \mathbf{d}(\eta) \in \mathcal{F} \cap \mathrm{H}^n(U_\eta)$ such that $\mathbf{d}(\eta) - \mathbf{d}(\zeta) = \mathbf{u}_\zeta - \mathbf{u}_\eta$ on $U_\zeta \cap U_\eta$. Hence $\mathbf{d}(\eta) + \mathbf{u}_\eta = \mathbf{d}(\zeta) + \mathbf{u}_\zeta$ on $U_\zeta \cap U_\eta$. Since $A\mathbf{d}_\zeta = \mathbf{0} \in \mathrm{H}(U_\zeta)^m$ it follows that $A(\mathbf{d}_\zeta + \mathbf{u}_\zeta) = \mathbf{b} \in \mathrm{H}(U_\zeta)^m$. As $\mathbf{d}(\eta) + \mathbf{u}_\eta = \mathbf{d}(\zeta) + \mathbf{u}_\zeta$ on $U_\zeta \cap U_\eta$ it follows that all these section can be patched to the vector $\mathbf{v} \in \mathrm{H}(\Omega)^n$ which is a global solution of (1.15.1). □

Problems

1. Consider a system of one equation over $\mathbb{C}^p, p > 1$

   $$\sum_{i=1}^{p} z_i u_i = 1, \quad \mathbf{u} = (u_1, \ldots, u_p)^\top,\ \mathbf{z} = (z_1, \ldots, z_p).$$

   Let $\Omega := \mathbb{C}^p \backslash \{\mathbf{0}\}$.

   (a) Show that Condition 1.15.1 holds for Ω.

   (b) Show that the system is not solvable at $\mathbf{z} = \mathbf{0}$. (Hence it dose not have a solution $\mathbf{u}(\mathbf{z}) \in \mathrm{H}_{\mathbf{0}}^p$.)

   (c) Show that the system does not have a solution $\mathbf{u}(\mathbf{z}) \in \mathrm{H}(\mathit{\Omega})^p$. (**Hint**: Prove by contradiction using Hartog's theorem.)

2. Let the assumptions of Definition 1.15.11 hold. Show for any $p \geq 0$.

   (a) $\delta_{p+1}\delta_p = 0$.

   (b) Range $\delta_p \subseteq$ Ker δ_{p+1}.

3. Let $\mathcal{U} = \{U_i, i \in \mathcal{I}\}, \mathcal{V} = \{V_j, j \in \mathcal{J}\}$ be two open covers of an open set $\Omega \subset \mathbb{C}^p$. $\mathcal{V}$ is called a refinement of $\mathcal{U}$, denoted $\mathcal{V} \prec \mathcal{U}$, if each V_j is contained in some U_i. For each V_j we fix an arbitrary U_i with $V_j \subseteq U_i$, and write it as $U_{i(j)} :\ V_j \subseteq U_{i(j)}$. Let $\mathcal{F}$ be a subsheaf as in Definition 1.15.11. Show

   (a) Define $\phi : \mathrm{C^p}(\Omega, \mathcal{F}, \mathcal{U}) \rightarrow \mathrm{C^p}(\Omega, \mathcal{F}, \mathcal{V})$ as follows. For $c \in \mathrm{C^p}(\Omega, \mathcal{F}, \mathcal{U})$ let $(\phi(c))(j_0, \ldots, j_p) \in \mathrm{C^p}(\Omega, \mathcal{F}, \mathcal{V})$ be the restriction of the section $c(i(j_0), \ldots, i(j_p))$ to $V_{j_0 \ldots j_p}$. Then ϕ is a homomorphism.

   (b) ϕ is induces a homomorphism $\tilde{\phi} : \mathrm{H}^p(\Omega, \mathcal{F}, \mathcal{U}) \rightarrow \mathrm{H}^p(\Omega, \mathcal{F}, \mathcal{V})$. Furthermore, $\tilde{\phi}$ depends only on the covers $\mathcal{U}, \mathcal{V}$. (I.e., the choice of $i(j)$ is irrelevant.)

   (c) By refining the covers one obtains the $p - th$ cohomology group $\mathrm{H}^p(\Omega, \mathcal{F})$ with the following property. The homomorphism $\tilde{\phi}$ described in 3b induces an injective homomorphism that $\tilde{\phi} : \mathrm{H}^p(\Omega, \mathcal{F}, \mathcal{U}) \rightarrow \mathrm{H}^p(\Omega, \mathcal{F})$ for $p \geq 1$. (Recall that $\mathrm{H}^0(\Omega, \mathcal{F}, U) \equiv \mathcal{F} \cap \mathrm{H}^n(\Omega)$.) In particular, $\mathrm{H}^p(\Omega, \mathcal{F})$ is trivial, i.e. $\mathrm{H}^p(\Omega, \mathcal{F}) = \{0\}$, if and only if each $\mathrm{H}^p(\Omega, \mathcal{F}, \mathcal{U})$ is trivial.

1.16 Historical remarks

Most of the material in Sections 1.1-1.10 are standard. See [Lan58], [Lan67] and [vdW59] for the algebraic concepts. Consult [GuR65] and [Rud74] for the concepts and results concerning the analytic functions. See [Kap49] for the properties of elementary divisor domains. It is not known if there exists a Bezout domain which is not an elementary divisor domain. Theorem 1.5.3 for $\Omega = \mathbb{C}$ is due to [Hel40]. For §1.10 see [CuR62] or [McD33]. Most of §1.11 is well known, e.g. [McD33]. §1.12 seems to be new since the underlying ring is assumed to be only a Bezout domain. Theorems 1.12.2 and 1.12.3 are well known for an elementary divisor domain, since A is equivalent to a diagonal matrix. It would be interesting to generalize Theorem 1.12.2 for $\mathbb{D} = \mathbb{F}[x_1, ..., x_p]$ for $p \geq 2$. The fact that the Smith normal form can be achieved for $\mathbb{D}_{ED}$ is due to Helmer [Hel43]. More can be found in [Kap49].

Most of the results of §1.14 are from [Fri80b]. I assume that Theorem 1.15.6 is known to the experts.

Chapter 2

Canonical Forms for Similarity

2.1 Strict equivalence of pencils

Definition 2.1.1 *A matrix $A(x) \in \mathbb{D}[x]^{m\times n}$ is a pencil if*

$$A(x) = A_0 + xA_1, \quad A_0, A_1 \in \mathbb{D}^{m\times n}. \tag{2.1.1}$$

A pencil $A(x)$ is regular if $m = n$ and $\det A(x) \neq 0$*. Otherwise $A(x)$ is a singular pencil. Two pencils $A(x), B(x) \in \mathbb{D}[x]^{m\times n}$ are strictly equivalent if*

$$A(x)\overset{s}{\sim}B(x) \iff B(x) = QA(x)P,\ P \in \mathbf{GL}(n,\mathbb{D}),\ Q \in \mathbf{GL}(m,\mathbb{D}). \tag{2.1.2}$$

The classical works of Weierstrass [Wei67] and Kronecker [Kro90] classify the equivalence classes of pencils under the strict equivalence relation in the case $\mathbb{D}$ is a field $\mathbb{F}$. We give a short account of their main results.

First note that the strict equivalence of $A(x), B(x)$ implies the equivalence of $A(x), B(x)$ over the domain $\mathbb{D}[x]$. Furthermore let

$$B(x) = B_0 + xB_1. \tag{2.1.3}$$

Then the condition (2.1.2) is equivalent to

$$B_0 = QA_0P,\ B_1 = QA_1P, \quad P \in \mathbf{GL}(n,\mathbb{D}),\ Q \in \mathbf{GL}(m,\mathbb{D}). \tag{2.1.4}$$

Thus we can interchange A_0 with A_1 and B_0 with B_1 without affecting the strict equivalence relation. Hence it is natural to consider a *homogeneous* pencil

$$A(x_0, x_1) = x_0 A_0 + x_1 A_1. \tag{2.1.5}$$

Assume that $\mathbb{D}$ is $\mathbb{D}_U$. Then $\mathbb{D}_U[x_0, x_1]$ is also $\mathbb{D}_U$ (Problem 1.4.6.) In particular $\mathbb{D}_U[x_0, x_1]$ is $\mathbb{D}_G$. Let $\delta_k(x_0, x_1),\ i_k(x_0, x_1)$ be the invariant determinants and factors of $A(x_0, x_1)$ respectively for $k = 1, ..., \mathrm{rank\ A(x_0, x_1)}$.

Lemma 2.1.2 *Let $A(x_0, x_1)$ be a homogeneous pencil over $\mathbb{D}_U[x_0, x_1]$. Then its invariant determinants and the invariant polynomials $\delta_k(x_0, x_1),\ i_k(x_0, x_1),\ k = 1, ..., \mathrm{rank\ A(x_0, x_1)}$ are homogeneous polynomials. Moreover, if $\delta_k(x)$ and $i_k(x)$ are the invariant determinants and factors of the pencil $A(x)$ for $k = 1, ..., \mathrm{rank\ A(x)}$, then*

$$\delta_k(x) = \delta_k(1, x),\ i_k(x) = i_k(1, x), \quad k = 1, ..., \mathrm{rank\ A(x)}. \tag{2.1.6}$$

Proof. Clearly any $k \times k$ minor of $A(x_0, x_1)$ is either zero or a homogeneous polynomial of degree k. In view of Problem 1 we deduce that the g.c.d. of all nonvanishing $k \times k$ minors is a homogeneous polynomial $\delta_k(x_0, x_1)$. As $i_k(x_0, x_1) = \frac{\delta_k(x_0,x_1)}{\delta_{k-1}(x_0,x_1)}$ Problem 1 implies that $i_k(x_0, x_1)$ is a homogeneous polynomial. Consider the pencil $A(x) = A(1, x)$. So $\delta_k(x)$ - the g.c.d. of $k \times k$ minors of $A(x)$ is obviously divisible by $\delta_k(1, x)$. On the other hand we have the following relation between the minors of $A(x_0, x_1)$ and $A(x)$

$$\det A(x_0, x_1)[\alpha, \beta] = x_0^k \det A(\frac{x_1}{x_0})[\alpha, \beta], \quad \alpha, \beta \in [n]_k. \tag{2.1.7}$$

This shows that $x_0^{\rho_k}\delta_k(\frac{x_1}{x_0})$ ($\rho_k = \deg \delta_k(x)$) divides any $k \times k$ minor of $A(x_0, x_1)$. So $x_0^{\rho_k}\delta_k(\frac{x_1}{x_0})|\delta_k(x_0, x_1)$. This proves the first part of (2.1.6). So

$$\delta_k(x_0, x_1) = x_0^{\phi_k}(x_0^{\rho_k}\delta_k(\frac{x_1}{x_0})), \quad \rho_k = \deg \delta_k(x),\ \phi_k \geq 0. \tag{2.1.8}$$

The equality

$$i_k(x_0, x_1) = \frac{\delta_k(x_0, x_1)}{\delta_{k-1}(x_0, x_1)}$$

implies

$$i_k(x_0, x_1) = x_0^{\psi_k}(x_0^{\sigma_k} i_k(\frac{x_1}{x_0})), \quad \sigma_k = \deg i_k(x),\ \psi_k \geq 0. \tag{2.1.9}$$

□

$\delta_k(x_0, x_1)$ and $i_k(x_0, x_1)$ are called the invariant *homogeneous* determinants and the invariant *homogeneous* polynomials (factors) respectively. The classical result due to Weierstrass [Wei67] states:

Theorem 2.1.3 *Let $A(x) \in \mathbb{F}[x]^{n\times n}$ be a regular pencil. Then a pencil $B(x) \in \mathbb{F}[x]^{n\times n}$ is strictly equivalent to $A(x)$ if and only if $A(x)$ and $B(x)$ have the same invariant homogeneous polynomials.*

Proof. The necessary part of the theorem holds for any $A(x),\ B(x)$ which are strictly equivalent. Suppose now that $A(x)$ and $B(x)$ have the same invariant homogeneous polynomials. According to (1.4.4) the pencils $A(x)$ and $B(x)$ have the same invariant polynomials. So $A(x) \sim B(x)$ over $\mathbb{F}[x]$. Therefore

$$W(x)B(x) = A(x)U(x),\ U(x), W(x) \in \mathbf{GL}(n, \mathbb{F}[x]). \tag{2.1.10}$$

Assume first that A_1 and B_1 are nonsingular. Then (see Problem 2) it is possible to divide $W(x)$ by $A(x)$ from the right and to divide $U(x)$ by $B(x)$ from the left

$$W(x) = A(x)W_1(x) + R, \quad U(x) = U_1(x)B(x) + P, \tag{2.1.11}$$

where P and R are constant matrices. So

$$A(x)(W_1(x) - U_1(x))B(x) = A(x)P - RB(x).$$

Since $A_1, B_1 \in \mathbf{GL}(n, \mathbb{F})$ we must have that $W_1(x) = U_1(x)$, otherwise the left-hand side of the above equality would be of degree 2 at least (see Definition 2.1.5), while the right-hand side of this equality is at most 1. So

$$W_1(x) = U_1(x), \quad RB(x) = A(x)P. \tag{2.1.12}$$

It is left to show that P and Q are nonsingular. Let $V(x) = W(x)^{-1} \in \mathbf{GL}(n, \mathbb{F}[x])$. Then $I = W(x)V(x)$. Let $V(x) = B(x)V_1(x) + S$. Use the second identity of (2.1.12) to obtain

$$\begin{aligned}
&I = (A(x)W_1(x) + R)V(x) = A(x)W_1(x)V(x) + RV(x) = \\
&A(x)W_1(x)V(x) + RB(x)V_1(x) + RS = \\
&A(x)W_1(x)V(x) + A(x)PV_1(x) + RS = \\
&A(x)(W_1(x)V(x) + PV_1(x)) + RS.
\end{aligned}$$

Since $A_1 \in \mathbf{GL}(n, \mathbb{F})$ the above equality implies

$$W_1(x)V(x) + PV_1(x) = 0, \quad RS = I.$$

Hence R is invertible. Similar arguments show that P is invertible. Thus $A(x) \overset{s}{\sim} B(x)$ if $\det A_1, \det B_1 \neq 0$.

Consider now the general case. Introduce new variables y_0, y_1:

$$y_0 = ax_0 + bx_1,\ y_1 = cx_0 + dx_1, \quad ad - cb \neq 0.$$

Then

$$A(y_0, y_1) = y_0A'_0 + y_1A'_1,\ B(y_0, y_1) = y_0B'_0 + y_1B'_1.$$

Clearly $A(y_0, y_1)$ and $B(y_0, y_1)$ have the same invariant homogeneous polynomials. Also $A(y_0, y_1)\overset{s}{\sim}B(y_0, y_1) \iff A(x_0, x_1)\overset{s}{\sim}B(x_0, x_1)$. Since $A(x_0, x_1)$ and $B(x_0, x_1)$ are regular pencils it is possible to choose a, b, c, d such that A'_1 and B'_1 are nonsingular. This shows that $A(y_0, y_1)\overset{s}{\sim}B(y_0, y_1)$. Hence $A(x)\overset{s}{\sim}B(x)$. □

Using the proof of Theorem 2.1.3 and Problem 2 we obtain:

Theorem 2.1.4 *Let $A(x), B(x) \in \mathbb{D}[x]^{n\times n}$. Assume that $A_1, B_1 \in$ $\mathbf{GL}(n, \mathbb{D})$. Then $A(x)\overset{s}{\sim}B(x) \iff A(x) \sim B(x)$.*

For singular pencils the invariant homogeneous polynomials alone do not determine the class of strictly equivalent pencils. We now introduce the notion of column and row indices for $A(x) \in \mathbb{F}[x]^{m\times n}$. Consider the system (1.14.15). The set of all solutions $\mathbf{w}(x)$ is an $\mathbb{F}[x]$-module $\mathbf{M}$ with a a finite basis $\mathbf{w}_1(x), ..., \mathbf{w}_s(x)$. (Theorem 1.12.3.) To specify a choice of a basis we need the following definition.

Definition 2.1.5 *Let $A \in \mathbb{D}[x_1, ..., x_k]^{m\times n}$. So*

$$A(x_1, ..., x_k) = \sum_{|\alpha|\le d} A_\alpha x^\alpha,\ A_\alpha \in \mathbb{D}^{m\times n},$$

$$\alpha = (\alpha_1, ..., \alpha_k) \in \mathbb{Z}_+^k,\ |\alpha| = \sum_{i=1}^k \alpha_i,\ x^\alpha = x_1^{\alpha_1}...x_k^{\alpha_k}. \tag{2.1.13}$$

Then the degree of $A(x_1, ..., x_k) \neq 0$ (deg A) is d if there exists $A_\alpha \neq 0$ with $|\alpha| = d$. Let deg $0 = 0$.

Definition 2.1.6 *Let $A \in \mathbb{F}[x]^{m\times n}$ and consider the module $\mathbf{M} \subset \mathbb{F}[x]^n$ of all solutions of* (1.14.15)*. Choose a basis $\mathbf{w}_1(x), ..., \mathbf{w}_s(x),\ s = n-$rank A in $\mathbf{M}$ such that $\mathbf{w}_k(x) \in \mathbf{M}$ has the lowest degree among all $\mathbf{w}(x) \in \mathbf{M}$ which are linearly independent over $\mathbb{F}(x)$ of $\mathbf{w}_1, ..., \mathbf{w}_{k-1}(x)$ for $k = 1, ..., s$. Then the column indices $0 \le \alpha_1 \le \alpha_2 \le ... \le \alpha_s$ of $A(x)$ are given as*

$$\alpha_k = \deg \mathbf{w}_k(x), \quad k = 1, ..., s. \tag{2.1.14}$$

The row indices $0 \le \beta_1 \le \beta_2 \le ... \le \beta_t,\ t = m -$ rank A, of $A(x)$ are the column indices of $A(x)^\top$.

It can be shown [Gan59] that the column (row) indices are independent of a particular allowed choice of a basis $\mathbf{w}_1(x), ..., \mathbf{w}_s(x)$. We state the Kronecker result [Kro90]. (See [Gan59] for a proof.)

Theorem 2.1.7 *The pencils $A(x), B(x) \in \mathbb{F}[x]^{m\times n}$ are strictly equivalent if and only if they have the same invariant homogeneous polynomials and the same row and column indices.*

For a canonical form of a singular pencil under the strict equivalence see Problems 8– 12.

Problems

1. Using the fact that $\mathbb{D}_U[x_1, ..., x_n]$ is $\mathbb{D}_U$ and the equality (1.13.5) show that if $a \in \mathbb{D}_U[x_1, ..., x_n]$ is a homogeneous polynomial then in the decomposition (1.3.1) each p_i is a homogeneous polynomial.

2. Let

$$W(x) = \sum_{k=0}^{q} W_k x^k, \; U(x) = \sum_{k=0}^{p} U_k x^k \in \mathbb{D}[x]^{n\times n}. \tag{2.1.15}$$

Assume that $A(x) = A_0 + xA_1$ such that $A_0 \in \mathbb{D}^{n\times n}$ and $A_1 \in \mathbf{GL}(n, \mathbb{D})$. Show that if $p, q \geq 1$ then

$$W(x) = A(x)A_1^{-1}(W_q x^{q-1}) + \tilde{W}(x), \quad U(x) = (U_p x^{p-1})A_1^{-1}A(x) + \tilde{U}(x),$$

where

$$\deg \tilde{W}(x) < q, \quad \deg \tilde{U}(x) < p.$$

Prove the equalities (2.1.11) where R and P are constant matrices. Suppose that $A_1 = I$. Show that R and P in (2.1.11) can be given as

$$R = \sum_{k=0}^{q} (-A_0)^k W_k, \quad P = \sum_{k=0}^{p} U_k (-A_0)^k. \tag{2.1.16}$$

3. Let $A(x) \in \mathbb{D}_U[x]^{n\times n}$ be a regular pencil such that $\det A_1 \neq 0$. Prove that in (2.1.8) and (2.1.9) $\phi_k = \psi_k = 0$ for $k = 1, ..., n$. (Use equality (1.12.12) for $A(x)$ and $A(x_0, x_1)$.)

4. Consider the following two pencils

$$A(x) = \begin{bmatrix} 2+x & 1+x & 3+3x \\ 3+x & 2+x & 5+2x \\ 3+x & 2+x & 5+2x \end{bmatrix}, \; B(x) = \begin{bmatrix} 2+x & 1+x & 1+x \\ 1+x & 2+x & 2+x \\ 1+x & 1+x & 1+x \end{bmatrix}$$

over $\mathbb{R}[x]$. Show that $A(x)$ and $B(x)$ are equivalent but not strictly equivalent.

5. Let
$$A(x) = \sum_{k=0}^{p} A_k x^k \in \mathbb{F}[x]^{m\times n}.$$
Put
$$A(x_0, x_1) = \sum_{k=0}^{q} A_k x_0^{q-k} x_1^k, \quad q = \deg A(x).$$
Prove that $i_k(x_0, x_1)$ is a homogeneous polynomial for $k = 1, ..., \mathrm{rank\ A(x)}$. Show that $i_1(1, x), ..., i_k(1, x)$ are the invariant factors of $A(x)$.

6. Let $A(x), B(x) \in \mathbb{F}[x]^{m\times n}$. $A(x)$ and $B(x)$ are called strictly equivalent $(A(x)\overset{s}{\sim}B(x))$ if
$$B(x) = PA(x)Q, \quad P \in \mathbf{GL}(m, \mathbb{F}),\ Q \in \mathbf{GL}(n, \mathbb{F}).$$
Show that if $A(x)\overset{s}{\sim}B(x)$ then $A(x_0, x_1)$ and $B(x_0, x_1)$ have the same invariant factors.

7. Let $A(x), B(x) \in \mathbb{F}[x]^{m\times n}$. Show that $A(x)\overset{s}{\sim}B(x) \iff A(x)^\top\overset{s}{\sim}B(x)^\top$.

8. (a) Let $L_m(x) \in \mathbb{F}[x]^{m\times(m+1)}$ be matrix with 1 on the main diagonal and x on the diagonal above it, and all other entries 0:
$$L_m(x) = \begin{bmatrix} 1 & x & 0 & \ldots & 0 \\ 0 & 1 & x & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & x \end{bmatrix}.$$
Show that $\mathrm{rank\ L_m = m}$ and $\alpha_1 = m$.

   (b) Let $1 \le \alpha_1 \le \ldots \le \alpha_s,\ 1 \le \beta_1 \le \ldots \le \beta_t$ be integers. Assume that $B(x) = B_0 + xB_1 \in \mathbb{F}[x]^{l\times l}$ is a regular pencil. Show that $A(x) = B(x) \oplus_{i=1}^{s} L_{\alpha_i} \oplus_{j=1}^{t} L_{\beta_j}^{\top}$ has the column and the row indices $1 \le \alpha_1 \le \ldots \le \alpha_s,\ 1 \le \beta_1 \le \ldots \le \beta_t$ respectively.

9. Show if a pencil $A(x)$ is a direct sum of pencils of the below form, where one of the summands of the form 9a-9b appears, it is a singular pencil.

   (a) $L_m(x)$.

   (b) $L_m(x)^\top$.

   (c) $B(x) = B_0 + xB_1 \in \mathbb{F}[x]^{l\times l}$ is a regular pencil.

10. Show that a singular pencil $A(x)$ is strictly similar to the singular pencil given in Problem 9, if and only if there are no column and row indices equal to 0.

11. Assume that $A(x) \in \mathbb{F}[x]^{m\times n}$ is a singular pencil.

    (a) Show that $A(x)$ has exactly k column indices equal to 0, if and only if it is strict equivalent to $[0_{m\times k}\ A_1(x)], A_1(x) \in \mathbb{F}[x]^{m\times(n-k)}$, where either $A_1(x)$ is regular or singular. If $A_1(x)$ is singular then the row indices of $A_1(x)$ are the row indices of $A(x)$, and the column indices of $A_1(x)$ are the nonzero column indices of of $A(x)$.

    (b) By considering $A(x)^\top$ state and prove similar result for the row indices of $A(x)$.

12. Use Problems 8–11 to find a canonical from for a singular pencil $A(x)$ under the strict equivalence.

2.2 Similarity of matrices

Definition 2.2.1 *Let $A, B \in \mathbb{D}^{m\times m}$. Then A and B are similar ($A \approx B$) if*

$$B = QAQ^{-1}, \tag{2.2.1}$$

for some $Q \in \mathbf{GL}(m, \mathbb{D})$.

Clearly the similarity relation is an equivalence relation. So $\mathbb{D}^{m\times m}$ is divided into equivalences classes which are called the similarity classes. For a $\mathbb{D}$ module $\mathbf{M}$ we let Hom $(\mathbf{M})$:= Hom $(\mathbf{M}, \mathbf{M})$. It is a standard fact that each similarity class corresponds to all possible representations of some $T \in$ Hom $(\mathbf{M})$, where $\mathbf{M}$ is a $\mathbb{D}$-module having a basis of m elements. Indeed, let $[\mathbf{u}_1, ..., \mathbf{u}_m]$ be a basis in $\mathbf{M}$. Then T is represented by $A = [a_{ij}] \in \mathbb{D}^{m\times m}$, where

$$T\mathbf{u}_j = \sum_{i=1}^{m} a_{ij}\mathbf{u}_i, \quad j = 1, ..., m. \tag{2.2.2}$$

Let $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ be another basis in $\mathbf{M}$. Assume that $Q \in \mathbf{GL}(m, \mathbb{D})$ is given by (1.10.5). According to (2.2.2) and the arguments of §1.10, the representation of T in the basis $[\tilde{\mathbf{u}}_1, ..., \tilde{\mathbf{u}}_m]$ is given by the matrix B of the form (2.2.1).

The similarity notion of matrices is closely related to strict equivalence of certain regular pencils.

Lemma 2.2.2 *Let $A, B \in \mathbb{D}^{m\times m}$. Associate with these matrices the following regular pencils*

$$A(x) = -A + xI, \quad B(x) = -B + xI. \tag{2.2.3}$$

Then A and B are similar if and only if the pencils $A(x)$ and $B(x)$ are strictly equivalent.

Proof. Assume first that $A \approx B$. Then (2.2.1) implies (2.1.2) where $P = Q^{-1}$. Suppose now that $A(x)\overset{s}{\sim}B(x)$. So $B = QAP$, $I = QP$. That is $P = Q^{-1}$ and $A \approx B$. □

Clearly $A(x)\overset{s}{\sim}B(x) \Rightarrow A(x) \sim B(x)$.

Corollary 2.2.3 *Let $A, B \in \mathbb{D}_U^{m\times m}$. Assume that A and B are similar. Then the corresponding pencils $A(x)$, $B(x)$ given by* (2.2.3) *have the same invariant polynomials.*

In the case $\mathbb{D}_U = \mathbb{F}$ the above condition is also a sufficient condition in view of Lemma 2.2.2 and Corollary 2.1.4

Theorem 2.2.4 *Let $A, B \in \mathbb{F}^{m\times m}$. Then A and B are similar if and only if the pencils $A(x)$ and $B(x)$ given by (2.2.3) have the same invariant polynomials.*

It can be shown (see Problem 1) that even over Euclidean domains the condition that $A(x)$ and $B(x)$ have the same invariant polynomials does not imply in general that $A \approx B$.

Problems

1. Let
$$A = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}, \; B = \begin{bmatrix} 1 & 1 \\ 0 & 5 \end{bmatrix} \in \mathbb{Z}^{2\times 2}.$$
Show that $A(x)$ and $B(x)$ given by (2.2.3) have the same invariant polynomials over $\mathbb{Z}[x]$. Show that A and B are not similar over $\mathbb{Z}$.

2. Let $A(x) \in \mathbb{D}_U[x]^{n\times n}$ be given by (2.2.3). Let $i_1(x), ..., i_n(x)$ be the invariant polynomial of $A(x)$. Using the equality (1.12.12) show that each $i_k(x)$ can be assumed to be normalized polynomial and
$$\sum_{k=1}^{n} \deg i_k(x) = n. \tag{2.2.4}$$

3. Let $A \in \mathbb{F}^{n\times n}$. Show that $A \approx A^\top$.

2.3 The companion matrix

Theorem 2.2.4 shows that if $A \in \mathbb{F}^{n\times n}$ then the invariant polynomials determine the similarity class of A. We now show that any set of normalized polynomials $i_1(x), ..., i_n(x) \in \mathbb{D}_U[x]$, such that $i_j(x)|i_{j+1}(x),\ j = 1, ..., n-1$ and which satisfy (2.2.4), are invariant polynomials of $xI - A$ for some $A \in \mathbb{D}_U^{n\times n}$. To do so we introduce the notion of a companion matrix.

Definition 2.3.1 *Let $p(x) \in \mathbb{D}[x]$ be a normalized polynomial*

$$p(x) = x^m + a_1x^{m-1} + ... + a_m.$$

Then $C(p) = [c_{ij}]_1^m \in \mathbb{D}^{m\times m}$ is the companion matrix of $p(x)$ if

$$c_{ij} = \delta_{(i+1)j}, \quad i = 1, ..., m-1,\ j = 1, ..., m,$$

$$c_{mj} = -a_{m-j+1}, \quad j = 1, ..., m, \tag{2.3.1}$$

$$C(p) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_m & -a_{m-1} & -a_{m-2} & \dots & -a_2 & -a_1 \end{bmatrix}.$$

Lemma 2.3.2 *Let $p(x) \in \mathbb{D}_U[x]$ be a normalized polynomial of degree m. Consider the pencil $C(x) = xI - C(p)$. Then the invariant polynomials of $C(x)$ are*

$$i_1(C) = ... = i_{m-1}(C) = 1, \quad i_m(C) = p(x). \tag{2.3.2}$$

Proof. For $k < m$ consider a minor of $C(x)$ composed of the rows $1, ..., k$ and the columns $2, ..., k+1$. Since this minor is the determinant of a lower triangular matrix with -1 on the main diagonal we deduce that its value is $(-1)^k$. So $\delta_k(C) = 1,\ k = 1, ..., m-1$. This establishes the first equality in (2.3.2). Clearly $\delta_m(C) = \det\ (xI - C)$. Expand the determinant of $C(x)$ by the first row and use induction to prove that $\det\ (xI - C) = p(x)$. This shows that $i_m(C) = \frac{\delta_m(C)}{\delta_{m-1}(C)} = p(x)$. □

Using the results of Problem 2.1.13 and Lemma 2.3.2 we get:

Theorem 2.3.3 *Let $p_j(x) \in \mathbb{D}_U[x]$ be normalized polynomials of positive degrees such that $p_j(x)|p_{j+1}(x),\ j = 1, ..., k-1$. Consider the matrix*

$$C(p_1, ..., p_k) = \oplus_{j=1}^k C(p_j). \tag{2.3.3}$$

Then the nontrivial invariant polynomials of $xI - C(p_1, ..., p_k)$ (i.e. those polynomials which are not the identity element) are $p_1(x), ..., p_k(x)$.

Combining Theorems 2.2.4 and 2.3.3 we obtain a canonical representation for the similarity class in $\mathbb{F}^{n\times n}$.

Theorem 2.3.4 *Let $A \in \mathbb{F}^{n\times n}$ and assume that $p_j(x) \in \mathbb{F}[x]$, $j = 1, ..., k$ are the nontrivial normalized invariant polynomials of $xI - A$. Then A is similar to $C(p_1, ..., p_k)$.*

Definition 2.3.5 *For $A \in \mathbb{F}^{n\times n}$ the matrix $C(p_1, ..., p_k)$ is called the rational canonical form of A.*

Let $\mathbb{F}$ be the quotient field of $\mathbb{D}$. Assume that $A \in \mathbb{D}^{n\times n}$. Let $C(p_1, ..., p_k)$ be the rational canonical form of A in $\mathbb{F}^{n\times n}$. We now discuss the case when $C(p_1, ..., p_k) \in \mathbb{D}^{n\times n}$. Assume that $\mathbb{D}$ is $\mathbb{D}_U$. Let δ_k be the g.c.d of $k \times k$ minors of $xI - A$. So δ_k divides the minor $p(x) = \det\,(xI - A)[\alpha, \alpha]$, $\alpha = \{1, ..., k\}$. Clearly $p(x)$ is normalized polynomial in $\mathbb{D}_U[x]$. Recall that $\mathbb{D}_U[x]$ is also $\mathbb{D}_U$ (§1.4).

According to Theorem 1.4.8 the decomposition of $p(x)$ into irreducible factors in $\mathbb{D}_U[x]$ is of the form (1.4.4), where $a = 1$ and each $q_i(x)$is a nontrivial normalized polynomial in $\mathbb{D}_U[x]$. Hence $i_k = \frac{\delta_k}{\delta_{k-1}}$ is either identity or a nontrivial polynomial in $\mathbb{D}_U[x]$. Thus

Theorem 2.3.6 *Let $A \in \mathbb{D}_U^{n\times n}$. Then the rational canonical form $C(p_1, ..., p_k)$ of A over the quotient field $\mathbb{F}$ of $\mathbb{D}_U$ belongs to $\mathbb{D}_U^{n\times n}$.*

Corollary 2.3.7 *Let $A \in \mathbb{C}[x_1, ..., x_m]^{n\times n}$. Then the rational canonical form of A over $\mathbb{C}(x_1, ..., x_n)$ belongs to $\mathbb{C}[x_1, ..., x_m]^{n\times n}$.*

Using the results of Theorem 1.4.9 we deduce that Theorem 2.3.6 applies to the ring of analytic functions in several variables although this ring is not $\mathbb{D}_U$ (§1.3).

Theorem 2.3.8 *Let $A \in \mathrm{H}(\Omega)^{n\times n}$ $(\Omega \subset \mathbb{C}^m)$. Then the rational canonical form of A over the field of meromorphic functions in Ω belongs to $\mathrm{H}(\Omega)^{n\times n}$.*

Problems

1. Let $p(x) \in \mathbb{D}_U[x]$ be a normalized nontrivial polynomial. Assume that $p(x) = p_1(x)p_2(x)$, where $p_i(x)$ is a normalized nontrivial polynomial in $\mathbb{D}_U[x]$ for $i = 1, 2$. Using Problem 1.13.1 and 2 show that $xI - C(p_1, p_2)$ given by (2.3.3) have the same invariant polynomials as $xI - C(p)$ if and only if $(p_1, p_2) = 1$.

2. Let $A \in \mathbb{D}_U^{n\times n}$ and assume that $p_1(x),...,p_k(x)$ are the nontrivial normalized invariant polynomials of $xI - A$. Let

$$p_j(x) = (\phi_1(x))^{m_{1j}}...(\phi_l(x))^{m_{lj}}, \quad j = 1,...,k, \tag{2.3.4}$$

where $\phi_1(x),...,\phi_l(x)$ are nontrivial normalized irreducible polynomials in $\mathbb{D}_U[x]$ such that $(\phi_i, \phi_j) = 1$ for $i \neq j$. Prove that

$$m_{ik} \geq 1,\ m_{ik} \geq m_{i(k-1)} \geq ... \geq m_{i1} \geq 0,\ \sum_{i,j=1}^{l,k} m_{ij} = n. \tag{2.3.5}$$

The polynomials $\phi_i^{m_{ij}}$ for $m_{ij} > 0$ are called the *elementary divisors* of $xI - A$. Let

$$E = \oplus_{m_{ij}>0} C(\phi_i^{m_{ij}}). \tag{2.3.6}$$

Show that $xI - A$ and $xI - E$ have the same invariant polynomials. Hence $A \approx E$ over the quotient field $\mathbb{F}$ of $\mathbb{D}_U$. (In some references E is called the rational canonical form of A.)

2.4 Splitting to invariant subspaces

Let $\mathbf{V}$ be an m dimensional vector space over $\mathbb{F}$ and let $T \in \text{Hom}\,(\mathbf{V})$. In §2.2 we showed that the set of all matrices $\mathcal{A} \subset \mathbb{F}^{m\times m}$, which represents T in different bases, is an equivalence class of matrices with respect to the similarity relation. Theorem 2.2.4 shows that the class $\mathcal{A}$ is characterized by the invariant polynomials of $xI - A$ for some $A \in \mathcal{A}$. Since $xI - A$ and $xI - B$ have the same invariant polynomials if and only if $A \approx B$ we define:

Definition 2.4.1 *Let $T \in \text{Hom}\,(\mathbf{V})$ and let $A \in \mathbb{F}^{m\times m}$ be a representation matrix of T given by the equality* (2.2.2) *in some basis $\mathbf{u}_1, ..., \mathbf{u}_m$ of $\mathbf{V}$. Then the invariant polynomials $i_1(x), ..., i_m(x)$ of T are defined as the invariant polynomials of $xI - A$. The characteristic polynomial of T is the polynomial* $\det\,(xI - A)$.

The fact that the characteristic polynomial of T is independent of a representation matrix A follows from the identity (1.12.12)

$$\det\,(xI - A) = p_1(x)...p_k(x), \tag{2.4.1}$$

where $p_1(x),...,p_k(x)$ are the nontrivial invariant polynomials of $xI - A$. In §2.3 we showed that the matrix $C(p_1,...,p_k)$ is a representation matrix of T. In this section we consider another representation of T which is closely related to the matrix E given in (2.3.6). This form is achieved by splitting $\mathbf{V}$ to a direct sum

$$\mathbf{V} = \oplus_{i=1}^{l} \mathbf{U}_i, \tag{2.4.2}$$

where each $\mathbf{U}_i$ is an invariant subspace of T defined as follows:

Definition 2.4.2 *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$ and $T \in \mathrm{Hom}\,(\mathbf{V})$. A subspace $\mathbf{U} \subseteq \mathbf{V}$ is an invariant subspace of T (T-invariant) if*

$$T\mathbf{U} \subseteq \mathbf{U}. \tag{2.4.3}$$

$\mathbf{U}$ is called trivial if $\mathbf{U} = \{\mathbf{0}\}$ or $\mathbf{U} = \mathbf{V}$. $\mathbf{U}$ is called nontrivial, (proper), if $\{\mathbf{0}\} \neq \mathbf{U} \neq \mathbf{V}$. $\mathbf{U}$ is called irreducible if $\mathbf{U}$ can not be expressed a direct sum of two nontrivial invariant subspaces of T. The restriction of T to a T-invariant subspace $\mathbf{U}$ is denoted by $T|\mathbf{U}$.

Thus if $\mathbf{V}$ splits into a direct sum of nontrivial invariant subspaces of T, then a direct sum of matrix representations of T on each $\mathbf{U}_j$ gives a representation of T. So, a simple representation of T can be achieved by splitting $\mathbf{V}$ into a direct sum of irreducible invariant subspaces. To do so we need to introduce the notion of the minimal polynomial of T. Consider the linear operators $I = T^0, T, T^2, ..., T^{m^2}$, where I is the identity operator ($I\mathbf{v} = \mathbf{v}$). As $\dim \mathrm{Hom}\,(\mathbf{V}) = m^2$, these $m^2 + 1$ operators are linearly dependent. So there exists an integer $q \in [0, m^2]$ such that $I, T, ..., T^{q-1}$ are linearly independent and $I, T, ..., T^q$ are linearly dependent. Let $0 \in \mathrm{Hom}\,(\mathbf{V})$ be the zero operator: $0\mathbf{v} = \mathbf{0}$. For $\phi \in \mathbb{F}[x]$ let $\phi(T)$ be the operator

$$\phi(T) = \sum_{i=0}^{l} c_i T^i, \quad \phi(x) = \sum_{i=1}^{l} c_i x^i.$$

ϕ is annihilated by T if $\phi(T) = 0$.

Definition 2.4.3 *A polynomial $\psi(x) \in \mathbb{F}[x]$ is a minimal polynomial of $T \in \mathrm{Hom}\,(\mathbf{V})$ if $\psi(x)$ is a normalized polynomial of the smallest degree annihilated by T.*

Lemma 2.4.4 *Let $\psi(x) \in \mathbb{F}[x]$ be the minimal polynomial $T \in \mathrm{Hom}\,(\mathbf{V})$. Assume that T annihilates ϕ. Then $\psi|\phi$.*

Proof. Divide ϕ by ψ:

$$\phi(x) = \chi(x)\psi(x) + \rho(x), \quad \deg \rho < \deg \psi.$$

As $\phi(T) = \psi(T) = 0$ it follows that $\rho(T) = 0$. From the definition of $\psi(x)$ it follows that $\rho(x) = 0$. □

Since $\mathbb{F}[x]$ is a unique factorization domain, let

$$\begin{aligned} &\psi(x) = \phi_1^{s_1}...\phi_l^{s_l}, \\ &(\phi_i, \phi_j) = 1 \text{ for } 1 \le i < j \le l, \quad \deg \phi_i \ge 1, \ i = 1, ..., l, \end{aligned} \tag{2.4.4}$$

where each ϕ_i is a normalized irreducible polynomial if $\mathbb{F}[x]$.

Theorem 2.4.5 *Let ψ be the minimal polynomial of $T \in \mathrm{Hom}\,(\mathbf{V})$. Assume that ψ splits to a product of coprime factors given in* (2.4.4). *Then the vector space $\mathbf{V}$ splits to a direct sum* (2.4.2), *where each $\mathbf{U}_j$ is a nontrivial invariant subspace of $T|\mathbf{U}_j$. Moreover $\phi_j^{s_j}$ is the minimal polynomial of $T|\mathbf{U}_j$.*

The proof of the theorem follows immediately from the lemma below.

Lemma 2.4.6 *Let ψ be the minimal polynomial of $T \in \mathrm{Hom}\,(\mathbf{V})$. Assume that ψ splits to a product of two nontrivial coprime factors*

$$\psi(x) = \psi_1(x)\psi_2(x), \quad \deg \psi_i \ge 1, \ i = 1, 2, \quad (\psi_1, \psi_2) = 1, \tag{2.4.5}$$

where each ψ_i is normalized. Then

$$\mathbf{V} = \mathbf{U}_1 \oplus \mathbf{U}_2, \tag{2.4.6}$$

where each $\mathbf{U}_j$ is a nontrivial T-invariant subspace and ψ_j is the minimal polynomial of $T_j := T|\mathbf{U}_j$.

Proof. The assumptions of the lemma imply the existence of polynomials $\theta_1(x)$ and $\theta_2(x)$ such that

$$\theta_1(x)\psi_1(x) + \theta_2(x)\psi_2(x) = 1. \tag{2.4.7}$$

Define

$$\mathbf{U}_j = \{\mathbf{u} \in \mathbf{V} : \quad \psi_j(T)\mathbf{u} = \mathbf{0}\}, \quad j = 1, 2. \tag{2.4.8}$$

Since any two polynomials in T commute (i.e. $\mu(T)\nu(T) = \nu(T)\mu(T)$) it follows that each $\mathbf{U}_j$ is T-invariant. The equality (2.4.7) implies

$$I = \psi_1(T)\theta_1(T) + \psi_2(T)\theta_2(T).$$

Hence for any $\mathbf{u} \in \mathbf{V}$ we have

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2, \quad \mathbf{u}_1 = \psi_2(T)\theta_2(T)\mathbf{u} \in \mathbf{U}_1,\ \mathbf{u}_2 = \psi_1(T)\theta_1(T)\mathbf{u} \in \mathbf{U}_2.$$

So $\mathbf{V} = \mathbf{U}_1 + \mathbf{U}_2$. Suppose that $\mathbf{u} \in \mathbf{U}_1 \cap \mathbf{U}_2$. Then $\psi_1(T)\mathbf{u} = \psi_2(T)\mathbf{u} = \mathbf{0}$. Hence $\theta_1(T)\psi_1(T)\mathbf{u} = \theta_2(T)\psi_2(T)\mathbf{u} = \mathbf{0}$. Thus

$$\mathbf{u} = \psi_1(T)\mathbf{u} + \psi_2(T)\mathbf{u} = 0.$$

So $\mathbf{U}_1 \cap \mathbf{U}_2 = \{\mathbf{0}\}$ and (2.4.6) holds. Clearly T_j annihilates ψ_j. Let $\bar\psi_j$ be the minimal polynomial of T_j. So $\bar\psi_j|\psi_j,\ j = 1,2$. Now

$$\bar\psi_1(T)\bar\psi_2(T)\mathbf{u} = \bar\psi_1(T)\bar\psi_2(T)(\mathbf{u}_1+\mathbf{u}_2) = \bar\psi_2(T)\bar\psi_1(T)\mathbf{u}_1+\bar\psi_1(T)\bar\psi_2(T)\mathbf{u}_2 = 0.$$

Hence T annihilates $\bar\psi_1\bar\psi_2$. Since ψ is the minimal polynomial of T we have $\psi_1\psi_2|\bar\psi_1\bar\psi_2$. Therefore $\bar\psi_j = \psi_j,\ j = 1,2$. As deg $\psi_j \geq 1$ it follows that dim $\mathbf{U}_j \geq 1$. □

Problems

1. Assume that (2.4.6) holds, where $T\mathbf{U}_j \subseteq \mathbf{U}_j,\ j = 1,2$. Let ψ_j be the minimal polynomial of $T_j = T|\mathbf{U}_j$ for $j = 1,2$. Show that the minimal polynomial ψ of T is equal to $\frac{\psi_1\psi_2}{(\psi_1,\psi_2)}$.

2. Let the assumptions of Problem 1 hold. Assume furthermore that $\psi = \phi^s$, where ϕ is irreducible over $\mathbb{F}[x]$. Then either $\psi_1 = \psi$ or $\psi_2 = \psi$.

3. Let $C(p) \in \mathbb{D}^{m\times m}$ be the companion matrix given by (2.3.1). Let $\mathbf{e}_i = (\delta_{i1}, ..., \delta_{im})^\top,\ i = 1, ..., m$ be the standard basis in $\mathbb{D}^m$. Show

   (2.4.9) $$C(p)\mathbf{e}_i = \mathbf{e}_{i-1} - a_{m-i+1}\mathbf{e}_m, \quad i = 1, ..., m, \quad (\mathbf{e}_0 = \mathbf{0}).$$

   Prove that $p(C) = 0$ and that any polynomial $0 \neq q \in \mathbb{D}[x]$, deg $q < m$ is not annihilated by C. (Consider $q(C)\mathbf{e}_i$ and use (2.4.9).) That is: p is the minimal polynomial of $C(p)$.

   Hint: Use the induction on m as follows. Set $\mathbf{f}_i = \mathbf{e}_{m-i+1}$ for $i = 1, \ldots, m$. Let $q = x^{m-1} + a_1x^{m-1} + \ldots + a_{m-1}$. Set $Q = \begin{bmatrix} 0 & \mathbf{0}_{m-1}^\top \\ \mathbf{0}_{m-1} & C(q) \end{bmatrix}$. Use the induction hypothesis on $C(q)$, i.e. $q(C(q)) = 0$, and the facts that $C(p)\mathbf{f}_i = Q\mathbf{f}_i$ for $i = 1, \ldots, m-1$, $C(p)\mathbf{f}_m = \mathbf{f}_{m+1} + Q\mathbf{f}_m$ to obtain that $p(C(p))\mathbf{f}_1 = \mathbf{0}$. Now use (2.4.9) to show that $p(C(p))\mathbf{f}_i = \mathbf{0}$ for $i = 2, \ldots, m+1$.

4. Let $A \in \mathbb{F}^{m\times m}$. Using Theorem 2.3.4 and Problems 1 and 3 show that the minimal polynomial ψ of A is the last invariant polynomial of $xI - A$. That is:

$$\psi(x) = \frac{\det(xI - A)}{\delta_{m-1}(x)}, \tag{2.4.10}$$

where $\delta_{m-1}(x)$ is the g.c.d. of all $(m-1)\times(m-1)$ minors of $xI - A$.

5. Show that the results of Problem 4 apply to $A \in \mathbb{D}_U^{m\times m}$. In particular, if $A \approx B$ then A and B have the same minimal polynomials.

6. Deduce from Problem 4 the *Cayley-Hamilton* theorem which states that $T \in \mathrm{Hom}\,(\mathbf{V})$ annihilates its characteristic polynomial.

7. Let $A \in \mathbb{D}^{m\times m}$. Prove that A annihilates its characteristic polynomial. (Consider the quotient field $\mathbb{F}$ of $\mathbb{D}$.)

8. Use Problem 6 and Lemma 2.4.4 to show

$$\deg \psi \le \dim \mathbf{V}. \tag{2.4.11}$$

9. Let $\psi = \phi^s$, where ϕ is irreducible in $\mathbb{F}[x]$. Assume that $\deg \psi = \dim \mathbf{V}$. Use Problems 2 and 8 to show that $\mathbf{V}$ is an irreducible invariant subspace of T.

10. Let $p(x) \in \mathbb{F}[x]$ be a nontrivial normalized polynomial such that $p = \phi^s$, where ϕ is a normalized irreducible in $\mathbb{F}[x]$. Let $T \in \mathrm{Hom}\,(\mathbf{V})$ be represented by $C(p)$. Use Problem 9 to show that $\mathbf{V}$ is an irreducible invariant subspace of T.

11. Let $T \in \mathrm{Hom}\,(\mathbf{V})$ and let E be the matrix given by (2.3.6), which is determined by the elementary divisors of T. Using Problem 10 show that the representation E of T corresponds to a splitting of V to a direct sum of irreducible invariant subspaces of T.

12. Deduce from Problem 9 and 11 that $\mathbf{V}$ is an irreducible invariant subspace of T if and only if the minimal polynomial ψ of T satisfies the assumptions of Problem 9

2.5 An upper triangular form

Definition 2.5.1 *Let $\mathbf{M}$ be a $\mathbb{D}$-module and assume that $T \in \mathrm{Hom}\,(\mathbf{M})$. $\lambda \in \mathbb{D}$ is an eigenvalue of T if there exists $\mathbf{0} \neq \mathbf{u} \in \mathbf{M}$ such that*

$$T\mathbf{u} = \lambda\mathbf{u}. \tag{2.5.1}$$

The element, (vector), $\mathbf{u}$ *is an eigenelement, (eigenvector), corresponding to* λ*. An element* $\mathbf{0} \neq \mathbf{u}$ *is a generalized eigenelement, (eigenvector), if*

$$(\lambda I - T)^k \mathbf{u} = \mathbf{0} \tag{2.5.2}$$

for some positive integer k*, where* λ *is an eigenvalue of* T*. For* $T \in \mathbb{D}^{m\times m}$ λ *is an eigenvalue if* (2.5.1) *holds for some* $\mathbf{0} \neq \mathbf{u} \in \mathbb{D}^m$*. The element* $\mathbf{u}$ *is eigenelement, (eigenvector), or generalized eigenelement, (eigenvector), if either* (2.5.1) *or* (2.5.2) *holds respectively.*

Lemma 2.5.2 *Let* $T \in \mathbb{D}^{m\times m}$*. Then* λ *is an eigenvalue of* T *if and only if* λ *is a root of the characteristic polynomial* $\det\,(xI - T)$*.*

Proof. Let $\mathbb{F}$ be the quotient field of $\mathbb{D}$. Assume first that λ is an eigenvalue of T. As (2.5.1) is equivalent to $(\lambda I - T)\mathbf{u} = \mathbf{0}$ and $\mathbf{u} \neq \mathbf{0}$, then above system has a nontrivial solution. Therefore $\det\,(\lambda I - T) = 0$. Vice versa, if $\det\,(\lambda I - T) = 0$ then the system $(\lambda I - T)\mathbf{v} = \mathbf{0}$ has a nontrivial solution $\mathbf{v} \in \mathbb{F}^m$. Then there exists $0 \neq a \in \mathbb{D}$ such that $\mathbf{u} := a\mathbf{v} \in \mathbb{D}^m$ and $T\mathbf{u} = \lambda\mathbf{u}$. □

Definition 2.5.3 *A matrix* $A = [a_{ij}] \in \mathbb{D}^{m\times m}$ *is an upper, (lower), triangular if* $a_{ij} = 0$ *for* $j < i$*,* $(j > i)$*. Let* $\mathbf{UT}(m, \mathbb{D})$*,* $(\mathbf{LT}(m, \mathbb{D})) \subset \mathbb{D}^{m\times m}$ *be the ring of upper, (lower), triangular* $m\times m$ *matrices. Let* $\mathbf{UTG}(m, \mathbb{D}) = \mathbf{UT}(m, \mathbb{D}) \cap \mathbf{GL}(m, \mathbb{D})$*,* $(\mathbf{LTG}(m, \mathbb{D}) = \mathbf{LT}(m, \mathbb{D}) \cap \mathbf{GL}_m(\mathbb{D}))$*.*

Theorem 2.5.4 *Let* $T \in \mathbb{D}^{m\times m}$*. Assume that the characteristic polynomial of* T *splits to linear factors over* $\mathbb{D}$

$$\det\,(xI - T) = \prod_{i=1}^{m}(x - \lambda_i), \quad \lambda_i \in \mathbb{D},\ i = 1, ..., m. \tag{2.5.3}$$

Assume furthermore that $\mathbb{D}$ *is a Bezout domain. Then*

$$T = QAQ^{-1}, \quad Q \in \mathbf{GL}(m, \mathbb{D}), \quad A = [a_{ij}]_1^m \in \mathbf{UT}(m, \mathbb{D}), \tag{2.5.4}$$

such that $a_{11}, ..., a_{mm}$ *are the eigenvalues* $\lambda_1, ..., \lambda_m$ *appearing in any specified order.*

Proof. Let λ be an eigenvalue of T and consider the set of all $\mathbf{u} \in \mathbb{D}^m$ which satisfies (2.5.1). Clearly this set is a $\mathbb{D}$-module $\mathbf{M}$. Lemma 2.5.2 yields that $\mathbf{M}$ contains nonzero vectors. Assume that $\mathbb{D}$ is $\mathbb{D}_B$. According to Theorem 1.12.3 $\mathbf{M}$ has a basis $\mathbf{u}_1, ..., \mathbf{u}_k$ which can be completed to a basis $\mathbf{u}_1, ..., \mathbf{u}_m$ in $\mathbb{D}^m$. Let

$$T\mathbf{u}_i = \sum_{j=1}^{m} b_{ji}\mathbf{u}_j, \quad i = 1, ..., m, \quad B = [b_{ij}] \in \mathbb{D}^{m\times m}. \tag{2.5.5}$$

A straightforward computation shows that $T \approx B$. As $T\mathbf{u}_i = \lambda\mathbf{u}_i$, $i = 1, ..., k$ we have that $b_{j1} = 0$ for $j > 1$. So

$$\det\,(xI - T) = \det\,(xI - B) = (x - \lambda)\det\,(xI - \tilde{B}),$$

where $\tilde{B} = [b_{ij}]_{i,j=2}^{n} \in \mathbb{D}^{(m-1)\times(m-1)}$. Here the last equality is achieved by expanding $\det\,(xI - B)$ by the first column. Use the induction hypothesis to obtain that $\tilde{B} \approx A_1$, where $A_1 \in \mathbf{UT}(m-1, \mathbb{D})$, with the eigenvalues of $\tilde{B}$ on the main diagonal of A_1 appearing in any prescribed order. Hence $T \approx C = [c_{ij}]_1^n$, where $C \in \mathbf{UT}(m, \mathbb{D})$ with $c_{11} = \lambda$, $[c_{ij}]_{i,j=2}^{n} = A_1$. □

The upper triangular form of A is not unique unless A is a *scalar matrix*: $A = aI$. See Problem 1.

Definition 2.5.5 *Let $T \in \mathbb{D}^{m\times m}$ and assume that* (2.5.3) *holds. Then the eigenvalue multiset of T is the set* $\mathrm{S(T)} = \{\lambda_1, ..., \lambda_\mathrm{m}\}$. *The multiplicity of $\lambda \in \mathrm{S(T)}$, denoted by $m(\lambda)$, is the number of elements in* $\mathrm{S(T)}$ *which are equal to λ. λ is called a simple eigenvalue if $m(\lambda) = 1$. The spectrum of T, denoted by* spec (T), *is the set of all distinct eigenvalues of T:*

$$\sum_{\lambda\in\mathrm{spec\ (T)}} m(\lambda) = m. \tag{2.5.6}$$

For $T \in \mathbb{C}^{m\times m}$ arrange the eigenvalues of T in the decreasing order of their absolute values (*unless otherwise stated*)*:*

$$|\lambda_1| \geq \cdots \geq |\lambda_m| \geq 0, \tag{2.5.7}$$

The spectral radius of T, denoted by $\rho(T)$, is equal to $|\lambda_1|$.

Problems

1. Let Q correspond to the elementary row operation described in Definition 1.11.6(iii). Assume that $A \in \mathbf{UT}(m, \mathbb{D})$. Show that if $j < i$ then $QAQ^{-1} \in \mathbf{UT}(m, \mathbb{D})$ with the same diagonal as A. More general, for any $Q \in \mathbf{UTG}_m(\mathbb{D})$ $QAQ^{-1} \in \mathbf{UT}(m, \mathbb{D})$ with the same diagonal as A.

2. Show that if $T \in \mathbb{D}^{m\times m}$ is similar to $A \in \mathbf{UT}(m, \mathbb{D})$ then the characteristic polynomial of T splits to linear factors over $\mathbb{D}[x]$.

3. Let $T \in \mathbb{D}^{m\times m}$ and put

$$\det\,(xI - T) = x^m + \sum_{j=1} a_{m-j}x^j. \tag{2.5.8}$$

Assume that the assumptions of Theorem 2.5.4 holds. Show that (2.5.9)

$$(-1)^k a_k = \sum_{\alpha\in[m]_k} \det T[\alpha,\alpha] = s_k(\lambda_1, ..., \lambda_m), \quad k = 1, ..., m.$$

Here $s_k(x_1, ..., x_m)$ is the $k-th$ elementary symmetric polynomial of $x_1, ..., x_m$. The coefficient $-a_1$ is called the *trace* of A:

$$\mathrm{tr}\, A = \sum_{i=1}^{m} a_{ii} = \sum_{i=1}^{m} \lambda_i. \tag{2.5.10}$$

4. Let $T \in \mathbb{D}^{m\times m}$ and assume the assumptions of Theorem 2.5.4. Suppose furthermore that $\mathbb{D}$ is $\mathbb{D}_U$. Using the results of Theorem 2.5.4 and Problem 2.4.5 show that the minimal polynomial $\psi(x)$ of T is of the form

$$\psi(x) = \prod_{i=1}^{l}(x-\alpha_i)^{s_i},$$
$$\alpha_i \neq \alpha_j \text{ for } i \neq j, \quad 1 \le s_i \le m_i := m(\alpha_i), \quad i = 1, ..., l, \tag{2.5.11}$$

where spec (T) $= \{\alpha_1, ..., \alpha_\mathrm{l}\}$. (*Hint*: Consider the diagonal elements of $\psi(A)$.)

5. Let $T \in \mathbb{D}_U^{m\times m}$ and assume that the minimal polynomial of T is given by (2.5.11). Using Problem 2.4.4 and the equality (2.4.1) show

$$\det\,(xI - T) = \prod_{i=1}^{l}(x-\alpha_i)^{m_i}. \tag{2.5.12}$$

2.6 Jordan canonical form

Theorem 2.5.4 and Problem 2.5.2 shows that $T \in \mathbb{D}^{m\times m}$ is similar to an upper triangular matrix if and only if the characteristic polynomial of T

splits to linear factors. Unfortunately, the upper triangular form of T is not unique. If $\mathbb{D}$ is a field then there is a special upper triangular form in the similarity class of T which is essentially unique. For convenience we state the theorem for an operator $T \in \mathrm{Hom}\,(\mathbf{V})$.

Theorem 2.6.1 *Let* $\mathbf{V}$ *be a vector space over the field* $\mathbb{F}$. *Let* $T \in \mathrm{Hom}\,(\mathbf{V})$ *and assume that the minimal polynomial* $\psi(x)$ *of* T *splits to a product of linear factors as given by* (2.5.11). *Then* $\mathbf{V}$ *splits to a direct sum of nontrivial irreducible invariant subspaces of* T

$$\mathbf{V} = \mathbf{W}_1 \oplus ... \oplus \mathbf{W}_q. \tag{2.6.1}$$

In each invariant subspace $\mathbf{W}(= \mathbf{W}_j)$ *it is possible to choose a basis consisting of generalized eigenvectors* $\mathbf{x}_1, ..., \mathbf{x}_r$ *such that*

$$\begin{aligned} &T\mathbf{x}_1 = \lambda_0\mathbf{x}_1, \\ &T\mathbf{x}_{k+1} = \lambda_0\mathbf{x}_{k+1} + \mathbf{x}_k, \quad k = 1, ..., r-1, \end{aligned} \tag{2.6.2}$$

where λ_0 *is equal to some* α_i *and* $r \le s_i$. *(For* $r = 1$ *the second part of* (2.6.2) *is void.) Moreover for each* α_i *there exists an invariant subspace* $\mathbf{W}$ *whose basis satisfies* (2.6.2) *with* $\lambda_0 = \alpha_i$ *and* $r = s_i$.

Proof. Assume first that the minimal polynomial of T is

$$\psi(x) = x^s. \tag{2.6.3}$$

Recall that $\psi(x)$ is the last invariant polynomial of T. Hence each nontrivial invariant polynomial of T is of the form x^r for $1 \le r \le s$. Theorem 2.3.4 implies that $\mathbf{V}$ has a basis in which T is presented by its rational canonical form

$$C(x^{r_1}) \oplus ... \oplus C(x^{r_k}), \quad 1 \le r_1 \le r_2 \le ... \le r_k = s.$$

Hence $\mathbf{V}$ splits to a direct sum of T-invariant subspaces (2.6.1). Let $\mathbf{W}$ be an invariant subspace in the decomposition (2.6.1). Then $\tilde{T} := T|\mathbf{W}$ has the minimal polynomial x^r, $1 \le r \le s$. Furthermore, $\mathbf{W}$ has a basis $\mathbf{x}_1, ..., \mathbf{x}_r$ so that $\tilde{T}$ is represented in this basis by the companion matrix $C(x^r)$. It is straightforward to show that $x_1, ..., x_r$ satisfies (2.6.2) with $\lambda_0 = 0$. As $\mathbf{W}$ is spanned by $\mathbf{x}_r, \tilde{T}\mathbf{x}_r, ..., \tilde{T}^{r-1}\mathbf{x}_r$ it follows that $\mathbf{W}$ is an irreducible invariant subspace of T. Assume now that the minimal polynomial of T is $(x - \lambda_0)^s$. Let $T_0 = T - \lambda_0 I$. Clearly x^s is the minimal polynomial of T_0. Let (2.6.1) be the decomposition of $\mathbf{V}$ to invariant subspaces of T_0 as above. In each invariant subspace $\mathbf{W}$ choose a basis for T_0 as above. Then our theorem holds in this case too.

Assume now that the minimal polynomial of T is given by (2.5.11). Use Theorem 2.4.5 and the above arguments to deduce the theorem. □

Let

$$H_n := C(x^n) = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{bmatrix}. \tag{2.6.4}$$

Sometimes we denote H_n by H when the dimension of H is well defined.

Let $\mathbf{W} = \text{span}\,(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_\mathrm{r})$. Let $T \in \text{Hom}\,(\mathbf{W})$ be given by (2.6.2). Then T is presented in the basis $\mathbf{x}_1, ..., \mathbf{x}_r$ by the *Jordan block* $\lambda_0 I_r + H_r$. Theorem 2.6.1 yields:

Theorem 2.6.2 *Let $A \in \mathbb{F}^{n\times n}$. Assume that the minimal polynomial $\psi(x)$ of A splits to linear factors as in* (2.5.11)*. Then there exists $P \in \mathbf{GL}(n, \mathbb{F})$ such that*

$$P^{-1}AP = J,$$

$$J = \oplus_{i=1}^{l} \oplus_{j=1}^{q_i} (\alpha_i I_{m_{ij}} + H_{m_{ij}}), \tag{2.6.5}$$

$$1 \le m_{iq_i} \le m_{iq_{i-1}} \le ... \le m_{i1} = s_i, \quad i = 1, ..., l. \tag{2.6.6}$$

Definition 2.6.3 *Let $A \in \mathbb{F}^{n\times n}$ satisfy the assumptions of Theorem 2.6.2. The matrix J in* (2.6.5) *is called the Jordan canonical form of A. Let $T \in$* Hom (**V**) *and assume that its minimal polynomial splits over $\mathbb{F}$. Then a representation matrix J* (2.6.5) *is called the Jordan canonical form of T.*

Remark 2.6.4 *Let $A \in \mathbb{F}^{n\times n}$ and suppose that the minimal polynomial ψ of A does not split over $\mathbb{F}$. Then there exits a finite extension $\mathbb{K}$ of $\mathbb{F}$ such that ψ splits over $\mathbb{K}$. Then (2.6.5) holds for some $P \in \mathbf{GL}(n, \mathbb{K})$. J is referred as the Jordan canonical form of A.*

Corollary 2.6.5 *Let $A \in \mathbb{F}^{n\times n}$. Assume that the minimal polynomial of A is given by* (2.5.11)*. Let J be the Jordan canonical form of A given by* (2.6.5)*. Set*

$$m_{iq_i+1} = ... = m_{in} = 0, \quad i = 1, ..., l. \tag{2.6.7}$$

Then the elementary polynomials of $xI - A$, which are the elementary divisors of $xI - A$ defined in Problem 2, are

$$\phi_{ij} = (x - \alpha_i)^{m_{ij}}, \quad j = 1, ..., n,\ i = 1, ..., l. \tag{2.6.8}$$

Hence the invariant polynomials $i_1(x), ..., i_n(x)$ of $xI - A$ are

$$i_r(x) = \prod_{i=1}^{l} (x - \alpha_i)^{m_{i(n-r+1)}}, \quad r = 1, ..., n. \tag{2.6.9}$$

The above Corollary shows that the Jordan canonical form is unique up to a permutation of Jordan blocks.

Problems

1. Show directly that to each eigenvalue λ_0 of a companion matrix $C(p) \in \mathbb{F}^{n\times n}$ corresponds one dimensional eigenvalues subspace spanned by the vector $(1, \lambda_0, \lambda_0^2, ..., \lambda_0^{n-1})^\top$.

2. Let $A \in \mathbb{F}^{n\times n}$ and assume that the minimal polynomial of A splits in $\mathbb{F}$. Let $\mathbf{U}_1, \mathbf{U}_2 \subset \mathbb{F}^n$ be the subspaces of all generalized eigenvectors of A, $A^\top$ respectively corresponding to $\lambda \in$ spec (A). Show that there exists bases $\mathbf{x}_1, ..., \mathbf{x}_m$ and $\mathbf{y}_1, ..., \mathbf{y}_m$ in $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively so that
$$\mathbf{y}_i^\top \mathbf{x}_j = \delta_{ij}, \quad i, j = 1, ..., m.$$
(*Hint*: Assume first that A is in its Jordan canonical form.)

3. Let $A \in \mathbb{F}^{n\times n}$. Let $\lambda, \mu \in \mathbb{F}$ be two distinct eigenvalues of A. Let $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ be two generalized eigenvectors of $A, A^\top$ corresponding to λ, μ respectively. Show that $\mathbf{y}^\top \mathbf{x} = 0$.

4. Verify directly that J (given in (2.6.5)) annihilates its characteristic polynomial. Using the fact that any $A \in \mathbb{F}^{n\times n}$ is similar to its Jordan canonical form over the finite extension field $\mathbb{K}$ of $\mathbb{F}$ deduce the Cayley-Hamilton theorem.

5. Let $A, B \in \mathbb{F}^{n\times n}$. Show that $A \approx B$ if and only if A and B have the same Jordan canonical form.

2.7 Some applications of Jordan canonical form

Definition 2.7.1 *Let $A \in \mathbb{F}^{n\times n}$ and assume that* det $(xI - A)$ *splits in $\mathbb{F}$. Let λ_0 be an eigenvalue of A. Then the number of factors of the form $x - \lambda_0$ appearing in the minimal polynomial $\psi(x)$ of A is called the index of λ_0 and is denoted by* index λ_0. *The dimension of the eigenvalue subspace of A corresponding to λ_0 is called the geometric multiplicity of λ_0.*

Using the results of the previous section we obtain.

Lemma 2.7.2 *Let the assumptions of Definition* 2.7.1 *hold. Then* index λ_0 *is the size of the largest Jordan block corresponding to* λ_0*, and the geometric multiplicity of* λ_0 *is the number of the Jordan blocks corresponding to* λ_0*.*

Let $T \in \mathrm{Hom}\,(\mathbf{V}),\ \lambda_0 \in \mathrm{spec}\ (\mathrm{T})$ and consider the invariant subspaces

$$\begin{aligned} &\mathbf{X}_r = \{\mathbf{x} \in \mathbf{V} : \quad (\lambda_0 I - T)^r \mathbf{x} = \mathbf{0}\}, \quad r = 0, 1, \ldots, \\ &\mathbf{Y}_r = (\lambda_0 I - T)^r \mathbf{V}, \quad r = 0, 1, \ldots. \end{aligned} \tag{2.7.1}$$

Theorem 2.7.3 *Let* $T \in \mathrm{Hom}\,(\mathbf{V})$ *and assume that* λ_0 *is the eigenvalue of* T*. Let* index $\lambda_0 = m_1 \geq m_2 \geq ... \geq m_p \geq 1$ *be the dimensions of all Jordan blocks corresponding to* λ_0 *which appear in the Jordan canonical form of* T*. Then*

$$\begin{aligned} &\dim \mathbf{X}_r = \sum_{i=1}^{p} \min(r, m_i), \quad r = 0, 1, ..., \\ &\dim \mathbf{Y}_r = \dim \mathbf{V} - \dim \mathbf{X}_r, \quad r = 0, 1, ... \end{aligned} \tag{2.7.2}$$

In particular

$$\begin{aligned} &[\mathbf{0}] = \mathbf{X}_0 \subsetneqq \mathbf{X}_1 \subsetneqq \mathbf{X}_2 \subsetneqq ... \subsetneqq \mathbf{X}_m, \\ &\mathbf{X}(\lambda_0) := \mathbf{X}_m = \mathbf{X}_{m+1} = ..., \quad m = \mathrm{index}\ \lambda_0. \\ &\mathbf{V} = \mathbf{Y}_0 \supsetneqq \mathbf{Y}_1 \supsetneqq \mathbf{Y}_2 \supsetneqq ... \supsetneqq \mathbf{Y}_m, \\ &\mathbf{Y}(\lambda_0) := \mathbf{Y}_m = \mathbf{Y}_{m+1} = ... \\ &\mathbf{V} = \mathbf{X}(\lambda_0) \oplus \mathbf{Y}(\lambda_0). \end{aligned} \tag{2.7.3}$$

Let

$$\nu_i = \dim \mathbf{X}_i - \dim \mathbf{X}_{i-1}, \quad i = 1, ..., m+1, \quad m := \mathrm{index}\ \lambda_0. \tag{2.7.4}$$

Then ν_i *is the number of Jordan block of size* i *at least corresponding to* λ_0*. In particular*

$$\nu_1 \geq \nu_2 \geq ... \geq \nu_m > \nu_{m+1} = 0. \tag{2.7.5}$$

Furthermore

$$\begin{aligned} &\nu_i - \nu_{i+1} \quad \text{is the number of Jordan blocks of order} \\ &i \text{ in the Jordan canonical form of } T \text{corresponding to } \lambda_0. \end{aligned} \tag{2.7.6}$$

Proof. Assume first that det $(xI - T) = \psi(x) = (x - \lambda_0)^m$. That is T has one Jordan block of order m corresponding to λ_0. Then the theorem follows straightforward. Observe next that for

$$\mathrm{Ker}\ (\lambda I - T) = 0, \quad \mathrm{Range\ Ker}\ (\lambda I - T) = \mathbf{V}, \quad \lambda \neq \lambda_0.$$

Assume now that det $(xI - T)$ splits in $\mathbb{F}$ and $\mathbf{V}$ has the decomposition (2.6.1). Apply the above arguments to each $T|\mathbf{W}_i$ for $i = 1, ..., q$ to deduce the theorem in this case. In the general case, where det $(xI - T)$ does not split to linear factors, use the rational canonical form of T to deduce the theorem. □

Thus (2.7.3) gives yet another characterization of the index λ_0. Note that in view of Definition 2.5.1 each $0 \neq \mathbf{x} \in \mathbf{X}_k$ is a generalized eigenvector of T. The sequence (2.7.4) is called the Weyr sequence corresponding to λ_0.

Definition 2.7.4 *A transformation $T \in$ Hom $(\mathbf{V})$ is diagonable if there exists a basis in $\mathbf{V}$ which consists entirely of eigenvectors of T. That is any representation matrix A of T is diagonable, i.e. A is similar to a diagonal matrix.*

For such T we have that $\mathbf{X}_1 = \mathbf{X}_{m_1}$ for each $\lambda_0 \in$ spec (T). Theorem 2.6.1 yields.

Theorem 2.7.5 *Let $T \in$ Hom $(\mathbf{V})$. Then T is diagonable if and only if the minimal polynomial ψ of T splits to linear, pairwise different factors. That is the index of any eigenvalue of T equals to* 1.

Definition 2.7.6 *Let $\mathbf{M}$ be a $\mathbb{D}$-module and let $T \in$ Hom $(\mathbf{M})$. T is nilpotent if $T^s = 0$ for some positive integer s.*

Let $T \in$ Hom $(\mathbf{V})$ and assume that det $(xI - T)$ splits in $\mathbb{F}$. For $\lambda_0 \in$ spec (T) let $\mathbf{X}(\lambda_0) \subset \mathbf{V}$ be the T-invariant subspace defined in (2.7.3). Then the decomposition (2.6.1) yields the spectral decomposition of $\mathbf{V}$:

$$\mathbf{V} = \oplus_{\lambda \in \text{spec (T)}} \mathbf{X}(\lambda). \tag{2.7.7}$$

The above decomposition is courser then the fine decomposition (2.6.1). The advantage of the spectral decomposition is that it is uniquely defined. Note that each $\mathbf{X}(\lambda)$, $\lambda \in$ spec (T) is direct sum of irreducible T-invariant subspaces corresponding to the eigenvalue λ in the decomposition (2.6.1). Clearly $T - \lambda I|\mathbf{X}(\lambda)$ is a nilpotent operator. In the following theorem we address the problem of the choices of irreducible invariant subspaces in the decomposition (2.6.1) for a nilpotent transformation T.

Theorem 2.7.7 *Let $T \in$ Hom $(\mathbf{V})$ be nilpotent. Let* index $0 = m = m_1 \geq m_2 \geq ... \geq m_p \geq 1$ *be the dimensions of all Jordan blocks appearing in the Jordan canonical form of T. Let* (2.6.1) *be a decomposition of $\mathbf{V}$ to a direct sum of irreducible T-invariant subspaces such that*

$$\begin{aligned} &\dim \mathbf{W}_1 = m_1 \geq \dim \mathbf{W}_2 = m_2 \geq ... \geq \dim \mathbf{W}_q = m_q \geq 1, \\ &m_1 = ... = m_{i_1} > m_{i_1+1} = ... = m_{i_2} > ... > m_{i_{p-1}+1} = ... = m_{i_p} = m_q. \end{aligned} \tag{2.7.8}$$

Assume that each $\mathbf{W}_i$ *has a basis* $\mathbf{y}_{i,1},...,\mathbf{y}_{i,m_i}$ *satisfying* (2.6.2)*, with* $\lambda_0 = 0$*. Let* $\mathbf{X}_i, \mathbf{Y}_i, i = 0, ...$ *be defined as in* (2.7.1) *for* $\lambda_0 = 0$*. Then the above bases in* $\mathbf{W}_1, ..., \mathbf{W}_q$ *can be chosen recursively as follows:*
(a) $\mathbf{y}_{1,1},...,\mathbf{y}_{i_1,1}$ *is an arbitrary basis in* $\mathbf{Y}_{m-1}$*.*
(b) *Let* $1 \le k < m$*. Assume that* $\mathbf{y}_{l,j}$ *are given for all* l *such that* $m_l \ge m-k+1$ *and all* j *such that* $1 \le j \le m_l - m + k$*. Then each* $\mathbf{y}_{l,(k+1)}$ *is any element in* $T^{-1}\mathbf{y}_{l,k} \cap \mathbf{Y}_{m-k-1}$*, which is a coset of the subspace* $\mathbf{Y}_{m-k-1} \cap \mathbf{X}_1 = \operatorname{Ker} T|\mathbf{Y}_{m-k-1}$*. If* $m - k = m_t$ *for some* $1 < t \le i_p$ *then* $\mathbf{y}_{i_{t-1}+1,1},...,\mathbf{y}_{i_t,1}$ *is any set of linearly independent vectors in* $\mathbf{Y}_{m-k-1} \cap \mathbf{X}_1$*, which complements the above chosen vectors* $\mathbf{y}_{l,j}$*,* $m_l \ge m-k+1$*,* $m_l - m + k + 1 \ge j$ *to a basis in* $\mathbf{Y}_{m-k-1}$*.*

See Problem 1 for the proof of the Theorem.

Corollary 2.7.8 *Let the assumptions of Theorem 2.7.7 hold. Suppose furthermore that* $\mathbf{Z} \subset \mathbf{V}$ *is an eigenspace of* T*. Then there exists a decomposition* (2.6.1) *of* $\mathbf{V}$ *to a direct sum of irreducible* T*-invariant subspaces such that* $\mathbf{Z}$ *has a basis consisting of* $l = \dim \mathbf{Z}$ *eigenvectors of the restrictions of* $T|\mathbf{W}_{j_1}, \ldots, T|\mathbf{W}_{j_l}$ *for* $1 \le j_1 < \ldots < j_l \le q$*.*

Proof. Let $\mathbf{Z}_1 := \mathbf{Z} \cap \mathbf{Y}_{m-1} \subset \ldots \subset \mathbf{Z}_m := \mathbf{Z} \cap \mathbf{Y}_0$ and denote $l_i = \dim \mathbf{Z}_i$ for $i = 1, \ldots, m$. We then construct bases in $\mathbf{W}_1, \ldots, \mathbf{W}_q$ as in Theorem 2.7.7 in the following way. If $l_1 = \dim \mathbf{Z}_1 > 0$ we pick $\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{l_1,1}$ to be from $\mathbf{Z}_1$. In general, for each $k = 1, \ldots, m-1$, $1 < t \le i_p$ and $m_t = m - k$ such that $l_{k+1} > l_k$ we let $\mathbf{y}_{i_{t-1}+1,1},...,\mathbf{y}_{i_{t-1}+l_{k+1}-l_k,1}$ be any set of linearly independent vectors in $\mathbf{Z}_{k+1}$, which form a basis in $\mathbf{Z}_{k+1}/\mathbf{Z}_k$. □

Problems

1. Prove Theorem 2.7.7

2.8 The matrix equation $AX - XB = 0$

Let $A, B \in \mathbb{D}^{n\times n}$. A possible way to determine if A and B are similar over $\mathbf{GL}(n, \mathbb{D})$ is to consider the matrix equation

$$AX - XB = 0. \tag{2.8.1}$$

Then $A \approx B$ if and only if there exists a solution $X \in \mathbb{D}^{n\times n}$ such that $\det X$ is an invertible element in $\mathbb{D}$. For $X \in \mathbb{D}^{m\times n}$ let $\hat{X} \in \mathbb{D}^{mn}$ be the column vector composed of the n columns of X:

$$\hat{X} = (x_{11}, ..., x_{m1}, x_{12}, ..., x_{m2}, ..., x_{m(n-1)}, x_{1n}, ..., x_{mn})^\top, \tag{2.8.2}$$

where $X = [x_{ij}] \in \mathbb{D}^{m\times n}$. Then the equation (2.8.1), where $A \in \mathbb{D}^{m\times m}, B \in \mathbb{D}^{n\times n}$, has a simple form in tensor notation [MaM64]. (See also Problems 1 and 2.8.13.)

$$(I \otimes A - B^\top \otimes I)\hat{X} = 0. \tag{2.8.3}$$

Assume that $\mathbb{D}$ is a Bezout domain. Then the set of all $X \in \mathbb{D}^{n\times n}$ satisfying (2.8.1) forms a $\mathbb{D}$-module with a basis $X_1, ..., X_\nu$, (Theorem 1.12.3). So any matrix X which satisfies (2.8.1) is of the form

$$X = \sum_{i=1}^{\nu} x_i X_i, \quad x_i \in \mathbb{D}, \quad i = 1, ..., \nu.$$

It is "left" to find whether a function

$$\delta(x_1, ..., x_\nu) := \det\,(\sum_{i=1}^{\nu} x_i X_i)$$

has an invertible value. In such a generality this is a difficult problem. A more modest task is to find the value of ν and to determine if $\delta(x_1, ..., x_\nu)$ vanish identically. For that purpose it is enough to assume that $\mathbb{D}$ is actually a field $\mathbb{F}$ (for example the quotient field of $\mathbb{D}$). Also we may replace $\mathbb{F}$ by a finite extension field $\mathbb{K}$ in which the characteristic polynomial of A and B split. Finally we are going to study the equation (2.8.1) where

$$A \in \mathbb{K}^{m\times m}, \quad B \in \mathbb{K}^{n\times n}, \quad X \in \mathbb{K}^{m\times n}.$$

Let $\psi(x)$, $\phi(x)$ and J, K be the minimal polynomials and the Jordan canonical forms of A, B respectively.

$$\begin{aligned}
&\psi(x) = \prod_{i=1}^{l} (x - \lambda_i)^{s_i}, \quad \text{spec (A)} = \{\lambda_1, ..., \lambda_l\},\\
&\phi(x) = \prod_{j=1}^{k} (x - \mu_j)^{t_j}, \quad \text{spec (B)} = \{\mu_1, ..., \mu_k\},\\
&P^{-1}AP = J = \oplus_{i=1}^{l} J_i,\\
&J_i = \oplus_{r=1}^{q_i} (\lambda_i I_{m_{ir}} + H_{m_{ir}}), \quad 1 \le m_{iq_i} \le ... \le m_{i1} = s_i, \quad i = 1, ..., l,\\
&Q^{-1}BQ = K = \oplus_{j=1}^{k} K_j,\\
&K_j = \oplus_{r=1}^{p_j} (\mu_j I_{n_{jr}} + H_{n_{jr}}), \quad 1 \le n_{jp_j} \le ... \le n_{j1} = t_j, \quad j = 1, ..., k.
\end{aligned} \tag{2.8.4}$$

Let $Y = P^{-1}XQ$. Then the system (2.8.1) is equivalent to $JY - YK = 0$. Partition Y according to the partitions of J and K as given in (2.8.4). So

$$\begin{aligned}
&Y = (Y_{ij}), \quad Y_{ij} \in \mathbb{K}^{m_i \times n_j},\\
&m_i = \sum_{r=1}^{q_i} m_{ir},\ n_j = \sum_{r=1}^{p_j} n_{jr}, \quad i = 1, ..., l,\ j = 1, ..., k.
\end{aligned}$$

Then the matrix equation for Y reduces to lk matrix equations

$$J_i Y_{ij} - Y_{ij} K_j = 0, \quad i = 1, ..., l,\ j = 1, ..., k. \tag{2.8.5}$$

The following two lemmas analyze the above matrix equations.

Lemma 2.8.1 *Let $i \in [l]$, $j \in [k]$. If $\lambda_i \neq \mu_j$ then the corresponding matrix equation in* (2.8.5) *has the unique trivial solution $Y_{ij} = 0$.*

Proof. Let

$$\begin{aligned}
J_i = \lambda_i I_{m_i} + \bar{J}_i, \quad \bar{J}_i = \oplus_{r=1}^{q_i} H_{m_{ir}},\\
K_j = \mu_j I_{n_j} + \bar{K}_j, \quad \bar{K}_j = \oplus_{r=1}^{p_j} H_{n_{jr}}.
\end{aligned}$$

Note that $\bar{J}^u = \bar{K}^v = 0$ for $u \geq m_i$ and $v \geq n_j$. Then (2.8.5) becomes

$$(\lambda_i - \mu_j) Y_{ij} = -\bar{J} Y_{ij} + Y_{ij} \bar{K}.$$

Thus

$$\begin{aligned}
&(\lambda_i - \mu_j)^2 Y_{ij} = -\bar{J}_i (\lambda_i - \mu_j) Y_{ij} + (\lambda_i - \mu_j) Y_{ij} \bar{K}_j =\\
&-\bar{J}_i(-\bar{J}_i Y_{ij} + Y_{ij} \bar{K}_j) + (-\bar{J}_i Y_{ij} + Y_{ij} \bar{K}_j) \bar{K}_j =\\
&(-\bar{J}_i)^2 Y_{ij} + 2(-\bar{J}_i) Y_{ij} \bar{K}_j + Y_{ij} \bar{K}_j^2.
\end{aligned}$$

Continuing this procedure we get

$$(\lambda_i - \mu_j)^r Y_{ij} = \sum_{u=0}^{r} \binom{r}{u} (-\bar{J}_i)^u Y_{ij} \bar{K}_j^{r-u}.$$

Hence for $r = m_i + n_j$ either $\bar{J}_i^u$ or $\bar{K}_j^{r-u}$ is a zero matrix. Since $\lambda_i \neq \mu_j$ we deduce that $Y_{ij} = 0$. □

Lemma 2.8.2 *Let $Z = [z_{\alpha\beta}] \in \mathbb{F}^{m \times n}$ satisfy the equation*

$$H_m Z = Z H_n. \tag{2.8.6}$$

Then the entries of Z *are of the form*

$$\begin{aligned} z_{\alpha\beta} &= 0 && \text{for } \beta < \alpha + n - \min(m,n), \\ z_{\alpha\beta} &= z_{(\alpha+1)(\beta+1)} && \text{for } \beta \geq \alpha + n - \min(m,n). \end{aligned} \tag{2.8.7}$$

In particular, the subspace of all $m \times n$ *matrices* Z *satisfying* (2.8.6) *has dimension* $\min(m,n)$.

Proof. Note that the first column and the last row of H_l are equal to zero. Hence the first column and the last row of $ZH_n = H_m Z$ are equal to zero. That is

$$z_{\alpha 1} = z_{m\beta} = 0, \quad \alpha = 2, ..., m, \ \beta = 1, ..., n-1.$$

In all other cases, equating the (α, β) entries of $H_m Z$ and ZH_n we obtain

$$z_{(\alpha+1)\beta} = z_{\alpha(\beta-1)}, \quad \alpha = 1, ..., m-1, \ \beta = 2, ..., n.$$

The above two sets of equalities yield (2.8.7). □

Combine the above two lemmas to obtain.

Theorem 2.8.3 *Consider the system of* (2.8.5). *If* $\lambda_i \neq \mu_j$ *then* $Y_{ij} = 0$. *Assume that* $\lambda_i = \mu_j$. *Partition* Y_{ij} *according to the partitions of* J_i *and* K_j *as given in* (2.8.4):

$$Y_{ij} = [Y_{ij}^{(uv)}], \ Y_{ij}^{(uv)} \in \mathbb{K}^{m_{iu} \times n_{jv}}, \ u = 1, ..., q_i, \ v = 1, ..., p_j.$$

Then each $Y_{ij}^{(uv)}$ *is of the form given in Lemma 2.8.2 with* $m = m_{iu}$ *and* $n = n_{jv}$. *Assume that*

$$\begin{aligned} &\lambda_i = \mu_i, \quad i = 1, ..., t, \\ &\lambda_i \neq \mu_j, \quad i = t+1, ..., l, \ j = t+1, ..., k. \end{aligned} \tag{2.8.8}$$

Then the dimension of the subspace $\mathcal{Y} \subset \mathbb{K}^{m \times n}$ *of block matrices* $Y = [Y_{ij}]_{i,j=1}^{lk}$ *satisfying* (2.8.5) *is given by the formula*

$$\dim \mathcal{Y} = \sum_{i=1}^{t} \sum_{u,v=1}^{q_i, p_i} \min(m_{iu}, n_{iv}). \tag{2.8.9}$$

Consider a special case of (2.8.1)

$$C(A) = \{X \in \mathbb{D}^{n\times n} : \quad AX - XA = 0\}. \tag{2.8.10}$$

Then $C(A)$ is an *algebra* over $\mathbb{D}$ with the identity I. In case $\mathbb{D}$ is a field $\mathbb{F}$, or more generally $\mathbb{D}$ is a Bezout domain, $C(A)$ has a finite basis. Theorem 2.8.3 yields

$$\dim C(A) = \sum_{i=1}^{l} \sum_{u,v=1}^{q_i} \min(m_{iu}, m_{iv}).$$

(Note that the dimension of $C(A)$ does not change if we pass from $\mathbb{F}$ to an finite extension field $\mathbb{K}$ in which the characteristic polynomial of A splits.) As $\{m_{iu}\}_{u=1}^{p_i}$ is a decreasing sequence we have

$$\sum_{v=1}^{q_i} \min(m_{iu}, m_{iv}) = um_{iu} + \sum_{v=u+1}^{q_i} m_{iv}.$$

So

$$\dim C(A) = \sum_{i=1}^{l} \sum_{u=1}^{q_i} (2u-1)m_{iu}. \tag{2.8.11}$$

Let $i_1(x), ..., i_n(x)$ be the invariant polynomials of $xI - A$. Use (2.6.7-2.6.9) to deduce

$$\dim C(A) = \sum_{u=1}^{n} (2u-1)\mathrm{deg}\ i_{n-u+1}(x). \tag{2.8.12}$$

The above formula enables us to determine when any commuting matrix with A is a polynomial in A. Clearly, the dimension of the subspace spanned by the powers of A is equal to the degree of the minimal polynomial of A.

Corollary 2.8.4 *Let $A \in \mathbb{F}^{n\times n}$. Then each commuting matrix with A can be expressed as a polynomial in A if and only if the minimal and the characteristic polynomial of A are equal. That is, A is similar to a companion matrix $C(p)$, where $p(x) = \det\,(xI - A)$.*

A matrix for which the minimal and characteristic polynomial coincide is called *nonderogatory.* If the minimal polynomial of A is a strict factor of the characteristic polynomial of A, i.e. the degree of the minimal polynomial is strictly less than the degree of the characteristic polynomial, then A is called *derogatory.*

Problems

1. For $A = [a_{ij}] \in \mathbb{D}^{m\times p},\ B = (b_{kl}) \in \mathbb{D}^{n\times q}$ let

$$A \otimes B := [a_{ij}B] \in \mathbb{D}^{mn\times pq}, \tag{2.8.13}$$

   be the tensor (Kronecker) product of A and B. Show

$$(A_1 \otimes A_2)(B_1 \otimes B_2) = (A_1B_1)\otimes(A_2B_2),\ A_i \in \mathbb{D}^{m_i\times n_i},\ B_i \in \mathbb{D}^{n_i\times p_i},\ i = 1,2.$$

2. Let $\mu : \mathbb{D}^{m\times n} \to \mathbb{D}^{mn}$ be given by $\mu(X) = \hat{X}$, where $\hat{X}$ is defined be (2.8.2). Show that

$$\mu(AX) = (I_n \otimes A)\mu(X), \quad \mu(XB) = (B^\top \otimes I_m)\mu(X), \quad A \in \mathbb{D}^{m\times m},\ B \in \mathbb{D}^{n\times n}.$$

3. Let $P \in \mathbb{F}^{m\times m},\ Q \in \mathbb{F}^{n\times n},\ R \in \mathbb{F}^{m\times n}$. Let

$$A = \begin{bmatrix} P & R \\ 0 & Q \end{bmatrix},\ B = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \in \mathbb{F}^{(m+n)\times(m+n)}.$$

   Assume that the characteristic polynomials of P and Q are coprime. Show that there exists $X = \begin{bmatrix} I_m & Y \\ 0 & I_m \end{bmatrix}$ which satisfies (2.8.1). Hence $A \approx B$.

4. Let $A = \sum_{i=1}^{\ell} \oplus A_i \in \mathbb{F}^{n\times n}$. Show that

$$\dim C(A) \geq \sum_{i=1}^{\ell} \dim C(A_i), \tag{2.8.14}$$

   and the equality holds if and only if

$$(\det(xI - A_i), \det(xI - A_j)) = 1 \quad \text{for } i = 1,...,\ell,\ j = 1,...,\ell - 1.$$

5. Let $A \in \mathbb{D}^{n\times n}$. Show that the ring $C(A)$ is a commutative ring if and only if A satisfies the conditions of Corollary 2.8.4, where $\mathbb{F}$ is the quotient field of $\mathbb{D}$.

6. Let $A \in \mathbb{D}^{n\times n},\ B \in C(A)$. Then B is an invertible element in the ring $C(A)$ if and only if B is a unimodular matrix.

7. Let $A \in \mathbb{D}^{m\times m},\ B \in \mathbb{D}^{n\times n}$. Define

$$C(A, B) := \{X \in \mathbb{D}^{m\times n} : \quad AX - XB = 0\}. \tag{2.8.15}$$

   Show that $C(A, B)$ is a left (right) module of $C(A)$ $(C(B))$ under the matrix multiplication.

8. Let $A, B \in \mathbb{D}^{n\times n}$. Show that $A \approx B$ if and only if the following two conditions hold:

   (a) $C(A,B)$ is a $C(A)$-module with a basis consisting of one element U;

   (b) any basis element U is a unimodular matrix.

2.9 A criterion for similarity of two matrices

Definition 2.9.1 *Let $A \in \mathbb{D}^{m\times m}$, $B \in \mathbb{D}^{n\times n}$. Denote by $r(A,B)$ and $\nu(A,B)$ the rank and the nullity of the matrix $I_n \otimes A - B^\top \otimes I_m$ viewed as a matrix acting on the vector space $\mathbb{F}^{m\times n}$, where $\mathbb{F}$ is the quotient field of $\mathbb{D}$.*

According to Theorem 2.8.3 we have

$$\nu(A,B) = \sum_{i=1}^{t} \sum_{u,v=1}^{q_i,p_i} \min(m_{iu}, n_{iv}),$$

(2.9.1)

$$r(A,B) = mn - \sum_{i=1}^{t} \sum_{u,v=1}^{q_i,p_i} \min(m_{iu}, n_{iv}).$$

Theorem 2.9.2 *Let $A \in \mathbb{D}^{m\times m}$, $B \in \mathbb{D}^{n\times n}$. Then*

$$\nu(A,B) \le \frac{1}{2}(\nu(A,A) + \nu(B,B)).$$

Equality holds if and only if $m = n$ and A and B are similar over the quotient field $\mathbb{F}$.

Proof. Without loss of generality we may assume that $\mathbb{D} = \mathbb{F}$ and the characteristic polynomials of A and B split over $\mathbb{F}[x]$. For $x, y \in \mathbb{R}$ let $\min(x,y)$ ($\max(x,y)$) be the minimum (maximum) of the values of x and y. Clearly $\min(x,y)$ is a homogeneous concave function on $\mathbb{R}^2$. Hence

$$\min(a+b, c+d) \ge \frac{\min(a,c) + \min(b,d) + \min(a,d) + \min(b,c)}{2}. \tag{2.9.2}$$

A straightforward calculation shows that if $a = c$ and $b = d$ then equality holds if and only if $a = b$. Let

$$N = \max(m,n), \quad m_{iu} = n_{jv} = 0,$$
$$\text{for } q_i < u \le N,\ p_i < v \le N,\ i = 1, ..., \ell,\ j = 1, ..., k.$$

Then

$$\nu(A,A)+\nu(B,B)=\sum_{i,u=1}^{\ell,N}(2u-1)m_{iu}+\sum_{j,u=1}^{k,N}(2u-1)n_{ju}\geq$$
$$\sum_{i,u=1}^{t,N}(2u-1)(m_{iu}+n_{iu}).$$

and the equality holds if and $\ell=k=t$. Next consider the inequality

$$\sum_{i,u=1}^{t,N}(2u-1)(m_{iu}+n_{iu})=\sum_{i=1}^{t}\sum_{u,v=1}^{N}\min(m_{iu}+n_{iu},m_{iv}+n_{iv})\geq$$
$$\frac{1}{2}\sum_{i=1}^{t}\sum_{u,v=1}^{N}(\min(m_{iu},m_{iv})+\min(m_{iu},n_{iv})+$$
$$\min(n_{iu},m_{iv})+\min(n_{iu},n_{iv}))=$$
$$\frac{1}{2}\sum_{i,u=1}^{t,N}(2u-1)(m_{iu}+n_{iu})+\sum_{i}^{t}\sum_{u,v=1}^{q_i,p_i}\min(m_{iu},n_{iv}).$$

Combine the above results to obtain the inequality (2.9.2). Equality sign holds in (2.9.2) if and only if A and B have the same Jordan canonical forms. That is $m=n$ and A is similar to B over $\mathbb{F}$. □

Suppose that $A\approx B$. Hence (2.2.1) holds. The rules for the tensor product (Problem 2.8.1) imply

$$I\otimes A-B^\top\otimes I=((Q^\top)^{-1}\otimes I)(I\otimes A-A^\top\otimes I)(Q^\top\otimes I),$$
(2.9.3)
$$I\otimes A-B^\top\otimes I=((Q^\top)^{-1}\otimes Q)(I\otimes A-A^\top\otimes I)(Q^\top\otimes Q^{-1}).$$

Hence the three matrices

(2.9.4) $$I\otimes A-A^\top\otimes I,\quad I\otimes A-B^\top\otimes I,\quad I\otimes B-B^\top\otimes I$$

are similar. In particular, these matrices are equivalent. Over a field $\mathbb{F}$ the above matrices are equivalent if and only if the have the same nullity. Hence Theorem 2.9.2 yields.

Theorem 2.9.3 *Let $A,B\in\mathbb{F}^{n\times n}$. Then A and B are similar if and only if the three matrices in* (2.9.4) *are equivalent.*

The obvious part of Theorem 2.9.3 extends trivially to any integral domain $\mathbb{D}$.

Proposition 2.9.4 *Let $A, B \in \mathbb{D}^{n\times n}$. If A and B are similar over $\mathbb{D}$ then the three matrices in* (2.9.4) *are equivalent over $\mathbb{D}$.*

However, this condition is not sufficient for the similarity of A and B even in the case $\mathbb{D} = \mathbb{Z}$. (See Problem 1.) The disadvantage of the similarity criterion stated in Theorem 2.9.3 is due to the appearance of the matrix $I \otimes A - B^\top \otimes I$, which depends on A and B. It is interesting to note that the equivalence of just two matrices in (2.9.4) does not imply the similarity of A and B. Indeed

$$I \otimes A - A^\top \otimes I = I \otimes (A + \lambda I) - (A + \lambda I)^\top \otimes I$$

for any $\lambda \in \mathbb{F}$. If $\mathbb{F}$ has an infinite characteristic then $A \not\approx A + \lambda I$ for any $\lambda \neq 0$. (Problem 2.) Also if $A = H_n$ and $B = 0$ then $\nu(A, A) = \nu(A, B) = n$. (Problem 3.) however, under certain assumptions the equality $\nu(A, A) = \nu(A, B)$ implies $A \approx B$.

Theorem 2.9.5 *Let $A \in \mathbb{C}^{n\times n}$. Then there exists a neighborhood of $A = [a_{ij}]$*

$$D(A, \rho) := \{B = [b_{ij}] \in \mathbb{C}^{n\times n} : \quad \sum_{i,j=1}^{n} |b_{ij} - a_{ij}|^2 < \rho^2\}, \tag{2.9.5}$$

for some positive ρ depending on A, such that if

$$\nu(A, A) = \nu(A, B), \quad B \in D(A, \rho), \tag{2.9.6}$$

then B is similar to A.

Proof. Let r be the rank of $I \otimes A - A^\top \otimes I$. So there exist indices

$$\alpha = \{(\alpha_{11}, \alpha_{21}), ..., (\alpha_{1r}, \alpha_{2r})\},\ \beta = \{(\beta_{11}, \beta_{21}), ..., (\beta_{1r}, \beta_{2r})\} \subset [n] \times [n],$$

viewed as elements of $[n^2]_r$, such that $\det\,(I \otimes A - A^\top \otimes I)[\alpha, \beta] \neq 0$. Also $\det\,(I \otimes A - A^\top \otimes I)[\gamma, \delta] = 0$ for any $\gamma, \delta \in [n^2]_{r+1}$. First choose a positive ρ' such that

$$\det\,(I \otimes A - B^\top \otimes I)[\alpha, \beta] \neq 0, \text{for all } B \in D(A, \rho'). \tag{2.9.7}$$

Consider the system (2.8.1) as a system in n^2 variables, which are the entries of $X = [x_{ij}]_1^n$. In the system (2.8.1) consider the subsystem of r equations corresponding to the set α:

$$\sum_{k=1}^{n} a_{ik}x_{kj} - x_{ik}b_{kj} = 0, \quad i = \alpha_{1p},\ j = \alpha_{2p},\ p = 1, ..., r. \tag{2.9.8}$$

Let

$$x_{kj} = \delta_{kj} \quad \text{for} \quad (k,j) \neq (\beta_{1p}, \beta_{2p}),\ p = 1, ..., r. \tag{2.9.9}$$

The condition (2.9.7) yields that the system (2.9.8)-(2.9.9) has a unique solution $X(B)$ for any $B \in D(A, \rho')$. Also $X(A) = I$. Use the continuity argument to deduce the the existence of $\rho \in (0, \rho']$ so that $\det X(B) \neq 0$ for all $B \in D(A, \rho)$. Let $\mathcal{V}$ be the algebraic variety of all matrices $B \in \mathbb{C}^{n\times n}$ satisfying

$$\det\, (I \otimes A - B^\top \otimes I)[\gamma, \delta] = 0 \quad \text{for any } \gamma, \delta \in Q_{(r+1),n^2}. \tag{2.9.10}$$

We claim that $\mathcal{V} \cap D(A, \rho)$ is the set of matrices of the form (2.9.6). Indeed, let $B \in \mathcal{V} \cap D(A, \rho)$. Then (2.9.7) and (2.9.10) yield that $\nu(A, B) = \nu(A, A) = n^2 - r$. Assume that B satisfies (2.9.6). Hence (2.9.10) holds and $B \in \mathcal{V} \cap D(A, \rho)$. Assume that $B \in \mathcal{V} \cap D(A, \rho)$. Then

$$AX(B) - X(B)B = 0, \quad \det X(B) \neq 0 \quad \Rightarrow \quad A \approx B.$$

□

Problems

1. Show that for A and B given in Problem 2.2.1 the three matrices in (2.9.4) are equivalent over $\mathbb{Z}$, but A and B are not similar over $\mathbb{Z}$. (See Problem 2.2.1.)

2. Show that if $\mathbb{F}$ has an infinite characteristic then for any $A \in \mathbb{F}^{n\times n}$ $A \approx A + \lambda I$ if and only if $\lambda = 0$. (Compare the traces of A and $A + \lambda I$.)

3. Show that if $A = H_n$ and $B = 0$ then $\nu(A, A) = \nu(A, B) = n$.

4. Let $A, B \in \mathbb{D}^{n\times n}$. Assume that the three matrices in (2.9.4) are equivalent. Let $\mathcal{I}$ be a maximal ideal in $\mathbb{D}$. Let $\mathbb{F} = \mathbb{D}/\mathcal{I}$ and view A, B as matrices over $\mathbb{F}$. Prove that A and B similar over $\mathbb{F}$. (Show that the matrices in (2.9.4) are equivalent over $\mathbb{F}$.)

2.10 The matrix equation $AX - XB = C$

A related equation to (2.8.1) is the nonhomogeneous equation

$$AX - XB = C, \quad A \in \mathbb{F}^{m\times m},\ B \in \mathbb{F}^{n\times n},\ C, X \in \mathbb{F}^{m\times n}. \tag{2.10.1}$$

This equation can be written in the tensor notation as

$$(I_n \otimes A - B^\top \otimes I_m)\hat{X} = \hat{C}. \tag{2.10.2}$$

The necessary and sufficient condition for the solvability of (2.10.2) can be stated in the dual form as follows. Consider the homogenous system whose coefficient matrix is the transposed coefficient matrix of (2.10.2), (see Problem 1),

$$(I_n \otimes A^\top - B \otimes I_m)\hat{Y} = 0.$$

Then (2.10.2) is solvable if and only if any solution $\hat{Y}$ of the above system is orthogonal to $\hat{C}$ (e.g. Problem 2). In matrix form the above equation is equivalent to

$$A^\top Y - YB^\top = 0, \quad Y \in \mathbb{F}^{m\times n}.$$

The orthogonality of $\hat{Y}$ and $\hat{C}$ are written as $\operatorname{tr} Y^\top C = 0$. (See Problem 3.) Thus we showed:

Theorem 2.10.1 *Let $A \in \mathbb{F}^{m\times m}$, $B \in \mathbb{F}^{n\times n}$. Then* (2.10.1) *is solvable if and only if*

$$\operatorname{tr} ZC = 0 \tag{2.10.3}$$

for all $Z \in \mathbb{F}^{m\times n}$ satisfying

$$ZA - BZ = 0. \tag{2.10.4}$$

Using the above Theorem we can obtain a stronger version of Problem 4.

Theorem 2.10.2 *Let*

$$G = [G_{ij}]_1^\ell, \quad G_{ij} \in \mathbb{F}^{n_i\times n_j},\ G_{ij} = 0\ for\ j < i,\ i,j = 1,...,\ell.$$

Then

$$\dim C(G) \ge \sum_{i=1}^{\ell} \dim C(G_{ii}). \tag{2.10.5}$$

Proof. Consider first the case $\ell = 2$. Let $G = \begin{bmatrix} A & E \\ 0 & B \end{bmatrix}$. Assume that $T = \begin{bmatrix} U & X \\ 0 & V \end{bmatrix}$ commutes with G. So

$$AU - UA = 0, \quad BV - VB = 0, \quad AX - XB = UE - EV. \tag{2.10.6}$$

Theorem 2.10.1 implies that $U \in C(A)$, $V \in C(B)$ satisfy the last equation of (2.10.6) if and only if $\operatorname{tr} Z(UE - EV) = 0$ for all Z satisfying (2.10.4). Thus the dimension of pairs (U, V) satisfying (2.10.6) is at least

$$\dim C(A) + \dim C(B) - \dim C(B, A).$$

On the other hand, for a fixed (U, V) satisfying (2.10.6), the set of all X satisfying the last equation of (2.10.6) is of the form $X_0 + C(A, B)$. The equality (2.8.9) yields $\dim C(A, B) = \dim C(B, A)$. Hence (2.10.5) holds for $\ell = 2$. The general case follows straightforward by induction on ℓ. □

We remark that contrary to the results given in Problem 2.8.4 the equality in (2.10.5) may occur even if $G_{ii} = G_{jj}$ for some $i \neq j$. (See Problem 4.)

Theorem 2.10.3 *Let $A \in \mathbb{F}^{m\times m}$, $B \in \mathbb{F}^{n\times n}$, $C \in \mathbb{F}^{m\times n}$. Let*

$$F = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \; G = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \in \mathbb{F}^{(m+n)\times(m+n)}.$$

Show that $F \approx G$ if and only if the matrix equation (2.10.1) is solvable.

Proof. Assume that (2.10.1) solvable. Then $U = \begin{bmatrix} I_m & X \\ 0 & I_n \end{bmatrix} \in \mathbf{GL}(m+n, \mathbb{F})$ and $G = U^{-1}FU$.

Assume now that $F \approx G$. We prove the solvability of (2.10.1)) by induction on $m + n$, where $m, n \geq 1$. Let $\mathbb{K}$ be a finite extension of $\mathbb{F}$ such that the characteristic polynomial of A and B split to linear factors. Clearly it is enough to prove the solvability of (2.10.1) for the field $\mathbb{K}$. Suppose first that A and B do not have common eigenvalues. Then Problem 2.8.3 yields that $F \approx G$. Assume now that A and B have a common eigenvalue λ_1. For $m = n = 1$ it means that $A = B = \lambda_1 \in \mathbb{F}$. Then the assumption that $F \approx G$ implies that $C = 0$ and (2.10.1) is solvable with $X = 0$.

Assume now that the theorem holds for all $2 \leq m + n < L$. Let $m + n = L$. The above arguments yield that it is enough to consider the case where the characteristic polynomials of A and B split to linear factors and λ_1 is a common eigenvalue of A and B. By considering the matrices $F - \lambda_1 I_{m+n}$, $G - \lambda_1 I_{m+n}$ we may assume that 0 is an eigenvalue of A and B. By considering the similar matrices $U^{-1}FU$, $U^{-1}GU$ where $U = U_1 \oplus U_2$, $U_1 \in \mathbf{GL}(m, \mathbb{F})$, $U_2 \in \mathbf{GL}(n, \mathbb{F})$ we may assume that A and B are in a Jordan canonical form of the form

$$A = A_1 \oplus A_2, \; B = B_1 \oplus B_2, \quad A_1^m = 0, \; B_1^n = 0, \; 0 \notin (\text{spec (A}_2) \cup \text{spec (B}_2)).$$

(It is possible that either $A = A_1$ or $B = B_1$.) Let

$$U = \begin{bmatrix} I_m & X \\ 0 & I_n \end{bmatrix}, \quad X = \begin{bmatrix} 0 & X_{12} \\ X_{21} & 0 \end{bmatrix}, \; C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Use Problem 2.8.3 to deduce that one can choose X_{12}, X_{21} such that $G' = U^{-1}GU = G = \begin{bmatrix} A & C' \\ 0 & B \end{bmatrix}$ and $C'_{12} = 0,\ C'_{21} = 0$. For simplicity of notation we will assume that $C_{12} = 0,\ C_{21} = 0$. Permute second and third blocks in F, G to obtain that F, G are permutationally similar to

$$\hat{F} = \begin{bmatrix} A_1 & 0 & 0 & 0 \\ 0 & B_1 & 0 & 0 \\ 0 & 0 & A_2 & 0 \\ 0 & 0 & 0 & B_2 \end{bmatrix}, \ \hat{G} = \begin{bmatrix} A_1 & C_{11} & 0 & 0 \\ 0 & B_1 & 0 & 0 \\ 0 & 0 & A_2 & C_{22} \\ 0 & 0 & 0 & B_2 \end{bmatrix},$$

respectively. So the Jordan canonical form of $\hat{F}, \hat{G}$ corresponding to 0 are determined by the Jordan canonical forms of $\begin{bmatrix} A_1 & 0 \\ 0 & B_1 \end{bmatrix}, \begin{bmatrix} A_1 & C_{11} \\ 0 & B_1 \end{bmatrix}$ respectively. The Jordan canonical form of $\hat{F}, \hat{G}$ corresponding to other eigenvalues are determined by the Jordan canonical forms of $\begin{bmatrix} A_2 & 0 \\ 0 & B_2 \end{bmatrix}, \begin{bmatrix} A_2 & C_{22} \\ 0 & B_2 \end{bmatrix}$ respectively. Hence

$$\begin{bmatrix} A_1 & 0 \\ 0 & B_1 \end{bmatrix} \approx \begin{bmatrix} A_1 & C_{11} \\ 0 & B_1 \end{bmatrix}, \quad \begin{bmatrix} A_2 & 0 \\ 0 & B_2 \end{bmatrix} \approx \begin{bmatrix} A_2 & C_{22} \\ 0 & B_2 \end{bmatrix}.$$

Thus if either A or B are not nilpotent the theorem follows by induction.

It is left to consider the case where A and B are nilpotent matrices, which are in their Jordan canonical form. If $A = 0,\ B = 0$ then $C = 0$ and the theorem follows. So we assume that either at least one of the matrices in $\{A, B\}$ is not a zero matrix. Since $\dim \ker F = \dim \ker G$ Problem 6 yields that (after the upper triangular similarity applied to G) we may assume that $\ker F = \ker G$. Let

$$A = \oplus_{i=1}^{p} A_i, \quad B = \oplus_{j=1}^{q} B_j,$$

where each A_i, B_j is an upper triangular Jordan block of dimension m_i, n_j respectively. Let

$$C = [C_{ij}], \quad C_{ij} \in \mathbb{C}^{m_i \times n_j},\ i = 1, ..., p, j = 1, ..., q,$$

be the block partition of C induced by the block partition of A and B respectively. The assumption that $\ker F = \ker G$ is equivalent to the assumption that the first column of each C_{ij} is zero. Consider $\mathbf{V} = \mathbb{F}^{m+n}/\ker F$. Then F, G induce the operators $\hat{F}, \hat{G}$ on $\mathbf{V}$ which are obtained from F, G by deleting the rows and columns corresponding to the vectors in the kernels of A and B. (These vectors are formed by some of the vectors in the canonical basis of $\mathbb{F}^{m+n}$.) Note that the Jordan canonical forms of $\hat{F}, \hat{G}$ are direct sums of reduced Jordan blocks (obtained by deleting the first row and column in each Jordan block) corresponding to $\hat{F}, \hat{G}$ respectively. As F and G have the same Jordan blocks it follows that $\hat{F}, \hat{G}$ have the same Jordan blocks, i.e. $\hat{F} \approx \hat{G}$. It is easy to see that

$$\begin{aligned}
&\hat{F} = \begin{bmatrix} \hat{A} & 0 \\ 0 & \hat{B} \end{bmatrix}, \quad \hat{G} = \begin{bmatrix} \hat{A} & \hat{C} \\ 0 & \hat{B} \end{bmatrix}, \\
&\hat{A} = \oplus_{i=1}^{p} \hat{A}_i, \quad \hat{B} = \oplus_{j=1}^{q} \hat{B}_j, \quad \hat{C} = (\hat{C}_{ij}), \\
&\hat{A}_i \in \mathbb{F}^{(m_i-1)\times(m_i-1)}, \ \hat{B}_j \in \mathbb{F}^{(n_j-1)\times(n_j-1)}, \ \hat{C}_{ij} \in \mathbb{F}^{(m_i-1)\times(n_j-1)}.
\end{aligned}$$

Here $\hat{A}_i, \hat{B}_j, \hat{C}_{ij}$ obtained from A_i, B_j, C_{ij} be deleting the first row and column respectively. Since $\hat{F} \approx \hat{G}$ we can use the induction hypothesis. That is there exists $X = (X_{ij}) \in \mathbb{F}^{m\times n}$ partitioned as C with the following properties: The first row and the column of each X_{ij} is zero. $A_i X_{ij} - X_{ij} B_j - C_{ij}$ have zero entries in the last $m_i - 1$ rows and in the first column. By considering $U^{-1}GU$ with $U = \begin{bmatrix} I_m & X \\ 0 & I_n \end{bmatrix}$ we already may assume that the last $m_i - 1$ rows and the first column of each C_{ij} are zero. Finally we observe that if A_i and B_j are Jordan blocks that the equation (2.10.1) is solvable by letting X_{ij} be a corresponding matrix with the last $m_i - 1$ rows equal to zero. □

Problems

1. Let $A \otimes B$ be defined as in (2.8.13). Prove that $(A \otimes B)^\top = A^\top \otimes B^\top$.

2. Consider the system

$$Ax = b, \quad A \in \mathbb{F}^{m\times n}, \ b \in \mathbb{F}^n.$$

Show the above system is solvable if and only any solution of $A^\top y = 0$ satisfies $y^\top b = 0$. (Change variables to bring A to its diagonal form as in §1.12.)

3. Let $X, Y \in \mathbb{D}^{m\times n}$. Let $\mu(X), \mu(Y) \in \mathbb{D}^{mn}$ be defined as in Problem 2.8.1. Show that

$$\mu(X)^\top \mu(Y) = \operatorname{tr} Y^\top X.$$

4. Assume in Theorem 2.10.2 $\ell = 2$, $G_{11} = G_{22} = 0$, $G_{12} = I$. Show that in this case the equality sign holds in (2.10.5).

5. Let $A_i \in \mathbb{F}^{n_i \times n_i},\ i = 1, 2$ and suppose that A_1 and A_2 do not have a common eigenvalue. Assume that $A = A_1 \oplus A_2$. Let
$$C = [C_{ij}]_1^2,\ X = [X_{ij}]_1^2, \quad C_{ij}, X_{ij} \in \mathbb{F}^{n_i \times n_j},\ i, j = 1, 2.$$
Using Problem 2.8.4 prove that the equation $AX - XA = C$ is solvable if and only if the equations $A_i X_{ii} - X_{ii} A_i = C_{ii},\ i = 1, 2$ are solvable.

6. Let $A \in \mathbb{F}^{m \times m},\ B \in \mathbb{F}^{n \times n}$ be two nilpotent matrix. Let $C \in \mathbb{F}^{m \times n}$ and define the matrices $F, G \in \mathbb{F}^{(m+n)\times(m+n)}$ as in Theorem 2.10.3. Show that dim $\ker F \ge$ dim $\ker G$. Equality holds if and only if $C \ker B \subset \mathrm{Range}\ A$. Equivalently, equality holds if and only if there exists $X \in \mathbb{F}^{m \times n}$ such that
$$\ker F = \ker U^{-1} G U, \quad U = \begin{bmatrix} I_m & X \\ 0 & I_n \end{bmatrix}.$$

2.11 A case of two nilpotent matrices

Theorem 2.11.1 *Let $T \in$ Hom $(\mathbf{V})$ be nilpotent. Let* index $0 = m = m_1 \ge m_2 \ge \ldots \ge m_p \ge 1$ *be the dimensions of all Jordan blocks appearing in the Jordan canonical form of T. Let $\mathbf{Z} \subset \mathbf{V}$ be an eigenspace of T corresponding to the eigenvalue 0. Denote by $\mathbf{W} = \mathbf{V}/\mathbf{Z}$. Then T induces a nilpotent operator $T' \in$ Hom $(\mathbf{W})$. The dimension of Jordan blocks of T' correspond to the positive integers in the sequence $m_1', m_2', \ldots, m_p'$, where m_i' is either m_i or $m_i - 1$. Furthermore, exactly* dim $\mathbf{Z}$ *of indices of m_i' are equal to $m_i - 1$.*

Proof. Suppose first that $p = 1$, i.e. $\mathbf{W}$ is an irreducible invariant subspace of T. Then $\mathbf{Z}$ is the eigenspace of T and the theorem is straightforward. Use Corollary 2.7.8 in the general case to deduce the theorem. □

Theorem 2.11.2 *Let $A \in \mathbb{F}^{n \times n}$ be a nilpotent matrix. Put*
$$\mathbf{X}_k = \{\mathbf{x} \in \mathbb{F}^n : \quad A^k \mathbf{x} = \mathbf{0}\}, \quad k = 0, \ldots$$
Then
$$\mathbf{X}_0 = \{\mathbf{0}\} \quad \text{and} \quad \mathbf{X}_i \subset \mathbf{X}_{i+1},\ i = 0, \ldots \tag{2.11.1}$$
Assume that
$$\mathbf{X}_i \ne \mathbf{X}_{i+1} \quad \text{for } i = 0, \ldots, p-1, \quad \text{and } \mathbf{X}_p = \mathbb{F}^n,\ p = \text{index } 0.$$

Suppose that $B \in \mathbb{F}^{n\times n}$ *satisfies*

$$B\mathbf{X}_{i+1} \subset \mathbf{X}_i, \quad \text{for } i = 1, ..., p-1. \tag{2.11.2}$$

Then B *is nilpotent and*

$$\nu(A, A) \le \nu(B, B). \tag{2.11.3}$$

Equality holds if and only if B *is similar to* A.

Proof. Clearly (2.11.2) holds for any $A \in \mathbb{F}^{n\times n}$. As $B^p\mathbb{F}^n = B^p\mathbf{X}_p \subset \mathbf{X}_0 = \{\mathbf{0}\}$, it follows that B is nilpotent. We prove the claim by induction on p. For $p = 1$ $A = B = 0$ and equality holds in (2.11.3). Suppose that the theorem holds for $p = q-1$. Let $p = q$.

Assume that the Jordan blocks of A and B are the sizes $q = m_1 \ge \ldots \ge m_j \ge 1$ and $l_1 \ge \ldots \ge l_k \ge 1$ respectively. Recall that $\mathbf{X}_1$ is the eigenspace of A corresponding to $\lambda = 0$. Hence $j = \dim \mathbf{X}_1$. Since $B\mathbf{X}_1 = \mathbf{X}_0 = \{0\}$ it follows that the dimension of the eigenspace of B is at least j. Hence $k \ge j$.

Let $\mathbf{W} := \mathbf{V}/\mathbf{X}_1$. Since $A\mathbf{X}_1 = \{0\}$ A induces a nilpotent operator $A' \in \mathrm{Hom}\,(\mathbf{W})$. Let $\mathbf{X}'_i = \ker(A')^i, i = 1, \ldots$. Then $\mathbf{X}'_i = \mathbf{X}_{i+1}/\mathbf{X}_1, i = 0, 1, \ldots$. Hence the index of $A' = q-1$. Furthermore the Jordan blocks of A' correspond to the positive numbers in the sequence $m'_1 = m_1 - 1 \ge \ldots \ge m'_j = m_j - 1$. Since $B\mathbf{X}_1 = \{\mathbf{0}\}$ it follows that B induces the operator $B' \in \mathrm{Hom}\,(\mathbf{W})$. The equality $\mathbf{X}'_i = \mathbf{X}_{i+1}/\mathbf{X}_1$ implies that $B'\mathbf{X}'_i \subset \mathbf{X}'_{i-1}$ for $i = 1, \ldots$.

Theorem 2.11.1 implies that the Jordan blocks of B' correspond to nonzero $l'_1, \ldots, l'_k$, where l'_i is either l_i or $l_i - 1$. Furthermore exactly j of l'_i are equal to $l_i - 1$. Recall (2.8.11) that

$$\nu(A, A) = \sum_{i=1}^{j}(2i-1)m_i = \sum_{i=1}^{j}(2i-1)m'_i + \sum_{i=1}^{j}(2i-1) = \nu(A', A') + j^2.$$

Assume that

$$l_1 = \ldots = l_{k_1} > l_{k_1+1} = \ldots = l_{k_2} > l_{k_2+1} = \ldots > l_{k_{r-1}+1} = \ldots = l_{k_r},$$

where $k_r = k$. let $k_0 = 0$. Suppose that in the set of $\{k_{s-1}+1, \ldots, k_s\}$ we have exactly $i \le k_s - k_{s-1}$ indices such that $l'_r = l_r - 1$ for $r \in \{k_{s-1}+1, \ldots, k_s\}$. We then assume that $l'_r = l_r - 1$ for $r = k_s, k_s - 1, \ldots, k_s - i + 1$. Hence $l'_1 \ge \ldots \ge l'_k \ge 0$. Thus $\nu(B', B') = \sum_{i=1}^{k}(2i-1)l'_i$. So

$$\nu(B, B) = \sum_{i=1}^{k}(2i-1)l_i = \nu(B', B') + \sum_{i=1}^{k}(2i-1)(l_i - l'_i) \ge \nu(B', B') + j^2.$$

equality holds if and only if $l'_i = l_i - 1$ for $i = 1, \ldots, j$. The induction hypothesis implies that $\nu(A', A') \le \nu(B', B')$ and equality holds if and only if $A' \sim B'$, i.e A' and B' have the same Jordan blocks. Hence $\nu(A, A) \le \nu(B, B)$ and equality holds if and only if $A \sim B$. □

2.12 Historical remarks

The exposition of §2.1 is close to [Gan59]. The content of §2.2 is standard. Theorem 2.3.4 is well known [Gan59]. Other results of §2.3 are not common and some of them may be new. §2.4 is standard and its exposition is close to [Gan59]. Theorem 2.5.4 is probably known for $\mathbb{D}_{ed}$ (see [Lea48] for the case $\mathbb{D} = \mathrm{H}(\Omega)$, $\Omega \subset \mathbb{C}$.) Perhaps it is new for Bezout domains. The results of §2.6 are standard. Most of §2.7 is standard. The exposition of §2.8 is close to [Gan59]. For additional properties of tensor product see [MaM64]. Problem 2.8.8 is close to the results of [Fad66]. See also [Gur80] for an arbitrary integral domain $\mathbb{D}$. Theorems 2.9.2 and 2.9.3 are taken from [Fri80b]. See [GaB77] for a weaker version of Theorem 2.9.3. Some of the results of §2.10 may be new. Theorem 2.10.1 was taken from [Fri80a]. Theorem 2.10.3 is called Roth's theorem [Rot52]. Theorem 2.11.2 is taken from [Fri80b].

Chapter 3

Functions of Matrices and Analytic Similarity

3.1 Components of a matrix and functions of matrices

In this Chapter we assume that all the matrices are complex valued ($\mathbb{F} = \mathbb{C}$) unless otherwise stated. Let $\phi(x)$ be a polynomial ($\phi \in \mathbb{C}[x]$). The following relations are easily established

$$\begin{aligned} &\phi(B) = P\phi(A)P^{-1}, \quad B = PAP^{-1},\ A, B \in \mathbb{C}^{n\times n},\ P \in \mathbf{GL}_n(\mathbb{C}), \\ &\phi(A_1 \oplus A_2) = \phi(A_1) \oplus \phi(A_2). \end{aligned} \tag{3.1.1}$$

It often pays to know the explicit formula for $\phi(A)$ in terms of the Jordan canonical form of A. In view of (3.1.1) it is enough to consider the case where J is composed of one Jordan block.

Lemma 3.1.1 *Let $J = \lambda_0 I + H \in \mathbb{C}^{n\times n}$, where $H = H_n$. Then for any $\phi \in \mathbb{C}[x]$*

$$\phi(J) = \sum_{k=0}^{n-1} \frac{\phi^{(k)}(\lambda_0)}{k!} H^k.$$

Proof. For any ϕ we have the Taylor expansion

$$\phi(x) = \sum_{k=0}^{N} \frac{\phi^{(k)}(\lambda_0)}{k!}(x - \lambda_0)^k, \quad N = \max(\deg \phi, n).$$

As $H^\ell = 0$ for $\ell \geq n$ from the above equality we deduce the lemma. □

Using the Jordan canonical form of A we obtain.

Theorem 3.1.2 *Let $A \in \mathbb{C}^{n\times n}$. Assume that the Jordan canonical form of A is given by* (2.6.5). *Then for $\phi \in \mathbb{C}[x]$ we have*

$$\phi(A) = P(\oplus_{i=1}^{\ell} \oplus_{j=1}^{q_i} \sum_{k=0}^{m_{ij}-1} \frac{\phi^{(k)}(\lambda_i)}{k!} H_{m_{ij}}^k)P^{-1}. \tag{3.1.2}$$

Definition 3.1.3 *Let the assumptions of Theorem 3.1.2 hold. Then $Z_{ik} = Z_{ik}(A)$ is called the (i,k) component of A and is given by*

$$\begin{gathered} Z_{ik} = P(0 \oplus ... \oplus 0 \oplus_{j=1}^{q_i} H_{m_{ij}}^k \oplus 0... \oplus 0)P^{-1}, \\ k = 0, ..., s_i - 1, \quad s_i = m_{i1},\ i = 1, ..., \ell. \end{gathered} \tag{3.1.3}$$

Compare (3.1.2) with (3.1.3) to deduce

$$\phi(A) = \sum_{i=1}^{\ell} \sum_{j=0}^{s_i-1} \frac{\phi^{(j)}(\lambda_i)}{j!} Z_{ij}. \tag{3.1.4}$$

Definition 3.1.4 *Let $A \in \mathbb{C}^{n\times n}$ and assume that $\Omega \subset \mathbb{C}$ contains* spec (A). *Then for $\phi \in \mathrm{H}(\Omega)$ define $\phi(A)$ by* (3.1.4).

Using (3.1.3) it is easy verify that the components of A satisfy

$$\begin{aligned} & Z_{ij},\ i = 1, ..., \ell,\ j = 1, ..., s_i - 1, \quad \text{are linearly independent}, \\ & Z_{ij}Z_{pq} = 0 \quad \text{if either } i \neq p, \text{ or } i = p \text{ and } j + q \geq s_i, \\ & Z_{ij}Z_{iq} = Z_{i(j+q)}, \quad \text{for } j + q \leq s_i - 1, \\ & A = P(\sum_{i=1}^{\ell} \lambda_i Z_{i0} + Z_{i1})P^{-1}. \end{aligned} \tag{3.1.5}$$

Consider the component $Z_{i(s_i-1)}$. The above relations imply

$$AZ_{i(s_i-1)} = Z_{i(s_i-1)}A = \lambda_i Z_{i(s_i-1)}. \tag{3.1.6}$$

Thus the nonzero columns of $Z_{i(s_i-1)}$, $Z_{i(s_i-1)}^\top$ are the eigenvectors of A, $A^\top$ respectively corresponding to λ_i. (Note that $Z_{i(s_i-1)} \neq 0$.)

Lemma 3.1.5 *Let $A \in \mathbb{C}^{n\times n}$. Assume that λ_i is an eigenvalue of A. Let $\mathbf{X}_i$ be the generalized eigenspace of A corresponding to λ_i:*

$$\mathbf{X}_i = \{\mathbf{x} \in \mathbb{C}^n : \quad (\lambda_i I - A)^{s_i}\mathbf{x} = \mathbf{0}\}. \tag{3.1.7}$$

Then

$$\mathrm{rank\ Z_{i(s_i-1)} = dim\ \ (\lambda_i I - A)^{s_i-1}}\mathbf{X}_\mathrm{i}. \tag{3.1.8}$$

Proof. It is enough to assume that A is in its Jordan form. Then $\mathbf{X}_i$ is the subspace of all $\mathbf{x} = (x_1, ..., x_n)^\top$, where the first $\sum_{p=1}^{i-1}\sum_{j=1}^{q_p} m_{pj}$ coordinates and the last $\sum_{p=i+1}^{\ell}\sum_{j=1}^{q_p} m_{pj}$ coordinates vanish. So $(\lambda_i I - A)^{s_i-1}\mathbf{X}_i$ contains only those eigenvectors which correspond to Jordan blocks of the length s_i. Clearly, the rank of $Z_{i(s_i-1)}$ is exactly the number of such blocks. □

Definition 3.1.6 *Let $A \in \mathbb{C}^{n\times n}$. Then the spectral radius $\rho(A)$, the peripheral spectrum* $\mathrm{spec}_{\ \mathrm{peri}}(\mathrm{A})$ *and the index A of A are given by*

$$\begin{aligned} &\rho(A) = \max_{\lambda\in \mathrm{spec\ (A)}} |\lambda|, \\ &\mathrm{spec}_{\ \mathrm{peri}}(\mathrm{A}) = \{\lambda \in \mathrm{spec\ (A)} : \quad |\lambda| = \rho(\mathrm{A})\}, \\ &\mathrm{index}\ A = \max_{\lambda\in \mathrm{spec}_{\ \mathrm{peri}}(\mathrm{A})} \mathrm{index}\ \lambda. \end{aligned} \tag{3.1.9}$$

Problems

1. Let $A \in \mathbb{C}^{n\times n}$ and let $\psi \in \mathbb{C}[x]$ be the minimal polynomial of A. Assume that $\Omega \subset \mathbb{C}$ is an open set in $\mathbb{C}$ such that $\mathrm{spec\ (A)} \subset \Omega$. Let $\phi \in \mathrm{H}(\Omega)$. Then the values

   $$\phi^{(k)}(\lambda), \quad k = 0, ..., \mathrm{index}\ \lambda - 1,\ \lambda \in \mathrm{spec\ (A)} \tag{3.1.10}$$

   are called the *values* of ϕ on the *spectrum* of A. Two functions $\phi, \theta \in \mathrm{H}(\Omega)$ are said to *coincide* on spec (A) if they have the same values on spec (A). Assume that $\phi \in \mathbb{C}[x]$ and let

   $$\phi = \omega\psi + \theta, \quad \deg\theta < \deg\psi.$$

   Show that θ coincide with ϕ on spec (A). Let

   $$\frac{\phi(x)}{\psi(x)} = \omega(x) + \frac{\theta(x)}{\psi(x)} = \omega(x) + \sum_{i=1}^{\ell}\sum_{j=1}^{s_i}\frac{\alpha_{ij}}{(x-\lambda_i)^j},\ s_i = \mathrm{index}\ \lambda_i,\ i = 1, ..., \ell,$$

where ψ is given by (2.5.11). Show that $\alpha_{ij},\ j = s_i, ..., s_i - p$ are determined recursively by $\phi^{(j)},\ j = 0, ..., p$. (Multiply the above equality by $\psi(x)$ and evaluate this identity at λ_i.) For any $\phi \in \mathrm{H}(\Omega)$ define θ by the equality

$$\theta(x) = \psi(x) \sum_{i=1}^{\ell} \sum_{j=1}^{s_i} \frac{\alpha_{ij}}{(x - \lambda_i)^j}. \tag{3.1.11}$$

The polynomial θ is called the *Lagrange-Sylvester* (L-S) *interpolation* polynomial of ϕ (corresponding to ψ). Prove that

$$\phi(A) = \theta(A). \tag{3.1.12}$$

Let θ_j be the L-S polynomials of $\phi_j \in \mathrm{H}(\Omega)$ for $j = 1, 2$. Show that $\theta_1\theta_2$ coincides with L-S polynomial of $\phi_1\phi_2$ on spec (A). Use this fact to prove the identity

$$\phi_1(A)\phi_2(A) = \phi(A), \quad \phi = \phi_1\phi_2. \tag{3.1.13}$$

2. Prove (3.1.13) by using the definition (3.1.4) and the relation (3.1.5).

3. Let the assumptions of Problem 1 hold. Assume that a sequence $\{\phi_m\}_1^\infty \subset \mathrm{H}(\Omega)$ converges to $\phi \in \mathrm{H}(\Omega)$. That is $\{\phi_m\}_1^\infty$ converges uniformly on any compact set of Ω. Hence

$$\lim_{m\to\infty} \phi_m^{(j)}(\lambda) = \phi^{(j)}(\lambda), \quad \text{for any } j \in \mathbb{Z}_+ \text{ and } \lambda \in \Omega.$$

Use the definition (3.1.4) to show

$$\lim_{m\to\infty} \phi_m(A) = \phi(A). \tag{3.1.14}$$

Apply this result to prove

$$e^A = \sum_{m=0}^{\infty} \frac{A^m}{m!} \quad (= \lim_{N\to\infty} \sum_{m=0}^{N} \frac{A^m}{m!}). \tag{3.1.15}$$

$$(\lambda I - A)^{-1} = \sum_{m=0}^{\infty} \frac{A^m}{\lambda^{m+1}} \quad \text{for } |\lambda| > \rho(A). \tag{3.1.16}$$

3.2 Cesaro convergence of matrices

Let

$$A_k = [a_{ij}^{(k)}] \in \mathbb{C}^{m\times n}, \quad k = 0, 1, ... \tag{3.2.1}$$

be a sequence of matrices. The $p-th$ *Cesaro sequence* is defined as follows. First $A_{k,0} = A_k$ for each $k \in \mathbb{Z}_+$. Then for $p \in \mathbb{N}$ $A_{k,p}$ defined recursively by

$$A_{k,p} = [a_{ij}^{(k,p)}] := \frac{1}{k+1}\sum_{j=0}^{k} A_{j,p-1}, \quad k \in \mathbb{Z}_+,\ p \in \mathbb{N}. \tag{3.2.2}$$

Definition 3.2.1 *A sequence* $\{A_k\}_0^\infty$ *converges to* $A = [a_{ij}] \in \mathbb{C}^{m\times n}$ *if*

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij}, \quad i = 1, ..., m,\ j = 1, ..., n \quad \iff \quad \lim_{k\to\infty} A_k = A.$$

A sequence $\{A_k\}_0^\infty$ *converges p-Cesaro to* $A = [a_{ij}]$ *if* $\lim_{k\to\infty} A_{k,p} = A$ *for* $p \in \mathbb{Z}_+$. *A sequence* $\{A_k\}_0^\infty$ *converges p-Cesaro exactly to* $A = [a_{ij}]$ *if* $\lim_{k\to\infty} A_{k,p} = A$ *and* $\{A_{k,p-1}\}_{k=0}^\infty$ *does not converge.*

It is known (e.g. [Har49]) that if $\{A_k\}$ is p-Cesaro convergent then $\{A_k\}$ is also $p+1$-Cesaro convergent. A simple example of exact 1-Cesaro convergent sequence is the sequence $\{\lambda^k\}$, where $|\lambda| = 1,\ \lambda \neq 1$. More generally, see [Har49] or Problem 1:

Lemma 3.2.2 *Let* $|\lambda| = 1,\ \lambda \neq 1$. *Then for* $p \in \mathbb{N}$ *the sequence* $\{\binom{k}{p-1}\lambda^k\}_{k=0}^\infty$ *is exactly p-Cesaro convergent.*

We now show how to recover the component $Z_{\alpha(s_\alpha-1)}(A)$ for $0 \neq \lambda_\alpha \in$ spec $_{\text{peri}}$(A) using the notion of Cesaro convergence.

Theorem 3.2.3 *Let* $A \in \mathbb{C}^{n\times n}$. *Assume that* $\rho(A) > 0$ *and* $\lambda_\alpha \in$ spec $_{\text{peri}}$(A). *Let*

$$A_k = \frac{(s_\alpha - 1)!}{k^{s_\alpha - 1}}(\frac{\bar{\lambda}_\alpha A}{|\lambda_\alpha|^2})^k, \quad s_\alpha = \text{index } \lambda_\alpha. \tag{3.2.3}$$

Then

$$\lim_{k\to\infty} A_{k,p} = Z_{\alpha(s_\alpha-1)}, \quad p = \text{index } A - \text{index } \lambda_\alpha + 1. \tag{3.2.4}$$

The sequence A_k *is exactly p-Cesaro convergent unless either* spec $_{\text{peri}}$(A) $=$ $\{\lambda_\alpha\}$ *or* index $\lambda <$ index λ_α *for any* $\lambda \neq \lambda_\alpha$ *in* spec $_{\text{peri}}$(A). *In these exceptional cases* $\lim_{k\to\infty} A_k = Z_{\alpha(s_\alpha-1)}$.

Proof. It is enough to consider the case where $\lambda_\alpha = \rho(A) = 1$. By letting $\phi(x) = x^k$ in (3.1.4) we get

$$A^k = \sum_{i=1}^{\ell} \sum_{j=0}^{s_i-1} \binom{k}{j} \lambda_i^{k-j} Z_{ij}. \tag{3.2.5}$$

So

$$A_k = \sum_{i=1}^{\ell} \sum_{j=0}^{s_i-1} \frac{(s_\alpha-1)!}{k^{s_\alpha-1}} \frac{k(k-1)...(k-j+1)}{j!} \lambda_i^{k-j} Z_{ij}.$$

Since the components $Z_{ij},\ i = 1, ..., \ell,\ j = 0, ..., s_i - 1$ are linearly independent it is enough to analyze the sequence $\frac{(s_\alpha-1)!}{k^{s_\alpha-1}}\binom{k}{j}\lambda_i^{k-j},\ k = j, j+1, ...$ Clearly for $|\lambda| < 1$ and any j or for $|\lambda| = 1$ and $j < s_\alpha - 1$ this sequence converges to zero. For $\lambda_i = 1$ and $j = s_\alpha - 1$ the above sequence converges to 1. For $|\lambda_i| = 1,\ \lambda_i \neq 1$ and $j \geq s_\alpha - 1$ the given sequence is exactly $j - s_\alpha + 2$ convergent to 0 in view of Lemma 3.2.2. From these arguments the theorem easily follows. □

The proof of Theorem 3.2.3 yields:

Corollary 3.2.4 *Let the assumptions of Theorem 3.2.3 hold. Then*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} \frac{(s-1)!}{k^{s-1}} (\frac{A}{\rho(A)})^k = Z, \quad s = \text{index}\, A. \tag{3.2.6}$$

If $\rho(A) \in \text{spec (A)}$ *and* $\text{index}\,\rho(A) = s$ *then* $Z = Z_{\rho(A)(s-1)}$. *Otherwise* $Z = 0$.

Problems

1. Let $|\lambda| = 1,\ \lambda \neq 1$ be fixed. Differentiate the formula

$$\sum_{j=0}^{k-1} \lambda^j = \frac{\lambda^k - 1}{\lambda - 1}$$

r times with respect to λ and divide by $r!$ to obtain

$$\sum_{j=0}^{k-1} \binom{j}{r} \lambda^j = k \sum_{\ell=0}^{r-1} f(\lambda, r, \ell) \binom{k-1}{\ell} \lambda^{k-1} + (-1)^r \frac{\lambda^k - 1}{(\lambda-1)^{r+1}},$$

where $f(\lambda, r, \ell)$ are some fixed nonzero functions. Use the induction on r to prove Lemma 3.2.2.

2. Let $\phi(x)$ be a normalized polynomial of degree $p-1$. Prove that the sequence $\{\phi(k)\lambda^k\}_{k=0}^{\infty}$ for $|\lambda| = 1,\ \lambda \ne 1$ is exactly p-Cesaro convergent.

3. Let $A \in \mathbb{C}^{n\times n}$. For $\lambda_i \in$ spec (A) let

   (3.2.7) $$Z_{ij}(A) = [z_{\mu\nu}^{(ij)}]_{\mu,\nu=1}^n,\ j = 0, ..., \text{index } \lambda_i - 1.$$

   Let

   (3.2.8) $$\text{index}_{\mu\nu}\lambda_i := 1 + \max\{j :\ z_{\mu\nu}^{(ij)} \ne 0,\ j = 0, ..., \text{index } \lambda_i - 1\},$$

   where $\text{index}_{\mu\nu}\lambda_i = 0$ if $z_{\mu\nu}^{(ij)} = 0$ for $j = 0, ..., \text{index;}\lambda_i - 1$.

   (3.2.9) $$\rho_{\mu\nu}(A) = \max\{|\lambda_i| :\ \text{index}_{\mu\nu}\lambda_i > 0\},$$

   where $\rho_{\mu\nu}(A) = -\infty$ if $\text{index}_{\mu\nu}\lambda_i = 0$ for all $\lambda_i \in$ spec (A). The quantities index ${}_{\mu\nu}\lambda_i,\ \rho_{\mu\nu}(A)$ are called the (μ,ν) index of λ_i and the (μ,ν) spectral radius respectively. Alternatively these quantities are called the local index and the local spectral radius respectively. Show that Theorem 3.2.3 and Corollary 3.2.4 could be stated in a local form. That is for $1 \le \mu,\nu \le n$ assume that

   $$\lambda_\alpha = \rho_{\mu\nu}(A),\ s_\alpha = \text{index}_{\mu\nu}\lambda_\alpha,\ A^k = (a_{\mu\nu}^{(k)}),\ A_k = [a_{\mu\nu,k}],\ A_{k,p} = [a_{\mu\nu,kp}],$$

   where A_k and $A_{k,p}$ are given by (3.2.3) and (3.2.2) respectively. Prove

   $$\lim_{k\to\infty} a_{\mu\nu,kp} = z_{\mu\nu}^{(\alpha(s_\alpha-1))}, \quad p = \text{index}_{\mu\nu}A - \text{index}_{\mu\nu}\lambda_\alpha + 1,$$

   $$\lim_{N\to\infty} \frac{1}{N+1}\sum_{k=0}^{N} \frac{(s-1)!}{k^{s-1}}(\frac{a_{\mu\nu,k}}{\rho_{\mu\nu}(A)})^k = z_{\mu\nu}, \quad s = \text{index}_{\mu\nu}A,\ \rho_{\mu\nu}(A) > 0,$$

   where $z_{\mu\nu} = 0$ unless $\lambda_1 = \rho_{\mu\nu}(A) \in$ spec (A) and $\text{index}_{\mu\nu}\lambda_1 = \text{index}_{\mu\nu} = s$. In this exceptional case $z_{\mu\nu} = z_{\mu\nu}^{(1(s-1))}$.

   Finally A is called irreducible if $\rho_{\mu\nu}(A) = \rho(A)$ for each $\mu,\nu = 1, ..., n$. Thus for an irreducible A the local and the global versions of Theorem 3.2.3 and Corollary 3.2.4 coincide.

3.3 An iteration scheme

Consider an iteration given by

(3.3.1) $$\mathbf{x}^{i+1} = A\mathbf{x}^i + \mathbf{b}, \quad i = 0, 1, ...,$$

where $A \in \mathbb{C}^{n\times n}$ and $\mathbf{x}^i, \mathbf{b} \in \mathbb{C}^n$. Such an iteration can be used to solve a system

$$\mathbf{x} = A\mathbf{x} + \mathbf{b}. \tag{3.3.2}$$

Assume that $\mathbf{x}$ is the unique solution of (3.3.2) and let $\mathbf{y}^i := \mathbf{x}^i - \mathbf{x}$. Then

$$\mathbf{y}^{i+1} = A\mathbf{y}^i, \quad i = 0, 1, ... \tag{3.3.3}$$

Definition 3.3.1 *The system* (3.3.3) *is called stable if the sequence* $\mathbf{y}^i,\ i = 0, 1, ...$ *converges to zero for any choice of* $\mathbf{y}^0$. *The system* (3.3.3) *is called bounded if the sequence* $\mathbf{y}^i,\ i = 0, 1, ...$ *is bounded for any choice of* $\mathbf{y}^0$.

Clearly, the solution to (3.3.3) is $\mathbf{y}^i = A^i\mathbf{y}^0,\ i = 0, 1, ...$ So (3.3.3) is stable if and only if

$$\lim_{i\to\infty} A^i = 0. \tag{3.3.4}$$

Furthermore (3.3.3) is bounded if and only if

$$\|A^i\| \le M, \quad i = 0, 1, ..., \tag{3.3.5}$$

for some (or any) vector norm $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}_+$ and some $M > 0$. For example one can choose the l_∞ norm on $\mathbb{C}^{m\times n}$ to obtain the induced matrix norm:

$$\|B\| = \max_{1\le i\le m, 1\le j\le n} |b_{ij}|, \quad B = [b_{ij}] \in \mathbb{C}^{m\times n}. \tag{3.3.6}$$

See §7.4 and §7.7 for definitions and properties of vector and operator norms.

Theorem 3.3.2 *Let* $A \in \mathbb{C}^{n\times n}$. *Then conditions* (3.3.4) *holds if and only if* $\rho(A) < 1$. *Conditions* (3.3.5) *hold if either* $\rho(A) < 1$ *or* $\rho(A) = 1$ *and* index $A = 1$.

Proof. Consider the identity (3.2.5). Since all the components of A are linearly independent (3.3.4) is equivalent to

$$\lim_{k\to\infty} \binom{k}{j} \lambda_i^{k-j} = 0, \quad \lambda_i \in \mathrm{spec\ (A)},\ \mathrm{j} = 0, 1, ..., \mathrm{index}\ \lambda_\mathrm{i} - 1.$$

Clearly the above conditions are equivalent to $\rho(A) < 1$.

Since all vector norms on $\mathbb{C}^{m\times n}$ are equivalent, the condition (3.3.5) is equivalent to the statement that the sequence $\binom{k}{j}\lambda_i^{k-j},\ k = 0, ...,$ is bounded for each $\lambda_i \in \mathrm{spec}\,(\mathrm{A})$ and each $j \in [0, \mathrm{index}\,\lambda_i - 1]$. Hence $\rho(A) \le 1$. Furthermore if $|\lambda_i| = 1$ then index $\lambda_i = 1$. □

Problems

1. Let $A \in \mathbb{C}^{n\times n}$ and ψ be the minimal polynomial of A given by (2.5.11). Verify

$$e^{At} = \sum_{i=1}^{\ell} \sum_{j=0}^{s_i-1} \frac{t^j e^{\lambda_i t}}{j!} Z_{ij}. \tag{3.3.7}$$

Use (3.1.5) or (3.1.15) to show

$$\frac{d}{dt} e^{At} = A e^{At} = e^{At} A. \tag{3.3.8}$$

(In general t may be complex valued, but in this problem we assume that t is real.) Verify that the system

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}, \quad x(t) \in \mathbb{C}^n \tag{3.3.9}$$

has the unique solution

$$\mathbf{x}(t) = e^{A(t-t_o)} \mathbf{x}(t_o). \tag{3.3.10}$$

The system (3.3.9) is called stable if $\lim_{t\to\infty} \mathbf{x}(t) = o$ for any solution (3.3.10). The system (3.3.9) is called bounded if any solution $\mathbf{x}(t)$ (3.3.10) is bounded on $[t_0, \infty)$. Prove that (3.3.9) is stable if and only if

$$\Re\lambda < 0 \quad \text{for each } \lambda \in \mathrm{spec}\,(\mathrm{A}). \tag{3.3.11}$$

Furthermore (3.3.9) is bounded if and only if each $\lambda \in \mathrm{spec}\,(\mathrm{A})$ satisfies

$$\Re\lambda \le 0 \quad \text{and} \quad \mathrm{index}\,\lambda = 1 \text{ if } \Re\lambda = 0. \tag{3.3.12}$$

3.4 Cauchy integral formula for functions of matrices

Let $A \in \mathbb{C}^{n\times n}$, $\phi \in \mathrm{H}(\Omega)$, where Ω is an open set in $\mathbb{C}$. If $\mathrm{spec}\,(\mathrm{A}) \subset \Omega$ it is possible to define $\phi(A)$ by (3.1.4). The aim of this section is to give an integral formula for $\phi(A)$ using the Cauchy integration formula for $\phi(\lambda)$. The resulting expression is simply looking and very useful in theoretical studies of $\phi(A)$. Moreover, this formula remains valid for bounded operators in Banach spaces (e.g. [Kat80]-[Kat82]).

Consider the function $\phi(x,\lambda) = (\lambda - x)^{-1}$. The domain of analyticity of $\phi(x,\lambda)$ (with respect to x) is the punctured complex plane $\mathbb{C}$ at λ. Thus if $\lambda \notin$ spec (A) (3.1.4) yields

$$(\lambda I - A)^{-1} = \sum_{i=1}^{\ell}\sum_{j=0}^{s_i-1} (\lambda - \lambda_i)^{-(j+1)} Z_{ij}. \tag{3.4.1}$$

Definition 3.4.1 *The function $(\lambda I - A)^{-1}$ is called the resolvent of A and is denoted by*

$$R(\lambda, A) = (\lambda I - A)^{-1}. \tag{3.4.2}$$

Let $\Gamma = \{\Gamma_1, ..., \Gamma_k\}$ be a set of disjoint simply connected rectifiable curves such that Γ forms the boundary ∂D of an open set D and

$$D \cup \Gamma \subset \Omega, \quad \Gamma = \partial D. \tag{3.4.3}$$

For $\phi \in \mathrm{H}(\Omega)$ the classical Cauchy integration formula states (e.g. [Rud74])

$$\phi(\zeta) = \frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} (\lambda - \zeta)^{-1}\phi(\lambda)d\lambda, \quad \zeta \in D. \tag{3.4.4}$$

Differentiate the above equality j times to obtain

$$\frac{\phi^{(j)}(\zeta)}{j!} = \frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} (\lambda - \zeta)^{-(j+1)}\phi(\lambda)d\lambda, \quad \zeta \in D,\ j = 0, 1, 2, ... \tag{3.4.5}$$

Theorem 3.4.2 *Let Ω be an open set in $\mathbb{C}$. Assume that $\Gamma = \{\Gamma_1, ..., \Gamma_k\}$ is a set of disjoint simple, connected, rectifiable curves such that Γ is a boundary of an open set D, and $\Gamma \cup D \subset \Omega$. Assume that $A \in \mathbb{C}^{n\times n}$ and* spec (A) $\subset$ D. *The for any $\phi \in$* H(Ω)

$$\phi(A) = \frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} R(\lambda, A)\phi(\lambda)d\lambda. \tag{3.4.6}$$

Proof. Insert the expression (3.4.1) into the above integral to obtain

$$\frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} R(\lambda, A)\phi(\lambda)d\lambda = \sum_{i=1}^{\ell}\sum_{j=0}^{s_i-1} \big(\frac{1}{2\pi\sqrt{-1}} \int_{\Gamma} (\lambda - \lambda_i)^{-(j+1)}\phi(\lambda)d\lambda\big) Z_{ij}.$$

Use the identity (3.4.5) to deduce

$$\frac{1}{2\pi\sqrt{-1}}\int_\Gamma R(\lambda,A)\phi(\lambda)d\lambda = \sum_{i=1}^{\ell}\sum_{j=0}^{s_i-1}\frac{\phi^{(j)}(\lambda_i)}{j!}Z_{ij}.$$

The definition (3.1.4) yields the equality (3.4.6). □

We generalize the above theorem as follows.

Theorem 3.4.3 *Let* Ω *be an open set in* $\mathbb{C}$. *Assume that* $\Gamma = \{\Gamma_1, ..., \Gamma_k\}$ *is a set of disjoint simple, connected, rectifiable curves such that* Γ *is a boundary of an open set* D, *and* $\Gamma \cup D \subset \Omega$. *Assume that* $A \in \mathbb{C}^{n\times n}$ *and* $\mathrm{spec}\,(\mathrm{A}) \cap \Gamma = \emptyset$. *Let* $\mathrm{spec}_{\mathrm{D}}(\mathrm{A}) := \mathrm{spec}\,(\mathrm{A}) \cap \mathrm{D}$. *The for any* $\phi \in \mathrm{H}(\Omega)$

$$\sum_{\lambda_i\in\mathrm{spec}_{\mathrm{D}}(\mathrm{A})}\sum_{j=0}^{s_i-1}\frac{\phi^{(j)}(\lambda_i)}{j!}Z_{ij} = \frac{1}{2\pi\sqrt{-1}}\int_\Gamma R(\lambda,A)\phi(\lambda)d\lambda. \tag{3.4.7}$$

If $\mathrm{spec}_{\mathrm{D}}(\mathrm{A}) = \emptyset$ *then the left-hand side of the above identity is zero.*

See Problem 1.

We illustrate the usefulness of Cauchy integral formula by two examples.

Theorem 3.4.4 *Let* $A \in \mathbb{C}^{n\times n}$ *and assume that* $\lambda_p \in \mathrm{spec}\,(\mathrm{A})$. *Suppose that* D *and* Γ *satisfy the assumptions of Theorem 3.4.3* $(\Omega = \mathbb{C})$. *Assume furthermore that* $\mathrm{spec}\,(\mathrm{A}) \cap \mathrm{D} = \{\lambda_{\mathrm{p}}\}$. *Then the* (p,q) *component of* A *is given by*

$$Z_{pq}(A) = \frac{1}{2\pi\sqrt{-1}}\int_\Gamma R(\lambda,A)(\lambda-\lambda_p)^q d\lambda. \tag{3.4.8}$$

$(Z_{pq} = 0$ *for* $q > s_p - 1.)$

See Problem 2.

Our next examples generalizes the first part of Theorem 3.3.2 to a compact set of matrices.

Definition 3.4.5 *A set* $\mathcal{A} \subset \mathbb{C}^{n\times n}$ *is called power stable if*

$$\lim_{k\to\infty}\big(\sup_{A\in\mathcal{A}}||A^k||\big) = 0, \tag{3.4.9}$$

for some vector norm on $\mathbb{C}^{n\times n}$. *A set* $\mathcal{A} \subset \mathbb{C}^{n\times n}$ *is called power bounded if*

$$||A^k|| \le K, \quad \text{for any } A \in \mathcal{A} \text{ and } k = 0,1,..., \tag{3.4.10}$$

for some positive K *and some vector norm on* $\mathbb{C}^{n\times n}$.

Theorem 3.4.6 *Let $\mathcal{A} \subset \mathbb{C}^{n\times n}$ be a compact set. Then $\mathcal{A}$ is power stable if and only if $\rho(A) < 1$ for any $A \in \mathcal{A}$.*

To prove the theorem we need a well known result on the roots of normalized polynomials in $\mathbb{C}[x]$ (e.g. [Ost66]).

Lemma 3.4.7 *Let $p(x) = x^m + \sum_{i=1}^m a_i x^{m-i} \in \mathbb{C}[x]$. Then the zeros $\xi_1, ..., \xi_m$ of $p(x)$ are continuous functions of its coefficients. That is for a given $a_1, ..., a_n$ and $\epsilon > 0$ there exists $\delta(\epsilon)$, depending on $a_1, ..., a_n$, such that if $|b_i - a_i| < \delta(\epsilon)$, $i = 1, ..., m$ it is possible to enumerate the zeros of $q(x) = x^m + \sum_{i=1}^m b_i x^{m-i}$ by $\eta_1, ..., \eta_m$, such that $|\eta_i - \xi_i| < \epsilon$. $i = 1, ..., m$. In particular the function*

$$\rho(p) = \max_{1\le i\le m} |\xi_i| \tag{3.4.11}$$

is a continuous function of $a_1, ..., a_m$.

Corollary 3.4.8 *The function $\rho : \mathbb{C}^{n\times n} \to \mathbb{R}_+$, which assigns to $A \in \mathbb{C}^{n\times n}$ its spectral radius $\rho(A)$ is a continuous function.*

Proof of Theorem 3.4.6. Suppose that (3.4.9) holds. Then by Theorem 3.3.2 $\rho(A) < 1$ for each $A \in \mathcal{A}$. Assume that $\mathcal{A}$ is compact and $\rho(A) < 1$. Corollary 3.4.8 yields

$$\rho := \max_{A\in\mathcal{A}} \rho(A) = \rho(\tilde{A}) < 1, \quad \tilde{A} \in \mathcal{A}.$$

Recall that $(\lambda I - A)^{-1} = [\frac{p_{ij}(\lambda)}{\det\,(\lambda I - A)}]_1^n$, where $p_{ij}(\lambda)$ is the (j, i) cofactor of $\lambda I - A$. Let $\lambda_1, ..., \lambda_n$ be the eigenvalues of A counted with multiplicities. Then for $|\lambda| > \rho$

$$|\det\,(\lambda I - A)| = |\prod_{i=1}^n (\lambda - \lambda_i)| \ge \prod_{i=1}^n (|\lambda| - \rho)^n.$$

Let $\rho < r < 1$. Since $\mathcal{A}$ is a bounded set, the above arguments yield that there exists a positive constant K such that $||(\lambda I - A)^{-1}|| \le K$ for each $A \in \mathcal{A}$, $|\lambda| = r$. Apply (3.4.6) to obtain

$$A^p = \frac{1}{2\pi\sqrt{-1}} \int_{|\lambda|=r} (\lambda I - A)^{-1} \lambda^p d\lambda, \tag{3.4.12}$$

for each $A \in \mathcal{A}$. Combine this equality with the estimate $||\lambda I - A)^{-1}|| \le K$ for $|\lambda| = r$ to obtain $||A^p|| \le K r^{p+1}$ for any $A \in \mathcal{A}$. As $r < 1$ the theorem follows. □

Theorem 3.4.9 *Let $\mathcal{A} \subset \mathbb{C}^{n\times n}$. Then $\mathcal{A}$ is power bounded if and only*

$$||(\lambda I - A)^{-1}|| \le \frac{K}{|\lambda| - 1}, \quad \text{for all } A \in \mathcal{A} \text{ and } |\lambda| > 1, \tag{3.4.13}$$

for some vector norm $|| \cdot ||$ on $\mathbb{C}^{n\times n}$ and $K \ge ||I_n||$.

Proof. For $|\lambda| > \rho(A)$ we have the Neumann series

$$(\lambda I - A)^{-1} = \sum_{i=0}^{\infty} \frac{A^i}{\lambda^{i+1}}. \tag{3.4.14}$$

Hence for any vector norm on $\mathbb{C}^{n\times n}$

$$||(\lambda I - A)^{-1}|| \le \sum_{i=0}^{\infty} \frac{||A^i||}{|\lambda|^{i+1}}, \quad |\lambda| > \rho(A). \tag{3.4.15}$$

(See Problem 3.) Assume first that (3.4.10) hold. As $A^0 = I_n$ it follows that $K \ge ||I_n||$. Furthermore as each $A \in \mathcal{A}$ is power bounded Theorem 3.3.2 yields that $\rho(A) \le 1$ for each $A \in \mathcal{A}$. Combine (3.4.10) and (3.4.15) to obtain (3.4.13).

Assume now that (3.4.13) holds. Since all vector norms on $\mathbb{C}^{n\times n}$ are equivalent we assume that the norm in (3.4.13) is the l_∞ norm given in (3.3.6). Let $A \in \mathcal{A}$. Note that $(\lambda I - A)$ in invertible for each $|\lambda| > 1$. Hence $\rho(A) \le 1$. Let $(\lambda I - A)^{-1} = [\frac{p_{ij}(\lambda)}{p(\lambda)}]$. Here $p(\lambda) = \det\,(\lambda I - A)$ is a polynomial of degree n and $p_{ij}(\lambda)$ (the (j,i) cofactor of $\lambda I - A$) is a polynomial of degree $n-1$ at most. Let $A^p = [a_{ij}^{(p)}]$, $p = 0, 1, ...$ Then for any $r > 1$ the equality (3.4.12) yields that

$$a_{ij}^{(p)} = \frac{1}{2\pi\sqrt{-1}} \int_{|\lambda|=r} \frac{p_{ij}(\lambda)}{p(\lambda)} \lambda^p d\lambda = \frac{1}{2\pi} \int_0^{2\pi} \frac{p_{ij}(re^{\sqrt{-1}\theta})}{p(re^{\sqrt{-1}\theta})} r^{p+1} e^{(p+1)\sqrt{-1}\theta} d\theta.$$

Problem 6 implies that

$$|a_{ij}^{(p)}| \le \frac{4(2n-1)r^{p+1}}{\pi(p+1)} \max_{|\lambda|=r} |\frac{p_{ij}(\lambda)}{p(\lambda)}| \le \frac{4(2n-1)r^{p+1}K}{\pi(p+1)(r-1)}.$$

Choose $r = 1 + \frac{1}{p+1}$ to obtain

$$|a_{ij}^{(p)}| \le \frac{4(2n-1)eK}{\pi}, \quad i,j = 1, ..., n, \ p = 0, 1, ..., \ A \in \mathcal{A}. \tag{3.4.16}$$

Hence $||A^p|| \le \frac{4(2n-1)eK}{\pi}$. $\Box$

Problems

1. Use the proof of Theorem 3.4.2 to prove Theorem 3.4.3.

2. Prove Theorem 3.4.4

3. Let $A \in \mathbb{C}^{n\times n}$. Show the Neumann series converge to the resolvent (3.4.14) for any $|\lambda| > \rho(A)$. (You may use (3.4.1).) Prove (3.4.15) for any vector norm on $\mathbb{C}^{n\times n}$.

4. Let $f(x)$ be a real continuous periodic function on $\mathbb{R}$ with period 2π. Assume furthermore that f' is a continuous function on $\mathbb{R}$. (f' is periodic of period 2π.) Then the Fourier series of f converge to f (e.g. [Pin09, Cor. 1.2.28]).

   $$f(\theta) = \sum_{k\in\mathbb{Z}} a_k e^{\sqrt{-1}k\theta},$$

   (3.4.17)

   $$a_k = \bar{a}_k = \frac{1}{2\pi}\int_{\theta_0}^{\theta_0+2\pi} f(\theta)e^{-\sqrt{-1}k\theta}d\theta,\ k \in \mathbb{Z},\ \theta_0 \in \mathbb{R}.$$

   Use integration by parts to conclude that

   $$a_k = \frac{1}{2\pi\sqrt{-1}k}\int_{\theta_0}^{\theta_0+2\pi} f'(\theta)e^{-\sqrt{-1}k\theta}d\theta,\ k \in \mathbb{Z}\backslash\{0\}. \tag{3.4.18}$$

   Assume that $f'(\theta)$ vanishes exactly at $m(\geq 2)$ points on the interval $[0, 2\pi)$. Show that

   $$|a_0| \leq \max_{\theta\in[0,2\pi)} |f(\theta)|,$$

   (3.4.19)

   $$|a_k| \leq \frac{m}{\pi|k|}\max_{\theta\in[0,2\pi)} |f(\theta)|, \text{ for all } k \in \mathbb{Z}\backslash\{0\}$$

   (*Hint.* The first inequality of (3.4.19) follows immediately from (3.4.17). Assume that f' vanishes at $0 \leq \theta_0 < ... < \theta_{m-1} < 2\pi \leq \theta_m = \theta_0 + 2\pi$. Then

   $$|\int_{\theta_{i-1}}^{\theta_i} f'(\theta)e^{-2\pi\sqrt{-1}k\theta}d\theta| \leq \int_{\theta_{i-1}}^{\theta_i} |f'(\theta)|d\theta =$$
   $$|f(\theta_i) - f(\theta_{i-1})| \leq 2\max_{\theta\in[0,2\pi)} |f(\theta)|,\ i = 1, ..., m.$$

   Use (3.4.18) to deduce the second part of (3.4.19).)

5. A real periodic function f is called a trigonometric polynomial of degree n if f has the expansion (3.4.17), where $a_k = 0$ for $|k| > n$ and $a_n \neq 0$. Show

   (a) A non zero trigonometric polynomial $f(\theta)$ of degree n vanishes at most $2n$ points on the interval $[0, 2\pi)$. (*Hint.* Let $z = e^{\sqrt{-1}\theta}$. Then $f = z^{-n}p(z)|_{|z|=1}$ for a corresponding polynomial p of degree $2n$.)

   (b) Let $f(\theta) = \frac{g(\theta)}{h(\theta)}$ be a nonconstant function, where g is a nonzero trigonometric polynomial of degree m at most and h is a nowhere vanishing trigonometric polynomial of degree n. Show that f' has at most $2(m+n)$ zeros on $[0, 2\pi)$.

6. Let $p(z), q(z)$ be nonconstant polynomials of degree m, n respectively. Suppose that $q(z)$ does not vanish on the circle $|z| = r > 0$. Let $M := \max_{|z|=r} \frac{|p(z)|}{|q(z)|}$. Show that for all $k \in \mathbb{Z}$

   $$(3.4.20) \quad |\frac{1}{2\pi}\int_0^{2\pi} \frac{p(re^{\sqrt{-1}\theta})}{q(re^{\sqrt{-1}\theta})} e^{\sqrt{-1}k\theta} d\theta| \leq \frac{4M\max(m+n, 2n-1)}{\pi\max(|k|, 1)}.$$

   Hint. Let $F(z) = \frac{p(z)}{q(z)} = \frac{p(z)q(\bar{z})}{q(z)q(\bar{z})}$ be a nonconstant rational function. Then $F(re^{\sqrt{-1}\theta}) = f_1(\theta) + \sqrt{-1}f_2(\theta)$, where f_1, f_2 as in Problem 5. Clearly $|f_1(\theta)|, |f_2(\theta)| \leq M$. Observe next that

   $$f_1'(\theta) + \sqrt{-1}f_2'(\theta) = \frac{\sqrt{-1}re^{\sqrt{-1}\theta}(p'(re^{\sqrt{-1}\theta})q(re^{\sqrt{-1}\theta}) - p(re^{\sqrt{-1}\theta})q'(re^{\sqrt{-1}\theta})(q(r\bar{e}^{\sqrt{-1}\theta}))^2}{|q(re^{\sqrt{-1}\theta})|^2}.$$

   Hence f_1', f_2' vanish at most $2\max(m+n, 2n-1)$ points on $[0, 2\pi)$. Use (3.4.19) for f_1, f_2 to deduce (3.4.20).

7. Let $\alpha \geq 0$ be fixed and assume that $\mathcal{A} \subset \mathbb{C}^{n\times n}$. Show that the following statements are equivalent:

   $$(3.4.21) \qquad ||A^k|| \leq k^\alpha K, \quad \text{for any } A \in \mathcal{A} \text{ and } k = 0, 1, ...,$$

   $$(3.4.22) \quad ||(\lambda I - A)^{-1}|| \leq \frac{K'|\lambda|^\alpha}{(|\lambda| - 1)^{1+\alpha}}, \quad \text{for all } A \in \mathcal{A} \text{ and } |\lambda| > 1.$$

   Hint. Use the fact that $(-1)^k\binom{-(1+\alpha)}{k}k^{-\alpha} \in [a, b],\ k = 1, ...$ for some $0 < a < b$, (e.g.[Olv74], p'119).)

8. Let $A \in \mathbb{C}^{n\times n}$. Using (3.4.1) deduce

$$Z_{i(s_i-1)} = \lim_{x\to\lambda_i} (x-\lambda_i)^{s_i}(xI-A)^{-1}, \quad i = 1,...,\ell. \tag{3.4.23}$$

let $R(x,A) = [r_{\mu\nu}]$. Using the definitions of Problem 3 show

$$z_{\mu\nu}^{i(s_i-1)} = \lim_{x\to\lambda_i} (x-\lambda_i)^s \, r_{\mu\nu}(x), \quad \text{if } s = \text{index}_{\ \mu\nu}\lambda_i > 0. \tag{3.4.24}$$

9. A set $\mathcal{A} \subset \mathbb{C}^{n\times n}$ is called *exponentially* stable if

$$\lim_{T\to\infty} \sup_{t\geq T} ||e^{At}|| = 0. \tag{3.4.25}$$

Show that a compact set $\mathcal{A}$ is exponentially stable if and only if $\Re\lambda < 0$ for each $\lambda \in$ spec (A) and each $A \in \mathcal{A}$.

10. A matrix $B \in \mathbb{C}^{n\times n}$ is called projection (idempotent) if $B^2 = B$. Let Γ be a set of simply connected rectifiable curves such that Γ from a boundary of an open bounded set $D \subset \mathbb{C}$. Let $A \in \mathbb{C}^{n\times n}$ and assume that $\Gamma \cap$ spec (A) $= \emptyset$. Define

$$\begin{aligned} P_D(A) &:= \frac{1}{2\pi\sqrt{-1}} \int_\Gamma R(x,A)dx, \\ A(D) &:= \frac{1}{2\pi\sqrt{-1}} \int_\Gamma R(x,A)xdx. \end{aligned} \tag{3.4.26}$$

Show that $P_D(A)$ is a projection. $P_D(A)$ is called the projection of A on D, and $A(D)$ is called the restriction of A to D. Prove
(3.4.27)
$$P_D(A) = \sum_{\lambda_i\in\text{spec }_{\text{D}}(\text{A})} Z_{i0}, \quad A(D) = \sum_{\lambda_i\in\text{spec }_{\text{D}}(\text{A})} (\lambda_i Z_{i0} + Z_{i1}).$$

Show that the rank of $P_D(A)$ is equal to the number of eigenvalues of A in D counted with their multiplicities. Prove that there exists a neighborhood of A such that $P_D(B)$ and $B(D)$ are analytic functions in B in this neighborhood. In particular, if D satisfies the assumptions of Theorem 3.4.4 then $P_D(A)$ is called the projection of A on λ_p: $P_D(A) = Z_{p0}$.

11. Let $B = QAQ^{-1} \in \mathbb{C}^{n\times n}$. Assume that D satisfies the assumptions of Problem 10. Show that $P_D(B) = QP_D(A)Q^{-1}$.

12. Let $A \in \mathbb{C}^{n\times n}$ and assume that the minimal polynomial $\psi(x)$ of A is given by (2.5.11). Let $\mathbb{C}^n = \mathbf{U}_1 \oplus ... \oplus \mathbf{U}_\ell$, where each $\mathbf{U}_p$

is an invariant subspace of A ($A\mathbf{U}_p \subset \mathbf{U}_p$), such that the minimal polynomial of $A|\mathbf{U}_p$ is $(x-\lambda_p)^{s_p}$. Show that

$$\mathbf{U}_p = Z_{p0}\mathbb{C}^n. \tag{3.4.28}$$

Hint. It is enough to consider the case where A is in the Jordan canonical form.

13. Let D_i satisfy the assumptions of Problem 10 for $i = 1, ...k$. Assume that $D_i \cap D_j = \emptyset$ for $i \neq j$. Show that $P_{D_i}(A)\mathbb{C}^n \cap P_{D_j}(A)\mathbb{C}^n = \{0\}$ for $i \neq j$. Assume furthermore that $D_i \cap \mathrm{spec\ (A)} \neq \emptyset,\ \mathrm{i} = 1, ..., \mathrm{k}$, and $\mathrm{spec\ (A)} \subset \cup_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{D_i}$. Let

$$P_{D_i}(A)\mathbb{C}^n = \mathrm{span}\ (\mathbf{y}_1^{(\mathrm{i})}, ..., \mathbf{y}_{\mathrm{n_i}}^{(\mathrm{i})}),\ \mathrm{i} = 1, ..., \mathrm{k},$$
$$X = [\mathbf{y}_1^{(1)}, ..., \mathbf{y}_{n_1}^{(1)}, ..., \mathbf{y}_{n_k}^{(k)}] \in \mathbb{C}^{n\times n}.$$

Show that
(3.4.29)

$$X^{-1}AX = \sum_{i=1}^{k} \oplus B_i, \quad \mathrm{spec\ (B_i) = D_i \cap spec\ (A),\ i = 1, ..., k.}$$

14. Let $A \in \mathbb{C}^{n\times n}$ and $\lambda_p \in \mathrm{spec\ (A)}$. Show that if index $\lambda_p = 1$ then

$$Z_{p0} = \prod_{\lambda_j \in \mathrm{spec\ (A)}, \lambda_j \neq \lambda_p} \frac{(A-\lambda_j I)^{s_j}}{(\lambda_p - \lambda_j)^{s_j}}, \quad s_j = \mathrm{index}\ \lambda_j.$$

Hint. Use the Jordan canonical form of A.

3.5 A canonical form over H_A

Consider the space $\mathbb{C}^{n\times n}$. Clearly $\mathbb{C}^{n\times n}$ can be identified with $\mathbb{C}^{n^2}$. As in Example 1.1.3 denote by H_A the set of analytic functions $f(B)$, where B ranges over a neighborhood $D(A,\rho)$ of the form (2.9.5) ($\rho = \rho(f) > 0$). Thus $B = [b_{ij}]$ is an element in $\mathrm{H}_A^{n\times n}$. Let $C \in \mathrm{H}_A^{n\times n}$ and assume that $C = C(B)$ is similar to B over H_A. Then

$$C(B) = X^{-1}(B)BX(B), \tag{3.5.1}$$

where $X(B) \in \mathrm{H}_A^{n\times n}$ and $\det X(A) \neq 0$. We want to find a "simple" from for $C(B)$ (simpler than B!). Let $\mathcal{M}_A$ be the quotient field of H_A (the set of meromorphic functions in the neighborhood of A). If we let $X \in \mathcal{M}_A^{n\times n}$

then we may take $C(B)$ to be $R(B)$ - the rational canonical form of B (2.3.3). According to Theorem 2.3.8 $R(B) \in \mathrm{H}_A^{n\times n}$. However B and $R(B)$ are not similar over H_A in general. (We shall give below the necessary and sufficient conditions for $B \approx R(B)$ over H_A.) For $C(B) = [c_{ij}(B)]$ we may ask how many independent variables are among $c_{ij}(B),\ i,j = 1,...,n$. For $X(B) = I$ the number of independent variables in $C(B) = B$ is n^2. Thus we call $C(B)$ to be simpler than B if $C(B)$ contains less independent variable than B. For a given $C(B)$ we can view $C(B)$ as a map

$$C(\cdot) : D(A,\rho) \to \mathbb{C}^{n\times n}, \tag{3.5.2}$$

where $D(A,\rho)$ is given by (2.9.5), for some $\rho > 0$. It is well known, e.g. [GuR65], that the number of independent variables is equal to the rank of the Jacobian matrix $DC(\cdot)$ over $\mathcal{M}_A$

$$DC(B) := (\frac{\partial \mu(C)}{\partial b_{ij}}(B)) \in \mathrm{H}_A^{n^2\times n^2}, \tag{3.5.3}$$

where μ is the map given in Problem 2.8.2.

Definition 3.5.1 *Let* rank DC, rank DC(A) *be the ranks of* $DC(\cdot)$, $DC(A)$ *over the fields* $\mathcal{M}_A$, $\mathbb{C}$ *respectively.*

Lemma 3.5.2 *Let* $C(B)$ *be similar to* B *over* H_A. *Then*

$$\mathrm{rank\,DC(A)} \geq \nu(\mathrm{A},\mathrm{A}). \tag{3.5.4}$$

Proof. Differentiating the relation $X^{-1}(B)X(B) = I$ with respect to b_{ij} we get

$$\frac{\partial X^{-1}}{\partial b_{ij}} = -X^{-1}\frac{\partial X}{\partial b_{ij}}X^{-1}.$$

So

$$\frac{\partial C}{\partial b_{ij}} = X^{-1}(-(\frac{\partial X}{\partial b_{ij}}X^{-1})B + B(\frac{\partial X}{\partial b_{ij}}X^{-1}) + E_{ij})X, \tag{3.5.5}$$

where

$$E_{ij} = [\delta_{i\alpha}\delta_{j\beta}]_{\alpha,\beta=1}^{m,n} \in \mathbb{C}^{m\times n}, \quad i = 1,...,m,\ j = 1,...,n, \tag{3.5.6}$$

and $m = n$. So

$$X(A)\frac{\partial C}{\partial b_{ij}}(A)X^{-1}(A) = AP_{ij} - P_{ij}A + E_{ij}, \quad P_{ij} = \frac{\partial X}{\partial b_{ij}}(A)X^{-1}(A).$$

Clearly, $AP_{ij} - P_{ij}A$ is in Range $\hat{A}$, where

$$\hat{A} = (I \otimes A - A^\top \otimes I) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}. \tag{3.5.7}$$

According to Definition 2.9.1 dim Range $\hat{A} = r(A, A)$. Let

$$\mathbb{C}^{n\times n} = \text{Range } \hat{A} \oplus \text{span } (\Gamma_1, ..., \Gamma_{\nu(\mathrm{A},\mathrm{A})}). \tag{3.5.8}$$

As $E_{ij},\ i, j = 1, ..., n$ is a basis in $\mathbb{C}^{n\times n}$

$$\Gamma_p = \sum_{i,j=1}^{n} \alpha_{ij}^{(p)} E_{ij}, \quad p = 1, ..., \nu(A, A).$$

Let

$$T_p := \sum_{i,j=1}^{n} \alpha_{ij}^{(p)} \frac{\partial C}{\partial b_{ij}}(A) = X^{-1}(A)(Q_p + \Gamma_p)X(A), \quad Q_p \in \text{Range } (\hat{A}),$$
$$p = 1, ..., \nu(A, A).$$

According to (3.5.8) $T_1, ..., T_{\nu(A,A)}$ are linearly independent. Hence (3.5.4) holds. □

Clearly rank DC $\geq$ rank DC(A) $\geq \nu(\mathrm{A}, \mathrm{A})$.

Theorem 3.5.3 *Let $A \in \mathbb{C}^{n\times n}$ and assume that $\Gamma_1, ..., \Gamma_{\nu(A,A)}$ be any $\nu(A, A)$ matrices satisfying (3.5.8). Then for any nonsingular matrix $P \in \mathbb{C}^{n\times n}$ it is possible to find $X(B) \in \mathrm{H}_A^{n\times n}$, $X(A) = P$, such that*

$$X^{-1}(B)BX(B) = P^{-1}AP + \sum_{i=1}^{\nu(A,A)} f_i(B)P^{-1}\Gamma_i P,$$
$$f_i \in \mathrm{H}_A,\ f_i(A) = 0, \quad i = 1, ..., \nu(A, A). \tag{3.5.9}$$

Proof. Let $R_1, ..., R_{r(A,A)}$ be a basis in Range $\hat{A}$. So there exist T_i such that $AT_i - T_iA = R_i$ for $i = 1, ..., r(A, A)$. Assume that $X(B)$ is of the form

$$X(B)P^{-1} = I + \sum_{j=1}^{r(A,A)} g_j(B)T_j, \quad g_j \in \mathrm{H}_A,\ g_j(A) = 0, \quad j = 1, ..., r(A, A). \tag{3.5.10}$$

The theorem will follow if we can show that the system

$$B(I + \sum_{j=1}^{r(A,A)} g_jT_j) = (I + \sum_{j=1}^{r(A,A)} g_jT_j)(A + \sum_{i=1}^{\nu(A,A)} f_i\Gamma_i) \tag{3.5.11}$$

is solvable for some $g_1, ..., g_{r(A,A)}, f_1, ..., f_{\nu(A,A)} \in \mathrm{H}_A$ which vanish at A. Clearly, the above system is trivially satisfied at $B = A$. The implicit function theorem implies that the above system is solved uniquely if the Jacobian of this system is nonsingular. Let $B = A + F$, $F = [f_{ij}] \in \mathbb{C}^{n\times n}$. Let $\alpha_i(F), \beta_j(F)$ be the linear terms of the Taylor expansions of $f_i(A+F), g_j(A+F)$. The linear part of (3.5.11) reduces to

$$F + \sum_{j=1}^{r(A,A)} \beta_j A T_j = \sum_{j=1}^{r(A,A)} \beta_j T_j A + \sum_{i=1}^{\nu(A,A)} \alpha_i \Gamma_i.$$

That is

$$F = \sum_{j=1}^{r(A,A)} \beta_j R_j + \sum_{i=1}^{\nu(A,A)} \alpha_i \Gamma_i.$$

In view of (3.5.8) $\alpha_1, ..., \alpha_{\nu(A,a)}, \beta_1, ..., \beta_{r(A,A)}$ are uniquely determined by F. □

Note that if $A = aI$ then the form (3.5.9) is not simpler than B. Also by mapping $T \to P^{-1}TP$ we get

$$\text{(3.5.12)} \qquad \mathbb{C}^{n\times n} = \text{Range } P^{-\hat{1}}AP \oplus \text{span } (\mathrm{P}^{-1}\Gamma_1\mathrm{P}, ..., \mathrm{P}^{-1}\Gamma_{\nu(\mathrm{A,A})}\mathrm{P}).$$

Lemma 3.5.4 *Let $B \in \mathrm{H}_A^{n\times n}$. Then the rational canonical form of B over $\mathcal{M}_A$ is a companion matrix $C(p)$, where $p(x) = \det\,(xI - B)$.*

Proof. The rational canonical form of B is $C(p_1, ..., p_k)$ is given by (2.3.3). We claim that $k = 1$. Otherwise $p(x)$ and $p'(x)$ have a common factor over $\mathcal{M}_A$. In view of Theorem 2.1.9 implies that $p(x)$ and $p'(x)$ have a common factor over H_A. That is any $B \in D(A, \rho)$ has at least one multiple eigenvalue. Evidently this is false. Consider $C = P^{-1}BP$ where $P \in \mathbb{C}^{n\times n}$ and $J = P^{-1}AP$ is the Jordan canonical form of A. So $C \in D(J, \rho')$. Choose C to be an upper diagonal. (This is possible since J is an upper diagonal matrix.) So the eigenvalues of C are the diagonal elements of C, and we can choose them to be pairwise distinct. Thus $p(x)$ and $p'(x)$ are coprime over $\mathcal{M}_A$, hence $k = 1$. Furthermore $p_1(x) = \det\,(xI - C(p)) = \det\,(xI - B)$. □

Theorem 3.5.5 *Let $A \in \mathbb{C}^{n\times n}$. Then $B \in \mathrm{H}_A^{n\times n}$ is similar to the companion matrix $C(p)$, $p(x) = \det\,(xI - B)$ over H_A if and only if $\nu(A, A) = n$. That is the minimal and the characteristic polynomial of A coincide, i.e. A is nonderogatory.*

Proof. Assume first that $C(B)$ in (3.5.1) can be chosen to be $C(p)$. Then for $B = A$ we obtain that A is similar to the companion matrix.

Corollary 2.8.4 yields $\nu(A,A)=n$. Assume now that $\nu(A,A)=n$. According to (2.8.12) we have that $i_1(x)=i_2(x)=...=i_{n-1}(x)=1$. That is, the minimal and the characteristic polynomials of A coincide, i.e. A is similar to a companion matrix. Use (3.5.9) to see that we may assume that A is a companion matrix. Choose $\Gamma_i=E_{ni},\ i=1,...,n$, where E_{ni} are defined in (3.5.6).

It is left to show that Range $\hat{A}\cap\mathrm{span}\,(\mathrm{E_{n1},...,E_{nn}})=\{0\}$. Suppose that $\Gamma=\sum_{i=1}^n\alpha_iE_{ni}\in$ Range $(\hat{A})$. Theorem 2.10.1 and Corollary 2.8.4 yield that $\mathrm{tr}\,\Gamma A^k=0,\ k=0,1,...,n-1$. Let $\alpha=(\alpha_1,...,\alpha_n)$. Since the first $n-1$ rows of Γ are zero rows we have

$$0=\mathrm{tr}\,\Gamma A^k=\alpha A^k\mathbf{e}_n,\quad \mathbf{e}_j=(\delta_{j1},...,\delta_{jn})^\top,\ j=1,...,n.$$

For $k=0$ the above equality implies that $\alpha_n=0$. Suppose that we already proved that these equalities for $k=0,...,\ell$ imply that $\alpha_n=...=\alpha_{n-l}=0$. Consider the equality $\mathrm{tr}\,\Gamma A^{\ell+1}=0$. Use Problem 2.4.9 to deduce

$$A^{\ell+1}\mathbf{e}_n=\mathbf{e}_{n-\ell-1}+\sum_{j=0}^{\ell}f_{(\ell+1)j}\mathbf{e}_{n-j}.$$

So $\mathrm{tr}\,\Gamma A^{\ell+1}=\alpha_{n-\ell-1}$ as $\alpha_n=...=\alpha_{n-\ell}=0$. Thus $\alpha_{n-\ell-1}=0$, which implies that $\Gamma=0$.

Theorem 3.5.3 yields that

$$C(B)=X^{-1}(B)BX(B)=A+\sum_{i=1}^n f_i(B)E_{ni}.$$

So $C(B)$ is a companion matrix. As $\det\,(xI-C(B))=\det\,(xI-B)$ it follows that $C(B)=C(p)$. $\Box$.

Problem 5 yields.

Lemma 3.5.6 *Let $A_i\in\mathbb{C}^{n_i\times n_i}$ $i=1,2$ and assume that*

$$\mathbb{C}^{n_i\times n_i}=\mathrm{Range}\ \hat{A}_i\oplus\mathrm{span}\,(\Gamma_1^{(i)},...,\Gamma^{(i)}_{\nu(\mathrm{A_i,A_i})}),\quad \mathrm{i}=1,2.$$

Suppose that A_1 and A_2 do not have a common eigenvalue. Then

$$\mathbb{C}^{(n_1+n_2)\times(n_1+n_2)}=\mathrm{Range}\ A_1\tilde{\oplus}A_2\oplus$$
$$\mathrm{span}\,(\Gamma_1^{(1)}\oplus 0,...,\Gamma^{(1)}_{\nu(\mathrm{A_1,A_1})}\oplus 0,0\oplus\Gamma_1^{(2)},...,0\oplus\Gamma^{(2)}_{\nu(\mathrm{A_2,A_2})}).$$

Theorem 3.5.7 *Let $A\in\mathbb{C}^{n\times n}$. Assume that* spec (A) *consists of ℓ distinct eigenvalues $\lambda_1,...,\lambda_\ell$, where the multiplicity of λ_i is n_i for $i=1,...,\ell$. Then B is similar over* H_A *to the matrix*

$$C(B) = \sum_{i=1}^{\ell} \oplus C_i(B),\ C_i(B) \in \mathrm{H}_A^{n_i \times n_i},\ (\lambda_i I_{n_i} - C_i(A))^{n_i} = 0,\ i = 1, ..., \ell. \tag{3.5.13}$$

Moreover $C(A)$ is the Jordan canonical form of A.

Proof. Choose P in the equality (3.5.9) such that $P^{-1}AP$ is the Jordan canonical of A and each $P^{-1}\Gamma_i P$ is of the form $\sum_{j=1}^{\ell} \Gamma_i^{(j)}$ as follows from Lemma 3.5.6. Then (3.5.9) yields the theorem. □

Problems

1. Let $A = \sum_{i=1}^{k} \oplus H_{n_i},\ n = \sum_{i=1}^{k} n_i$. Partition any $B \in \mathbb{C}^{n\times n}$ as a block matrix as A: $B = [B_{ij}],\ B_{ij} \in \mathbb{C}^{n_i \times n_j},\ i, j = 1, ..., k$. Using the results of Theorem 2.8.3 and Theorem 2.10.1 show that the matrices

$$\begin{aligned}
&\Gamma_{\alpha,\beta,\gamma} = (\Gamma_{ij}^{(\alpha,\beta,\gamma)})_1^k \in \mathbb{C}^{n\times n},\\
&\Gamma_{ij}^{(\alpha,\beta,\gamma)} = 0 \in \mathbb{C}^{n_i\times n_j}, \quad \text{if } (\alpha,\beta) \neq (i,j),\\
&\Gamma_{\alpha\beta}^{(\alpha,\beta,\gamma)} = E_{n_\alpha \gamma} \in \mathbb{C}^{n_\alpha \times n_\beta},\\
&\gamma = 1, ..., \min(n_\alpha, n_\beta),\ \alpha, \beta = 1, ..., k,
\end{aligned}$$

satisfy (3.5.8).

2. Let A be a matrix given by (2.8.4). Use Theorem 3.5.7 and Problem 1 to find a set of matrices $\Gamma_1, ..., \Gamma_{\nu(A,A)}$ which satisfy (3.5.8).

3. Let $A \in \mathbb{C}^{n\times n}$ and assume that λ_i is a simple eigenvalue of A, i.e. λ_i is a simple root of the characteristic polynomial of A. Use Theorem 2.8.3 to show the existence of $\lambda(B) \in \mathrm{H}_A$ such that $\lambda(B)$ is an eigenvalue of B and $\lambda(A) = \lambda_i$.

4. Let A satisfy the assumptions of Theorem 3.5.7. Denote by D_p an open set satisfying the assumptions of Theorem 3.4.4 for $p = 1, ..., \ell$. Let $P_k(B)$ be the projection of $B \in \mathrm{H}_A^{n\times n}$ on $D_k,\ k = 1, ..., \ell$. Problem 10 implies that $P_k(B) \in \mathrm{H}_A^{n\times n},\ k = 1, ..., \ell$. Let $P_k(A)\mathbb{C}^n = \mathrm{span}\,(\mathbf{x}^{\mathrm{k}1}, ..., \mathbf{x}^{\mathrm{kn_k}}),\ \mathrm{k} = 1, ..., \ell,\ \mathrm{B} \in \mathrm{D(A}, \rho)$, where ρ is some positive number. Let $X(B) \in \mathrm{H}_A^{n\times n}$ be formed by the columns $P_k(B)\mathbf{x}^{k1}, ..., P_k(B)\mathbf{x}^{kn_k},\ k = 1, ..., \ell$. Show that $C(B)$ given by (3.5.1) satisfies (3.5.13). (This yields another proof of Theorem 3.5.7.)

3.6 Analytic, pointwise and rational similarity

Definition 3.6.1 *Let* $\Omega \subset \mathbb{C}^m$ *and* $A, B \in \mathrm{H}(\Omega)^{n\times n}$. *Then*

(a) *A and B are called analytically similar, denoted by* $A\overset{a}{\approx}B$, *if A and B are similar over* $\mathrm{H}(\Omega)$.

(b) *A and B are called pointwise similar, denoted by* $A\overset{p}{\approx}B$, *if* $A(x)$ *and* $B(x)$ *are similar over* $\mathbb{C}$ *for all* $x \in \Omega_0$, *for some open set* $\Omega_0 \supset \Omega$.

(c) *A and B are called rationally similar, denoted by* $A\overset{r}{\approx}B$, *if A and B are similar over the field of meromorphic functions* $\mathcal{M}(\Omega)$.

Theorem 3.6.2 *Let* $\Omega \subset \mathbb{C}^m$ *and assume that* $A, B \in \mathrm{H}(\Omega)^{n\times n}$. *Then*

$$A\overset{a}{\approx}B \Rightarrow A\overset{p}{\approx}B \Rightarrow A\overset{r}{\approx}B.$$

Proof. Suppose that

$$B(x) = P^{-1}(x)A(x)P(x), \tag{3.6.1}$$

where $P, P^{-1} \in \mathrm{H}(\Omega)^{n\times n}$. Let $x_0 \in \Omega$. Then (3.6.1) holds in some neighborhood of x_0. So $A\overset{p}{\approx}B$. Assume now that $A\overset{p}{\approx}B$. Let $C(p_1, ..., p_k)$ and $C(q_1, ..., q_\ell)$ be the rational canonical forms of A and B respectively over $\mathcal{M}(\Omega)$. Then

$$C(p_1, ..., p_k) = S(x)^{-1}A(x)S(x), \quad C(q_1, ..., q_\ell) = T(x)^{-1}B(x)T(x),$$
$$S(x), T(x) \in \mathrm{H}(\Omega)^{n\times n}, \quad \det A(x) \not\equiv 0, \ \det B(x) \not\equiv 0.$$

Theorem 2.3.8 yields that $C(p_1, ..., p_k), C(q_1, ..., q_\ell) \in \mathrm{H}(\Omega)^{n\times n}$. Let $\Omega_0 \supset \Omega$ be an open set such that $A, B, S, T \in \mathrm{H}(\Omega_0)^{n\times n}$ and $A(x)$ and $B(x)$ are similar over $\mathbb{C}$ for any $x \in \Omega_0$. Let $x_0 \in \Omega_0$ be a point such that $\det S(x_0)T(x_0) \neq 0$. Then for all $x \in D(x_0, \rho)$ $C(p_1, ..., p_k) = C(q_1, ..., q_\ell)$. The analyticity of $C(p_1, ..., p_k)$ and $C(q_1, ..., q_\ell)$ imply that these matrices are identical in $\mathrm{H}(\Omega)$, i.e. $A\overset{r}{\approx}B$. □

Assume that $A\overset{a}{\approx}B$. Then according to Lemma 2.9.4 the three matrices

$$I \otimes A(x) - A(x)^\top \otimes I, \quad I \otimes A(x) - B(x)^\top \otimes I, \quad I \otimes B(x) - B(x)^\top \otimes I \tag{3.6.2}$$

are equivalent over $\mathrm{H}(\Omega)$. Theorem 2.9.3 yields.

Theorem 3.6.3 *Let* $A, B \in \mathrm{H}(\Omega)^{n\times n}$. *Assume that the three matrices in* (3.6.2) *are equivalent over* $\mathrm{H}(\Omega)$. *Then* $A\overset{p}{\approx}B$.

Assume that $\Omega \subset \mathbb{C}$ is a domain. Then $\mathrm{H}(\Omega)$ is $\mathbb{EDD}$. Hence we can determine when these matrices are equivalent.

The problem of finding a canonical form of $A \in \Omega^{n\times n}$ under analytic similarity is a very hard problem. This problem for the ring of local analytic functions in one variables will be discussed in the next sections. We now determine when A is analytically similar to its rational canonical form over H_ζ, the ring of local analytic functions in the neighborhood of $\zeta \in \mathbb{C}^m$.

For $A, B \in \mathrm{H}(\Omega)^{n\times n}$ denote by $r(A,B)$ and $\nu(A,B)$ the rank and the nullity of the matrix $C = I \otimes A - B^\top \otimes I$ over the field $\mathcal{M}(\Omega)$. Denote by $r(A(x), B(x))$ and $\nu(A(x), B(x))$ the rank of $C(x)$ over $\mathbb{C}$. As the rank of $C(x)$ is the largest size of a nonvanishing minor, we deduce

$$r(A(\zeta), B(\zeta)) \le r(A(x), B(x)) \le r(A,B)$$

(3.6.3)

$$\nu(A,B) \le \nu(A(x), B(x)) \le \nu(A(\zeta), B(\zeta)), \quad x \in D(\zeta,\rho)$$

for some positive ρ. Moreover for any $\rho > 0$ there exists at least one $x_0 \in D(\zeta,\rho)$ such that

$$r(A(x_0), B(x_0)) = r(A,B), \quad \nu(A(x_0, B(x_0)) = \nu(A,B). \tag{3.6.4}$$

Theorem 3.6.4 *Let $\zeta \in \mathbb{C}^m$ and $A \in H_\zeta^{n\times n}$. Assume that $C(p_1, ..., p_k)$ is the rational canonical form of A over $\mathcal{M}_\zeta$ and $C(\sigma_1, ..., \sigma_\ell)$ is the rational canonical form of $A(\zeta)$ over $\mathbb{C}$. That is $p_i = p_i(\lambda, x)$ and $\sigma_j(\lambda)$ are normalized polynomials in λ belonging to $\mathrm{H}_\zeta[\lambda]$ and $\mathbb{C}[\lambda]$ respectively for $i = 1, ..., k$ and $j = 1, ..., \ell$. Then*
(a) $\ell \ge k$;
(b) $\prod_{i=0}^q \sigma_{\ell-i}(\lambda) | \prod_{i=0}^q p_{k-i}(\lambda,\zeta)$ for $q = 0, 1, ..., k-1$.
Moreover $\ell = k$ and $p_i(\lambda,\zeta) = \sigma_i(\lambda)$ for $i = 1, ..., k$ if and only if

$$r(A(\zeta), B(\zeta)) = r(A,B), \quad \nu(A(\zeta), B(\zeta)) = \nu(A,B), \tag{3.6.5}$$

which is equivalent to the condition

$$\begin{aligned} r(A(\zeta), B(\zeta)) &= r(A(x), B(x)), \\ \nu(A(\zeta), B(\zeta)) &= \nu(A(x), B(x)), \quad x \in D(\zeta,\rho) \end{aligned} \tag{3.6.6}$$

for some positive ρ.

Proof. Let

$$u_{n-k+i}(\lambda, x) = \prod_{\alpha=1}^{i} p_\alpha(\lambda, x), \ v_{n-\ell+j}(\lambda) = \prod_{\beta=1}^{j} \sigma_\beta(\lambda),$$
$$i = 1, ..., k, \ j = 1, ..., \ell,$$
$$u_\alpha(\lambda, x) = v_\beta(\lambda) = 1, \quad \text{for} \quad \alpha \le n-k, \ \beta \le n-\ell.$$

So $u_i(\lambda, x)$ and $v_i(\lambda)$ are the g.c.d. of all minors of order i of matrices $\lambda I - A$ and $\lambda I - A(\zeta)$ over the rings $\mathcal{M}_\zeta[\lambda]$ and $\mathbb{C}[\lambda]$ respectively. As $u_i(\lambda, x) \in \mathrm{H}_\zeta[\lambda]$ it is clear that $u_i(\lambda, \zeta)$ divides all the minors of $I - A(\zeta)$ of order i. So $u_i(\lambda, \zeta)|v_i(\lambda)$ for $i = 1, ..., n$. Since $v_{n-\ell} = 1$ it follows that $u_{n-\ell}(\lambda, x) = 1$. Hence $k \le \ell$. Furthermore

$$u_n(\lambda, x) = \det\,(\lambda I - A(x)), \quad v_n(\lambda) = \det\,(\lambda I - A(\zeta)).$$

Therefore $u_n(\lambda, \zeta) = v_n(\lambda)$ and $\frac{v_n(\lambda)}{v_i(\lambda)}|\frac{u_n(\lambda,\zeta)}{u_i(\lambda,\zeta)}$. This establishes claims (a) and (b) of the theorem. Clearly if $C(q_1, ..., q_\ell) = C(p_1, ..., p_k)(\zeta)$ then $k = \ell$ and $p_i(\lambda, \zeta) = q_i(\lambda)$ for $i = 1, ..., \ell$. Assume now that (3.6.5) holds. According to (2.8.12)

$$\nu(A, A) = \sum_{i=1}^{k}(2i-1)\deg\, p_{k-i+1}(\lambda, x),$$

$$\nu(A(\zeta), A(\zeta)) = \sum_{j=1}^{\ell}(2j-1)\deg\, q_{\ell-j+1}(\lambda).$$

Note that the degrees of the invariant polynomials of $\lambda I - A$ and $\lambda I - A(\zeta)$ satisfy the assumptions of Problem 2. From the results of Problem 2 it follows that the second equality in (3.6.5) holds if and only if $k = \ell$ and $\deg p_i(\lambda, x) = \deg q_i(\lambda)$ for $i = 1, ..., k$. Finally (3.6.3-3.6.4) imply the equivalence of the conditions of (3.6.5) and (3.6.6). □

Corollary 3.6.5 *Let* $A \in \mathrm{H}_\zeta^{n\times n}$. *Assume that* (3.6.6) *holds. Then* $A\overset{a}{\approx}B$ *if and only if* $A\overset{p}{\approx}B$.

Proof. According to Theorem 3.6.2 it is enough to show that $A\overset{p}{\approx}B$ implies that $A\overset{a}{\approx}B$. Since A satisfies (3.6.6) the assumption that $A\overset{p}{\approx}B$ implies that B satisfies (3.6.6) too. According to Theorem 3.6.4 A and B are analytically similar to their canonical rational form. From Theorem 3.6.2 it follows that A and B have the same rational canonical form. □

Problems

1. Let
$$A(x) = \begin{bmatrix} 0 & x \\ 0 & 0 \end{bmatrix}, \quad B(x) = \begin{bmatrix} 0 & x^2 \\ 0 & 0 \end{bmatrix}.$$

Show that $A(x)$ and $B(x)$ are rationally similar over $\mathbb{C}(x)$ to $H_2 = A(1)$. Prove that

$$A \overset{p}{\not\approx} H_2,\ B \overset{p}{\not\approx} H_2,\ A \overset{p}{\approx} B,\ A \overset{a}{\not\approx} B$$

over $\mathbb{C}[x]$.

2. Let n be a positive integer and assume that $\{\ell_i\}_1^n,\ \{m_i\}_1^n$ are two nonincreasing sequences of nonnegative integers satisfying

$$\sum_{i=1}^{k} \ell_i \le \sum_{i=1}^{k} m_i, \quad k = 1, ..., n-1,$$
$$\sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} m_i.$$

Show (by induction) that

$$\sum_{i=1}^{n} (2i-1)m_i \le \sum_{i=1}^{n} (2i-1)\ell_i$$

and equality holds if and only if $\ell_i = m_i,\ i = 1, ..., n$.

3. Let $\zeta_n \in \mathbb{C},\ n = 1, ...,$ and $\lim_{n\to\infty} \zeta_n = \zeta$. Suppose that $\Omega \subset \mathbb{C}$ is a connected set and $\zeta_n \in \Omega,\ n = 1, ...,\ \zeta \in \Omega$. Recall that if $f \in \mathrm{H}(\Omega)$ and $f(\zeta_n) = 0,\ n = 1, ...,$ then $f = 0$. Show that for $A, B \in \mathrm{H}(\Omega)^{n\times n}$ the assumption that $A(\zeta_n) \approx B(\zeta_n),\ n = 1, ...,$ implies that $A \overset{r}{\approx} B$.

3.7 A Global Splitting

From this section to the end of the chapter we assume that Ω is a domain in $\mathbb{C}$. We now give a global version of Theorem 3.5.7.

Theorem 3.7.1 *Let $A \in \mathrm{H}(\Omega)^{n\times n}$. Suppose that*

$$\det\,(\lambda I - A(x)) = \phi_1(\lambda, x)\phi_2(\lambda, x), \tag{3.7.1}$$

where ϕ_1, ϕ_2 are two nontrivial normalized polynomials in $\mathrm{H}(\Omega)[\lambda]$ of positive degrees n_1 and n_2 respectively. Assume that $(\phi_1(\lambda, x_0), \phi_2(\lambda, x_0)) = 1$ for each $x_0 \in \Omega$. Then there exists $X \in \mathbf{GL}(n, \mathrm{H}(\Omega))$ such that

$$\begin{gathered} X^{-1}(x)C(x)X(x) = C_1(x) \oplus C_2(x), \\ C_i(x) \in \mathrm{H}(\Omega)^{n_i\times n_i}, \quad \det\,(\lambda I - C_i(x)) = \phi_i(\lambda, x), \quad i = 1, 2. \end{gathered} \tag{3.7.2}$$

Proof. Let $P_i(x)$ be the projection of $A(x)$ on the eigenvalues of $A(x)$ satisfying $\phi_i(\lambda, x) = 0$. Since $(\phi_1(\lambda, x_0), \phi_2(\lambda, x_0)) = 1$ it follows that $P_i(x) \in \mathrm{H}(\Omega)^{n\times n}$ for $i = 1, 2$. (See Problem 3.4.10.) Also for any x_0 the rank of $P_i(x_0)$ is n_i. Since $\mathrm{H}(\Omega)$ is $\mathbb{EDD}$ each $P_i(x)$ can be brought to the Smith normal form

$$P_i(x) = U_i(x)\,\mathrm{diag}(\epsilon_1^{(i)}(x), ..., \epsilon_{n_i}^{(i)}(x), 0, ..., 0)V_i(x)),$$
$$U_i, V_i \in \mathbf{GL}(n_i, \mathrm{H}(\Omega)),\ i = 1, 2.$$

As $\mathrm{rank}\ \mathrm{P_i(x_0) = n_i}$ for any $x_0 \in \Omega$ we deduce that $\epsilon_j^{(i)} = 1,\ j = 1, ..., n,\ i = 1, 2$. Let $\mathbf{u}_1^{(i)}(x), ..., \mathbf{u}_n^{(i)}(x)$ be the columns of $U_i(x)$ for $i = 1, 2$. As $V \in \mathbf{GL}(n, \mathrm{H}(\mathit{\Omega})$ we obtain

$$P_i(x)\mathbb{C}^n = \mathrm{span}\ (\mathbf{u}_1^{(\mathrm{i})}(\mathrm{x}), ..., \mathbf{u}_{\mathrm{n_i}}^{(\mathrm{i})}(\mathrm{x})), \tag{3.7.3}$$

for any $x \in \Omega$. Let

$$X(x) = [\mathbf{u}_1^{(1)}(x), ..., \mathbf{u}_{n_1}^{(1)}(x), \mathbf{u}_1^{(2)}(x), ..., \mathbf{u}_{n_2}^{(2)}(x)] \in \mathrm{H}(\Omega)^{n\times n}.$$

According to Problem 3.4.13 det $X(x_0) \neq 0$ for any $x_0 \in \mathrm{H}(\Omega)$. So $X(x) \in \mathbf{GL}(n, \mathrm{H}(\mathit{\Omega}))$. Then (3.7.2) follows from (3.4.29). □

3.8 First variation of a geometrically simple eigenvalue

Theorem 3.8.1 *Let $A(x)$ be a continuous family of $n \times n$ complex valued matrices for $|x - x_0| < \delta$, where the parameter x is either real or complex. Suppose that*

$$A(x) = A_0 + (x - x_0)A_1 + |x - x_0|o(1). \tag{3.8.1}$$

Assume furthermore that λ_0 is a geometrically simple eigenvalue of A_0 of multiplicity m. Let $\mathbf{x}_1, ..., \mathbf{x}_m$ and $\mathbf{y}_1, ..., \mathbf{y}_m$ be eigenvectors of A_0 and $A_0^\top$ respectively corresponding to λ_0, which form a biorthonormal system $\mathbf{y}_i^\top \mathbf{x}_j = \delta_{ij},\ i, j = 1, ..., m$. Then it is possible to enumerate the eigenvalues of $A(x)$ by $\lambda_i(x),\ i = 1, ..., n$, such that

$$\lambda_i(x) = \lambda_0 + (x - x_0)\mu_i + |x - x_0|o(1), \quad i = 1, ..., m, \tag{3.8.2}$$

where $\mu_1, ..., \mu_m$ are the eigenvalues of the matrix

$$S = [s_{ij}] \in \mathbb{C}^{m\times m}, \quad s_{ij} = \mathbf{y}_i{}^\top A_1 \mathbf{x}_j,\ i, j = 1, ..., m. \tag{3.8.3}$$

Proof. By considering the matrix $P^{-1}A(x)P$, for an appropriate $P \in \mathbf{GL}(n,\mathbb{C})$, we can assume that A_0 is in the Jordan canonical form such that the first m diagonal entries of A_0 are λ_0. The proofs of Theorems 3.5.3 and 3.5.7 implies the existence of

$$X(B) = I + Z(B), \quad Z \in \mathrm{H}_0^{n\times n}, \quad Z(0) = 0,$$

such that

$$X^{-1}(B)(A_0 + B)X(B) = \sum_{i=1}^{\ell} \oplus C_i(B), \quad C_1(0) = \lambda_0 I_m. \tag{3.8.4}$$

Substituting

$$\begin{aligned} B(x) &= A(x) - A_0 = (x-x_0)A_1 + |x-x_0|o(1), \\ X(x) &= X(B(x)) = I + (x-x_0)X_1 + |x-x_0|o(1) \end{aligned}$$

we get

$$C(X) = X^{-1}A(x)X(x) = A_0 + (x-x_0)(A_1 + A_0X_1 - X_1A_0) + |x-x_0|o(1).$$

According (3.8.4) $\lambda_1(x), ..., \lambda_m(x)$ are the eigenvalues of $C_1(B(x))$. As $C_1(B(x_0)) = \lambda_0 I_m$, by considering $(C_1(B(x)) - \lambda_0 I_m)/(x-x_0)$ we deduce that $(\lambda_i(x) - \lambda_0)/(x-x_0)$ are continuous functions at x_0. Also

$$(C_1(B(x)) - \lambda_0 I_m)/(x-x_0) = [\mathbf{v}_i{}^\top (A_1 + A_0X_1 - X_1A_0)\mathbf{u}_j]_{i,j=1}^m + o(1),$$

where $\mathbf{u}_i = \mathbf{v}_i = (\delta_{i1}, ..., \delta_{in})^\top$ for $i = 1, ..., m$. Since $\mathbf{u}_i$ and $\mathbf{v}_i$ are the eigenvectors of A_0 and $A_0^\top$ respectively corresponding to λ_0 for $i = 1, ..., m$, it follows that $\mathbf{v}_i^\top(A_0X_1 - X_1A_0)\mathbf{u}_j = 0$ for $i, j = 1, ..., m$. This establishes the result for a particular choice of eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_m$ and $\mathbf{v}_1, ..., \mathbf{v}_m$. It is left to note that any other choice of the eigenvectors $\mathbf{x}_1, ..., \mathbf{x}_m$ and $\mathbf{y}_1, ..., \mathbf{y}_m$, which form a biorthonormal system amounts to a new matrix S_1 which is similar to S. In particular S and S_1 have the same eigenvalues. □

Problems

1. Let $A(x) = \begin{bmatrix} 0 & 1 \\ x & 0 \end{bmatrix}$. Find the eigenvalues and the eigenvectors of $A(x)$ in terms of $\sqrt{x}$. Show that (3.8.2) does not apply for $x_0 = 0$ in this case. Let $B(x) = A(x^2)$. Show that (3.8.2) holds for x_0 even though $\lambda_0 = 0$ is not geometrically simple for $B(0)$.

3.9 Analytic similarity over H_0

Let $A, B \in \mathrm{H}_0^{n\times n}$. That is

$$A(x) = \sum_{k=0}^{\infty} A_k x^k, \quad |x| < r(A),$$

(3.9.1)

$$B(x) = \sum_{k=0}^{\infty} B_k x^k, \quad |x| < r(B).$$

Definition 3.9.1 *For $A, B \in \mathrm{H}_0^{n\times n}$ let $\eta(A,B)$ and $\kappa_p(A,B)$ be the index and the number of local invariant polynomials of degree p of the matrix $I_n \otimes A(x) - B(x)^\top \otimes I_n$ respectively.*

Theorem 3.9.2 *Let $A, B \in \mathrm{H}_0^{n\times n}$. Then A and B are analytically similar over H_0 if and only if A and B are rationally similar over H_0 and there exists $\eta(A,A)+1$ matrices $T_0, ..., T_\eta \in \mathbb{C}^{n\times n}$ $(\eta = \eta(A,A))$, such that* $\det T_0 \neq 0$ *and*

$$\sum_{i=0}^{k} A_i T_{k-i} - T_{k-i} B_i = 0, \quad k = 0, ..., \eta(A,A). \tag{3.9.2}$$

Proof. The necessary part of the theorem is obvious. Assume now that $A(x) \overset{r}{\approx} B(x)$ and the matrices $T_0, ..., T_\eta$ satisfy (3.9.2), where $T_0 \in \mathbf{GL}(n, \mathbb{C})$. Put

$$C(x) = T(x)B(x)T^{-1}(x), \quad T(x) = \sum_{k=0}^{\eta} T_k x^k.$$

As $\det T_0 \neq 0$ we deduce that $B(x) \overset{a}{\approx} C(x)$. Hence $A(x) \overset{r}{\approx} C(x)$. In particular $r(A,A) = r(A,C)$. Also (3.9.2) is equivalent to $A(x) - C(x) = x^{\eta+1}O(1)$. Thus

$$(I_n \otimes A(x) - A(x)^\top \otimes I_n) - (I_n \otimes A(x) - C(x)^\top \otimes I_n) = x^{\eta+1}O(1).$$

In view of Lemma 1.14.2 the matrices $(I_n \otimes A(x) - A(x)^\top \otimes I_n), (I_n \otimes A(x) - C(x)^\top \otimes I_n)$ are equivalent over H_0. In particular $\eta(A,A) = \eta(A,C)$. Also $I, 0, ..., 0$ satisfy the system (3.9.2) where $B_i = C_i, i = 0, 1, ..., \eta$. Theorem 1.14.3 yields the existence $P(x) \in \mathrm{H}_0^{n\times n}$ such that

$$A(x)P(x) - P(x)C(x) = 0, \quad P(0) = I.$$

Hence $A(x)\overset{a}{\approx}C(x)$. By the definition $C(x)\overset{a}{\approx}B(x)$. Therefore $A(x)\overset{a}{\approx}B(x)$. □

Note that if $\eta(A,A)=0$ the assumptions of Theorem 3.9.2 are equivalent to $A(x)\overset{r}{\approx}B(x)$. Then the implication that $A(x)\overset{a}{\approx}B(x)$ follows from Corollary 3.6.5.

Suppose that the characteristic polynomial of $A(x)$ splits over H_0. That is

$$(3.9.3) \qquad \det\,(\lambda I - A(x)) = \prod_{i=1}^{n}(\lambda-\lambda_i(x)), \quad \lambda_i(x)\in \mathrm{H}_0,\ i=1,...,n.$$

As H_0 is $\mathbb{ED}$ Theorem 2.5.4 yields that $A(x)$ is similar to an upper triangular matrix. Using Theorem 3.5.7 and Theorem 2.5.4 we obtain that $A(x)$ is analytically similar to

$$\begin{aligned} &C(x)=\oplus_{i=1}^{\ell}C_i(x), \quad C_i(x)\in \mathrm{H}_0^{n_i\times n_i},\\ (3.9.4)\qquad &\\ &(\alpha_i I_{n_i}-C_i(0))^{n_i}=0,\ \alpha_i=\lambda_{n_i}(0),\ \alpha_i\neq\alpha_j \text{ for } i\neq j,\ i,j=1,...,\ell. \end{aligned}$$

Furthermore each $C_i(x)$ is an upper triangular matrix. In what follows we are more specific on the form of the upper triangular matrix.

Theorem 3.9.3 *Let $A(x)\in \mathrm{H}_0^{n\times n}$. Assume that the characteristic polynomial of $A(x)$ splits in H_0.Then $A(x)$ is analytically similar to a block diagonal matrix $C(x)$ of the form* (3.9.4) *such that each $C_i(x)$ is an upper triangular matrix whose off-diagonal entries are polynomial in x. Moreover, the degree of each polynomial entry above the diagonal in the matrix $C_i(x)$ does not exceed $\eta(C_i,C_i)$ for $i=1,...,\ell$.*

Proof. In view of Theorem 3.5.7 we may assume that $\ell=1$. That is, $A(0)$ has one eigenvalue α_0. Furthermore, by considering $A(x)-\alpha_0 I$ we may assume that $A(0)$ is nilpotent. Also in view of Theorem 3 we may assume that $A(x)$ is already in the upper triangular form. Suppose in addition to all the above assumptions $A(x)$ is nilpotent. Define

$$\mathbf{X}_k=\{\mathbf{y}:\quad A^k\mathbf{y}=\mathbf{0},\ \mathbf{y}\in\mathrm{H}_0^n\},\quad k=0,1,...,.$$

Then

$$\{\mathbf{0}\}=\mathbf{X}_0\subsetneqq\mathbf{X}_1\subsetneqq\mathbf{X}_2\subsetneqq\ldots\subsetneqq\mathbf{X}_p=\mathrm{H}_0^n.$$

Using Theorem 1.12.3 one can show the existence of a basis $\mathbf{y}_1(x),...,\mathbf{y}_n(x)$ in H_0^n, such that $\mathbf{y}_1(x),...,\mathbf{y}_{\psi_k}(x)$ is a basis in $\mathbf{X}_k$ for $k=1,...,p$. As

$A(x)\mathbf{X}_{k+1} \subset \mathbf{X}_k$ we have

$$A(x)\mathbf{y}_j = \sum_{i=1}^{\psi_k} g_{ij}\mathbf{y}_i(x), \quad \psi_k < j \le \psi_{k+1}.$$

Define $g_{ij} = 0$ for $i > \psi_k$ and $\psi_k < j \le \psi_{k+1}$. Put

$$G(x) = [g_{ij}]_1^n,\ T(x) = [\mathbf{y}_1(x), ..., \mathbf{y}_n(x)] \in \mathrm{H}_0^{n\times n}.$$

Since $\mathbf{y}_1(x), ..., \mathbf{y}_n(x)$ is a basis in H_0^n we deduce that $T(x) \in \mathbf{GL}(n, \mathrm{H}_0)$. Hence

$$G(x) = T^{-1}(x)A(x)T(x), \quad s = \eta(A, A) = \eta(G, G).$$

Let

$$G(x) = \sum_{j=0}^{\infty} G_j x^j, \quad G^{(k)} = \sum_{j=0}^{k} G_j x^j, \quad k = 0, 1, ..., .$$

We claim that $G^{(s)} \overset{a}{\approx} G(x)$. First note that

$$(I_n \otimes G(x) - G(x)^\top \otimes I_n) - (I_n \otimes G^{(s)}(x) - G^{(s)}(x)^\top \otimes I_n) = x^{s+1}O(1).$$

Lemma 1.14.2 implies that the matrices $(I_n \otimes G(x) - G(x)^\top \otimes I_n)$, $(I_n \otimes G^{(s)}(x) - G^{(s)}(x)^\top \otimes I_n)$ have the same local invariant polynomial up to the degree s. So $r(G, G) \le r(G^{(s)}, G^{(s)})$ which is equivalent to

$$\nu(G^{(s)}, G^{(s)}) \le \nu(G, G). \tag{3.9.5}$$

Let

$$\mathbf{Y}_k = \{\mathbf{y} = (y_1, ..., y_n)^\top : \quad y_j = 0 \text{ for } j > \psi_k\}, \quad k = 0, ..., p.$$

Clearly if $g_{ij} = 0$ then $(i, j) - th$ entry of $G^{(s)}$ is also equal to zero. By the definition $g_{ij}(x) = 0$ for $i > \psi_k$ and $\psi_k < j < \psi_{k+1}$. So $G^{(s)}(x)\mathbf{Y}_{k+1} \subset \mathbf{Y}_k$ for $k = 0, ..., p-1$. Theorem 2.11.2 implies

$$\nu(G(x_0), G(x_0)) \le \nu(G^{(s)}(x_0), G^{(s)}(x_0)) \tag{3.9.6}$$

for all x_0 in the neighborhood of the origin. Hence $\nu(G, G) \le \nu(G^{(s)}, G^{(s)})$. This establishes equality in (3.9.5), which in return implies equality in (3.9.6) for $0 < |x_0| < \rho$. Theorem 2.11.2 yields that $G(x_0) \approx G^{(s)}(x_0)$ for $0 < |x_0| < \rho$. From Theorem 3.6.2 we deduce that $G \overset{r}{\approx} G^{(s)}$. As $G(x)I - IG^{(s)} = x^{s+1}O(1)$ Theorem 3.9.2 implies that $G \overset{a}{\approx} G^{(s)}$. This establishes the theorem in case that $A(x)$ is a nilpotent matrix.

We now consider the general case where $A(x)$ is an upper triangular matrix. Without loss of generality we may assume that $A(x)$ is of the form

$$\begin{aligned} &A(x) = [A_{ij}]_1^\ell, \quad A_{ij} \in \mathrm{H}_0^{n_i \times n_j}, \\ &A_{ij}(x) = 0 \text{ for } j < i, \quad (A_{ii}(x) - \lambda_i(x) I_{n_i})^{n_i} = 0, \\ &\lambda_i \not\equiv \lambda_j(x), \text{ for } i \neq j, \quad i, j = 1, ..., \ell. \end{aligned} \tag{3.9.7}$$

We already showed that

$$A_{ii}(x) = T_i(x)^{-1} F_{ii}(x) T_i(x), \quad T_i \in \mathbf{GL}(n, \mathrm{H_o}),$$

and each $F_{ii}(x) - \lambda_i(x) I_{n_i}$ is a nilpotent upper triangular matrix with polynomial entries of the form described above. Let

$$T(x) = \sum_{i=1}^{\ell} T_i(x), \quad G(x) = [G_{ij}(x)]_1^\ell = T(x)^{-1} A(x) T(x).$$

As $\lambda_i(x) \not\equiv \lambda_j(x)$ for $i \neq j$ Problem 3 implies $\nu(G, G) = \sum_{i=1}^{\ell} \nu(G_{ii}, G_{ii})$. Let $G^{(k)}(x) = [G_{ij}^{(k)}]$ be defined as above. Theorem 2.10.2 implies

$$\nu(G^{(k)}, G^{(k)}) \geq \sum_{i=1}^{\ell} \nu(G_{ii}^{(k)}, G_{ii}^{(k)}).$$

Using Theorem 2.11.2 as above we obtain $\nu(G_{ii}, G_{ii}) \leq \nu(G_{ii}^{(k)}, G_{ii}^{(k)})$. Combine the above inequalities we obtain $\nu(G, G) \leq \nu(G^{(s)}, G^{(s)})$. Compare this inequality with the inequality (3.9.5) to deduce equality in (3.9.5). Hence

$$\nu(G_{ii}^{(s)}, G_{ii}^{(s)}) = \nu(G_{ii}, G_{ii}), \quad i = 1, ..., \ell. \tag{3.9.8}$$

Let

$$\begin{aligned} &D_i(x) = \lambda_i(x) I_{n_i} = \sum_{j=0}^{\infty} D_{ij} x^j, \quad D_i^{(k)} = \sum_{j=0}^{k} D_{ij} x^j, \\ &D(x) = \oplus_{i=1}^{\ell} D_i(x), \quad D^{(k)}(x) = \oplus_{i=1}^{\ell} D_i^{(k)}(x). \end{aligned} \tag{3.9.9}$$

Then (3.9.8) is equivalent to

$$\nu(G_{ii}^{(s)} - D_i^{(s)}, G_{ii}^{(s)} - D_i^{(s)}) = \nu(G_{ii} - D_{ii}, G_{ii} - D_{ii}), \quad i = 1, ..., \ell.$$

As above Theorem 2.11.2 yields that $G_{ii}^{(s)} - D_i^{(s)} \overset{r}{\approx} G_{ii} - D_i \Rightarrow G_{ii}^{(s)} - D_i^{(s)} + D_i \overset{r}{\approx} G_{ii}$. Since $\lambda_i(x) \not\equiv \lambda_j(x)$ for $i \neq j$ we finally deduce that

$G\overset{r}{\approx}G^{(s)}-D^{(s)}+D$. Also $GI-I(G^{(s)}-D^{(s)}+D)=x^{s+1}O(1)$. Theorem 3.9.2 yields $G\overset{a}{\approx}G^{(s)}-D^{(s)}+D$. The proof of the theorem is completed. □

Theorem 3.9.4 *Let $P(x)$ and $Q(x)$ be matrices of the form* (3.9.4)

$$\begin{aligned}&P(x)=\sum_{i=1}^{p}\oplus P_i(x),\ P_i(x)\in\mathrm{H}_0^{m_i\times m_i},\\&(\alpha_iI_{m_i}-P_i(0))^{m_i}=0,\ \alpha_i\ne\alpha_j \text{ for } i\ne j,\ i,j=1,...,p,\\&Q(x)=\oplus_{j=1}^{q}Q_j(x),\ Q_j(x)\in\mathrm{H}_0^{n_j\times n_j},\\&(\beta_jI_{n_j}-Q_j(0))^{n_j}=0,\ \beta_i\ne\beta_j \text{ for } i\ne j,\ i,j=1,...,q.\end{aligned}\tag{3.9.10}$$

Assume furthermore that

$$\alpha_i=\beta_i,\ i=1,...,t,\ \alpha_j\ne\beta_j,\quad i=t+1,...,p,\ j=t+1,...,q,\ 0\le t\le\min(p,q).\tag{3.9.11}$$

Then the nonconstant local invariant polynomials of $I\otimes P(x)-Q(x)^\top\otimes I$ are the nonconstant local invariant polynomials of $I\otimes P_i(x)-Q_i(x)^\top\otimes I$ for $i=1,...,t$. That is

$$\kappa_p(P,Q)=\sum_{i=1}^{t}\kappa_p(P_i,Q_i),\quad p=1,...,.\tag{3.9.12}$$

In particular if $C(x)$ is of the form (3.9.4) *then*

$$\eta(C,C)=\max_{1\le i\le\ell}\eta(C_i,C_i).\tag{3.9.13}$$

Proof. Theorem 1.14.3 implies $\kappa_p(P,Q)=\dim\mathbf{W}_{p-1}-\dim\mathbf{W}_p$, where $\mathbf{W}_p\subset\mathbb{C}^{n\times n}$ is the subspace of $n\times n$ matrices X_0 such that

$$\sum_{j=0}^{k}P_{k-j}X_j-X_jQ_{k-j}=0,\quad k=0,...,p.\tag{3.9.14}$$

Here

$$P(x)=\sum_{j=0}^{\infty}P_jx^j,\ P_i(x)=\sum_{j=1}^{\infty}P_j^{(i)}x^j,\ P_j=\oplus_{i=1}^{p}P_j^{(i)},$$
$$Q(x)=\sum_{j=0}^{\infty}Q_jx^j,\ Q_i(x)=\sum_{j=1}^{\infty}Q_j^{(i)}x^j,\ Q_j=\oplus_{i=1}^{q}Q_j^{(i)}.$$

Partition X_j to $[X_{\alpha\beta}^{(j)}]$, $X_{\alpha\beta}^{(j)} \in \mathbb{C}^{m_\alpha \otimes n_\beta}$, $\alpha = 1, ..., p$, $\beta = 1, ..., q$. We claim that $X_{\alpha\beta}^{(j)} = 0$ if either $\alpha > t+1$, or $\beta > t+1$, or $\alpha \neq \beta$. Indeed in view of Lemma 2.8.1 the equation $P_0^{(\alpha)} Y - Y Q_0^{(\beta)} = 0$ has only the trivial solution for α, β satisfying the above conditions. Then the claim that $X_{\alpha\beta}^{(j)} = 0$ follows by induction. Thus (3.9.14) splits to the system

$$\sum_{j=0}^{k} P_{k-j}^{(i)} X_{ii}^{(j)} - X_{ii}^{(j)} Q_{k-j}^{(i)} = 0, \quad i = 1, ..., t.$$

Apply the characterizations of $\kappa_p(P, Q)$ and $\kappa_p(P_i, Q_i)$ for $i = 1, ..., t$ to deduce (3.9.12). Clearly (3.9.12) implies (3.9.13). □

We conclude this section by remarking that main assumptions of Theorem 3.9.3, the splitting of the characteristic polynomial of $A(x)$ in H_0, is not a heavy restriction in view of the Weierstrass preparation theorem (Theorem 1.7.4). That is the eigenvalues of $A(y^m)$ split in H_0 for some value of m. Recall that m can be always be chosen $n!$, i.e. the minimal m divides $n!$. Problem 1 claims $A(x) \overset{a}{\approx} B(x) \iff A(y^m) \overset{a}{\approx} B(y^m)$. In view of Theorem 3.9.3 the classification problem of analytic similarity classes reduces to the description of the polynomial entries which are above the diagonal (in the matrix C in Theorem 3.9.3). Thus given the rational canonical form of $A(x)$ and the index $\eta(A, A)$ the set of all possible analytic similarity classes which correspond to A is a certain finite dimensional variety.

The case $n = 2$ is classified completely (Problem 2). In this case to a given rational canonical form there are at most countable number of analytic similarity classes. For $n = 3$ we have an example in which to a given rational canonical form there the family of distinct similarity classes correspond to a finite dimensional variety (Problem 3).

Problems

1. Let $A(x), B(x) \in \mathrm{H}_0^{n \times n}$ and let m be a positive integer. Assume that $A(y^m)T(y) = T(y)A(y^m)$ where $T(y) \in \mathrm{H}_0^{n \times n}$. Show

$$A(x)Q(x) = Q(x)B(x), \quad Q(y^m) = \frac{1}{m}\sum_{k=1}^{m} T(ye^{\frac{2\pi\sqrt{-1}k}{m}}), \; Q(x) \in \mathrm{H}_0^{n \times n}.$$

Prove $A(x) \overset{a}{\approx} B(x) \iff A(y^m) \overset{a}{\approx} B(y^m)$.

2. Let $A(x) \in \mathrm{H}_0^{2\times 2}$ and assume that

$$\begin{aligned}
&\det\,(\lambda I - A(x)) = (\lambda - \lambda_1(x))(\lambda - \lambda_2(x),\\
&\lambda_i(x) = \sum_{j=0}^{\infty} \lambda_j^{(i)} x^j \in \mathrm{H}_0, \quad i = 1, 2,\\
&\lambda_j^{(1)} = \lambda_j^{(2)},\ j = 0, ..., p, \quad \lambda_{p+1}^{(1)} \neq \lambda_{p+1}^{(2)}, \quad -1 \le p \le \infty,\\
&p = \infty \iff \lambda_1(x) = \lambda_2(x).
\end{aligned}$$

Show that $A(x)$ is analytically similar either to a diagonal matrix or to

$$B(x) = \begin{bmatrix} \lambda_1(x) & x^k \\ 0 & \lambda_2(x) \end{bmatrix}, \quad k = 0, ..., p\ (p \ge 0).$$

Furthermore if $A(x) \overset{a}{\approx} B(x)$ then $\eta(A, A) = k$. (*Hint*: Use a similarity transformation of the form DAD^{-1}, where D is a diagonal matrix.)

3. Let $A(x) \in \mathrm{H}_0^{3\times 3}$. Assume that

$$A(x) \overset{r}{\approx} C(p), \quad p(\lambda, x) = \lambda(\lambda - x^{2m})(\lambda - x^{4m}), \quad m \ge 1.$$

Show that $A(x)$ is analytically similar to a matrix

$$B(x, a) = \begin{bmatrix} 0 & x^{k_1} & a(x) \\ 0 & x^{2m} & x^{k_2} \\ 0 & 0 & x^{4m} \end{bmatrix}, \quad 0 \le k_1, k_2 \le \infty\ (x^{\infty} = 0),$$

where $a(x)$ is a polynomial of degree $4m - 1$ at most. (Use Problem 2.) Assume that $k_1 = k_2 = m$. Show that $B(x, a) \overset{a}{\approx} B(x, b)$ if and only if

(1) if $a(0) \neq 1$ then $b - a$ is divisible by x^m.

(2) if $a(0) = 1$ and $\frac{d^i a}{dx^i} = 0,\ i = 1, ..., k - 1,\ \frac{d^k a}{dx^k} \neq 0$ for $1 \le k < m$ then $b - a$ is divisible by x^{m+k}.

(3) if $a(0) = 1$ and $\frac{d^i a}{dx^i} = 0,\ i = 1, ..., m$ then $b - a$ is divisible by x^{2m}.

Then for $k_1 = k_2 = m$ and $a(0) \in \mathbb{C}\backslash\{1\}$ we can assume that $a(x)$ is a polynomial of degree less than m. Furthermore the similarity classes of $A(x)$ is uniquely determined by such $a(x)$. These similarity classes are parameterized by $\mathbb{C}\backslash\{1\} \times \mathbb{C}^{m-1}$ (the Taylor coefficients of $a(x)$).

4. Let P and Q satisfy the assumptions of Theorem 3.9.4. Show that P and Q are analytically similar if and only if

$$p = q = t, \quad m_i = n_i,\ P_i(x) \overset{a}{\approx} Q_i(x), \quad i = 1, ..., t.$$

3.10 Similarity to diagonal matrices

Theorem 3.10.1 *Let $A(x) \in H_0^{n\times n}$ and assume that the characteristic polynomial of $A(x)$ splits in H_0 as in* (3.9.3)*. Let*

$$B(x) = \mathrm{diag}(\lambda_1(x), ..., \lambda_n(x)). \tag{3.10.1}$$

Then $A(x)$ and $B(x)$ are not analytically similar if and only if there exists a nonnegative integer p such that

$$\begin{aligned} &\kappa_p(A,A) + \kappa_p(B,B) < 2\kappa_p(A,B),\\ &\kappa_j(A,A) + \kappa_j(B,B) = 2\kappa_j(A,B),\ j = 0, ..., p-1, \quad \text{if } p \geq 1. \end{aligned} \tag{3.10.2}$$

In particular $A(x)\overset{a}{\approx}B(x)$ if and only if the three matrices given in (2.9.4) *are equivalent over* H_0.

Proof. Suppose first that (3.10.2) holds. Then the three matrices in (2.9.4) are not equivalent. Hence $A(x)\overset{a}{\not\approx}B(x)$. Assume now that $A(x)\overset{a}{\not\approx}B(x)$. Without a loss in generality we may assume that $A(x) = C(x)$ where $C(x)$ is given in (3.9.4). Let

$$B(x) = \oplus_{j=1}^{\ell} B_j(x),\ B_j(0) = \alpha_j I_{n_j},\ \ j = 1, ..., \ell.$$

We prove (3.10.2) by induction on n. For $n = 1$ (3.10.2) is obvious. Assume that the (3.10.2) holds for $n \leq N-1$. Let $n = N$. If $A(0) \not\approx B(0)$ then Theorem 2.9.2 implies the inequality (3.10.2) for $p = 0$. Suppose now $A(0) \approx B(0)$. That is $A_j(0) = B_j(0) = \alpha_j I_{n_j},\ j = 1, ..., \ell$. Suppose first that $\ell > 1$. Theorem 3.9.4 yields

$$\kappa_p(A,A) = \sum_{j=1}^{\ell} \kappa_p(A_j,A_j),\ \kappa_p(A,B) = \sum_{j=1}^{\ell} \kappa_p(A_j,B_j),$$
$$\kappa_p(B,B) = \sum_{j=1}^{\ell} \kappa_p(B_j,B_j).$$

Problem 4 implies that $A(x)\overset{a}{\not\approx}B(x) \iff A_j(x)\overset{a}{\not\approx}B_j(x)$ for some j. Use the induction hypothesis to deduce (3.10.2). It is left to consider the case

$$A(0) = B(0) = \alpha_0,\ \kappa_0(A,A) = \kappa_0(A,B) = \kappa_0(B,B) = 0.$$

Let

$$A^{(1)}(x) = \frac{A(x) - \alpha_0 I}{x}, \quad B^{(1)}(x) = \frac{B(x) - \alpha_0 I}{x}.$$

Clearly

$$\kappa_p(A,A)=\kappa_{p-1}(A^{(1)},A^{(1)}),\ \kappa_p(A,B)=\kappa_{p-1}(A^{(1)},B^{(1)}),$$
$$\kappa_p(B,B)=\kappa_{p-1}(B^{(1)},B^{(1)}).$$

Furthermore $A(x)\overset{a}{\approx}B(x) \iff A^{(1)}(x)\overset{a}{\approx}B^{(1)}(x)$. Continue this process. If at some (first) stage k either $A^{(k)}(0)\not\approx B^{(k)}(0)$ or $A^{(k)}(0)$ has at least two distinct eigenvalues we conclude (3.10.2) as above. Suppose finally that such k does not exist. Then $A(x)=B(x)=\lambda(x)I$, which contradicts the assumption $A(x)\overset{a}{\not\approx}B(x)$. □

3.11 Strict similarity of matrix polynomials

Definition 3.11.1 *Let $A(x),B(x)\in\mathbb{C}[x]^{n\times n}$. Then $A(x)$ and $B(x)$ are called strictly similar ($A\overset{s}{\approx}B$) if there exists $P\in\mathbf{GL}(n,\mathbb{C})$ such that $B(x)=PA(x)P^{-1}$.*

Definition 3.11.2 *Let ℓ be a positive integer and $(A_0,A_1,...,A_\ell),(B_0,...,B_\ell)\in(\mathbb{C}^{n\times n})^{\ell+1}$. Then $(A_0,A_1,...,A_\ell)$ and $(B_0,...,B_\ell)$ are called simultaneously similar $(A_0,A_1,...,A_\ell)\approx(B_0,...,B_\ell)$ if there exists $P\in\mathbf{GL}(n,\mathbb{C})$ such that $B_i=PA_iP^{-1},i=0,...,\ell$, i.e. $(B_0,B_1,...,B_\ell)=P(A_0,A_1,...,A_\ell)P^{-1}$.*

Clearly

Proposition 3.11.3 *Let*

$$A(x)=\sum_{i=0}^{\ell}A_ix^i,\ B(x)=\sum_{i=0}^{\ell}B_ix^i\in\mathbb{C}[x]^{n\times n}. \tag{3.11.1}$$

Then $(A\overset{s}{\approx}B)$ if and only if $(A_0,A_1,...,A_\ell)\approx(B_0,...,B_\ell)$.

The problem of simultaneous similarity of matrices, i.e. to describe the similarity class of a given m (≥ 2) tuple of matrices or to decide when a given two tuples of matrices are simultaneously similar, is a hard problem. See [Fri83]. There are some cases where this problem has a relatively simple solution.

Theorem 3.11.4 *Let $\ell\geq 1$ and $(A_0,...,A_\ell)\in(\mathbb{C}^{n\times n})^{\ell+1}$. Then $(A_0,...,A_\ell)$ is simultaneously similar to a diagonal tuple $(B_0,...,B_\ell)\in\mathbb{C}^{n\times n})^{\ell+1}$, i.e. each B_i is a diagonal matrix, if and only if $A_0,...,A_\ell$ are $\ell+1$ commuting diagonable matrices:*

$$A_iA_j=A_jA_i,\quad i,j=0,...,\ell. \tag{3.11.2}$$

Proof. Clearly if $(A_0, ..., A_\ell)$ is simultaneously similar to a diagonal tuple then $A_0, ..., A_\ell$ a set of commuting diagonal matrices. Assume that $A_0, ..., A_\ell$ a set of commuting diagonal matrices. We show that $(A_0, ..., A_\ell)$ is simultaneously similar to a diagonal tuple by the double induction on n and ℓ. It is convenient to let $\ell \geq 0$. For $n = 1$ the theorem trivially holds for any $\ell \geq 0$. For $\ell = 0$ the theorem trivially holds for any $n \geq 1$. Assume now that $p > 1,\ q \geq 1$ and assume that the theorem holds for $n \leq p - 1$ and all ℓ and for $n = p$ and $\ell \leq q - 1$. Assume that $A_0, ..., A_q \in \mathbb{C}^{p\times p}$ are $q + 1$ commuting diagonable matrices. Suppose first that $A_0 = aI_p$. The induction hypothesis yields that $(B_1, ..., B_q) = P(A_1, ..., A_q)P^{-1}$ is a diagonal q-tuple for some $P \in \mathbf{GL}(n, \mathbb{C})$. As $PA_0P^{-1} = A_0 = aI_p$ we deduce that $(A_0, B_1, ..., B_\ell) = P(A_0, A_1, ..., A_\ell)P^{-1}$.

Assume that A_0 is not a scalar matrix, i.e $A_0 \neq \frac{1}{p} \operatorname{tr} A\ I_p$. Let

$$\tilde{A}_0 = QA_0Q^{-1} = \oplus_{i=1}^k a_i I_{p_i},$$

$$1 \leq p_i,\ a_i \neq a_j \text{ for } i \neq j,\ i, j = 1, ..., k,\ \sum_{i=1}^k n_i = p.$$

Then the $q+1$ tuple $(\tilde{A}_0, ..., \tilde{A}_q) = Q(A_0, ..., A_q)Q^{-1}$ is a $q+1$ tuple of diagonable commuting matrices. The specific form of $\tilde{A}_0$ and the assumption that $\tilde{A}_0$ and $\tilde{A}_j$ commute implies

$$\tilde{A}_j = \oplus_{i=1}^k \tilde{A}_{j,i}, \quad \tilde{A}_{j,i} \in \mathbb{C}^{p_i\times p_i},\ i = 1, ..., k,\ j = 1, ..., q.$$

The assumption that $(\tilde{A}_0, ..., \tilde{A}_q)$ is a $q + 1$ tuple of diagonable commuting matrices implies that each i the tuple $(a_iI_{p_i}, \tilde{A}_{1,i}..., \tilde{A}_{q,i})$ is $q + 1$ tuple of diagonable commuting matrices. Hence the induction hypothesis yields that $(a_iI_{p_i}, \tilde{A}_{1,i}..., \tilde{A}_{q,i})$ is similar to a $q + 1$ diagonal tuple for $i = 1, ..., k$. It follows straightforward that $(\tilde{A}_0, \tilde{A}_1..., \tilde{A}_q)$ is simultaneously similar to a diagonal $q + 1$ tuple. □

The problem when $A(x) \in \mathbb{C}[x]^{n\times n}$ is strictly similar to an upper triangular matrix $B(x) \in \mathbb{C}[x]^{n\times n}$ is equivalent to the problem when an $\ell + 1$ tuple $(A_0, ..., A_\ell) \in (\mathbb{C}^{n\times n})^{\ell+1}$ is simultaneously an upper triangular tuple $(B_0, ..., B_\ell)$, i.e. each B_i is an upper triangular matrix, is solved in [DDG51]. We bring their result without a proof.

Definition 3.11.5 *Let $\mathbb{D}$ be a domain, let n, m be positive integers and let $C_1, ..., C_m \in \mathbb{D}^{n\times n}$. Then $\mathcal{A}(C_1, ..., C_m) \subset \mathbb{D}^{n\times n}$ denotes the minimal algebra in $\mathbb{D}^{n\times n}$ containing I_n and $C_1, ..., C_m$. That is every matrix $F \in \mathcal{A}(C_1, ..., C_m)$ is a noncommutative polynomial in $C_1, ..., C_m$.*

Theorem 3.11.6 *Let m, ℓ be positive integers and let $A_0, ..., A_\ell \in (\mathbb{C}^{n\times n})^{\ell+1}$. TFAE:*
(a) *$(A_0, ..., A_\ell)$ is simultaneously similar to an upper triangular tuple $(B_0, ..., B_\ell) \in$ $\mathrm{M_n}(\mathbb{C})^{\ell+1}$.*
(b) *For any $0 \le i < j \le \ell$ and $F \in \mathcal{A}(A_o, ..., A_\ell))$ the matrix $(A_iA_j - A_jA_i)F$ is nilpotent.*

The implication $(a) \Rightarrow (b)$ is trivial. (See Problem 2.) The verification of condition (b) can be done quite efficiently. (See Problem 3.)

Corollary 3.11.7 *Let m, ℓ be positive integers and assume that $A_0, ..., A_\ell \in \mathbb{C}^{n\times n}$ are commuting matrices. Then $(A_0, ..., A_\ell)$ is simultaneously similar to an upper triangular tuple $(B_0, ..., B_\ell)$.*

See Problem 4.

Problems

1. Let $\mathbb{F}$ be a field. View $\mathbb{F}^{n\times n}$ as an n^2 dimensional vector space over $\mathbb{F}$. Note that any $A \in \mathbb{F}^{n\times n}$ acts as a linear transformation on $\mathbb{F}^{n\times n}$ by left multiplication: $B \mapsto AB,\ B \in \mathbb{C}^{n\times n}$. Let $A_0, ..., A_\ell \in \mathbb{F}^{n\times n}$. Let $\mathbf{W}_0 = \mathrm{span}\ (\mathrm{I_n})$ and define

$$\mathbf{W}_k = \mathbf{W}_{k-1} + \sum_{j=0}^{\ell} A_j\mathbf{W}_{k-1}, \quad k = 1, ...,.$$

Show that $\mathbf{W}_{k-1} \subset \mathbf{W}_k$ for each $k \ge 1$. Let p be the minimal nonnegative integer for which the equality $\mathbf{W}_k = \mathbf{W}_{k+1}$ holds. Show that $\mathcal{A}(A_o, ..., A_\ell) = \mathbf{W}_p$. In particular $\mathcal{A}(A_o, ..., A_\ell)$ is a finite dimensional subspace of $\mathbb{F}^{n\times n}$.

2. Show the implication $(a) \Rightarrow (b)$ in Theorem 3.11.6.

3. Let the assumptions of Problem 1 hold. Let $\mathbf{X}_0 = \mathcal{A}(A_o, ..., A_\ell)$ and define recursively

$$\mathbf{X}_k = \sum_{0\le i<j\le\ell} (A_iA_j - A_jA_i)\mathbf{X}_{k-1} \subset \mathbb{F}^{n\times n}, \quad k = 1, ...,.$$

Show that the condition (a) of Theorem 3.11.6 to the following two conditions:

(c) $A_i\mathbf{X}_k \subset \mathbf{X}_k, \quad i = 0, ..., \ell,\ k = 0, ...,.$

(d) There exists $q \ge 1$ such that $\mathbf{X}_q = \{0\}$ and $\mathbf{X}_k$ is a strict subspace of $\mathbf{X}_{k-1}$ for $k = 1, ..., q$.

4. Let $A_0, ..., A_\ell \in \mathbb{F}^{n\times n}$. Assume that $\mathbf{0} \neq \mathbf{x} \in \mathbb{F}^n$ and $A_0\mathbf{x} = \lambda_0\mathbf{x}$. Suppose that $A_0A_i = A_iA_0,\ i = 1, ..., \ell$.

   (a) Show that any nonzero vector in $\mathcal{A}(A_1, ..., A_\ell)$span $(\mathbf{x})(\supset$ span $(\mathbf{x}))$ is an eigenvector of A_0 corresponding λ_0.

   (b) Assume in addition that $A_1, ..., A_\ell$ are commuting matrices whose characteristic polynomials split in $\mathbb{F}$ to linear factors. Show by induction that there exists $\mathbf{0} \neq \mathbf{y} \in \mathcal{A}(A_1, ..., A_\ell)$span $(\mathbf{x})$ such $A_i\mathbf{y} = \lambda_i\mathbf{y},\ i = 0, ..., \ell$.

   (c) Show that if $A_0, ..., A_\ell \in \mathbb{F}^{n\times n}$ are commuting matrices whose characteristic polynomials split in $\mathbb{F}$ to linear factors then $(A_0, ..., A_\ell)$ is simultaneously similar over $\mathbf{GL}(n, \mathbb{F})$ to an upper triangular $\ell + 1$ tuple.

3.12 Similarity to diagonal matrices

Let $A(x) \in \mathrm{H}_0^{n\times n}$. The Weierstrass preparation theorem (Theorem 1.7.4) implies that the eigenvalues of $A(y^s)$ are analytic in y for some $s|n!$. That is the eigenvalues $\lambda_1(x), ..., \lambda_n(x)$ are multivalued analytic functions in x which have the expansion

$$\lambda_j(x) = \sum_{k=0}^{\infty} \lambda_{jk} x^{\frac{k}{s}}, \quad j = 1, ..., n.$$

In particular each λ_j has s_j branches, where $s_j|m$. For more properties of the eigenvalues $\lambda_1(x), ..., \lambda_n(x)$ see for example [Kat80, Chap.2].

Let $A(x) \in \mathbb{C}[x]^{n\times n}$. Then

$$A(x) = \sum_{k=0}^{\ell} A_k x^k, \quad A_k \in \mathbb{C}^{n\times n},\ k = 0, ..., \ell. \tag{3.12.1}$$

The eigenvalues of $A(x)$ satisfy the equation

$$\det\,(\lambda I - A(x)) = \lambda^n + \sum_{j=1}^{n} a_j(x)\lambda^{n-j},\ a_j(x) \in \mathbb{C}[x],\ j = 1, ..., n. \tag{3.12.2}$$

Thus the eigenvalues $\lambda_1(x), ..., \lambda_n(x)$ are algebraic functions of x. (See for example [GuR65].) For each $\zeta \in \mathbb{C}$ we apply the Weierstrass preparation theorem in H_ζ to obtain the *Puiseaux* expansion of $\lambda_j(x)$ around $x = \zeta$:

$$\lambda_j(x) = \sum_{k=0}^{\infty} \lambda_{jk}(\zeta)(x - \zeta)^{\frac{k}{s}}, \quad j = 1, ..., n. \tag{3.12.3}$$

For simplicity of notation we choose $s \le n!$ for which the above expansion holds for each $\zeta \in \mathbb{C}$. (For example $s = n!$ is always a valid choice.) Since $A(x)$ is a polynomial matrix each $\lambda_j(x)$ has Puiseaux expansion at ∞. Let

$$A(x) = x^\ell B(\frac{1}{x}), \quad B(y) = \sum_{k=0}^{\ell} A_k y^{\ell-k}.$$

Then the Puiseaux expansion of the eigenvalues of $B(y)$ at $y = 0$ yields

$$\lambda_j(x) = x^\ell \sum_{k=0}^{\infty} \lambda_{jk}(\infty) x^{-\frac{k}{s}}, \quad j = 1, ..., n. \tag{3.12.4}$$

Equivalently, we view the eigenvalues $\lambda_j(x)$ as multivalued analytic functions over the Riemann sphere $\mathbb{P} = \mathbb{C} \cup \infty$. To view $A(x)$ as a matrix function over $\mathbb{P}$ we need to homogenize as in §2.1.

Definition 3.12.1 *Let $A(x)$ be given by* (3.12.1)*. Denote by $A(x_0, x_1)$ the corresponding homogeneous matrix*

$$A(x_0, x_1) = \sum_{k=0}^{\ell'} A_k e x_0^{\ell'-k} x_1^k \in \mathbb{C}[x_0, x_1]^{n\times n}, \tag{3.12.5}$$

where $\ell' = -1$ if $A(x) = 0$ and $A_{\ell'} \ne 0$ and $A_j = 0$ for $\ell' < j \le \ell$ if $A(x) \ne 0$.

Let $A(x), B(x) \in \mathbb{C}[x]^{n\times n}$. Then $A(x)$ and $B(x)$ are similar over $\mathbb{C}[x]$, denoted by $A(x) \approx B(x)$, if $B(x) = P(x)A(x)P^{-1}(x)$ for some $P(x) \in \mathbf{GL}(n, \mathbb{C}[x])$. Lemma 2.9.4 implies that if $A(x) \approx B(x)$ then the three matrices in (3.6.2) are equivalent over $\mathbb{C}[x]$. Assume a stronger condition $A \overset{s}{\approx} B$. Clearly if $B(x) = PA(x)P^{-1}$ then $B(x_0, x_1) = PA(x_0, x_1)P^{-1}$. According to Lemma 2.9.4 the matrices

$$\begin{aligned} I \otimes A(x_0, x_1) - A(x_0, x_1)^\top \otimes I, \ I \otimes A(x_0, x_1) - B(x_0, x_1)^\top \otimes I, \\ I \otimes B(x_0, x_1) - B(x_0, x_1)^\top \otimes I, \end{aligned} \tag{3.12.6}$$

are equivalent over $\mathbb{C}[x_0, x_1]$. Lemma 1.11.3 yields.

Lemma 3.12.2 *Let $A(x), B(x) \in \mathbb{C}[x]^{n\times n}$. Assume that $A(x) \overset{s}{\approx} B(x)$. Then the three matrices in* (3.12.6) *have the same invariant polynomials over* $\mathbb{C}[x_0, x_1]$.

Definition 3.12.3 *Let $A(x), B(x) \in \mathbb{C}[x]^{n\times n}$. Let $A(x_0, x_1), B(x_0, x_1)$ be the homogeneous matrices corresponding to $A(x), B(x)$ respectively. Denote by $i_k(A, B, x_0, x_1)$, $k = 1, ..., r(A, B)$ the invariant factors of $I \otimes A(x_0, x_1) - B(x_0, x_1)^\top \otimes I$.*

The arguments of the proof of Lemma 2.1.2 imply that $i_k(A, B, x_0, x_1)$ is a homogeneous polynomial for $k = 1, ..., r(A, B)$. Moreover $i_k(A, B, 1, x)$ are the invariants factors of $I \otimes A(x) - B(x)^\top \otimes I$. (See Problems 5-6.)

Theorem 3.12.4 *Let $A(x) \in \mathbb{C}[x]^{n\times n}$. Assume that the characteristic polynomial of $A(x)$ splits to linear factors over $\mathbb{C}[x]$. Let $B(x)$ be the diagonal matrix of the form* (3.10.1)*. Then $A(x) \approx B(x)$ if and only if the three matrices in* (3.6.2) *are equivalent over $\mathbb{C}[x]$. Furthermore $A(x)\overset{s}{\approx}B(x)$ if and only if the three matrices in* (1.34.8) *have the same invariant factors over $\mathbb{C}[x_0, x_1]$.*

Proof. Clearly if $A(x) \approx B(x)$ then the three matrices in (3.6.2) are equivalent over $\mathbb{C}[x]$. Similarly if $A(x)\overset{s}{\approx}B(x)$ then the three matrices in (1.34.8) have the same invariant factors over $\mathbb{C}[x_0, x_1]$. We now show the opposite implications.

Without loss of generality we may assume that $B(x)$ is of the form

$$B(x) = \oplus_{i=1}^m \lambda_i(x) I_{n_i} \in \mathbb{C}[x]^{n\times n}, \quad \lambda_i(x) \neq \lambda_j(x),\ i \neq j,\ i, j = 1, ..., m. \tag{3.12.7}$$

Thus for all but a finite number of points $\zeta \in \mathbb{C}$ we have that

$$\lambda_i(\zeta) \neq \lambda_j(\zeta) \text{ for } i \neq j, \quad i, j = 1, ..., m. \tag{3.12.8}$$

Assume first that $A(x) \approx B(x)$. Let $P_j(A)$ be the projection of $A(x)$ on $\lambda_j(x)$ for $j = 1, ..., m$. Suppose that (3.12.8) is satisfied at ζ. Problem 3.4.10 yields that each $P_j(x)$ is analytic in the neighborhood of ζ. Assume that (3.12.8) does not hold for $\zeta \in \mathbb{C}$. The assumptions that the three matrices in (3.12.6) have the same invariant polynomials imply that the matrices in (3.6.2) are equivalent over H_ζ. Now use Theorem 3.10.1 to get that $A(x) = Q(x)B(x)Q(x)^{-1},\ Q \in \mathbf{GL}(n, \mathrm{H}_\zeta)$. Clearly $P_j(B)$, the projection of $B(x)$ on $\lambda_j(x)$, is $0 \oplus I_{n_j} \oplus 0$. In particular $P_j(B)$ is analytic in the neighborhood of any $\zeta \in \mathbb{C}$ and its rank is always equal to n_j. Problem 3.4.11 yields that $P_j(A)(x) = Q(x)P_j(B)(x)Q(x)^{-1} \in \mathrm{H}_\zeta^{n\times n}$. Hence $\mathrm{rank\ P_j(A)(\zeta) = n_j}$ for all $\zeta \in \mathbb{C}$. Furthermore $P_j(A)(x) \in \mathrm{H}_\mathbb{C}^{n\times n}$, i.e. each entry of $P_j(A)$ is an entire function (analytic function on $\mathbb{C}$). Problem 3.4.14 yields that

$$P_j(A)(\zeta) = \prod_{k=1, k\neq j}^{n} \frac{A(\zeta) - \lambda_k(\zeta)I}{\lambda_j(\zeta) - \lambda_k(\zeta)}, \quad j = 1, ..., n. \tag{3.12.9}$$

Hence each entry of $P_j(A)(\zeta)$ is a rational function of ζ on $\mathbb{C}$. Since $P_i(A)(x)$ is analytic in the neighborhood of each $\zeta \in \mathbb{C}$ it follows that

$P_i(x) \in \mathbb{C}[x]^{n\times n}$. We also showed that its rank is locally constant, hence rank $\mathrm{P_i(x) = n_i,\ i = 1, ..., m}$. Therefore the Smith normal form of $P_i(x)$ over $\mathbb{C}[x]$ is $P_i(x) = U_i(x)(I_{n_i} \oplus 0)V_i(x)$, $U_i, V_i \in \mathbf{GL}(n, \mathbb{C}[x])$. Let $\mathbf{u}_{1,i}(x), ..., \mathbf{u}_{n_i,i}(x)$ be the first n_i columns of $U_i(x)$. Then $P_i(x)\mathbb{C}^n =$ span $(\mathbf{u}_{\mathrm{1,i}}(\mathrm{x}), ..., \mathbf{u}_{\mathrm{n_i,i}}(\mathrm{x}))$. Recall that $P_1(x) + ... + P_m(x) = I_n$. Hence $\mathbf{u}_{1,1}(x), ..., \mathbf{u}_{n_1,1}(x), ..., \mathbf{u}_{1,m}(x), ..., \mathbf{u}_{n_m,m}(x)$ is a basis for $\mathbb{C}^n$ for each $x \in \mathbb{C}$. Let $S(x)$ be the matrix with the columns

$$\mathbf{u}_{1,1}(x), ..., \mathbf{u}_{n_1,1}(x), ..., \mathbf{u}_{1,m}(x), ..., \mathbf{u}_{n_m,m}(x).$$

Then $S(x) \in \mathbf{GL}(n, \mathbb{C}[x])$. Let $D(x) = S^{-1}(x)A(x)S(x) \in \mathbb{C}[x]^{n\times n}$. Since $A(x)$ is pointwise diagonable $D(\zeta) = B(\zeta)$, where ζ satisfies (3.12.8) and $B(x)$ is of the form (3.12.7). Since only finite number of points $\zeta \in \mathbb{C}$ do not satisfy the condition (3.12.7) it follows that $D(x) = B(x)$. This proves the first part of the theorem.

Assume now that the three matrices in (1.34.8) have the same invariant factors over $\mathbb{C}[x_0, x_1]$. The same arguments imply that $A(x_0, 1)\overset{a}{\approx}B(x_0, 1)$ over the ring $\mathrm{H_0}$. That is $P_j(A)$ is also analytic at the neighborhood $\zeta = \infty$. So $P_j(A)$ is analytic on $\mathbb{P}$ hence bounded, i.e. each entry of $P_j(A)$ is bounded. Hence $P_j(A)$ is a constant matrix. Therefore $S(x)$ is a constant invertible matrix, i.e. $A(x)\overset{s}{\approx}B(x)$. □

Let $A(x) \in \mathbb{C}[x]^{n\times n}$ be of the form (3.12.1) with $\ell \geq 1$ and $A_\ell \neq 0$. Assume that $A(x)$ is strictly similar to a diagonal matrix $B(x)$. Then $A(x)$ is pointwise *diagonable*, i.e. $A(x)$ is similar to a diagonal matrix for each $x \in \mathbb{C}$, and $A_\ell \neq 0$ is diagonable. Equivalently, consider the homogeneous polynomial matrix $A(x_0, x_1)$. Then $A(x_0, x_1)$ is pointwise diagonable (in $\mathbb{C}^2$). However the assumption that any $A(x_0, x_1)$ is pointwise diagonable does not imply that $A(x)$ is strictly equivalent to a diagonal matrix. Consider for example

$$A(x) = \begin{bmatrix} x^2 & x \\ 0 & 1 + x^2 \end{bmatrix} \Rightarrow A(x_0, x_1) = \begin{bmatrix} x_1^2 & x_0x_1 \\ 0 & x_0^2 + x_1^2 \end{bmatrix}. \tag{3.12.10}$$

(See Problem 2.)

Definition 3.12.5 *Let $A(x) \in \mathbb{C}[x]^{n\times n}$ be of the form* (3.12.1) *with $\ell \geq 1$ and $A_\ell \neq 0$. Let $\lambda_p(x)$ and $\lambda_q(x)$ be two distinct eigenvalues of $A(x)$. ($\lambda_p(x)$ and $\lambda_q(x)$ have distinct Puiseaux expansion for any $\zeta \in \mathbb{P}$.) The eigenvalues $\lambda_p(x)$ and $\lambda_q(x)$ are said to be tangent at $\zeta \in \mathbb{P}$ if their Puiseaux expansion at ζ satisfy*

$$\lambda_{pk}(\zeta) = \lambda_{qk}(\zeta), \quad k = 0, ..., s. \tag{3.12.11}$$

(Note that two distinct eigenvalues are tangent at ∞ if the corresponding eigenvalues of $A(x,1)$ are tangent at 0*.)*

Note that for $A(x)$ given in (3.12.10) the two eigenvalues of $A(x)$ x^2 and $1+x^2$ are tangent at one point $\zeta=\infty$. (The eigenvalues of $A(x,1)$ are 1 and $1+x^2$.)

Theorem 3.12.6 *Let $A(x)\in\mathbb{C}[x]^{n\times n}$ be of the form (3.12.1) with $\ell\ge 1$ and $A_\ell\ne 0$. Then one of the following conditions imply that $A(x)=S(x)B(x)S^{-1}(x)$, where $S(x)\in\mathbf{GL}(n,\mathbb{C}[x])$ and $B(x)\in\mathbb{C}[x]^{n\times n}$ is a diagonal matrix of the form $\sum_{i=1}^m \lambda_i(x)I_{k_i}$, where $k_1,\dots,k_m\ge 1$. Furthermore $\lambda_1(x),...,\lambda_m(x)$ are m distinct polynomials satisfying the following conditions:*

(a) $\ell\ge\deg\lambda_i(x),\ i=1,...,m$.

(b) The polynomial $\lambda_i(x)-\lambda_j(x)$ has only simple roots in $\mathbb{C}$ for $i\ne j$. $(\lambda_i(\zeta)=\lambda_j(\zeta)\Rightarrow\lambda_i'(\zeta)\ne\lambda_j'(\zeta))$.

I. The characteristic polynomial of $A(x)$ splits in $\mathbb{C}[x]$, i.e. all the eigenvalues of $A(x)$ are polynomials. $A(x)$ is point-wise diagonable in $\mathbb{C}$ and no two distinct eigenvalues are tangent at any $\zeta\in\mathbb{C}$.

II. $A(x)$ is point-wise diagonable in $\mathbb{C}$ and A_ℓ is diagonable. No two distinct eigenvalues are tangent at any point $\zeta\in\mathbb{C}\cup\{\infty\}$. Then $A(x)$ is strictly similar to $B(x)$, i.e. $S(x)$ can be chosen in $\mathbf{GL}(n,\mathbb{C})$. Furthermore $\lambda_1(x),...,\lambda_m(x)$ satisfy the additional condition:

(c) $\deg\lambda_1(x)=l$. Furthermore, for $i\ne j$ either $\frac{d^\ell\lambda_i}{d^\ell x}(0)\ne\frac{d^\ell\lambda_j}{d^\ell x}(0)$ or $\frac{d^\ell\lambda_i}{d^\ell x}(0)=\frac{d^\ell\lambda_j}{d^\ell x}(0)$ and $\frac{d^{\ell-1}\lambda_i}{d^{\ell=1}x}(0)\ne\frac{d^{\ell-1}\lambda_j}{d^{\ell-1}x}(0)$.

Proof. View $A(x)$ as matrix in $\mathcal{M}^{n\times n}$, where $\mathcal{M}$ is field of rational functions. Let $\mathbb{K}$ be a finite extension of $\mathcal{M}$ such that det $(\lambda I-A(x))$ splits to linear factors over $\mathbb{K}$. Then $A(x)$ has m distinct eigenvalues $\lambda_1,...,\lambda_m\in\mathbb{K}$ of multiplicities $n_1,...,n_m$ respectively. We view these eigenvalues as multivalued functions $\lambda_1(x),...,\lambda_m(x)$. Thus for all but a finite number of points ζ (3.12.8) holds. Assume that ζ satisfies (3.12.8). Denote by $P_j(\zeta)$ the projection of $A(\zeta)$ on $\lambda_j(\zeta)$. Problem 10 implies that $P_j(x)$ is a multivalued analytic in the neighborhood of ζ and rank $P_j(\zeta)=n_j$. Problem 14 yields (3.12.9). We claim that in the neighborhood of any $\zeta\in\mathbb{C}$ each λ_j and P_j is multivalued analytic and rank $P_j(x)=n_j$. Let $\zeta\in\mathbb{C}$ for which (3.12.8) is violated. For simplicity of notation we consider $\lambda_1(x)$ and $P_1(x)$. Let

$$\lambda_1(\zeta)=...=\lambda_r(\zeta)\ne\lambda_k(\zeta),\quad k=r+1,...,m.$$

Theorem 3.5.7 implies the existence of $Q(x) \in \mathbf{GL}(n, \mathrm{H}_\zeta)$ such that

$$\begin{aligned}&Q^{-1}(x)A(x)Q(x) = C_1(x) \oplus C_2(x),\\ &C_j(x) \in \mathrm{H}_\zeta^{m_j \times m_j},\ j = 1, 2, \quad m_1 = \sum_{i=1}^{r} n_i,\ m_2 = n - m_1.\end{aligned}$$

The eigenvalues of $C_1(x)$ and $C_2(x)$ are $\lambda_1(x), ..., \lambda_r(x)$ and $\lambda_{r+1}(x), ..., \lambda_m(x)$ respectively in some neighborhood of ζ. Since $C(x)$ is pointwise diagonable in H_ζ it follows that $C_1(x)$ and $C_2(x)$ are pointwise diagonable in H_ζ. We claim that $\lambda_i(x) \in \mathrm{H}_\zeta$, the projection $\hat{P}_i(x)$ of $C_1(x)$ on $\lambda_i(x)$ is in $\mathrm{H}_\zeta^{m_1 \times m_1}$ and rank $\hat{P}_i(\zeta) = n_i$ for $i = 1, ..., r$. If $r = 1$ $\lambda_1(x) = \frac{1}{n_1} \operatorname{tr} C_1(x) \in \mathrm{H}_\zeta$ and $\hat{P}_1(x) = I_{n_1}$. Assume that $r > 1$. Since $C_1(\zeta)$ is diagonable and has one eigenvalue $\lambda_1(\zeta)$ of multiplicity m_1 it follows that $C_1(\zeta) = \lambda_1(\zeta)I_{m_1}$. Hence

$$C_1(x) = \lambda_1(\zeta)I_{m_1} + (x - \zeta)\hat{C}_1(x), \quad \hat{C}_1(x) \in \mathrm{H}_\zeta^{m_1 \times m_1}.$$

Clearly $\hat{C}_1(x)$ has r distinct eigenvalues $\hat{\lambda}_1(x), ..., \hat{\lambda}_r(x)$ such that

$$\lambda_i(x) = \lambda_1(\zeta) + (x - \zeta)\hat{\lambda}_i(x), \quad i = 1, ..., r.$$

Each $\hat{\lambda}_i(x)$ has Puiseaux expansion (3.12.3). The above equality shows that for $1 \le i < j \le r$ $\lambda_i(x)$ and $\lambda_j(x)$ are not tangent if and only if $\hat{\lambda}_i(\zeta) \ne \hat{\lambda}_j(\eta)$. By the assumption of theorem no two different eigenvalues of $A(x)$ are tangent in $\mathbb{C}$. Hence $\hat{\lambda}_i(\zeta) \ne \hat{\lambda}_j(\eta)$ for all $i \ne j \le r$. That is $\hat{C}_1(\zeta)$ has r distinct eigenvalues. Apply Theorem 3.5.7 to $\hat{C}_1(\zeta)$ to deduce that $\hat{C}_1(\zeta)$ is analytically similar $\tilde{C}_1 \oplus ... \oplus \tilde{C}_r$ such that $\tilde{C}_i$ has a unique eigenvalues $\hat{\lambda}_i(x)$ of multiplicity n_i for $i = 1, ..., r$. Hence $\hat{\lambda}_i(x) = \frac{1}{n_i} \operatorname{tr} \tilde{C}_i(x) \in \mathrm{H}_\zeta \Rightarrow \lambda_i(x) \in \mathrm{H}_\zeta$. Clearly the projection of $\tilde{C}_i(x)$ on $\hat{\lambda}_i(x)$ is I_{n_i}. Hence $\hat{P}_i(x)$ is analytically similar to the projection to $0 \oplus ... \oplus I_{n_i} ... \oplus 0$. So $\hat{P}_i(x) \in \mathrm{H}_\zeta^{m_1 \times m_1}$, rank $P_i(x) = n_i$ for $i = 1, .., r$. Hence $P_1(x) \in \mathrm{H}_\zeta^{n \times n}$, rank $P_1(x) = n_1$ as we claimed.

Assume now that $\lambda_1(x), \ldots, \lambda_n(x)$ are polynomials. Hence and $P_i(x)$ are entire functions on $\mathbb{C}$. (See for example [Rud74].) Since $\lim_{|x| \to \infty} \frac{A(x)}{x^\ell} = A_l$ it follows that $\limsup_{|x| \to \infty} \frac{|\lambda_i(x)|}{|x|^\ell} \le \rho(A)$, where $\rho(A)$ is the spectral radius of A_ℓ. Hence each $\lambda_j(x)$ is polynomial of degree ℓ at most. Since $A_\ell \ne 0$ it follows that at least one of $\lambda_j(x)$ is a polynomial of degree ℓ exactly. We may assume that deg $\lambda_1(x) = \ell$. This proves the condition (a) of the theorem. The condition (b) is equivalent to the statement that no two distinct eigenvalues of $A(x)$ are tangent in $\mathbb{C}$.

Define $P_i(\zeta)$ by (3.12.9). As in the proof of Theorem 3.12.4 it follows that $P_i(x) \in \mathbb{C}[x]^{n\times n}$ and rank $P_i(x) = n_i,\ i = 1, ..., m$. Furthermore we define $S(x) \in \mathbf{GL}(n, \mathbb{C}[x])$ as in the proof of Theorem 3.12.4 such that $B(x) = S^{-1}(x)A(x)S(x)$. This proves the first part of the theorem.

To prove the second part of the theorem observe that in view of our definition of tangency at ∞ the condition (c) is equivalent to the condition that no two distinct eigenvalues of A are tangent at infinity. Assume now that A_l is diagonable and no two distinct eigenvalues are tangent at ∞. Then the above arguments show that each $P_i(x)$ is also multivalued analytic at ∞. By considering $x^{-l}A(x)$ it follows that $P_i(x)$ is bounded at the neighborhood of ∞. Hence $P_i(x) = P_i(0)$ for $i = 1, ..., m$. Thus $S \in \mathbf{GL}(n, \mathbb{C})$. So $A(x)$ is diagonable by a constant matrix. In particular all the eigenvalues of $A(x)$ are polynomials. Sice no two distinct eigenvalues are tangent at ∞ we deduce the condition (c) holds. □

Problems

1. Let $A(x) \in \mathbb{C}[x]^{n\times n}$. Assume that there exists an infinite sequence of distinct points $\{\zeta_k\}_1^\infty$ such that $A(\zeta_k)$ is diagonable for $k = 1, ...,$. Show that $A(x)$ is diagonable for all but a finite number of points. (*Hint*: Consider the rational canonical form of $A(x)$ over the field of rational functions $\mathbb{C}(x)$.)

2. Consider the matrix $A(x)$ given in (3.12.10). Show

   (a) A(x) and $A(x_0, x_1)$ are pointwise similar to diagonal matrices in $\mathbb{C}$ and $\mathbb{C}^2$ respectively.

   (b) The eigenvalues of $A(x)$ are not tangent at any point in $\mathbb{C}$.

   (c) Find $S(x) \in \mathbf{GL}(2, \mathbb{C}[x])$ such that $S^{-1}(x)A(x)S(x) = \mathrm{diag}(x^2, 1+x^2)$.

   (d) Show that $A(x)$ is not strictly similar to $\mathrm{diag}(x^2, 1 + x^2)$.

   (c) Show that the eigenvalues of $A(x)$ are tangent at $\zeta = \infty$.

3.13 Property L

In this section and the next one we assume that all pencils $A(x) = A_0 + A_1 x$ are square pencils, i.e. $A(x) \in \mathbb{C}[x]^{n\times n}$, and $A_1 \neq 0$ unless stated otherwise. Then $A(x_0, x_1) = A_0x_0 + A_1x_1$.

Definition 3.13.1 *A pencil $A(x) \in \mathbb{C}[x]^{n\times n}$ has property L if all the eigenvalues of $A(x_0, x_1)$ are linear functions. That is $\lambda_i(x_0, x_1) = \alpha_i x_0 + \beta_i x_1$ is an eigenvalue of $A(x_0, x_1)$ of multiplicity n_i for $i = 1, ..., m$, where*

$$n = \sum_{i=1}^{m} n_i, \quad (\alpha_i, \beta_i) \neq (\alpha_j, \beta_j), \text{ for } 1 \leq i < j \leq m.$$

The proofs of the following propositions is left to the reader. (See Problems 1-2.)

Proposition 3.13.2 *Let $A(x) = A_0 + xA_1$ be a pencil in $\mathbb{C}[x]^{n\times n}$. TFAE:*
(a) $A(x)$ has property L.
(b) The eigenvalues of $A(x)$ are polynomials of degree 1 at most.
(c) The characteristic polynomial of $A(x)$ splits to linear factors over $\mathbb{C}[x]$.
(d) There is an ordering of the eigenvalues of A_0 and A_1, $a_1, ..., a_n$ and $b_1, ..., b_n$, respectively, such that the eigenvalues of $A_0x_0 + A_1x_1$ are $a_1x_0 + b_1x_1, ..., a_nx_0 + b_nx_1$.

Proposition 3.13.3 *Let $A(x)$ be a pencil in $\mathbb{C}[x]^{n\times n}$. Then $A(x)$ has property L if one of the following conditions hold:*
(a) $A(x)$ is similar over $\mathbb{C}(x)$ to an upper triangular matrix $U(x) \in \mathbb{C}(x)^{n\times n}$.
(b) $A(x)$ is strictly similar to an upper triangular pencil $U(x) = U_0 + U_1x$, i.e. U_0, U_1 are upper triangular.
(c) $A(x)$ is similar over $\mathbb{C}[x]$ to a diagonal matrix $B(x) \in \mathbb{C}[x]^{n\times n}$.
(d) $A(x)$ is strictly similar to diagonal pencil.

Note that for pencils with property L any two distinct eigenvalues are not tangent at any point of $\mathbb{P}$. For pencils one can significantly improve Theorem 3.12.6.

Theorem 3.13.4 *Let $A(x) = A_0 + A_1x \in \mathbb{C}[x]^{n\times n}$ be a nonconstant pencil ($A_1 \neq 0$). Assume that $A(x)$ is pointwise diagonable on $\mathbb{C}$. Then $A(x)$ has property L. Furthermore $A(x)$ is similar over $\mathbb{C}[x]$ to a diagonal pencil $B(x) = B_0 + B_1x$. Suppose furthermore that A_1 is diagonable, i.e. $A(x_0, x_1)$ is pointwise diagonable on $\mathbb{C}^2$. Then $A(x)$ is strictly similar to the diagonal pencil $B(x)$, i.e. A_0 and A_1 are commuting diagonable matrices.*

Proof. We follow the proof of Theorem 3.12.6. Let $\lambda_1(x), ..., \lambda_m(x)$ be the eigenvalues of $A(x)$ of multiplicities $n_1, ..., n_m$ respectively, where each $\lambda_j(x)$ is viewed as multivalued function of x. More precisely, there exists an irreducible polynomial

$$\begin{aligned}&\phi(x,\lambda) = \lambda^p + \sum_{q=1}^{p} \phi_q(x)\lambda^{p-q} \in \mathbb{C}[x,\lambda],\\ (3.13.1)\qquad & \\ &\phi(x,\lambda)|\det\,(\lambda I - A(x)),\end{aligned}$$

such that $\lambda_j(x)$ satisfies the algebraic equation

$$(3.13.2)\qquad \phi(x,\lambda) = 0.$$

Moreover all branches generated by $\lambda_j(x)$ on $\mathbb{C}$ will generate all the solutions $\lambda(x)$ of (3.13.2). Equivalently all pairs (x,λ) satisfying (3.13.2) form an affine algebraic variety $V_0 \subset \mathbb{C}^2$. If we compactify V_0 to a projective variety $V \subset \mathbb{P}^2$ then V is a compact Riemann surface. $V\backslash V_0$ consists of a finite number of points, the points of V_0 at infinity. The compactification of V_0 is equivalent to considering $\lambda_j(x)$ as a multivalued function on $\mathbb{P}$. See for example [GuR65]. Note that any local solution of (3.13.2) is some eigenvalue $\lambda_i(x)$ of $A(x)$. Since $A(\zeta)$ is diagonable at $\zeta \in \mathbb{C}$ Theorem 3.8.1 implies that the Puiseaux expansion of $\lambda_j(x)$ around ζ in (3.12.3) is of the form

$$\lambda_j(x) = \lambda_j(\zeta) + \sum_{k=s}^{\infty} \lambda_{jk}(\zeta)(x-\zeta)^{\frac{k}{s}}.$$

Then

$$\frac{d\lambda_j(x)}{dx} = \sum_{k=s}^{\infty} \frac{k}{s}\lambda_{jk}(\zeta)(x-\zeta)^{\frac{k-s}{s}}.$$

So $\frac{d\lambda_j(x)}{dx}$ is a multivalued locally bounded function on $\mathbb{C}$. Equivalently, using the fact that $\lambda_j(x)$ satisfy (3.13.2) we deduce

$$(3.13.3)\qquad \frac{d\lambda_j(x)}{dx} = -\frac{\frac{\partial\phi(x,\lambda)}{\partial x}}{\frac{\partial\phi(x,\lambda)}{\partial y}}.$$

Hence $\frac{d\lambda_j(x)}{dx}$ is a rational function on V, which is analytic on V_0 in view of the assumption that $A(x)$ is pointwise diagonable in $\mathbb{C}$. The Puiseaux expansion of $\lambda_j(x)$ it ∞ (3.12.4) is

$$\lambda_j(x) = x\sum_{k=0}^{\infty} \lambda_{jk}(\infty)x^{-\frac{k}{s}}.$$

Hence

$$\frac{d\lambda_j(x)}{dx} = \lambda_{j0}(\infty) + \sum_{k=1}^{\infty} \frac{s-k}{s}\lambda_{jk}(\infty)x^{-\frac{k}{s}}.$$

That is the multivalued function $\frac{d\lambda_j(x)}{dx}$ is bounded at the neighborhood of ∞. Equivalently the rational function in (3.13.3) is bounded at all points of $V\backslash V_0$. Thus the rational function in (3.13.3) is bounded on a compact Riemann surface (3.13.2). Hence it must be constant, i.e. $\frac{d\lambda_j(x)}{dx} = b_j \Rightarrow \lambda_j(x) = a_j + b_j x$. So we have property L by part (b) of Proposition 3.13.2. In particular two distinct eigenvalues of $A(x)$ are not tangent at any $\zeta \in \mathbb{P}$. The first part of Theorem 3.12.6 implies that $A(x)$ is similar to $B(x) = \sum_{j=1}^{m} \oplus(a_j + b_j x)I_{n_j}$ over $\mathbb{C}[x]$.

Assume now that A_1 is diagonable. Then the second part of Theorem 3.12.6 yields that $A(x)$ is strictly similar to $B(x)$, which is equivalent to the assumption that A_0, A_1 are commuting diagonable matrices (Theorem 3.11.4). □

Theorem 3.13.5 *Let $A(x) = A_0 + A_1 x \in \mathbb{C}[x]^{n\times n}$. Assume that A_1 and A_2 are diagonable and $A_0A_1 \neq A_1A_0$. Then exactly one of the following conditions hold:*
(a) $A(x)$ is not diagonable exactly at the points $\zeta_1, ..., \zeta_p$, where $1 \le p \le n(n-1)$.
(b) $A(x)$ is diagonable exactly at the points $\zeta_1 = 0, ..., \zeta_q$ for some $q \ge 1$.

Proof. Combine the assumptions of the theorem with Theorem 3.13.4 to deduce the existence of $0 \neq \zeta \in \mathbb{C}$ such that $A(\zeta)$ is not diagonable. Consider the homogenized pencil $A(x_0, x_1) = A_0x_0 + A_1x_1$. Let

$$C(p_1, ..., p_k)(x_0, x_1) = \oplus_{j=1}^{k} C(p_j) \in \mathbb{C}[x_0, x_1]^{n\times n},$$

$$\prod_{i=1}^{k} p_i(x_0, x_1, \lambda) = \det\,(\lambda I - A(x_0, x_1)),$$

$$p_i(x_0, x_1, \lambda) = \lambda^{m_i} + \sum_{j=1}^{m_i} \lambda^{m_i - j} p_{ij}(x) \in \mathbb{C}[x_0, x_1][\lambda], 1 \le m_i \quad i = 1, ..., k,$$

$$p_1 | p_2 | ... | p_k,$$

be the rational canonical form $A(x_0, x_1)$ over $\mathbb{C}(x_0, x_1)$. (See 2.3.) That is each $p_i(x_0, x_1, \lambda)$ is a nontrivial invariant polynomial of $\lambda I - A(x_0, x_1)$. Hence each $p_i(x_0, x_1, \lambda)$ is a homogeneous polynomial of degree m_i in x_0, x_1, λ. Furthermore $A(x_0, x_1) = S(x_0, x_1))C(p_1, ..., p_k)(x_0, x_1)S(x_0, x_1)^{-1}$ for some

$S(x_0, x_1) \in \mathbb{C}[x_0, x_1]^{n\times n} \cap \mathbf{GL}(n, \mathbb{C}(x_o, x_1))$. Choose $x_0 = \tau \neq 0$ such that $\det S(\tau, x_1)$ is not identically zero in x_1. Then $A(x) = A(1, x)$ is pointwise similar to
$\frac{1}{\tau}C(p_1, ..., p_k)(\tau, \tau x)$ at all point for which $\det S(\tau, \tau x) \neq 0$, i.e. at all but a finite number of points in $\mathbb{C}$.

Since $\mathbb{C}[x_0, x_1, \lambda]$ is $\mathbb{D}_u$ then $p_k(x_0, x_1, \lambda) = \prod_{i=1}^{r} \phi_i(x_0, x_1, \lambda)^{\ell_i}$, where each ϕ_i is a nonconstant irreducible (homogeneous) polynomial and ϕ_i is coprime with ϕ_j for $i \neq j$. Assume first that some $\ell_i > 1$. Then $C(p_k)(\tau, t)$ has a multiple eigenvalue for any $t \in \mathbb{C}$, hence it is not diagonable. That is the condition (b) of the theorem holds.

Assume now that $\ell_1 = ... = \ell_r = 1$. This is equivalent to the assumption that $p_k(x_0, x_1, \lambda) = 0$ does not have multiple roots for some (x_0, x_1). We claim that it is possible to choose $\tau \neq 0$ such that $p_k(\tau, x_1, \lambda)$ has m_k pairwise distinct roots (in λ) except in the points $\zeta_1, ..., \zeta_q$. Consider the discriminant $D(x_0, x_1)$ of $p_k(x_0, x_1, \lambda) \in \mathbb{C}[x_0, x_1][\lambda]$. See 1.8. Since p_{ki} is a homogeneous polynomial of degree i for $i = 1, ..., m_k$ it follows that $D(x_0, x_1)$ is homogeneous polynomial of degree $m_k(m_k - 1) \leq n(n-1)$. Since $p_k(x_0, x_1, \lambda) = 0$ does not have multiple roots for some (x_0, x_1) it follows that $D(x_0, x_1)$ is not a zero polynomial, and $p_k(x_0, x_1, \lambda) = 0$ has a multiple root if and only if $D(x_0, x_1) = 0$. Choose $\tau \neq 0$ such that $D(\tau, x_1)$ is not a zero polynomial. Let $\zeta_1, ..., \zeta_q$ be the distinct roots of $D(\tau, \tau x) = 0$. Since the degree of $D(x_0, x_1)$ is at most $n(n-1)$ it follows that the degree of $D(\tau, x)$ is at most $n(n-1)$. Hence $0 \leq q \leq n(n-1)$. By the definition of the invariant polynomials it follows that $p_k(x_0, x_1, A(x_0, x_1)) = 0$. Hence $p_k(\tau, \tau t, A(\tau, \tau t)) = 0$. Let $t \in X = \mathbb{C}\backslash\{\zeta_1, ..., \zeta_q\}$. Since $p_k(\tau, \tau t, \lambda)$ has m_k distinct roots, which are all eigenvalues of $A(\tau, \tau t)$ it follows that $A(\tau, \tau t) = \tau A(t)$ is a diagonable matrix. □

For $n = 2$ the case (b) in Theorem 3.13.5 does not arise. See Problem 4. We do not know if the case (b) of Theorem 3.13.5 arises. Recall that a hermitian matrix $A \in \mathbb{C}^{n\times n}$, $\bar{A}^\top = A$ is always diagonable.

Definition 3.13.6 *A pencil $A(x) = A_0 + A_1 x$ is called hermitian if $A_1, A_2 \in \mathbb{C}^{n\times n}$ are hermitian.*

Theorem 3.13.7 *Let $A(x) = A_0 + A_1 x \in \mathbb{C}[x]^{n\times n}$ be a hermitian pencil. Assume that $A_0 A_1 \neq A_1 A_0$. Then there exists $2q$ distinct complex points $\zeta_1, \bar{\zeta}_1 ..., \zeta_q, \bar{\zeta}_q \in \mathbb{C}\backslash\mathbb{R}$, $1 \leq q \leq \frac{n(n-1)}{2}$ such that $A(x)$ is not diagonable if and only if $x \in \{\zeta_1, \bar{\zeta}_1, ..., \zeta_q, \bar{\zeta}_q\}$.*

Proof. Clearly $A(x)$ is a hermitian matrix for any real x. Hence $A(x)$ is diagonable for $x \in \mathbb{R}$. Thus the condition (a) of Theorem 3.13.5 holds.

Assume that $A(\zeta)$ is not diagonable. Then $\zeta \in \mathbb{C}\backslash\mathbb{R}$. Let $A(\zeta) = QJQ^{-1}$, where J is a Jordan canonical form of $A(\zeta)$. Then

$$A(\bar{\zeta}) = \overline{A(\zeta)}^\top = (\bar{Q}^\top)^{-1}\bar{J}^\top\bar{Q}^\top.$$

Hence $A(\bar{\zeta})$ is not diagonable. Thus the number of distinct points ζ for which $A(\zeta)$ is not diagonable is $1 \le 2q \le n(n-1)$. □

The points $\zeta_1, ..., \zeta_q$ are called the resonance states of the hermitian pencil $A(x)$. They are important in certain chemical models [MoF80].

Problems

1. Prove Proposition 3.13.2.

2. (a) Show that property L is equivalent to the condition (a) of Proposition 3.13.3.

   (b) Prove the other conditions of Proposition 3.13.3.

3. Show that a pencil $A(x) = A_0 + A_1x \in \mathbb{C}[x]^{2\times 2}$ have property L if and only if $A(x)$ is strictly similar to an upper triangular pencil.

4. Let $A(x_0, x_1) = A_0x_0 + A_1x_1 \in \mathbb{C}[x_0, x_1]^{2\times 2}$. Then exactly one the following conditions hold.

   (a) $A(x_0, x_1)$ is strictly similar to a diagonal pencil. (Property L holds).

   (b) $A(x_0, x_1)$ is not diagonable except exactly for the points $(x_0, x_1) \ne (0, 0)$ lying on a line $ax_0 + bx_1 = 0$. (Property L holds, $A_0A_1 = A_1A_0$ but A_i is not diagonable for some $i \in \{1, 2\}$, $A(x_0, x_1)$ has a double eigenvalue.)

   (c) $A(x_0, x_1)$ is diagonable except exactly for the points $(x_0, x_1) \ne (0, 0)$ lying on a line $ax_0 + bx_1 = 0$. (Property L holds.)

   (d) $A(x_0, x_1)$ is diagonable except exactly the points $(x_0, x_1) \ne (0, 0)$ which lie on two distinct lines in $\mathbb{C}^2$. (Property L does not hold.)

5. Let
$$A_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ -1 & -1 & -2 \end{bmatrix}.$$

   (a) Show that A_0, A_1 are nilpotent while $A_0 + A_1$ is nonsingular.

   (b) Show that $A(x) = A_0 + A_1x$ does not have property L.

   (c) Show that $A(x)$ is diagonable for all $x \ne 0$.

3.14 Strict similarity of pencils and analytic similarity

Let $A(x) = A_0 + A_1x, B(x) = B_0 + B_1x \in \mathbb{C}[x]^{n\times n}$. Recall the notion of strict equivalence $A(x)\overset{s}{\sim}B(x)$ (2.1) and strict similarity $A(x)\overset{s}{\approx}B(x)$ (3.11). Clearly $A(x)\overset{s}{\approx}B(x) \Rightarrow A(x)\overset{s}{\sim}B(x)$. (2.9.3) yields.

Proposition 3.14.1 *Let $A(x), B(x) \in \mathbb{C}[x]^{n\times n}$ be two strictly similar pencils. Then the three pencils in* (3.6.2) *are strictly equivalent.*

Using Kronecker's result (Theorem 2.1.7) we can determine if the three pencils in (3.6.2) are strictly equivalent. We now study the implications of Proposition 3.14.1.

Lemma 3.14.2 *Let $A(x) = A_0 + A_1x, B(x) = B_0 + B_1x \in \mathbb{C}[x]^{n\times n}$ be two pencils such that*

$$I \otimes A(x) - A(x)^\top \otimes I \overset{s}{\sim} I \otimes A(x) - B(x)^\top \otimes I. \tag{3.14.1}$$

Then there exists two nonzero $U, V \in \mathbb{C}^{n\times n}$ such that

$$A(x)U - UB(x) = 0, \quad VA(x) - B(x)V = 0. \tag{3.14.2}$$

In particular

$$A_0 \ker V, A_1 \ker V \subset \ker V, \quad B_0 \ker U, B_1 \ker U \subset \ker U. \tag{3.14.3}$$

Proof. As $A(x)I - IA(x) = 0$ it follows that the kernels of $I \otimes A(x) - A(x)^\top \otimes I \in \mathbb{C}[x]^{n^2\times n^2}$ and its transpose contain a nonzero vector $\hat{I}_n \in \mathbb{C}^{n^2}$ which is induced by I_n. (See 2.8.) Hence the kernel of $I \otimes A(x) - B(x)^\top \otimes I$ contain nonzero constant vectors. This is equivalent to (3.14.2).

Assume that (3.14.2) holds. Let $\mathbf{x} \in \ker V$. Multiply the second equality in (3.14.2) from the right by $\mathbf{x}$ to deduce the first part (3.14.3). The second part of (3.14.3) is obtained similarly. □

Definition 3.14.3 $A_0, A_1 \in \mathbb{C}^{n\times n}$ *have a common invariant subspace if there exist a subspace* $\mathbf{U} \subset \mathbb{C}^n$, $1 \le \dim \mathbf{U} \le n-1$ *such that* $A_0\mathbf{U}, A_1\mathbf{U} \subset \mathbf{U}$.

The following claims are left to the reader (see Problems 1-2).

Proposition 3.14.4 *Let $A(x) = A_0 + xA_1 \in \mathbb{C}[x]^{n\times n}$. Then $A(x)$ is strictly similar to an upper triangular pencil*

$$B(x) = \begin{bmatrix} B_{11}(x) & B_{12}(x) \\ 0 & B_{22}(x) \end{bmatrix},$$

$$B_{11}(x) \in \mathbb{C}[x]^{n_1\times n_1},\ B_{12}(x) \in \mathbb{C}[x]^{n_1\times n_2},\ B_{22}(x) \in \mathbb{C}[x]^{n_2\times n_2}, \tag{3.14.4}$$
$$1 \le n_1, n_2,\ n_1 + n_2 = n,$$

if and only if A_0, A_1 have a common invariant subspace.

Proposition 3.14.5 *Assume that $A(x) \in \mathbb{C}[x]^{n\times n}$ is similar over $\mathbb{C}(x)$ to an upper triangular matrix $B(x)$ of the form* (3.14.4)*. Then* det $(\lambda I - A(x)) \in \mathbb{C}[x, \lambda]$ *is reducible.*

Theorem 3.14.6 *Let $A(x) = A_0 + A_1x, B(x) = B_0 + B_1x \in \mathbb{C}[x]^{n\times n}$. Assume that either* det $(\lambda I - A(x))$ *or* det $(\lambda I - B(x))$ *is irreducible over $\mathbb{C}[x, \lambda]$. Then $A(x)\overset{s}{\approx}B(x)$ if and only if* (3.14.1) *holds.*

Proof. Assume that (3.14.1) holds. Suppose that det $(\lambda I - A(x))$ is irreducible. Propositions 3.14.4-3.14.5 imply that A_0, A_1 do not have a common invariant subspace. Lemma 3.14.2 implies that the matrix V in (3.14.2) is invertible, i.e. $B(x) = VA(x)V^{-1}$. Similarly if det $(\lambda I - B(x))$ is irreducible then $B(x) = U^{-1}A(x)U$. □

Definition 3.14.7 *Let $\mathcal{I}_n \subset (\mathbb{C}^{n\times n})^2$ be the set of all pairs (A_0, A_1) such that* det $(\lambda I - (A_0 + A_1x))$ *irreducible.*

We will show later that $\mathcal{I}_n = (\mathbb{C}^{n\times n})^2\backslash X_n$ where X_n is a strict subvariety of $(\mathbb{C}^{n\times n})^2$. That is, for most of the pencils $A(x), (A_0, A_1) \in (\mathbb{C}^{n\times n})^2$ det $(\lambda I - A(x))$ is irreducible. Clearly if $(A_0, A_1)\overset{s}{\approx}(B_0, B_1)$ then either $(A_0, A_1), (B_0, B_1) \in \mathcal{I}_n$ or $(A_0, A_1), (B_0, B_1) \notin \mathcal{I}_n$.

Corollary 3.14.8 *Let $(A_0, A_1), (B_0, B_1) \in \mathcal{I}_n$. Then $A(x) = A_0 + A_1x$ is strictly similar to $B(x) = B_0 + B_1x$ if and only if* (3.14.1) *holds.*

We now discuss the connection between the notion of analytic similarity of matrices over H_0 and strict similarity of pencils. Let $A(x), B(x) \in \mathrm{H}_0^{n\times n}$ and assume that $\eta(A, A) = 1$. Suppose that $A\overset{r}{\approx}B(x)$. Theorem 3.9.2 claims that $A(x) \approx B(x)$ if and only if there exists two matrices $T_0 \in$ $\mathbf{GL}(n, \mathbb{C})$, $T_1 \in \mathbb{C}^{n\times n}$ such that

$$A_0T_0 = T_0B_0, \quad A_1T_0 + A_0T_1 = T_0B_1 + T_1B_0.$$

Let

$$F(A_0, A_1) = \begin{bmatrix} A_0 & A_1 \\ 0 & A_0 \end{bmatrix} \in \mathbb{C}^{2n\times 2n}. \tag{3.14.5}$$

Then (3.9.2) is equivalent in this case to

$$F(A_0, A_1)F(T_0, T_1) = F(T_0, T_1)F(B_0, B_1). \tag{3.14.6}$$

As det $F(T_0, T_1) = (\det T_0)^2$ it follows that T_0 is invertible if and only if $F(T_0, T_1)$ is invertible.

Definition 3.14.9 *Let $A_0, A_1, B_0, B_1 \in \mathbb{C}^{n\times n}$. Then $F(A_0, A_1)$ and $F(B_0, B_1)$ are called strongly similar ($F(A_0, A_1) \cong F(A_0, A_1)$) if there exists $F(T_0, T_1) \in \mathbf{GL}(2n, \mathbb{C})$ such that* (3.14.6) *holds.*

Clearly $F(A_0, A_1) \cong F(A_0, A_1) \Rightarrow F(A_0, A_1) \approx F(B_0, B_1)$. It can be shown that the notion of strong similarity is stronger that the notion of similarity. (Problem 10.)

Proposition 3.14.10 *The matrices $F(A_0, A_1)$ and $F(B_0, B_1)$ are strongly similar if and only if the pencils*

$$A(x) = F(0, I) + F(A_0, A_1)x, \quad B(x) = F(0, I) + F(B_0, B_1)x$$

are strictly similar.

Proof. Let $[P_{ij}]_1^2 \in \mathbb{C}^{2n\times 2n}$. Then $F(0, I)P = PF(0, I)$ if and only if $P_{11} = P_{22},\ P_{21} = 0$. That is $P = F(P_{11}, P_{12})$ and the proposition follows. □

Clearly $F(A_0, A_1) \cong F(A_0, B_1) \Rightarrow A_0 \approx B_0$. Without loss of generality we may assume that $A_0 = B_0$. (See Problem 5.) Consider all matrices T_0, T_1 satisfying (3.14.6). For $A_0 = B_0$ (3.14.6) reduces to

$$A_0T_0 = T_0A_0, \quad A_0T_1 - T_1A_0 = T_0B_1 - A_1T_0.$$

Theorem 2.10.1 yields that the set of matrices T_0 which satisfies the above conditions is of the form

$$\begin{aligned} \mathcal{P}(A_1, B_1) = \{T_0 \in C(A_0) :& \\ \operatorname{tr}(V(T_0B_1 - A_1T_0)) = 0,\ \text{for all } V \in C(A_0)\}.& \end{aligned} \tag{3.14.7}$$

Hence

Proposition 3.14.11 *Suppose that $F(A_0, A_1) \cong F(A_0, B_1)$. Then*

$$\dim \mathcal{P}(A_1, A_1) = \dim \mathcal{P}(A_1, B_1) = \dim \mathcal{P}(B_1, B_1). \tag{3.14.8}$$

As in Theorem 2.9.5 for a fixed A_0, A_1 there exists a neighborhood $D(A_1, \rho)$ such that the first two equalities in (3.12.7) imply that $F(A_0, A_1) \cong F(A_0, B_1)$ for all $B_1 \in D(A_1, \rho)$ (Problem 4).

We now considering a splitting result analogous to Theorem 3.5.7.

Theorem 3.14.12 *Assume that*

$$A_0 = \begin{bmatrix} A_{11}^{(0)} & 0 \\ 0 & A_{22}^{(0)} \end{bmatrix}, \quad A_{ii}^{(0)} \in \mathbb{C}^{n_i \times n_i},\ i = 1, 2, \tag{3.14.9}$$

where $A_{11}^{(0)}$ and $A_{22}^{(0)}$ do not have a common eigenvalue. Let

$$A_1 = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} \end{bmatrix}, \quad B_1 = \begin{bmatrix} B_{11}^{(1)} & B_{12}^{(1)} \\ B_{21}^{(1)} & B_{22}^{(1)} \end{bmatrix}$$

be the block partition of A_1, B_1 as the block partition of A_0. Then

$$\mathcal{P}(A_1, B_1) = \mathcal{P}(A_{11}^{(1)}, B_{11}^{(1)}) \oplus \mathcal{P}(A_{22}^{(1)}, B_{22}^{(1)}). \tag{3.14.10}$$

Moreover

$$F(A_0, A_1) \cong F(A_0, B_1) \iff F(A_{ii}^{(0)}, A_{ii}^{(1)}) \cong F(A_{ii}^{(0)}, B_{ii}^{(1)}) \text{ for } i = 1, 2.$$

Proof. According to Problem 4 $C(A_0) = C(A_{11}^{(0)}) \oplus C(A_{22}^{(0)})$. Then the trace condition in (3.14.7) reduces to

$$\operatorname{tr}(V_1(T_1^{(0)} B_{11}^{(1)} - A_{11}^{(1)} T_1^{(0)}) + V_2(T_2^{(0)} B_{22}^{(1)} - A_{22}^{(1)} T_2^{(0)})) = 0,$$

where

$$V = V_1 \oplus V_2,\ T_0^{(0)} = T_1^{(0)} \oplus T_2^{(0)} \in C(A_{11}^{(0)}) \oplus C(A_{22}^{(0)}).$$

Choosing either $V_0 = 0$ or $V_1 = 0$ we obtain (3.14.10). The right implication of the last claim of the theorem is straightforward. As $\det T_0 = \det T_1^{(0)} \det T_2^{(0)}$ it follows that $T_0 \in \mathbf{GL}(n, C) \iff T_i^{(0)} \in \mathbf{GL}(n_i, \mathbb{C})$, $i = 1, 2$. This establishes the left implication of the last claim of the theorem. □

Thus, the classification of strong similarity classes for matrices $F(A_0, A_1)$ reduces to the case where A_0 is nilpotent (Problem 6). In the case $A_0 = 0$ $F(0, A_1) \cong F(0, B_1) \iff A_1 \approx B_1$. In the case $A_0 = H_n$ the strong similarity classes of $F(H_n, A_1)$ classified completely (Problem 9). This case corresponds to the case discussed in Theorem 3.5.5. The case $A_0 = H_m \oplus H_m$ can be classified completely using the results of Problem 2 (Problem 3.14.11).

Problems

1. Prove Proposition 3.14.4.

2. Prove Proposition 3.14.5.

3. Let $A(x) \in \mathbb{C}[x]^{n\times n}$ and assume that $A(x)\mathbf{U} \subset \mathbf{U} \subset \mathbb{C}^n$ is a nontrivial invariant subspace of $A(x)$, i.e $1 \le \dim \mathbf{U} \le n-1$. Let $p(x,\lambda) \in \mathbb{C}[x,\lambda]$ be the minimal polynomial of $A(x)|\mathbf{U}$. Thus $1 \le \deg_{\lambda} p(x,\lambda) \le n-1$. Show that $p(x,\lambda)|\det(\lambda I - A(x))$. Hence $\det(\lambda I - A(x))$ is reducible over $\mathbb{C}[x,\lambda]$.

4. Modify the proof of Theorem 2.9.5 to show that for a fixed $A_0, A_1 \in \mathbb{C}^{n\times n}$ there exists $\rho > 0$ such that the first two equalities in (3.14.8) for $B_1 \in D(A_1,\rho)$ imply that $F(A_0,A_1) \cong F(A_0,B_1)$.

5. Show that for any $P \in \mathbf{GL}(n,\mathbb{C})$
$$F(A_0,A_1) \cong F(B_0,B_1) \iff F(A_0,A_1) \cong F(PB_0P^{-1}, PB_1P^{-1}).$$
Assume that $F(A_0,A_1) \cong F(B_0,B_1)$. Show that there exists $P \in \mathbf{GL}(n,\mathbb{C})$ such that $A_0 = PB_0P^{-1}$.

6. Show that for any $\lambda \in \mathbb{C}$
$$F(A_0,A_1) \cong F(B_0,B_1) \iff F(A_0-\lambda I, A_1) \cong F(B_0 - \lambda I, B_1).$$

7. Let $A_i \in \mathbb{C}^{n\times n},\ i = 0,...,s-1$. Define
$$F(A_0,...,A_{s-1}) = \begin{bmatrix} A_0 & A_1 & A_2 & ... & A_{s-1} \\ 0 & A_0 & A_1 & ... & A_{s-2} \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & ... & A_0 \end{bmatrix} \in \mathbb{C}^{sn\times sn}.$$
$F(A_0,...,A_{s-1})$ and $F(B_0,...,B_{s-1})$ are called strongly similar $(F(A_0,...,A_{s-1}) \cong F(B_0,...,B_{s-1}))$
$$F(A_0,...,A_{s-1}) = F(T_0,...,T_{s-1})F(B_0,...,B_{s-1})F(T_0,...,T_{s-1})^{-1},$$
$$F(T_0,...,T_{s-1}) \in \mathbf{GL}(sn,\mathbb{C}).$$
Show that $F(A_0,...,A_{s-1}) \cong F(B_0,...,B_{s-1})$ if and only if the equalities (3.9.2) hold for $k = 0,...,s-1$ and $T_0 \in \mathbf{GL}(n,\mathbb{C})$.

8. Let

$$Z = H_n \oplus ... \oplus H_n,\ X = [X_{pq}]_1^s,\ [Y_{pq}]_1^s \in \mathbb{C}^{sn\times sn},$$
$$X_{pq} = [x_{ij}^{(pq)}]_1^n,\ Y_{pq} = [y_{ij}^{(pq)}]_1^n \in \mathbb{C}^{n\times n}, \quad p, q = 1, ..., s.$$

Show that if each X_{pq} is an upper triangular matrix then

$$\det X = \prod_{r=1}^{n} \det\, [x_{rr}^{(pq)}]_{p,q=1}^s.$$

(Expand the determinant of X by the rows $n, 2n, ..., sn$ and use the induction.) Define

$$A_r = [a_{pq}^{(r)}]_1^s,\ B_r = [b_{pq}^{(r)}]_1^s \in \mathbb{C}^{s\times s},$$
$$a_{pq}^{(r)} = \sum_{i=1}^{r+1} x_{(n-r+i-1)i}^{(pq)},\ b_{pq}^{(r)} = \sum_{i=1}^{r+1} y_{(n-r+i-1)i}^{(pq)}, \quad r = 0, ..., n-1.$$

Using Theorem 2.8.3 show

$$F(Z, X) \cong F(Z, Y) \iff F(A_0, ..., A_{n-1}) \cong F(B_0, ..., B_{n-1}).$$

9. Let $X = [x_{ij}]_1^n,\ Y = [y_{ij}]_1^n \in \mathbb{C}^{n\times n}$. Using Problems 7-8 show that $F(H_n, X) \cong F(H_n, Y)$ if and only if

$$\sum_{i=1}^{r} x_{(n-r+i)i} = \sum_{i=1}^{r} y_{(n-r+i)i}, \quad \text{for } r = 1, ..., n.$$

10. Let $X = [x_{ij}]_1^2 \in \mathbb{C}^{2\times 2}$. Show that if $x_{21} \neq 0$ then $F(H_2, X) \cong H_4$. Combine this result with Problem 9 to show the existence of $Y \in \mathbb{C}^{2\times 2}$ such that $F(H_2, X)$ is similar to $F(H_2, Y)$ but $F(H_2, X)$ is not strongly similar to $F(H_2, Y)$.

(3.14.11)

Assume in Problem 8 $s = 2$. Let

$$A(x) = \sum_{i=0}^{n-1} A_i x^i,\ B(x) = \sum_{i=0}^{n-1} B_i x^i \in H_0^{2\times 2}.$$

Use the results of Problems 7-8, (3.9.2) and Problem 2 to show that $F(Z, X) \cong F(Z, Y)$ if and only if the three matrices in (3.6.2) have the same local invariant polynomials up to degree $n - 1$.

3.15 Historical remarks

The exposition of §3.1 is close to [Gan59]. The results of §3.2 were inspired by [Rot81]. The notion of local indices (Problem 3.2.3) can be found in [FrS80]. The content of §3.3 are standard. Theorem 3.3.2 can be found in [Wie67] and Problem §3.3.1 in [Gan59]. The use of Cauchy integration formula to study the properties of analytic functions of operators and matrices as in §3.4 is now common, e.g. [Kat80] and [Kat82]. Theorem 3.4.6 is standard. Theorem 3.4.9 is a part of the Kreiss matrix stability theorem [Kre62]. The inequality (3.4.16) is due to [Tad81]. The results of Problem 3.4.7 are from [Fri81]. The results of §3.5 influenced by Arnold [Arn71], in particular Theorem 3.5.3 is from [Arn71]. See also [Was77]. The subject of §3.6 and its applications in theory of differential equations in neighborhood of singularities was emphasized in works of Wasow [Was63],[Was77] and [Was78]. Theorem 3.6.4 for one complex variable appears in [Fri80b]. Corollary 3.6.5 is due to [Was63]. Theorem 3.7.1 for simply connected domain is due to [Gin78]. See [Was78] for the extension of Theorem 3.7.1 to certain domains $\Omega \subset \mathbb{C}^p$. It is shown there that Theorem 3.7.1 fails even for simply connected domains in $\mathbb{C}^3$.

Theorem 3.8.1 can be found in [Kat80] or [Fri78]. The results of §3.9-§3.10 were taken from [Fri80b]. It is worthwhile to mention that the conjecture stated in [Fri80b] that $A(x)$ and $B(x)$ are analytically similar over H_0 if the three matrices in (3.6.2) are equivalent over H_0 is false [Gur81, §6]. The contents of §3.11 are known to the experts. The nontrivial part of this section (Theorem 3.11.6) is due to [DDG51]. Some of the results in §3.12, in particular Theorem 3.12.4, seem to be new. Property L of §3.13 was introduced by Motzkin-Taussky [MoT52] and [MoT55]. Theorem 3.13.4 is a slight improvement of [MoT55]. Our proof of property L in Theorem 3.13.4 follows [Kat80]. Theorem 3.13.7 is taken from [MoF80]. Theorem 3.13.7 associates the "defective" points $\zeta_1, ..., \zeta_q$ with the resonance states of molecules. Many results in §3.14 are taken from [Fri80a] and [Fri80b]. It connects the analytic similarity of matrices with simultaneous similarity of certain pairs of matrices. Simultaneous similarity of matrices is discussed in [Fri83].

Chapter 4

Inner product spaces

4.1 Inner product

Definition 4.1.1 *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:*

$$
\begin{aligned}
&\text{(a)} \quad \langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle, \quad \text{for all } a, b \in \mathbb{F},\ \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V},\\
&\text{(br)} \quad \text{for } \mathbb{F} = \mathbb{R} \quad \langle \mathbf{y}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbf{V};\\
&\text{(bc)} \quad \text{for } \mathbb{F} = \mathbb{C} \quad \langle \mathbf{y}, \mathbf{x} \rangle = \overline{\langle \mathbf{x}, \mathbf{y} \rangle}, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbf{V};\\
&\text{(c)} \quad \langle \mathbf{x}, \mathbf{x} \rangle > 0 \quad \text{for all } \mathbf{x} \in \mathbf{V} \backslash \{\mathbf{0}\}.
\end{aligned}
$$

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

Other standard properties of inner products are mentioned in Problems 1-2. 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 4.1.2 *Let* $\mathbf{V}$ *be a vector space over* $\mathbb{R}$. *Identify* $\mathbf{V}_{\mathbb{C}}$ *with the set of pairs* $(\mathbf{x}, \mathbf{y}),\ \mathbf{x}, \mathbf{y} \in \mathbf{V}$. *Then* $\mathbf{V}_{\mathbb{C}}$ *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* $\mathbb{R}$ *then* $(\mathbf{e}_1, 0), ..., (\mathbf{e}_n, 0)$ *is a basis of* $\mathbf{V}_{\mathbb{C}}$ *over* $\mathbb{C}$. *Any inner product* $\langle \cdot, \cdot \rangle$ *on* $\mathbf{V}$ *over* $\mathbb{F}$ *induces the following inner product on* $\mathbf{V}_{\mathbb{C}}$:

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

We leave the proof of this proposition to the reader (Problem 3).

Definition 4.1.3 *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* $\mathbf{x} \in S, \ \mathbf{y} \in T$.
(c) For any $S \subset \mathbf{V}$ $S^{\perp} \subset \mathbf{V}$ *is the maximal orthogonal set to* S.
(d) $\mathbf{x}_1, ..., \mathbf{x}_m$ *is called an orthonormal set if*

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

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

Definition 4.1.4 *(***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* ($m \ge 0$)*. Then* $\tilde{S} = \{\mathbf{e}_1, ..., \mathbf{e}_p\}$ *is the orthonormal set* ($p \ge 1$) *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* $1 \le k < m$*. 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 4.1.5 *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:*

$$\begin{aligned} & \mathbf{y}_1 := \mathbf{x}_1, \ r_{11} := \|\mathbf{y}_1\|, \ \mathbf{e}_1 := \frac{1}{r_{11}}\mathbf{y}_1, \\ (4.1.1) \qquad & r_{ji} := \langle \mathbf{x}_i, \mathbf{e}_j \rangle, \ j = 1, ..., i-1, \\ & \mathbf{y}_i := \mathbf{x}_i - \sum_{j=1}^{i-1} r_{ji}\mathbf{e}_j, \ r_{ii} := \|y_i\|, \ \mathbf{e}_i := \frac{1}{r_{ii}}\mathbf{y}_i, \ i = 2, ..., n. \end{aligned}$$

In particular, $\mathbf{e}_i \in S_i$ *and* $\|\mathbf{y}_i\| = \mathrm{dist}(\mathbf{x}_i, S_{i-1})$*, where* $S_i = \mathrm{span}\ (\mathbf{x}_1, ..., \mathbf{x}_i)$ *for* $i = 1, ..., n$ *and* $S_0 = \{\mathbf{0}\}$*. (See Problem 4 for the definition of* $\mathrm{dist}(\mathbf{x}_i, S_{i-1})$*.)*

Corollary 4.1.6 *Any (ordered) basis in a finite dimensional IPS* $\mathbf{V}$ *induces an orthonormal basis by the Gram-Schmidt algorithm.*

See Problem 4 for some known properties related to the above notions.

Problems

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

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

2. 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:

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

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

(c) The triangle inequality

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

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

3. Prove Proposition 4.1.2.

4. 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

$$\mathbf{x} = \sum_{i=1}^{n}\langle \mathbf{x},\mathbf{e}_i\rangle\mathbf{e}_i. \tag{4.1.2}$$

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

$$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{n}\langle \mathbf{x},\mathbf{e}_i\rangle\langle \mathbf{y},\bar{}\mathbf{e}_i\rangle. \tag{4.1.3}$$

(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 $\tilde{S}\ne\emptyset$ 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 $0 \le m \le n$ such that

$$\begin{aligned} &\mathbf{e}_1, ..., \mathbf{e}_m \in S, \quad \text{span S} = \text{span} (\mathbf{e}_1, ..., \mathbf{e}_m),\\ &S^\perp = \text{span} (\mathbf{e}_{m+1}, ..., \mathbf{e}_n),\\ &(S^\perp)^\perp = \text{span S}. \end{aligned}$$

(e) Assume from here to the end of the problem that S is a subspace. 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

$$\begin{aligned} &\text{dist}(\mathbf{x}, S) := \|\mathbf{x} - P\mathbf{x}\| \le \|\mathbf{x} - \mathbf{w}\| \text{ for any } \mathbf{w} \in S\\ &\text{and equality} \iff \mathbf{w} = P\mathbf{x}. \end{aligned} \tag{4.1.4}$$

(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.

5. Let $X \in \mathbb{C}^{m\times n}$ and assume that $m \ge n$ and $\text{rank X} = \text{n}$. 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 $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{y}^*\mathbf{x}$. Perform the Gram-Schmidt algorithm (4.1.5) 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 $r_{ji},\ j \le i$ given by (4.1.1). Show that $\bar{Q}^T Q = I_n$ and $X = QR$. (This is the QR algorithm.) Show that if in addition $X \in \mathbb{R}^{m\times n}$ then Q and R are real valued matrices.

6. Let $C \in \mathbb{C}^{n\times n}$ and assume that $\{\lambda_1, ..., \lambda_n\}$ are n eigenvalues of C counted with their multiplicities. View C as an operator $C : \mathbb{C}^n \to \mathbb{C}^n$. View $\mathbb{C}^n$ as $2n$-dimensional vector space over $\mathbb{R}^{2n}$. Let $C = A + \sqrt{-1}B,\ A, B \in \mathbb{R}^{n\times n}$.

   a. Then $\hat{C} := \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \in \mathbb{R}^{(2n)\times(2n)}$ represents the operator $C : \mathbb{C}^n \to \mathbb{C}^n$ as an operator over $\mathbb{R}$ in suitably chosen basis.

   b. Show that $\{\lambda_1, \bar{\lambda}_1, ..., \lambda_n, \bar{\lambda}_n\}$ are the $2n$ eigenvalues of $\hat{C}$ counting with multiplicities.

   c. Show that the Jordan canonical form of $\tilde{C}$, 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 Special transformations in IPS

Proposition 4.2.1 *Let* $\mathbf{V}$ *be an IPS and* $T : \mathbf{V} \to \mathbf{V}$ *a linear transformation Then there exists a unique linear transformation* $T^* : \mathbf{V} \to \mathbf{V}$ *such that* $\langle T\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, T^*\mathbf{y}\rangle$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$.

See Problems 1-2.

Definition 4.2.2 *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 4.2.3 *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:

$$a_{ij} = \langle T\mathbf{e}_j, \mathbf{e}_i\rangle, \quad i, j = 1, ..., n. \tag{4.2.1}$$

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

$$\begin{array}{ll} (a) & T^* \text{ is represented by } A^\top, \\ (b) & T \text{ is selfadjoint } \iff A = A^\top, \\ (c) & T \text{ is anti selfadjoint } \iff A = -A^\top, \\ (d) & T \text{ is unitary } \iff A \text{ is orthogonal} \iff AA^\top = A^\top A = I, \\ (e) & T \text{ is normal } \iff A \text{ is normal} \iff AA^\top = A^\top A, \end{array}$$

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

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

See Problem 3.

Proposition 4.2.4 *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

Definition 4.2.5 *For a domain* $\mathbb{D}$ *with identity* 1 *let*

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

See Problem 5 for relations between these classes.

Theorem 4.2.6 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{C}$ *of dimension* n*. Then a linear transformation* $T : \mathbf{V} \to \mathbf{V}$ *is normal if and only if* $\mathbf{V}$ *has an orthonormal basis 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 2). 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 4.2.6 yields:

Corollary 4.2.7 *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 4.2.8 *Let* $\mathbf{V}$ *be an IPS over* $\mathbb{C}$. *Let* $T \in \mathbf{N}(\mathbf{V})$. *Then*

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

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 4.2.4 and Corollary 4.2.7 with the above proposition to deduce:

Corollary 4.2.9 *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 4.2.10 *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 $\mathbf{V}_i = 2$, spec $(\mathrm{T}|\mathbf{V}_i) \subset S^1 \backslash \{-1, 1\}$ *for* $i = 2, ..., k$.

See Problem 7.

Proposition 4.2.11 *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 $\mathbf{V}_0 \geq 0$,
(b) $T\mathbf{V}_i = \mathbf{V}_i$, dim $\mathbf{V}_i = 2$, spec $(\mathrm{T}|\mathbf{V}_i) \subset \sqrt{-1}\mathbb{R}\backslash\{0\}$ *for* $i = 2, ..., k$.

See Problem 8.

Theorem 4.2.12 *Let $\mathbf{V}$ be an IPS over $\mathbb{C}$ of dimension n. Let $T \in$ Hom $(\mathbf{V})$. 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:*
(4.2.2)

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

Let $\mathbf{V}$ be an IPS over $\mathbb{R}$ of dimension n. Let $T \in$ 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 (4.2.2) 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 $\mathbf{U} :=$ span $(\mathbf{g}_1)^\perp$. Let P be the orthogonal projection on $\mathbf{U}$. Let $T_1 := PT|_{\mathbf{U}}$. Then $T_1 \in$ 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

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

It is straightforward to show that Tspan $(\mathbf{g}_1, ..., \mathbf{g}_i) \subset$ 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 6 and 12 to deduce:

Corollary 4.2.13 *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 $B = (b_{ij})_1^n \in \mathbb{C}^{n\times n}$, such that $b_{ii} = \lambda_i$, $i = 1, ..., n$, and a unitary matrix $U \in \mathbf{U}_n$ such that $A = UBU^{-1}$. If $A \in \mathbf{N}(n, \mathbb{C})$ then B is a diagonal matrix.*

Let $A \in \mathbb{R}^{n\times n}$ and assume that 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 11).

Definition 4.2.14 *Let $\mathbf{V}$ be a vector space and assume that $T : \mathbf{V} \to \mathbf{V}$ is a linear operator. Let $0 \neq \mathbf{v} \in \mathbf{V}$. Then $\mathbf{W} = \mathrm{span}\,(\mathbf{v}, T\mathbf{v}, T^2\mathbf{v}, \ldots)$ 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 4.2.15 *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 $\mathbf{W} = \mathrm{span}\,(\mathbf{v}, T\mathbf{v}, ..., T^{r-1}\mathbf{v})$ be a cyclic T-invariant subspace of dimension r generated by $\mathbf{v}$. Let $\mathbf{u}_1, ..., \mathbf{u}_r$ be an orthonormal basis of $\mathbf{W}$ obtained by the Gram-Schmidt process from the basis $[\mathbf{v}, T\mathbf{V}, ..., T^{r-1}\mathbf{v}]$ of $\mathbf{W}$. Then $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = 0$ for $1 \le i \le j-2$, i.e. the representation matrix of $T|\mathbf{W}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ 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 $\mathbf{W}_j = \mathrm{span}\,(\mathbf{v}, \ldots, T^{j-1}\mathbf{v})$ 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 $\mathrm{span}\,(\mathbf{u}_1, ..., \mathbf{u}_j) = \mathbf{W}_j$ for $j = 1, \ldots, r$. Let $r \ge j \ge i+2$. Then $T\mathbf{u}_i \in T\mathbf{W}_i \subset \mathbf{W}_{i+1}$. As $\mathbf{u}_j \perp \mathbf{W}_{i+1}$ it follows that $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = 0$. Assume that $T^* = T$. Let $r \ge i \ge j+2$. Then $\langle T\mathbf{u}_i, \mathbf{u}_j\rangle = \langle \mathbf{u}_i, T\mathbf{u}_j\rangle = 0$. Hence the representation matrix of $T|\mathbf{W}$ in the basis $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ is a tridiagonal hermitian matrix. □

Problems

1. Prove Proposition 4.2.1.

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

3. Prove Proposition 4.2.3.

4. Prove Proposition 4.2.4 for finite dimensional $\mathbf{V}$. (*Hint*: Choose an orthonormal basis in $\mathbf{V}$.)

5. Show the following

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

6. Let $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$ be an orthonormal basis in IPS $\mathbf{V}$ over $\mathbb{F}$. Let $G = \{\mathbf{g}_1, ..., \mathbf{g}_n\}$ be another basis in $\mathbf{V}$. Show that F is an orthonormal basis if and only if the tranfer matrix either from E to G or from G to E is a unitary matrix.

7. Prove Proposition 4.2.10

8. Prove Proposition 4.2.11

9. a. Show that $A \in \mathbf{SO}(2,\mathbb{R})$ is of the form $A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \theta \in \mathbb{R}$.

b. Show that $\mathbf{SO}(2,\mathbb{R}) = e^{\mathbf{AS}(2,\mathbb{R})}$. That is for any $B \in \mathbf{AS}(2,\mathbb{R})$ $e^B \in \mathbf{SO}(2,\mathbb{R})$ and any $A \in \mathbf{SO}(n,\mathbb{R})$ 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 4.2.10 and 4.2.11 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 $\mathbf{z}_i(0) = \mathbf{x}_i, \mathbf{z}_i(1) = \mathbf{y}_i, i = 1, ..., p$.

10. a. Show that $\mathbf{U}_n = e^{\mathbf{AH}_n}$. (*Hint*: Use Proposition 4.2.8 and its proof.)

    b. Show that $\mathbf{U}_n$ is path connected.

    c. Prove Problem 9e for $\mathbb{F} = \mathbb{C}$.

11. 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) $A \in \mathbf{N}(n, \mathbb{R})$, $\sigma(A) \subset \mathbb{R} \iff A = UDU^\top$, $U \in \mathbf{SO}_n$, $D \in \mathbf{D}(n, \mathbb{R})$.

12. Show that an upper triangular or a lower triangular matrix $B \in \mathbb{C}^{n\times n}$ is normal if and only if B is diagonal. (**Hint**: consider the equality $(BB^*)_{11} = (B^*B)_{11}$.)

13. Let the assumptions of Theorem 4.2.15 hold. Show that instead of performing the Gram-Schmidt process on $\mathbf{v}, T\mathbf{v}, ..., T^{r-1}\mathbf{v}$ one can perform the following process. Let $\mathbf{w}_1 := \frac{1}{\|\mathbf{v}\|}\mathbf{v}$. Assume that one already obtained i orthonormal vectors $\mathbf{w}_1, ..., \mathbf{w}_i$. Let $\tilde{\mathbf{w}}_{i+1} := T\mathbf{w}_i - \sum_{j=1}^{i} \langle T\mathbf{w}_i, \mathbf{w}_j \rangle \mathbf{w}_j$. If $\tilde{\mathbf{w}}_{i+1} = 0$ then stop the process, i.e. one is left with i orthonormal vectors. If $\mathbf{w}_{i+1} \neq 0$ then $\mathbf{w}_{i+1} := \frac{1}{\|\tilde{\mathbf{w}}_{i+1}\|}\tilde{\mathbf{w}}_{i+1}$ and continue the process. Show that the process ends after obtaining r orthonormal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ and $\mathbf{u}_i = \mathbf{w}_i$ for $i = 1, ..., r$. (This is a version of *Lanczos tridiagonalization* process.)

4.3 Symmetric bilinear and hermitian forms

Definition 4.3.1 *Let* $\mathbf{V}$ *be a module over* $\mathbb{D}$ *and* $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{D}$. Q *is called a symmetric bilinear 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* $a, b \in \mathbb{D}$ *and* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$ *(bilinearity).*

For $\mathbb{D} = \mathbb{C}$ 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 1-2):

Proposition 4.3.2 *Let* $\mathbf{V}$ *be a module over* $\mathbb{D}$ *with a basis* $E = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$. *Then there is* $1-1$ *correspondence between a symmetric bilinear form* Q *on* $\mathbf{V}$ *and* $A \in \mathbf{S}(n, \mathbb{D})$*:*

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

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 4.3.3 *Let the assumptions of Proposition 4.3.2 hold. Then* A *is called the representation matrix of* Q *in the basis* E.

Proposition 4.3.4 *Let the assumptions of Proposition 4.3.2 Let* $F = \{\mathbf{f_1}, ..., \mathbf{f_n}\}$ *be another basis of the* $\mathbb{D}$ *module* $\mathbf{V}$. *Then the symmetric bilinear form* Q *is represented by* $B \in \mathbf{S}(n, \mathbb{D})$ *in the basis* F*, where* B *is congruent* A*:*

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

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

$$B = U^* AU, \quad U \in \mathbf{GL}(n, \mathbb{C})$$

and U *is the matrix corresponding to the basis change from* F *to* E.

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

Proposition 4.3.5 *Let* $\mathbf{V}$ *be an* n *dimensional vector space over* $\mathbb{R}$. *Let* $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ *be a symmetric bilinear form. 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 3

Normalization 4.3.6 *Let $\mathbf{V}$ is a finite dimensional IPS over $\mathbb{F}$. Let $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ be either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$. Then a representation matrix A of Q is chosen with respect to an orthonormal basis E.*

The following proposition is straightforward (see Problem 4).

Proposition 4.3.7 *Let $\mathbf{V}$ is an n-dimensional IPS over $\mathbb{F}$. Let $Q : \mathbf{V} \times \mathbf{V} \to \mathbb{F}$ be either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$. Then there exists a unique $T \in \mathbf{S}(\mathbf{V})$ such that $Q(\mathbf{x}, \mathbf{y}) = \langle T\mathbf{x}, \mathbf{y}\rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbf{V}$. In any orthonormal basis of $\mathbf{V}$ Q and T represented by the same matrix A. In particular the characteristic polynomial $p(\lambda)$ of T is called the characteristic polynomial of Q. 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 $D = \operatorname{diag}(\lambda_1(Q), ..., \lambda_n(Q))$ 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

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

is either a symmetric bilinear form for $\mathbb{F} = \mathbb{R}$ or a hermitian form for $\mathbb{F} = \mathbb{C}$.

In the rest of the book we use the following normalization unless stated otherwise.

Normalization 4.3.8 *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

1. Prove Proposition 4.3.2.
2. Prove Proposition 4.3.4.
3. Prove Proposition 4.3.5.
4. Prove Proposition 4.3.7.

4.4 Max-min characterizations of eigenvalues

Definition 4.4.1 *Let* $\mathbf{V}$ *be a finite dimensional space over the field* $\mathbb{F}$*. Denote by* $\mathrm{Gr}(m, \mathbf{V})$ *be the space of all* m*-dimensional subspaces in* $\mathbf{U}$ *of dimension* $m \in [0, n] \cap \mathbb{Z}_+$*.*

Theorem 4.4.2 *(The convoy principle) Let* $\mathbf{V}$ *be an* n*-dimensional IPS. Let* $T \in \mathbf{S}(\mathbf{V})$*. Then*

$$\lambda_k(T) = \max_{\mathbf{U} \in \mathrm{Gr}(k,\mathbf{V})} \min_{\mathbf{0} \neq \mathbf{x} \in \mathbf{U}} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \max_{\mathbf{U} \in \mathrm{Gr}(k,\mathbf{V})} \lambda_k(Q(T, \mathbf{U})), \quad k = 1, ..., n, \tag{4.4.1}$$

where the quadratic form $Q(T, \mathbf{U})$ *is defined in Proposition 4.3.7. 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

$$\lambda_1(T) = \max_{\mathbf{0} \neq \mathbf{x} \in \mathbf{V}} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}, \quad \lambda_n(T) = \min_{\mathbf{0} \neq \mathbf{x} \in \mathbf{V}} \frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} \tag{4.4.2}$$

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 (4.4.2) is called *convoy principle.*

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

$$T\mathbf{e}_i = \lambda_i(T)\mathbf{e}_i, \ < \mathbf{e}_i, \mathbf{e}_j >= \delta_{ij} \quad i, j = 1, ..., n. \tag{4.4.3}$$

Then

$$\frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\sum_{i=1}^n \lambda_i(T)|x_i|^2}{\sum_{i=1}^n |x_i|^2}, \quad \mathbf{x} = \sum_{i=1}^n x_i\mathbf{e}_i \neq \mathbf{0}. \tag{4.4.4}$$

The above equality yields straightforward (4.4.2) 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

$$\lambda_k(Q(T,\mathbf{U})) = \min_{\mathbf{0}\neq\mathbf{x}\in\mathbf{U}} \frac{\langle T\mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle} \quad \text{for any } \mathbf{U} \in \mathrm{Gr}(k,\mathbf{U}). \tag{4.4.5}$$

Next there exists $\mathbf{0} \neq \mathbf{x} \in \mathbf{U}$ such that $\langle \mathbf{x}, \mathbf{e}_i\rangle = 0$ for $i = 1, ..., k-1$. (For $k = 1$ this condition is void.) Hence

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

Let

$$\begin{aligned} &\lambda_1(T) = ... = \lambda_{n_1}(T) > \lambda(T)_{n_1+1}(T) = ... = \lambda_{n_2}(T) > ... > \\ (4.4.6)&\lambda_{n_{r-1}+1}(T) = ... = \lambda_{n_r}(T) = \lambda_n(T), \quad n_0 = 0 < n_1 < ... < n_r = n. \end{aligned}$$

Assume that $n_{j-1} < k \leq n_j$. Suppose that $\lambda_k(Q(T,\mathbf{U})) = \lambda_k(T)$. Then for $\mathbf{x} \in \mathbf{U}$ such that $\langle \mathbf{x}, \mathbf{e}_i\rangle = 0$ we have equality $\lambda_k(Q(T,\mathbf{U})) = \lambda_k(T)$ if and only if $\mathbf{x} = \sum_{i=k}^{n_j} x_i\mathbf{e}_i$. Thus $T\mathbf{x} = \lambda_k(T)\mathbf{x}$.

Let $\mathbf{U}_k = \mathrm{span}\,(\mathbf{e}_1, ..., \mathbf{e}_k)$. Let $\mathbf{0} \neq \mathbf{x} = \sum_{i=1}^{k} \in \mathbf{U}_k$. 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 (4.4.3). (See Problem 1.)

Corollary 4.4.3 *Let the assumptions of Theorem 4.4.2 hold. Let* $1 \leq \ell \leq n$. *Then*

$$\lambda_k(T) = \max_{\mathbf{W}\in\mathrm{Gr}(\ell,\mathbf{V})} \lambda_k(Q(T,\mathbf{W})), \quad k = 1, ..., \ell. \tag{4.4.7}$$

Proof. For $k \leq \ell$ apply Theorem 4.4.2 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

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

□

Theorem 4.4.4 *(Courant-Fisher principle) Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $T \in \mathbf{S}(\mathbf{V})$*. Then*

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

See Problem 2 for the proof of the theorem and the following corollary.

Corollary 4.4.5 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $T \in \mathbf{S}(\mathbf{V})$*. Let* $k, \ell \in [1, n]$ *be integers satisfying* $k \leq l$*. Then*

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

Theorem 4.4.6 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS and* $S, T \in \mathbf{S}(\mathbf{V})$*. Then for any* $i, j \in \mathbb{N}, i + j - 1 \leq n$ *the inequality* $\lambda_{i+j-1}(S + T) \leq \lambda_i(S) + \lambda_j(T)$ *holds.*

Proof. Let $\mathbf{U}_{i-1}, \mathbf{V}_{j-1} \subset \mathbf{V}$ be eigenspaces of S, T spanned by the first $i - 1, j - 1$ eigenvectors of S, T respectively. So

$$\langle S\mathbf{x}, \mathbf{x}\rangle \leq \lambda_i(S)\langle \mathbf{x}, \mathbf{x}\rangle, \ \langle T\mathbf{y}, \mathbf{y}\rangle \leq \lambda_j(T)\langle \mathbf{y}, \mathbf{y}\rangle \text{ for all } \mathbf{x} \in \mathbf{U}_{i-1}^\perp, \mathbf{y} \in \mathbf{V}_{j-1}^\perp.$$

Note that $\dim \mathbf{U}_{i-1} = i - 1, \dim \mathbf{V}_{j-1} = j - 1.$. Let $\mathbf{W} = \mathbf{U}_{i-1} + \mathbf{V}_{j-1}$. Then $\dim \mathbf{W} = l - 1 \leq i + j - 2$. Assume that $\mathbf{z} \in \mathbf{W}^\perp$. Then $\langle (S + T)\mathbf{z}, \mathbf{z}\rangle = \langle S\mathbf{z}, \mathbf{z}\rangle + \langle T\mathbf{z}, \mathbf{z}\rangle \leq (\lambda_i(S) + \lambda_j(T))\langle \mathbf{z}, \mathbf{z}\rangle$. Hence $\max_{\mathbf{0} \neq \mathbf{z} \in \mathbf{W}^\perp} \frac{\langle (S+T)\mathbf{z}, \mathbf{z}\rangle}{\langle \mathbf{z}, \mathbf{z}\rangle} \leq \lambda_i(S) + \lambda_j(T)$. Use Theorem 4.4.4 to deduce that $\lambda_{i+j-1}(S + T) \leq \lambda_l(S + T) \leq \lambda_i(S) + \lambda_j(T)$. □

Definition 4.4.7 *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* $< \mathbf{f}_i, \mathbf{f}_j >= \delta_{ij}$ *for* $i, j = 1, ..., k$*. Denote by* $\mathrm{Fr}(k, \mathbf{V})$ *the set of all orthonormal* k*-frames in* $\mathbf{V}$*.*

Note that each $F_k \in \mathrm{Fr}(k, \mathbf{V})$ induces $\mathbf{U} = \mathrm{span\ F_k} \in \mathrm{Gr(k}, \mathbf{V})$. Vice versa, any $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ induces the set $\mathrm{Fr}(k, \mathbf{U})$ of all orthonormal k-frames which span $\mathbf{U}$.

Theorem 4.4.8 *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 $\operatorname{span} \mathrm{F_k}$ *is spanned by $\mathbf{e}_1, ..., \mathbf{e}_k$ satisfying (4.4.3).*

Proof. Define

$$\operatorname{tr} Q(T, \mathbf{U}) := \sum_{i=1}^{k} \lambda_i(Q(T, \mathbf{U})) \quad \text{for } \mathbf{U} \in \mathrm{Gr}(k, \mathbf{V}),$$

(4.4.8)

$$\operatorname{tr}_k T := \sum_{i=1}^{k} \lambda_i(T).$$

Let $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{V})$. Set $\mathbf{U} = \operatorname{span} \mathrm{F_k}$. Then in view of Corollary 4.4.3

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

Let $E_k := \{\mathbf{e}_1, ..., \mathbf{e}_k\}$ where $\mathbf{e}_1, ..., \mathbf{e}_n$ are given by (4.4.3). Clearly $\operatorname{tr}_k T = \operatorname{tr} Q(T, \operatorname{span} \mathrm{E_k})$. This shows the maximal characterization of $\operatorname{tr}_k T$.

Let $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$ 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 $G_k = \{\mathbf{g}_1, ..., \mathbf{g}_k\} \in \mathrm{Fr}(k, \mathbf{U}))$ such that

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

Use Theorem 4.4.2 to deduce that $T\mathbf{g}_i = \lambda_i(T)\mathbf{g}_i$ for $i = 1, ..., k$. □

Theorem 4.4.9 *Let $\mathbf{V}$ be an n-dimensional IPS and $T \in \mathbf{S}(\mathbf{V})$. Then for any integer $k, l \in [1, n]$, such that $k + l \le n$*

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

Proof. Let $\mathbf{W}_j := \operatorname{span}(\mathbf{e}_1, ..., \mathbf{e}_j), j = 1, \ldots, n$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ are given by (4.4.3). Then $\mathbf{V}_1 := \mathbf{V} \cap \mathbf{W}_l$ is an invariant subspace of T. Let

$T_1 := T|\mathbf{V}_1$. Then $\lambda_i(T_1) = \lambda_{l+i}(T)$ for $i = 1, \ldots, n-l$. Theorem 4.4.8 for T_1 yields

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

Let $T_2 := T|\mathbf{W}_{l+k}$ and $\mathbf{W} \in \mathrm{Gr}(l, \mathbf{V})$. Set $\mathbf{U} := \mathbf{W}_{l+k} \cap \mathbf{W}^\perp$. Then $\dim \mathbf{U} \geq k$. Apply Theorem 4.4.8 to $-T_2$ to deduce

$$\sum_{i=1}^{k} \lambda_i(-T_2) \geq \sum_{i=1}^{k} \langle -T\mathbf{f}_i, \mathbf{f}_i\rangle \text{ for } \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}).$$

The above inequality is equal to the inequality

$$\sum_{i=l+1}^{l+k} \lambda_i(T) \leq \sum_{i=1}^{k} \langle T\mathbf{f}_i, \mathbf{f}_i\rangle \text{ for } \{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}) \leq$$

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

The above inequalities yield the theorem. □

Problems

1. 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 $\mathbf{W} = \mathrm{span}\,(\mathbf{e}_1, \mathbf{e}_3)$.

   (a) Show that for each $t \in [\lambda_3(T), \lambda_1(T)]$ there exists a unique $\mathbf{W}(t) \in \mathrm{Gr}(1, \mathbf{W})$ such that $\lambda_1(Q(T, \mathbf{W}(t))) = t$.

   (b) Let $t \in [\lambda_2(T), \lambda_1(T)]$. Let $\mathbf{U}(t) = \mathrm{span}\,(\mathbf{W}(t), \mathbf{e}_2) \in \mathrm{Gr}(2, \mathbf{V})$. Show that $\lambda_2(T) = \lambda_2(Q(T, \mathbf{U}(t))$.

2. (a) Let the assumptions of Theorem 4.4.4 hold. Let $\mathbf{W} \in \mathrm{Gr}(k-1, \mathbf{V})$. Show that there exists $\mathbf{0} \neq \mathbf{x} \in \mathbf{W}^\perp$ such that $\langle \mathbf{x}, \mathbf{e}_i\rangle = 0$ for $k+1, ..., n$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ satisfy (4.4.3). 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 $\mathbf{U}_\ell = \mathrm{span}\,(\mathbf{e}_1, ..., \mathbf{e}_\ell)$. Show that $\lambda_1(Q(T, \mathbf{U}_\ell^\perp)) = \lambda_{\ell+1}(T)$ for $\ell = 1, ..., n-1$.

   (c) Prove Theorem 4.4.4.

(d) Prove Corollary 4.4.5. (**Hint**: Choose $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{W})$ such that $\mathbf{U} \subset \mathbf{W} \cap \mathrm{span}\,(\mathbf{e}_{n-\ell+k+1}, ..., \mathbf{e}_n)^{\perp}$. Then $\lambda_{n-\ell+k}(T) \le \lambda_k(Q(T, \mathbf{U})) \le \lambda_k(Q(T, \mathbf{W}))$.)

3. Let $B = [b_{ij}]_{i,j=1}^n \in \mathbf{H}_n$ and denote by $A \in \mathbf{H}_{n-1}$ the matrix obtained from B by deleting the $j - th$ row and column.

   (a) Show the Cauchy interlacing inequalities

   $$\lambda_i(B) \ge \lambda_i(A) \ge \lambda_{i+1}(B), \text{ for } i = 1, ..., n-1.$$

   (b) Show that inequality $\lambda_1(B) + \lambda_n(B) \le \lambda_1(A) + b_{ii}$.
   Hint. Express the traces of B and A respectively in terms of eigenvalues to obtain

   $$\lambda_1(B) + \lambda_n(B) = b_{ii} + \lambda_1(A) + \sum_{i=2}^{n-1} (\lambda_i(A) - \lambda_i(B)).$$

   Then use the Cauchy interlacing inequalities.

4. Show the following generalization of Problem 3.b ([Big96, p.56]). Let $B \in \mathbf{H}_n$ be the following 2×2 block matrix $B = \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^* & B_{22} \end{bmatrix}$. Show that

   $$\lambda_1(B) + \lambda_n(B) \le \lambda_1(B_{11}) + \lambda_1(B_{22}).$$

   Hint. Assume that $B\mathbf{x} = \lambda_1(B)\mathbf{x}, \mathbf{x}^\top = (\mathbf{x}_1^\top, \mathbf{x}_2^\top)$, partitioned as B. Consider $\mathbf{U} = \mathrm{span}\,((\mathbf{x}_1^\top, \mathbf{0})^\top, (\mathbf{0}, \mathbf{x}_2^\top)^\top)$. Analyze $\lambda_1(Q(T, \mathbf{U})) + \lambda_2(Q(T, \mathbf{U}))$.

5. 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 $k = 1, ..., n$.

6. Let $T \in \mathbf{S}(\mathbf{V})$. Denote by $\iota_+(T), \iota_0(T), \iota_-(T)$ the number of positive, negative and zero eigenvalues among $\lambda_1(T) \ge ... \ge \lambda_n(T)$. The triple $\iota(T) := (\iota_+(T), \iota_0(T), \iota_-(T))$ is called the inertia of T. For $B \in \mathbf{H}_n$ let $\iota(B) := (\iota_+(B), \iota_0(B), \iota_-(B))$ be the inertia of B, where $\iota_+(B), \iota_0(B), \iota_-(B)$ is the number of positive, negative and zero eigenvalues of B respectively. Let $\mathbf{U} \in \mathrm{Gr}(k, \mathbf{V})$. Show

   (a) Assume that $\lambda_k(Q(T, \mathbf{U})) > 0$, i.e. $Q(T, \mathbf{U}) > 0$. Then $k \le \iota_+(T)$. If $k = \iota_+(T)$ then $\mathbf{U}$ is the unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to positive eigenvalues of T.

   (b) Assume that $\lambda_k(Q(T, \mathbf{U})) \ge 0$, i.e. $Q(T, \mathbf{U}) \ge 0$. Then $k \le \iota_+(T) + \iota_0(T)$. If $k = \iota_+(T) + \iota_0(T)$ then $\mathbf{U}$ is the unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to nonnegative eigenvalues of T.

(c) Assume that $\lambda_1(Q(T,\mathbf{U})) < 0$, i.e. $Q(T,\mathbf{U}) < 0$. Then $k \le \iota_-(T)$. If $k = \iota_-(T)$ then $\mathbf{U}$ is a unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to negative eigenvalues of T.

(d) Assume that $\lambda_1(Q(T,\mathbf{U})) \le 0$, i.e. $Q(T,\mathbf{U}) \le 0$. Then $k \le \iota_-(T) + \iota_0(T)$. If $k = \iota_-(T) + \iota_0(T)$ then $\mathbf{U}$ is a unique invariant subspace of $\mathbf{V}$ spanned by the eigenvectors of T corresponding to nonpositive eigenvalues of T.

7. Let $B \in \mathbf{H}_n$ and assume that $A = PBP^*$ for some $P \in \mathbf{GL}(n,\mathbb{C})$. Then $\iota(A) = \iota(B)$.

4.5 Positive definite operators and matrices

Definition 4.5.1 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F} = \mathbb{C},\mathbb{R}$. Let $S,T \in \mathbf{S}(\mathbf{V})$. Then $T > S$, $(T \ge S)$ if $\langle T\mathbf{x},\mathbf{x}\rangle > \langle S\mathbf{x},\mathbf{x}\rangle$, $(\langle T\mathbf{x},\mathbf{x}\rangle \ge \langle S\mathbf{x},\mathbf{x}\rangle)$ for all $\mathbf{0} \ne \mathbf{x} \in \mathbf{V}$. T is called positive (nonnegative) definite if $T > 0$ $(T \ge 0)$, where 0 is the zero operator in* Hom $(\mathbf{V})$.

Denote by $\mathbf{S}_+(\mathbf{V})^o \subset \mathbf{S}_+(\mathbf{V}) \subset \mathbf{S}(\mathbf{V})$ the open set of positive definite self adjoint operators and the closed set of nonnegative selfadjoint operators respectively.

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

For $A,B \in \mathbf{H}_n$ $B > A$ $(B \ge A)$ if $\mathbf{x}^*B\mathbf{x} > \mathbf{x}^*A\mathbf{x}$ $(\mathbf{x}^*B\mathbf{x} \ge \mathbf{x}^*A\mathbf{x})$ for all $\mathbf{0} \ne \mathbf{x} \in \mathbb{C}^n$. $B \in \mathbf{H}_n$ is called is called positive (nonnegative) definite if $B > 0$ $(B \ge 0)$. Denote by $\mathbf{H}_{n,+}^o \subset \mathbf{H}_{n,+} \subset \mathbf{H}_n$ the open set of positive definite $n \times n$ hermitian matrices and the closed set of $n \times n$ nonnegative hermitian matrices respectively. Let $\mathbf{S}_+(n,\mathbb{R}) := \mathbf{S}(n,\mathbb{R}) \cap \mathbf{H}_{n,+}$, $\mathbf{S}_+(n,\mathbb{R})^o := \mathbf{S}(n,\mathbb{R}) \cap \mathbf{H}_{n,+}^o$.

Use (4.4.1) to deduce.

Corollary 4.5.2 *Let $\mathbf{V}$ be n-dimensional IPS. Let $T \in \mathbf{S}(\mathbf{V})$. Then $T > 0$ $(T \ge 0)$ if and only if $\lambda_n(T) > 0$ $(\lambda_n(T) \ge 0)$. Let $S \in \mathbf{S}(\mathbf{V})$ and assume that $T > S$ $(T \ge S)$. Then $\lambda_i(T) > \lambda_i(S)$ $(\lambda_i(T) \ge \lambda_i(S))$ for $i = 1,...,n$.*

Proposition 4.5.3 *Let $\mathbf{V}$ be a finite dimensional IPS. Assume that $T \in \mathbf{S}(\mathbf{V})$. Then $T \ge 0$ if and only if there exists $S \in \mathbf{S}(\mathbf{V})$ such that $T =$*

S^2. Furthermore $T > 0$ if and only if S is invertible. For $0 \le T \in \mathbf{S}(\mathbf{V})$ there exists a unique $0 \le S \in \mathbf{S}(\mathbf{V})$ 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 $T \ge 0$. Let $\mathbf{e}_1, ..., \mathbf{e}_n$ be an orthonormal basis consisting of eigenvectors of T as in (4.4.3). Since $\lambda_i(T) \ge 0,\ i = 1, ..., n$ we can define $P \in \mathrm{Hom}\ (\mathbf{V})$ 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 $\langle T\mathbf{x}, \mathbf{x}\rangle = \langle S\mathbf{x}, S\mathbf{x}\rangle \ge 0$. Hence $T \ge 0$. Clearly $\langle T\mathbf{x}, \mathbf{x}\rangle = 0 \iff S\mathbf{x} = 0$. Hence $T > 0 \iff S \in \mathbf{GL}(\mathbf{V})$. Suppose that $S \ge 0$. 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. Clearly $T > 0$ if and only if $\sqrt{\lambda_n(T)} > 0$, i.e. if and only if S is invertible. □

Corollary 4.5.4 *Let $B \in \mathbf{H}_n$ $(\mathbf{S}(n, \mathbb{R}))$. Then $B \ge 0$ if and only there exists $A \in \mathbf{H}_n$ $(\mathbf{S}(n, \mathbb{R}))$ such that $B = A^2$. Furthermore $B > 0$ if and only if A is invertible. For $B \ge 0$ there exists a unique $A \ge 0$ such that $B = A^2$. This A is denoted by $B^{\frac{1}{2}}$.*

Theorem 4.5.5 *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.) $G(\mathbf{x}_1, ..., \mathbf{x}_n) > 0$ if and only $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. Furthemore for any integer $k \in [1, n-1]$*

$$\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \le \det G(\mathbf{x}_1, ..., \mathbf{x}_k) \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n). \tag{4.5.1}$$

Equality holds if and only if either $\det G(\mathbf{x}_1, ..., \mathbf{x}_k) \det G(\mathbf{x}_{k+1}, ..., \mathbf{x}_n) = 0$ or $\langle \mathbf{x}_i, \mathbf{x}_j\rangle = 0$ for $i = 1, ..., k$ and $j = k+1, ..., n$.

Proof. Clearly $G(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbf{H}_n$. If $\mathbf{V}$ is an IPS over $\mathbb{R}$ then $G(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbf{S}(n, \mathbb{R})$. Let $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{F}^n$. Then

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

Equality holds if and only if $\sum_{i=1}^n a_i \mathbf{x}_i = 0$. Hence $G(\mathbf{x}_1, ..., \mathbf{x}_n) \ge 0$ and $G(\mathbf{x}_1, ..., \mathbf{x}_n) > 0$ if and only if $\mathbf{x}_1, ..., \mathbf{x}_n$ are linearly independent. In particular $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) \geq 0$ 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 (4.5.1). Assume first that the right-hand side of (4.5.1) 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 (4.5.1) 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 (4.5.1). It is left to show the inequality (4.5.1) 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 (4.1.1). Let $S_j = \text{span}\,(\mathbf{x}_1, ..., \mathbf{x}_j)$ for $j = 1, ..., n$. Corollary 4.1.1 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 $b_1, ..., b_{n-1} \in \mathbb{F}$. Let G' be the matrix obtained from $G(\mathbf{x}_1, ..., \mathbf{x}_n)$ by subtracting from the n-th row b_j times j-th row. Thus the last row of G' is $(\langle \mathbf{y}_n, \mathbf{x}_1\rangle, ..., \langle \mathbf{y}_n, \mathbf{x}_n\rangle) = (0, ..., 0, \|\mathbf{y}_n\|^2)$. Clearly $\det G(\mathbf{x}_1, ..., \mathbf{x}_n) = \det G'$. Expand $\det G'$ by the last row to deduce

$$\begin{aligned} &\det G(\mathbf{x}_1, ..., \mathbf{x}_n) = \det G(\mathbf{x}_i, ..., \mathbf{x}_{n-1})\, \|\mathbf{y}_n\|^2 = ... = \\ (4.5.2)\qquad &\det G(\mathbf{x}_1, ..., \mathbf{x}_k) \prod_{i=k+1}^{n} \|\mathbf{y}_i\|^2 = \\ &\det G(\mathbf{x}_1, ..., \mathbf{x}_k) \prod_{i=k+1}^{n} \text{dist}(\mathbf{x}_i, S_{i-1})^2, \quad k = n-1, ..., 1. \end{aligned}$$

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

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

for $j = k+1, ..., n$, where $\hat{S}_k = \{\mathbf{0}\}$. Use (4.5.2) 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 (4.5.1). Assume now equality holds in (4.5.1). 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 $\text{dist}(\mathbf{x}_j, S_{j-1}) = \text{dist}(\hat{\mathbf{y}}_{\mathrm{j}}, \mathrm{S}_{\mathrm{j}-1}) = \|\mathbf{y}_j\|$. Hence $\|\hat{\mathbf{y}}_j\| = \text{dist}(\hat{\mathbf{y}}_{\mathrm{j}}, \mathrm{S}_{\mathrm{j}-1})$. Part (h) of Problem 4.1.4 yields that $\hat{\mathbf{y}}_j$ is orthogonal on S_{j-1}. In particular each $\hat{\mathbf{y}}_j$ is orthogonal to S_k for $j = k+1, ..., n$. Hence $\mathbf{x}_j \perp S_k$ for $j = k+1, ..., n$, i.e. $\langle \mathbf{x}_j, \mathbf{x}_i\rangle = 0$ for $j > k$ and $i \leq k$. 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 $\mathbf{0}$. That is Π is a convex hull spanned by the vectors $\mathbf{0}$ and $\sum_{i \in S} \mathbf{x}_i$ for all nonempty subsets $S \subset \{1, ..., n\}$. Then $\sqrt{\det G(\mathbf{x}_1, ..., \mathbf{x}_n)}$ is the n-volume of Π. The inequality (4.5.1) and equalities (4.5.2) are "obvious" from this geometrical point of view.

Corollary 4.5.6 *Let* $0 \le B = (b_{ij})_1^n \in \mathbf{H}_{n,+}$. *Then*

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

For a fixed k equality holds if and only if either the right-hand side of the above inequality is zero or $b_{ij} = 0$ for $i = 1, ..., k$ and $j = k+1, ..., n$.

Proof. From Corollary 4.5.4 it follows that $B = X^2$ for some $X \in \mathbf{H}_n$. Let $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{C}^n$ be the n-columns of $X^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Let $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^* \mathbf{x}$. Since $X \in \mathbf{H}_n$ we deduce that $B = G(\mathbf{x}_1, ..., \mathbf{x}_n)$. □

Theorem 4.5.7 *Let* $\mathbf{V}$ *be an n-dimensional IPS. Let* $T \in \mathbf{S}$. *TFAE:*
(a) $T > 0$.
(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) $\Rightarrow$ (b). According to Proposition 4.5.3 $T = S^2$ for some $S \in \mathbf{S}(\mathbf{V}) \cap \mathbf{GL}(\mathbf{V})$. 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 4.5.1 implies that $\det G(S\mathbf{g}_1, ..., S\mathbf{g}_k) > 0$ for $k = 1, ..., n$.
(b) $\Rightarrow$ (a). The proof is by induction on n. For $n = 1$ (a) is obvious. Assume that (a) holds for $n = m-1$. Let $\mathbf{U} :=$ span $(\mathbf{g}_1, ..., \mathbf{g}_{n-1})$ and $Q := Q(T, \mathbf{U})$. Then there exists $P \in \mathbf{S}(\mathbf{U})$ such that $< P\mathbf{x}, \mathbf{y} >= Q(\mathbf{x}, \mathbf{y}) =< T\mathbf{x}, \mathbf{y} >$ for any $\mathbf{x}, \mathbf{y} \in \mathbf{U}$. By induction $P > 0$. Corollary 4.4.3 yields that $\lambda_{n-1}(T) \ge \lambda_{n-1}(P) > 0$. Hence T has at least $n-1$ positive eigenvalues. Let $\mathbf{e}_1, ..., \mathbf{e}_n$ be given by (4.4.3). Then $\det\ (\langle T\mathbf{e}_i, \mathbf{e}_j \rangle)_{i,j=1}^n = \prod_{i=1}^n \lambda_i(T) > 0$. Let $A = (a_{pq})_1^n \in \mathbf{GL}(n, \mathbb{C})$ 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

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

(4.5.3)

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

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

Corollary 4.5.8 *Let $B = (b_{ij})_1^n \in \mathbf{H}_n$. Then $B > 0$ if and only if $\det\,(b_{ij})_1^k > 0$ for $k = 1, ..., n$.*

The following result is straightforward (see Problem 1):

Proposition 4.5.9 *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 $T > 0$ if and only if $(\mathbf{x},\mathbf{y}) := \langle T\mathbf{x},\mathbf{y}\rangle$ is an inner product on $\mathbf{V}$. Vice versa any inner product $(\cdot,\cdot) : \mathbf{V}\times\mathbf{V} \to \mathbb{R}$ is of the form $(\mathbf{x},\mathbf{y}) =< T\mathbf{x},\mathbf{y} >$ for a unique self-adjoint positive definite operator $T \in \mathrm{Hom}\,(\mathbf{V})$.*

Example 4.5.10 *Each $0 < B \in \mathbf{H}_n$ induces and inner product on $\mathbb{C}^n$: $(\mathbf{x},\mathbf{y}) = \mathbf{y}^*B\mathbf{x}$. Each $0 < B \in \mathbf{S}(n,\mathbb{R})$ induces and inner product on $\mathbb{R}^n$: $(\mathbf{x},\mathbf{y}) = \mathbf{y}^TB\mathbf{x}$. Furthermore any inner product on $\mathbb{C}^n$ or $\mathbb{R}^n$ is of the above form. In particular, the standard inner products on $\mathbb{C}^n$ and $\mathbb{R}^n$ are induces by the identity matrix I.*

Definition 4.5.11 *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 $(\cdot,\cdot)$.*

Problems

1. Show Proposition 4.5.9.
2. Recall the Hölder inequality

(4.5.4)
$$\sum_{l=1}^n x_l y_l a_l \le (\sum_{l=1}^n x_l^p a_l)^{\frac{1}{p}} (\sum_{l=1}^n y_l^q a_l)^{\frac{1}{q}}$$

for any $\mathbf{x} = (x_1,\ldots,x_n)^\top, \mathbf{y} = (y_1,\ldots,y_n)^\top, \mathbf{a} = (a_1,\ldots,a_n) \in \mathbb{R}_+^n$ and $p, q \in (1,\infty)$ such that $\frac{1}{p}+\frac{1}{q} = 1$. Show

(a) Let $A \in \mathbf{H}_{n,+}, \mathbf{x} \in \mathbb{C}^n$ and $0 \le i < j < k$ be three integers. Then

$$\mathbf{x}^* A^j \mathbf{x} \le (\mathbf{x}^* A^i \mathbf{x})^{\frac{k-j}{k-i}} (\mathbf{x}^* A^k \mathbf{x})^{\frac{j-i}{k-i}}. \tag{4.5.5}$$

Hint: Diagonalize A.

(b) Assume that $A = e^B$ for some $B \in \mathbf{H}_n$. Show that (4.5.5) holds for any three real numbers $i < j < k$.

4.6 Majorization and applications

Definition 4.6.1 *Let*

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

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 $\mathbf{x} = (x_1, ..., x_n)^T, \mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$. *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*

$$\sum_{i=1}^{k} \bar{x}_i \le \sum_{i=1}^{k} \bar{y}_i, \quad k = 1, ..., n. \tag{4.6.1}$$

$\mathbf{x}$ *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* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

Definition 4.6.2 $A \in \mathbb{R}_+^{n \times n}$ *is called doubly stochastic matrix if the sum of each row and column of* A *is equal to* 1. *Denote by* $\Omega_n \subset \mathbb{R}_+^{n\times n}$ *the set of doubly stochastic matrices. Denote by* $\frac{1}{n} J_n$ *the* $n \times n$ *doubly stochastic matrix whose all entries are equal to* $\frac{1}{n}$, *i.e.* $J_n \in \mathbb{R}_+^{n\times n}$ *is the matrix whose each entry is* 1.

Definition 4.6.3 $P \in \mathbb{R}_+^{n \times n}$ *is called a permutation matrix if each row and column of* P *a contains exactly one nonzero element which is equal to* 1. *Denote by* $\mathcal{P}_n$ *the set of* $n \times n$ *permutation matrices.*

Lemma 4.6.4 *The following properties hold.*

1. $A \in \mathbb{R}_+^{n\times n}$ *is double stochastic if and only if* $A\mathbf{1} = A^\top \mathbf{1} = \mathbf{1}$, *where* $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^n$.

2. $\Omega_1 = \{1\}$.

3. *$A, B \in \Omega_n \Rightarrow tA + (1-t)B \in \Omega_n$ for each $t \in [0,1]$.*

4. *$A, B \in \Omega_n \Rightarrow AB \in \Omega_n$.*

5. *$\mathcal{P}_n \subset \Omega_n$.*

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

7. *$A \in \Omega_l,\ B \in \Omega_m \to A \oplus B \in \Omega_{l+m}$.*

See Problem 1.

Theorem 4.6.5 *$A \in \mathbb{R}_+^{n\times n}$ is doubly stochastic if and only*

$$A = \sum_{P\in\mathcal{P}_n} a_P P \text{ for some } a_P \geq 0,\ P \in \mathcal{P}_n,\ \sum_{P\in\mathcal{P}_n} a_P = 1. \tag{4.6.2}$$

Proof. In view of properties 3 and 5 of Lemma 4.6.4 it follows that any A of the form (4.6.2) is doubly stochastic. We now show by induction on n that any $A \in \Omega_n$ is of the form (4.6.2). For $n = 1$ the result trivially holds. Assume that the result holds for $n = m-1$ and assume that $n = m$.

Assume that $A = (a_{ij}) \in \Omega_n$. Let $l(A)$ be the number of nonzero entries of A. Since each row sum of A is 1 it follows that $l(A) \geq n$. Suppose first $l(A) \leq 2n-1$. Then there exists a row i of A which has exactly one nonzero element, which must be 1. Hence there exists $i, j \in \langle n\rangle$ such that $a_{ij} = 1$. Then all other elements of A on the row i and column j are zero. Denote by $A_{ij} \in \mathbb{R}_+^{(n-1)\times(n-1)}$ the matrix obtained from A by deleting the row and column j. Clearly $A_{ij} \in \Omega_{n-1}$. Use the induction hypothesis on A_{ij} to deduce (4.6.2), where $a_P = 0$ if the entry (i,j) of P is not 1.

We now show by induction on $l(A) \geq 2n-1$ that A is of the form (4.6.2). Suppose that any $A \in \Omega_n$ such that $l(A) \leq l-1, l \geq 2n$ is of the form (4.6.2). Assume that $l(A) = l$. Let $\mathcal{S} \subset \langle n\rangle \times \langle n\rangle$ be the set of all indices $(i,j) \in \langle n\rangle \times \langle n\rangle$ where $a_{ij} > 0$. Note $\#\mathcal{S} = l(A) \geq 2n$. Consider the following system of equations in n^2 variables, which are the entries $X = (x_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$:

$$\sum_{j=1}^n x_{ij} = \sum_{j=1}^n x_{ji} = 0, \quad i = 1, \ldots, n.$$

Since the sum of all rows of X is equal to the sum of all columns of X we deduce that the above system has at most $2n-1$ linear independent equations. Assume furthermore the conditions $x_{ij} = 0$ for $(i,j) \notin \mathcal{S}$. Since we have at least $2n$ variables it follows that there exist $X \neq 0_{n\times n}$ satisfying the above conditions. Note that X has zero entry in the places where A has zero entry. Furthermore, X has at least one positive and one negative entry. Therefore the exists $b, c > 0$ such that $A - bX, A + cX \in \Omega_n$ and $l(A - bX), l(A + cX) < l$. So $A - bX, A + cX$ are of the form (4.6.2). As $A = \frac{c}{b+c}(A - bX) + \frac{b}{b+c}(A + cX)$ we deduce that A is of the form (4.6.2). □

Theorem 4.6.6 . *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. *Then* $\mathbf{x} \prec \mathbf{y}$ *if and only if there exists* $A \in \Omega_n$ *such that* $\mathbf{x} = A\mathbf{y}$.

Proof. Assume first that $\mathbf{x} = P\mathbf{y}$ for some $P \in \mathcal{P}_n$. Then it is straightforward to see that $\mathbf{x} \prec \mathbf{y}$. Assume that $\mathbf{x} = A\mathbf{y}$ for some $A \in \Omega_n$. Use Theorem 4.6.5 to deduce that $\mathbf{x} \prec \mathbf{y}$.

Assume now that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x} \prec \mathbf{y}$. Since $\mathbf{x} = P\bar{\mathbf{x}}, \mathbf{y} = Q\bar{\mathbf{y}}$ for some $P, Q \in \mathcal{P}_n$, it follows that $\bar{\mathbf{x}} \prec \bar{\mathbf{y}}$. In view of Lemma 4.6.4 it is enough to show that $\bar{\mathbf{x}} = B\bar{\mathbf{y}}$ some $B \in \Omega_n$. We prove this claim by induction on n. For $n = 1$ this claim is trivial. Assume that if $\bar{\mathbf{x}} \prec \bar{\mathbf{y}} \in \mathbb{R}^l$ then $\bar{\mathbf{x}} = B\bar{\mathbf{y}}$ for some $B \in \Omega_l$ for all $l \le m - 1$. Assume that $n = m$ and $\bar{\mathbf{x}} \prec \bar{\mathbf{y}}$. Suppose first that for some $1 \le k \le n - 1$ we have the equality $\sum_{i=1}^k \bar{x}_i = \sum_{i=1}^k \bar{y}_i$. Let

$$\mathbf{x}_1 = (\bar{x}_1, \ldots, \bar{x}_k)^\top, \mathbf{y}_1 = (\bar{y}_1, \ldots, \bar{y}_k)^\top \in \mathbb{R}^k,$$
$$\mathbf{x}_2 = (\bar{x}_{k+1}, \ldots, \bar{x}_n)^\top, \mathbf{y}_2 = (\bar{y}_{k+1}, \ldots, \bar{y}_n)^\top \in \mathbb{R}^{n-k}.$$

Then $\mathbf{x}_1 \prec \mathbf{y}_1, \mathbf{x}_2 \prec \mathbf{y}_2$. Use the induction hypothesis that $\mathbf{x}_i = B_i\mathbf{y}_i, i = 1, 2$ where $B_1 \in \Omega_k, B_2 \in \Omega_{n-k}$. Hence $\bar{\mathbf{x}} = (B_1 \oplus B_2)\mathbf{y}$ and $B_1 \oplus B_2 \in \Omega_n$.

It is left to consider the case where strict inequalities hold in (4.6.1) for $k = 1, \ldots, n - 1$. We now define a finite number of vectors

$$\bar{\mathbf{y}} = \mathbf{z}_1 \succ \mathbf{z}_2 = \bar{\mathbf{z}}_2 \succ \ldots \succ \mathbf{z}_N = \bar{\mathbf{z}}_N \succ \mathbf{x},$$

where $N \ge 2$, such that

1. $\mathbf{z}_{i+1} = B_i\mathbf{z}_i$ for $i = 1, \ldots, N - 1$.
2. $\sum_{i=1}^k \bar{x}_i = \sum_{i=1}^k w_i$ for some $k \in \langle n-1 \rangle$, where $\mathbf{z}_N = \mathbf{w} = (w_1, \ldots, w_n)^\top$.

Observe first that we can not have $\bar{y}_1 = \ldots = \bar{y}_n$. Otherwise $\bar{\mathbf{x}} = \bar{\mathbf{y}}$ and we have equalities in (4.6.1) for all $k \in \langle n \rangle$, which contradicts out assumptions. Assume that we defined

$$\bar{\mathbf{y}} = \mathbf{z}_1 \succ \mathbf{z}_2 = \bar{\mathbf{z}}_2 \succ \ldots \succ \mathbf{z}_r = \bar{\mathbf{z}}_r = (u_1, \ldots, u_n)^\top \succ \mathbf{x},$$

for $1 \le r$ such that $\sum_{i=1}^k \bar{x}_i < \sum_{i=1}^k u_i$ for $k = 1, \ldots, n - 1$. Assume that $u_1 = \ldots = u_p > u_{p+1} = \ldots = u_{p+q}$, where $u_{p+q} > u_{p+q+1}$ if $p + q < n$.

Let $C(t) = ((1-t)I_{p+q} + \frac{t}{p+q}J_{p+q}) \oplus I_{n-(p+q)})$ for $t \in [0,1]$ and define $\mathbf{u}(t) = C(t)\mathbf{z}_r$. We vary t continuously from $t=0$ to $t=1$. Note that $\mathbf{u}(t) = \bar{\mathbf{u}}(t)$ for all $t \in [0,1]$. We have two possibilities. First there exists exits $t_0 \in (0,1]$ such that $\mathbf{u}(t) \succ \mathbf{x}$ for all $t \in [0,t_0]$. Furthermore for $\mathbf{w} = \mathbf{u}(t_0) = (w_1,\ldots,w_n)^\top$ we have the equality $\sum_{i=1}^k \bar{x}_i = \sum_{i=1}^k w_i$ for some $k \in \langle n-1\rangle$. In that case $r = N-1$ and $\mathbf{z}_N = \mathbf{u}(t_0)$.

Otherwise let $\mathbf{z}_{r+1} = \mathbf{u}(1) = (v_1,\ldots,v_n)^\top$, where $v_1 = \ldots = v_{p+q} > v_{p+q+1}$. Repeat this process for $\mathbf{z}_{r+1}$ and so on until we deduce the conditions *1* and *2*. So $\mathbf{x} = B_N\mathbf{z}_N = B_NB_{N-1}\mathbf{z}_{N-1} = B_N\ldots B_1\mathbf{y}$. In view of 4 of Lemma 4.6.4 we deduce that $\mathbf{x} = A\mathbf{y}$ for some $A \in \Omega_n$.

□

Combine Theorems 4.6.6 and 4.6.5 to deduce.

Corollary 4.6.7 *Let* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. *Then* $\mathbf{x} \prec \mathbf{y}$ *if and only if*

$$\mathbf{x} = \sum_{P\in\mathcal{P}_n} a_P P\mathbf{y} \textit{ for some } a_P \ge 0, P \in \mathcal{P}_n \textit{ where } \sum_{P\in\mathcal{P}_n} a_P = 1. \tag{4.6.3}$$

Furthermore, if $\mathbf{x} \prec \mathbf{y}$ *and* $\mathbf{x} \ne P\mathbf{y}$ *for all* $P \in \mathcal{P}_n$ *then in (4.6.3) each* $a_P < 1$.

Definition 4.6.8 *Let* $I \subseteq \mathbb{R}$ *be an interval. A function* $\phi : I \to \mathbb{R}$ *is called convex if for any* $x,y \in I$ *and* $t \in [0,1]$ $\phi(tx+(1-t)y) \le t\phi(x) + (1-t)\phi(y)$. ϕ *is called strictly convex on* I *if for any* $x,y \in I, x \ne y$ *and* $t \in (0,1)$ $\phi(tx+(1-t)y) < t\phi(x)+(1-t)\phi(y)$.

A function $\psi : I \to \mathbb{R}$ *is called concave or strictly concave if the function* $-\psi$ *is convex or strictly convex respectively.*

Theorem 4.6.9 *Let* $\mathbf{x} = (x_1,...,x_n)^\top \prec \mathbf{y} = (y_1,...,y_n)^\top$. *Let* $\phi : [\bar{y}_n, \bar{y}_1] \to \mathbb{R}$ *be a convex function. Then*

$$\sum_{i=1}^n \phi(x_i) \le \sum_{i=1}^n \phi(y_i). \tag{4.6.4}$$

If ϕ *is strictly convex on* $[\bar{y}_n, \bar{y}_1]$ *and* $P\mathbf{x} \ne \mathbf{y}$ *for all* $P \in \mathcal{P}_n$ *then strict inequality holds in (4.6.4).*

Proof. Problem 3 implies that if $\mathbf{x} = (x_1,\ldots,x_n)^\top \prec \mathbf{y} = (y_1,\ldots,y_n)$ then $x_i \in [\bar{y}_n, \bar{y}_1]$ for $i = 1,\ldots,n$. Use Corollary 4.6.7 and the convexity of ϕ, see Problem 2, to deduce:

$$\phi(x_i) \le \sum_{P\in\mathcal{P}_n} a_P \phi((P\mathbf{y})_i), \quad i = 1,\ldots,n.$$

Observe next that $\sum_{i=1}^n \phi(y_i) = \sum_{i=1}^n \phi((P\mathbf{y})_i)$ for all $P \in \mathcal{P}_n$. Sum up the above inequalities to deduce (4.6.4).

Assume now that ϕ is strictly convex and $\mathbf{x} \neq P\mathbf{y}$ for all $P \in \mathcal{P}_n$. Then Corollary 4.6.7 and strict convexity of ϕ implies that at least in one the above $i - th$ inequality one has strict inequality. Hence strict inequality holds in (4.6.4). □

Corollary 4.6.10 *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 convex function. Then*

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

If ϕ *is strictly convex then* $\sum_{i=1}^n \phi(\lambda_i(T)) = \sum_{i=1}^n \phi(\langle T\mathbf{f}_i, \mathbf{f}_i\rangle)$ *if and only if* $\mathbf{f_1}, \ldots, \mathbf{f_n}$ *is a set of* n *orthonormal eigenvectors of* T*.*

See Problem 4.

Problems

1. Prove Lemma 4.6.4.

2. Let $I \subseteq \mathbb{R}$ be an interval and assume that $\phi : I \to \mathbb{R}$ is convex. Let $x_1, \ldots, x_m \in I, m \geq 3$. Show

   (a) Let $a_1, \ldots, a_m \in [0, 1]$ and assume that $\sum_{i=1}^m a_i = 1$. Then

   $$\phi(\sum_{i=1}^m a_i x_i) \leq \sum_{i=1}^m a_i \phi(x_i). \tag{4.6.5}$$

   (b) Assume in addition that ϕ is strictly convex, $x_i \neq x_j$ for $i \neq j$ and $a_1, \ldots, a_m > 0$. Then strict inequality holds in (4.6.5).

3. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Show that $\mathbf{x} \prec \mathbf{y} \iff -\mathbf{y} \prec -\mathbf{x}$.

4. Prove Corollary 4.6.10.

4.7 Spectral functions

Let $\mathcal{T} \subset \mathbf{S}(\mathbf{V})$, where $\mathbf{V}$ is an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Denote $\lambda(\mathcal{T}) := \{\lambda(T) \in \mathbb{R}^n_{\searrow} : \quad T \in \mathcal{T}\}$. A function $f : \mathcal{T} \to \mathbb{R}$ is called a spectral function if there exists a set $D \subset \mathbb{R}^n_{\searrow}$ and $h : D \to \mathbb{R}$ such that $\lambda(\mathcal{T}) \subset D$ and $f(T) = h(\lambda(T))$ for each $T \in \mathcal{T}$. D is called a Schur set if

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{\searrow},\ \mathbf{x} \prec \mathbf{y},\ \mathbf{y} \in D \Rightarrow \mathbf{x} \in D.$$

Let $D \subset \mathbb{R}^n_{\searrow}$ be a Schur set. A function $h : D \to \mathbb{R}$ is called Schur's order preserving if

$$h(\mathbf{x}) \le h(\mathbf{y}) \quad \text{for any } \mathbf{x}, \mathbf{y} \in D \text{ such } \mathbf{x} \prec \mathbf{y}.$$

h is called strict Schur's order preserving if a strict inequality holds in the above inequality whenever $\mathbf{x} \ne \mathbf{y}$. h is called strong Schur's order preserving if

$$h(\mathbf{x}) \le h(\mathbf{y}) \quad \text{for any } \mathbf{x}, \mathbf{y} \in D \text{ such } \mathbf{x} \preceq \mathbf{y}.$$

h is called strict strong Schur's order preserving if a strict inequality holds in the above inequality whenever $\mathbf{x} \ne \mathbf{y}$.

Note that $h((x_1, ..., x_n)) := \sum_{i=1}^n g(x_i)$ for some convex function $g : \mathbb{R} \to \mathbb{R}$ then Corollary 4.6.10 implies that $h : \mathbb{R}^n_{\searrow} \to \mathbb{R}$ is Schur's order preserving. The results of Section 4.4 yield:

Proposition 4.7.1 *Let $\mathbf{V}$ be an n-dimensional IPS, $D \subset \mathbb{R}^n_{\searrow}$ be a Schur set and $h : D \to R$ be Schur's order preserving function. Let $T \in \mathbf{S}(\mathbf{V})$ and assume that $\lambda(T) \in D$. Then*

$$h(\lambda(T)) = \max_{\mathbf{x} \in D, \mathbf{x} \prec \lambda(T)} h(\mathbf{x}).$$

Definition 4.7.2 *A set $D \subset \mathbb{R}^n$ is called a regular set if the interior of D, denoted by $D^o \subset \mathbb{R}^n$, is a nonempty set, and D is a subset of the closure D^o, denoted by $\mathrm{Cl}(D^o)$. For a regular set D a function $F : D \to \mathbb{R}$ is in the class $\mathrm{C^k(D)}$, i.e. F has k continuous derivatives, if $F \in \mathrm{C^k(D^o)}$ and F and any of its derivative of order not greater than k has a continuous extension to D.*

Definition 4.7.3 *Let $\mathbf{V}$ be a vector space over $\mathbb{R}$. For $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ denote*

$$[\mathbf{x}, \mathbf{y}] := \{\mathbf{z} : \quad \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \text{ for all } \alpha \in [0, 1]\}.$$

A set $C \subset \mathbf{V}$ is called convex if for each $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ $[\mathbf{x}, \mathbf{y}] \subset C$. Assume that $C \subset \mathbf{V}$ is a nonempty convex set and let $\mathbf{x} \in C$. Denote by $C - \mathbf{x}$ the set

$\{\mathbf{z} : \ \mathbf{z} = \mathbf{y} - \mathbf{x}, \ \mathbf{y} \in C\}$. *Let* $\mathbf{U} = \text{span}\,(\mathrm{C} - \mathbf{x})$. *Then the dimension* C*, denoted by* $\dim C$*, is the dimension of the vector space* $\mathbf{U}$*.* $C - \mathbf{x}$ *has interior as a subset of* $\mathbf{U}$ *called* relative interior *and denoted by* $\text{ri}\,(C - \mathbf{x})$*. Then the relative interior of* C *is defined as* $\text{ri}\,(C - \mathbf{x}) + \mathbf{x}$*.*

It is straightforward to show that $\dim C$, $\text{ri}\, C$ do not depend on the choice of $\mathbf{x} \in C$. Furthermore $\text{ri}\, C$ is convex. See Problem 7 or [Roc70].

Proposition 4.7.4 *Let* $\mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n_\searrow$*. Denote*

$$M(\mathbf{y}) := \{\mathbf{x} \in \mathbb{R}^n_\searrow : \quad \mathbf{x} \prec \mathbf{y}\}.$$

Then $M(\mathbf{y})$ *is a closed convex set.*

See Problem 1.

Theorem 4.7.5 *Let* $D \subset \mathbb{R}^n_\searrow$ *be a regular Schur set in* $\mathbb{R}^n$*. Let* $F \in \mathrm{C}^1(\mathrm{D})$*. Then* F *is Schur's order preserving if and only if*

$$\frac{\partial F}{\partial x_1}(\mathbf{x}) \ge ... \ge \frac{\partial F}{\partial x_n}(\mathbf{x}), \text{ for each } \mathbf{x} = (x_1, ..., x_n)^\top \in D. \tag{4.7.1}$$

If for any point $\mathbf{x} = (x_1, ..., x_n)^\top \in D$ *such that* $x_i > x_{i+1}$ *the inequality* $\frac{\partial F}{\partial x_i}(\mathbf{x}) > \frac{\partial F}{\partial x_{i+1}}(\mathbf{x})$ *holds then* F *is strict Schur's order preserving.*

Proof. Assume that $F \in \mathrm{C}^1(\mathrm{D})$ and F is Schur's order preserving. Let $\mathbf{x} = (x_1, ..., x_n)^\top \in D^o$. Hence $x_1 > ... > x_n$. Let $\mathbf{e}_i = (\delta_{i1}, ..., \delta_{in})^\top, i = 1, ..., n$. For $i \in [1, n-1] \cap \mathbb{Z}_+$ let $\mathbf{x}(t) := \mathbf{x} + t(\mathbf{e}_i - \mathbf{e}_{i+1})$. Then

$$\begin{aligned} &\mathbf{x}(t) \in R^n_\searrow \text{ for } |t| \le \tau := \frac{\min_{j \in [1, n-1] \cap \mathbb{Z}_+} x_j - x_{j+1}}{2}, \\ &\text{and } \mathbf{x}(t_1) \prec \mathbf{x}(t_2) \text{ for } -\tau \le t_1 \le t_2 \le \tau. \end{aligned} \tag{4.7.2}$$

See Problem 2. Since D^o is open there exists $\epsilon > 0$ such that $\mathbf{x}(t) \in D^o$ for $t \in [-\epsilon, \epsilon]$. Then $f(t) := F(\mathbf{x}(t))$ is an increasing function on $[-\epsilon, \epsilon]$. Hence $f'(0) = \frac{\partial F}{\partial x_i}(\mathbf{x}) - \frac{\partial F}{\partial x_{i+1}}(\mathbf{x}) \ge 0$. This proves (4.7.1) in D^o. The continuity argument yields (4.7.1) in D.

Assume now that (4.7.1) holds. Let $\mathbf{y} = (y_1, ..., y_n)^\top, \mathbf{z} = (z_1, ..., z_n)^\top \in D$ and define

$$\mathbf{y}(t) := (1-t)\mathbf{y} + t\mathbf{z}, \ g(t) := F((1-t)\mathbf{y} + t\mathbf{z}), \text{ for } t \in [0, 1].$$

Suppose that $\mathbf{y} \prec \mathbf{z}$. Then $\mathbf{y}(t_1) \prec \mathbf{y}(t_2)$ for $0 \le t_1 \le t_2 \le 1$. Since D is Schur set $[\mathbf{y}, \mathbf{z}] \subset D$. Then

$$g'(t) = \sum_{i=1}^{n-1} (\frac{\partial F(\mathbf{y}(t))}{\partial x_i} - \frac{\partial F(\mathbf{y}(t))}{\partial x_{i+1}}) \sum_{j=1}^{i} (z_j - y_j). \tag{4.7.3}$$

See Problem 3. Hence $g'(t) \ge 0$, i.e. $g(t)$ is a nondecreasing function on $[0, 1]$. Thus $F(\mathbf{y}) = g(0) \le g(1) = F(\mathbf{z})$. Assume that for any point $\mathbf{x} = (x_1, ..., x_n)^\top \in D$ such that $x_i > x_{i+1}$ the inequality $\frac{\partial F}{\partial x_i}(\mathbf{x}) > \frac{\partial F}{\partial x_{i+1}}(\mathbf{x})$. Suppose that $\mathbf{y} \neq \mathbf{z}$. Then $g'(t) > 0$ and $F(\mathbf{y}) = g(0) < g(1) = F(\mathbf{z})$. □

Theorem 4.7.6 *Let $D \subset \mathbb{R}^n_\searrow$ be a regular Schur set in $\mathbb{R}^n$. Let $F \in \mathrm{C}^1(\mathrm{D})$. If F is strong Schur's order preserving then*

$$\frac{\partial F}{\partial x_1}(\mathbf{x}) \ge ... \ge \frac{\partial F}{\partial x_n}(\mathbf{x}) \ge 0, \text{ for each } \mathbf{x} = (x_1, ..., x_n)^\top \in D. \tag{4.7.4}$$

Suppose that in addition to the above assumptions D is convex. If F satisfies the above inequalities then F is strong Schur's order preserving. If for any point $\mathbf{x} = (x_1, ..., x_n)^\top \in D$ $\frac{\partial F}{\partial x_n}(\mathbf{x}) > 0$ and $\frac{\partial F}{\partial x_i}(\mathbf{x}) > \frac{\partial F}{\partial x_{i+1}}(\mathbf{x})$ whenever $x_i > x_{i+1}$ holds then F is strict strong Schur's order preserving.

Proof. Assume that F is strong Schur's order preserving. Since F is Schur's order preserving (4.7.1) holds. Let $\mathbf{x} = (x_1, ..., x_n)^\top \in D^o$. Define $\mathbf{w}(t) = \mathbf{x} + t\mathbf{e}_n$. Then there exists $\epsilon > 0$ such that $\mathbf{w}(t) \in D^o$ for $t \in [-\epsilon, \epsilon]$. Clearly $\mathbf{w}(t_1) \preceq \mathbf{w}(t_2)$ for $-\epsilon \le t_1 \le t_2 \le \epsilon$. Hence the function $h(t) := F(\mathbf{w}(t))$ is not decreasing on the interval $[-\epsilon, \epsilon]$. Thus $\frac{\partial F}{\partial x_n}(\mathbf{x}) = h'(0) \ge 0$. Use the continuity argument to deduce that $\frac{\partial F}{\partial x_n}(\mathbf{x}) \ge 0$ for any $\mathbf{x} \in D$.

Assume that D is convex and (4.7.4) holds. Let $\mathbf{y}, \mathbf{z} \in D$ and define $\mathbf{y}(t)$ and $g(t)$ as in the proof of Theorem 4.7.5. Then

$$\begin{aligned} g'(t) = & \sum_{i=1}^{n-1} (\frac{\partial F(\mathbf{y}(t))}{\partial x_i} - \frac{\partial F(\mathbf{y}(t))}{\partial x_{i+1}}) \sum_{j=1}^{i} (z_j - y_j) \\ & + \frac{\partial F(\mathbf{y}(t))}{\partial x_n} \sum_{j=1}^{n} (z_j - y_j). \end{aligned} \tag{4.7.5}$$

See Problem 3. Assume that $\mathbf{y} \preceq \mathbf{z}$. Then $g'(t) \ge 0$. Hence $F(\mathbf{y}) \le F(\mathbf{z})$.

Assume now that for any point $\mathbf{x} = (x_1, ..., x_n)^\top \in D$ $\frac{\partial F}{\partial x_n}(\mathbf{x}) > 0$ and $\frac{\partial F}{\partial x_i}(\mathbf{x}) > \frac{\partial F}{\partial x_{i+1}}(\mathbf{x})$ whenever $x_i > x_{i+1}$. Let $\mathbf{y}, \mathbf{z} \in D$ and assume that

$\mathbf{y} \preceq \mathbf{z}$ and $\mathbf{y} \neq \mathbf{z}$. Define $g(t)$ on $[0,1]$ as above. Use (4.7.5) to deduce that $g'(t) > 0$ on $[0,1]$. Hence $F(\mathbf{y}) < F(\mathbf{z})$. □

Let C be a convex set. A function $f : C \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}), \text{ for any } \alpha \in [0,1], \tag{4.7.6}$$

and $\mathbf{x}, \mathbf{y} \in C$.

f is called *strictly* convex if for any $\mathbf{x}, \mathbf{y} \in C$, $\mathbf{x} \neq \mathbf{y}$ and $\alpha \in (0,1)$ the strict inequality holds in the above inequality. The following result is well known. (See Problems 4-9)

Theorem 4.7.7 *Let $C \subset \mathbb{R}^d$ be a regular convex set. Assume that $f \in \mathrm{C}^2(\mathrm{C})$. Then f is convex if and only if the symmetric matrix $H(f) := (\frac{\partial^2 f}{\partial x_i x_j})_{i,j=1}^d$ is nonnnegative definite for each $\mathbf{y} \in C$. Furthermore, if $H(f)$ is positive definite for each $\mathbf{y} \in C$ then f is strictly convex.*

For any set $T \in \mathbf{V}$ we let $\mathrm{Cl}\,T$ be the closure of T in the standard topology in $\mathbf{V}$ (which is identified with the standard topology of $\mathbb{R}^{\dim_{\mathbb{R}}\mathbf{V}}$).

Proposition 4.7.8 *Let $C \subset \mathbf{V}$ be convex. Then $\mathrm{Cl}\,C$ is convex. Assume that C is a regular set and $f \in \mathrm{C}^0(\mathrm{Cl}\,\mathrm{C})$. Then f is convex in $\mathrm{Cl}\,C$ if and only if f is convex in C.*

See Problem 8.

Denote by $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ the extended real line. Then $a + \infty = \infty + a = \infty$ for $a \in \mathbb{R} \cup \{\infty\}$, $a - \infty = -\infty + a = -\infty$ for $a \in \mathbb{R} \cup \{-\infty\}$ and $\infty - \infty$, $-\infty + \infty$ are not defined. For $a > 0$ we let $a\infty = \infty a = \infty$, $a(-\infty) = (-\infty)a = -\infty$ and $0\infty = \infty 0 = 0$, $0(-\infty) = (-\infty)0 = 0$. Clearly for any $a \in \mathbb{R}$ $-\infty < a < \infty$. Let C be a convex set. Then $f : C \to \overline{\mathbb{R}}$ is called an extended convex function if (4.7.6) holds. Let $f : C \to \mathbb{R}$ be a convex function. Then f has the following continuity and differentiability properties:

In the one dimensional case where $C = (a,b) \subset \mathbb{R}$ f is continuous on C and f has a derivative $f'(x)$ at all but a countable set of points. $f'(x)$ is an nondecreasing function (where defined). In particular f has left and right derivatives at each x, which is given as the left and the right limits of $f'(x)$ (where defined).

In the general case $C \subset \mathbf{V}$, f is continuous function in $\mathrm{ri}\,C$, has a differential Df in a dense set C_1 of $\mathrm{ri}\,C$, the complement of C_1 in $\mathrm{ri}\,C$ has a zero measure, and Df is continuous in C_1. Furthermore at each $\mathbf{x} \in \mathrm{ri}\,C$ f has a subdifferential $\phi \in \mathrm{Hom}(\mathbf{V}, \mathbb{R})$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \phi(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in C. \tag{4.7.7}$$

See for example [Roc70].

Proposition 4.7.9 *(The maximal principle) Let C be a convex set and let $f_\phi : C \to \bar{\mathbb{R}}$ be an extended convex function for each ϕ a set Φ. Then*

$$f(\mathbf{x}) := \sup_{\phi\in\Phi} f_\phi(\mathbf{x}), \quad \text{for each } \mathbf{x} \in \mathbf{V},$$

is an extended convex function on C.

Theorem 4.7.10 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Then the function $\phi_i : \mathbf{S}(\mathbf{V}) \to \mathbb{R}$ given by*

$$\phi_i(T) := \sum_{j=1}^{i} \lambda_i(T), \quad T \in \mathbf{S}(\mathbf{V}),\ i = 1, ..., n, \tag{4.7.8}$$

is a continuous homogeneous convex function for $i = 1, ..., n-1$. $\phi_n(T) = \operatorname{tr} T$ is a linear function on $\mathbf{S}(\mathbf{V})$.

Proof. Clearly $\phi_i(aT) = a\phi_i(T)$ for $a \in [0, \infty)$. Hence ϕ_i is a homogeneous function. Since the eigenvalues of T are continuous it follows that ϕ_i is a continuous function. Clearly ϕ_n is a linear function on the vector space $\mathbf{S}(\mathbf{V})$. Combine Theorem 4.4.8 with Proposition 4.7.9 to deduce that ϕ_i is convex. □

Corollary 4.7.11 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Then*

$$\lambda(\alpha A + (1-\alpha)B) \prec \alpha\lambda(A) + (1-\alpha)\lambda(B), \text{ for any } A, B \in \mathbf{S}(\mathbf{V}),\ \alpha \in [0, 1].$$

For $\alpha \in (0, 1)$ equality holds if an only if there exists an orthonormal basis $[\mathbf{v}_1, ..., \mathbf{v}_n]$ in $\mathbf{V}$ such that

$$A\mathbf{v}_i = \lambda_i(A)\mathbf{u}_i,\ B\mathbf{v}_i = \lambda_i(B)\mathbf{v}_i, \quad i = 1, ..., n.$$

See Problem 10.

Proposition 4.7.12 *Let $\mathbf{V}$ be n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. For $D \subset \mathbb{R}^n$ let*

$$\lambda^{-1}(D) := \{T \in \mathbf{S}(\mathbf{V}) : \quad \lambda(T) \in D\}. \tag{4.7.9}$$

If $D \subset \mathbb{R}^n$ is a regular convex Schur set then $\lambda^{-1}(D)$ is regular convex set in the vector space $\mathbf{S}(\mathbf{V})$.

Proof. The continuity of the function $\lambda : \mathbf{S}(\mathbf{V}) \to \mathbb{R}^n_{\searrow}$ implies that $\lambda^{-1}(D)$ is a regular set in $\mathbf{S}(\mathbf{V})$. Suppose that $A, B \in \lambda(D)^{-1}$ and $\alpha \in [0,1]$. Since D is convex $\alpha\lambda(A)+(1-\alpha)\lambda(B) \in D$. Since D is a Schur set Corollary 4.7.11 yields that $\lambda(\alpha A+(1-\alpha)B) \in D$. Hence $\alpha A+(1-\alpha)B \in \lambda^{-1}(D)$. □

Theorem 4.7.13 *Let $D \subset \mathbb{R}^n_{\searrow}$ be a regular convex Schur set and let $h : D \to \mathbb{R}$. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $f : \lambda^{-1}(D) \to \mathbb{R}$ be the spectral function given by $f(A) := h(\lambda(A))$. Then the following are equivalent:*
(a) f is (strictly) convex on $\lambda^{-1}(D)$.
(b) h is (strictly) convex and (strictly) Schur's order preserving on D.

Proof. Choose a fixed orthonormal basis $[\mathbf{u}_1, ..., \mathbf{u}_n]$. We then identify $\mathbf{S}(\mathbf{V})$ with $\mathbf{H}_n(\mathbb{F})$. Thus we view $\mathcal{T} := \lambda^{-1}(D)$ is a subset of $\mathbf{H}_n(\mathbb{F})$. Since D is a regular convex Schur set Proposition 4.7.12 yields that $\mathcal{T}$ is a regular convex set. For $\mathbf{x} = (x_1, ..., x_n)^\top \in \mathbb{R}^n$ let $D(\mathbf{x}) := \operatorname{diag}(x_1, ..., x_n)$. Then $\lambda(D(\mathbf{x})) = \bar{\mathbf{x}}$. Thus $D(\mathbf{x}) \in \mathcal{T} \iff \bar{\mathbf{x}} \in D$ and $f(D(\mathbf{x})) = h(\bar{\mathbf{x}})$ for $\bar{\mathbf{x}} \in D$.
(a) ⇒ (b). Assume that f convex on $\mathcal{T}$. By restricting f to $D(\mathbf{x}), \mathbf{x} \in D$ we deduce that h is convex on D. If f is strictly convex on $\mathcal{T}$ we deduce that h is strictly convex on D.

Let $\mathbf{x}, \mathbf{y} \in D$ and assume that $\mathbf{x} \prec \mathbf{y}$. Then (4.6.3) holds. Hence

$$D(\mathbf{x}) = \sum_{P \in \mathcal{P}_n} a_P P D(\mathbf{y}) P^\top.$$

Clearly $\lambda(PD(\mathbf{y})P^\top) = \lambda(D(\mathbf{y})) = \mathbf{y}$. The convexity of f yields

$$h(\mathbf{x}) = f(D(\mathbf{x})) \le \sum_{P \in \mathcal{P}_n} a_P f(PD(\mathbf{y})P^\top) = f(D(\mathbf{y})) = h(\mathbf{y}).$$

See Problem 6. Hence h is Schur's order preserving. If f is strictly convex on $\mathcal{T}$ then in the above inequality one has a strict inequality if $\mathbf{x} \ne \mathbf{y}$. Hence h is strictly Schur's order preserving.
(b) ⇒ (a). Assume that h is convex. Then for $A, B \in \mathcal{T}$

$$\alpha f(A)+(1-\alpha)f(B) = \alpha h(\lambda(A))+(1-\alpha)h(\lambda(B)) \ge h(\alpha\lambda(A)+(1-\alpha)\lambda(B)).$$

Use Corollary 4.7.11 and the assumption that and h is Schur's order preserving to deduce the convexity of f. Suppose that h is strictly convex and strictly Schur's order preserving. Assume that $f(\alpha A + (1-\alpha)B) = \alpha f(A) + (1-\alpha)f(B)$ for some $A, B \in \mathcal{T}$ and $\alpha \in (0,1)$. Hence $\lambda(A) = \lambda(B)$ and $\lambda(\alpha A + (1-\alpha)B) = \alpha\lambda(A) + (1-\alpha)\lambda(B)$. Use Corollary 4.7.11 to deduce that $A = B$. Hence f is strictly convex. □

Theorem 4.7.14 *Let $D \subset \mathbb{R}^n_{\searrow}$ be a regular convex Schur set and let $h \in \mathrm{C}^1(\mathrm{D})$. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $f : \lambda^{-1}(D) \to \mathbb{R}$ be the spectral function given by $f(A) := h(\lambda(A))$. Then the following are equivalent:*
(a) f is convex on $\lambda^{-1}(D)$ and $f(A) \le f(B)$ for any $A, B \in \lambda^{-1}(D)$ such that $A \le B$.
(b) h is convex and strongly Schur's order preserving on D.

Proof. We repeat the proof of Theorem 4.7.13 with the following modifications.
(b) $\Rightarrow$ (a). Since h is convex and Schur's order preserving Theorem 4.7.13 yields that f is convex on $\mathcal{T}$. Let $A, B \in \mathcal{T}$ and assume that $A \le B$. Then $\lambda(A) \preceq \lambda(B)$. As h is strongly Schur's order preserving $h(\lambda(A)) \le h(\lambda(B)) \Rightarrow f(A) \le f(B)$.
(a) $\Rightarrow$ (b). Since f is convex on $\mathcal{T}$ Theorem 4.7.13 implies that h is convex and Schur's order preserving. Since $h \in \mathrm{C}^1(\mathrm{D})$ Theorem 4.7.5 yields that h satisfies the inequalities (4.7.1). Let $\mathbf{x} \in D^o$ and define $\mathbf{x}(t) := \mathbf{x} + t\mathbf{e}_n$. Then for a small $a > 0$ $\mathbf{x}(t) \in D^o$ for $t \in (-a, a)$. Clearly $D(\mathbf{x}(t_1)) \le D(\mathbf{x}(t_2))$ for $t_1 \le t_2$. Hence $g(t) := f(D(\mathbf{x}(t)) = h(\mathbf{x}(t))$ is a nondecreasing function on $(-a, a)$. Hence $\frac{\partial h}{\partial x_n}(\mathbf{x}) = g'(0) \ge 0$. Use the continuity hypothesis to deduce that h satisfies (4.7.4). Theorem 4.7.6 yields that h is strong Schur's order preserving. □

Theorem 4.7.15 *Let $\mathbf{V}$ be an N-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. For $n, N \in \mathbb{N}$ and $n < N$ let $\lambda_{(n)} : \mathbf{S}(\mathbf{V}) \to \mathbb{R}^n_{\searrow}$ be the map $A \mapsto \lambda_{(n)}(A) := (\lambda_1(A), ..., \lambda_n(A))^\top$. Assume that $D \subset \mathbb{R}^n_{\searrow}$ is a regular convex Schur set and let $\mathcal{T} := \lambda_{(n)}^{-1}(D) \subset \mathbf{S}(\mathbf{V})$. Let $f : \mathcal{T} \to \mathbb{R}$ be the spectral function given by $f(A) := h(\lambda_{(n)}(A))$. Assume that $n < N$ Then the following are equivalent:*
(a) f is convex on $\mathcal{T}$.
(b) h is convex and strongly Schur's order preserving on D.
(c) f is convex on $\mathcal{T}$ and $f(A) \le f(B)$ for any $A, B \in \mathcal{T}$ such that $A \le B$.

Proof. Let $\pi : \mathbb{R}^N_{\searrow} \to \mathbb{R}^n_{\searrow}$ be the projection on the first n coordinates. Let $D_1 := \pi^{-1}(D) \subset \mathbb{R}^N_{\searrow}$. It is straightforward to show that D_1 is a regular convex set. Let $h_1 := h \circ \pi : D_1 \to \mathbb{R}$. Then $\frac{\partial h_1}{\partial x_i} = 0$ for $i = n+1, ..., N$.
(a) $\Rightarrow$ (b). Suppose that f is convex on $\mathcal{T}$. Then Theorem 4.7.13 yields that h_1 is convex and Schur's order preserving. Theorem 4.7.5 yields the inequalities (4.7.1). Hence $\frac{\partial h_1}{\partial x_n}(\mathbf{y}) \ge \frac{\partial h_1}{\partial x_{n+1}}(\mathbf{y}) = 0$ for any $\mathbf{y} \in D_1$. Clearly

h is convex and

$$\frac{\partial h}{\partial x_i}(\mathbf{x}) = \frac{\partial h_1}{\partial x_i}(\mathbf{y}),\ i = 1, ..., n,\ \text{where } \mathbf{x} \in D,\ \mathbf{y} \in D_1,\ \text{and } \pi(\mathbf{y}) = \mathbf{x}.$$

Thus h satisfies (4.7.4). Theorem 4.7.6 yields that h is strongly Schur's order preserving.

Other nontrivial implications follow as in the proof of Theorem 4.7.14.

Problems

1. Show Proposition 4.7.4

2. Let $\mathbf{x} = (x_1, ..., x_n)^\top \in \mathbb{R}^n$ and assume $x_1 > ... > x_n$. Let $\mathbf{x}(t)$ be defined as in the proof of Theorem 4.7.5. Prove (4.7.2).

3. Let $D \subset \mathbb{R}^n$ be a regular set and assume that $[\mathbf{y}, \mathbf{z}] \subset D, \mathbf{y} = (y_1, ..., y_n)^\top, \mathbf{z} = (z_1, ..., z_n)^\top$. Let $F \in \mathrm{C}^1(\mathrm{D})$ and assume that $g(t)$ is defined as in the proof of Theorem 4.7.5. Show the equality (4.7.5). Suppose furthermore that $\sum_{i=1}^n y_i = \sum_{i=1}^n z_i$. Show the equality (4.7.3).

4. (a) Let $f \in \mathrm{C}^1(\mathrm{a,b})$. Show that f is convex on (a, b) if and only if $f'(x)$ is nondecreasing on (a, b). Show that if $f'(x)$ is increasing on (a, b) then f is strictly convex on (a, b).

   (b) Let $f \in \mathrm{C[a,b]} \cap \mathrm{C}^1(\mathrm{a,b})$. Show that f is convex in $[a, b]$ if and only if f is convex in (a, b). Show that if $f'(x)$ is increasing on (a, b) then f is strictly convex on $[a, b]$.

   (c) Let $f \in \mathrm{C}^2(\mathrm{a,b})$. Show that f is convex on (a, b) if and only if f'' is a nonnegative function on (a, b). Show that if $f''(x) > 0$ for each $x \in (a, b)$ then f is strictly convex on (a, b).

   (d) Prove Theorem 4.7.7.

5. (*This problem offers an alternative proof of Theorem 4.6.9.*) Let $a < b$ and $n \in \mathbb{N}$. Denote

   $$[a, b]^n_{\searrow} := \{(x_1, ..., x_n) \in \mathbb{R}^n_{\searrow} : \quad x_i \in [a, b],\ i = 1, ..., n\}.$$

   (a) Show that $[a, b]^n_{\searrow}$ is a regular convex Schur domain.

   (b) Let $f \in \mathrm{C}^1[\mathrm{a,b}]$ be a convex function. Let $F : [a, b]^n \to \mathbb{R}$ be defined by $F((x_1, ..., x_n)^\top) := \sum_{i=1}^n f(x_i)$. Show that F satisfies the condition (4.7.1) on $[a, b]^n_{\searrow}$. Hence Theorem 4.6.9 holds for any $\mathbf{x}, \mathbf{y} \in [a, b]^n$ such that $\mathbf{x} \prec \mathbf{y}$. (c) Assume that any convex $f \in \mathrm{C[a,b]}$ can be uniformly approximated by as sequence of convex $f_k \in \mathrm{C}^1[\mathrm{a,b}], \mathrm{k} = 1, ...$ Show that Theorem 4.6.9 holds for $\mathbf{x}, \mathbf{y} \in [a, b]^n$ such that $\mathbf{x} \prec \mathbf{y}$.

6. Let $D \subset \mathbf{V}$ be convex set and assume that $f : D \to \mathbb{R}$ be a convex function. Show that for any $k \geq 3$
$$f(\sum_{j=1}^{k} \mathbf{u}_j) \leq \sum_{j=1}^{k} \alpha_j f(\mathbf{u}_j), \text{ for any } \mathbf{u}_1, ..., \mathbf{u}_k \in D, \ \alpha_1 \geq 0, ..., \alpha_k \geq 0, \ \sum_{j=1}^{k} \alpha_j = 1.$$

7. Let $\mathbf{V}$ be a finite dimensional space and $C \subset \mathbf{V}$ a nonempty convex set. Let $\mathbf{x} \in C$. Show

   a. The subspace $\mathbf{U} :=$ span $(\mathrm{C} - \mathbf{x})$ does not depend on $\mathbf{x} \in C$.

   b. $C - \mathbf{x}$ has a nonempty convex interior in $\mathbf{U}$ and the definition of ri C does not depend on $\mathbf{x} \in C$.

8. Prove Proposition 4.7.8.

9. Prove Proposition 4.7.9.

10. Use Theorem 4.4.8 to show the equality case in Corollary 4.7.11

11. For $p \in [1, \infty)$ let
$$\|\mathbf{x}\|_{p,\mathbf{w}} := (\sum_{i=1}^{n} w_i |x_i|^p)^{\frac{1}{p}}, \ \mathbf{x} = (x_1, ..., x_n)^\top \in \mathbb{R}^n, \ \mathbf{w} = (w_1, ..., w_n)^\top \in \mathbb{R}^n_+.$$

    (a) Show that $\| \cdot \|_{p,\mathbf{w}} : \mathbb{R}^n \to \mathbb{R}$ is a homogeneous convex function. Furthermore this function is strictly convex if and only if $p > 1$ and $w_i > 0$ for $i = 1, ..., n$. (*Hint:* First prove the case $\mathbf{w} = (1, ..., 1)$.)

    (b) For $q > 1$ show that $\| \cdot \|^q_{p,\mathbf{w}} : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Furthermore this function is strictly convex if and only if $w_i > 0$ for $i = 1, ..., n$. (*Hint:* Use the fact that $f(x) = x^q$ is strictly convex on $[0, \infty)$.)

    (c) Show that for $q > 0$ the function $\| \cdot \|^q_{p,\mathbf{w}} : \mathbb{R}^n_{+,\searrow} \to \mathbb{R}$ is strong Schur's order preserving if and only if $w_1 \geq ... \geq w_n \geq 0$. Furthermore this function is strictly strong Schur's order preserving if and only if $w_1 \geq ... \geq w_n > 0$.

    (d) Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Show that for $q \geq 1$, $w_1 \geq ... \geq w_n \geq 0$ the spectral function $T \to \|\lambda(T)\|^q_{p,\mathbf{w}}$ is a convex function on $\mathbf{S}(\mathbf{V})_+$ (the positive self-adjoint operators on $\mathbf{V}$.) If in addition $w_n > 0$ and $\max(p, q) > 1$ then the above function is strictly convex on $\mathbf{S}(\mathbf{V})_+$.

12. Use the differentiability properties of convex function to show that Theorems 4.7.14 and 4.7.15 holds under the lesser assumption $h \in C(D)$.

13. Show that on $\mathbf{H}_{n,+}^o$ the function $\log\det A$ is a strictly concave function, i.e. $\det\,(\alpha A+(1-\alpha)B) \geq (\det\,A)^\alpha(\det\,B)^{1-\alpha}$. (**Hint**: Observe that $-\log x$ is a strictly convex function on $(0,\infty)$.)

4.8 Inequalities for traces

Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F}=\mathbb{R},\mathbb{C}$. Let $T:\mathbf{V}\to\mathbf{V}$ be a linear operator. Then $\operatorname{tr} T$ is the trace of the representation matrix A of with respect to any orthonormal basis of $\mathbf{V}$. See Problem 1.

Theorem 4.8.1 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F}=\mathbb{R},\mathbb{C}$. Assume that $S,T\in\mathbf{S}(\mathbf{V})$. Then $\operatorname{tr} ST$ is bounded below and above by*

$$\sum_{i=1}^n \lambda_i(S)\lambda_{n-i+1}(T) \leq \operatorname{tr} ST \leq \sum_{i=1}^n \lambda_i(S)\lambda_i(T). \tag{4.8.1}$$

Equality for the upper bound holds if and only if $ST=TS$ and there exists an orthonormal basis $\mathbf{x}_1,...,\mathbf{x}_n\in\mathbf{V}$ such that

$$S\mathbf{x}_i=\lambda_i(S)\mathbf{x}_i,\quad T\mathbf{x}_i=\lambda_i(T)\mathbf{x}_i,\quad i=1,...,n. \tag{4.8.2}$$

Equality for the lower bound holds if and only if $ST=TS$ and there exists an orthonormal basis $\mathbf{x}_1,...,\mathbf{x}_n\in\mathbf{V}$ such that

$$S\mathbf{x}_i=\lambda_i(S)\mathbf{x}_i,\quad T\mathbf{x}_i=\lambda_{n-i+1}(T)\mathbf{x}_i,\quad i=1,...,n. \tag{4.8.3}$$

Proof. Let $\mathbf{y}_1,...,\mathbf{y}_n$ be an orthonormal basis of $\mathbf{V}$ such that

$$\begin{aligned}
&T\mathbf{y}_i=\lambda_i(T)\mathbf{y}_i,\quad i=1,...,n,\\
&\lambda_1(T)=...=\lambda_{i_1}(T)>\lambda_{i_1+1}(T)=...=\lambda_{i_2}(T)>...>\\
&\lambda_{i_{k-1}+1}(T)=...=\lambda_{i_k}(T)=\lambda_n(T),\ 1\leq i_1<...<i_k=n.
\end{aligned}$$

If $k=1 \iff i_1=n$ it follows that $T=\lambda_1 I$ and the theorem is trivial in this case. Assume that $k>1$. Then

$$\begin{aligned}
&\operatorname{tr} ST=\sum_{i=1}^n \lambda_i(T)\langle S\mathbf{y}_i,\mathbf{y}_i\rangle=\\
&\sum_{i=1}^{n-1}(\lambda_i(T)-\lambda_{i+1}(T))(\sum_{l=1}^{i}\langle S\mathbf{y}_l,\mathbf{y}_l\rangle)+\lambda_n(T)(\sum_{l=1}^{n}\langle S\mathbf{y}_l,\mathbf{y}_l\rangle)=\\
&\sum_{j=1}^{k-1}(\lambda_{i_j}(T)-\lambda_{i_{j+1}}(T))\sum_{l=1}^{i_j}\langle S\mathbf{y}_l,\mathbf{y}_l\rangle+\lambda_n(T)\operatorname{tr} S.
\end{aligned}$$

Theorem 4.4.8 yields that $\sum_{l=1}^{i_j}\langle S\mathbf{y}_l,\mathbf{y}_l\rangle \le \sum_{l=1}^{i_j}\lambda_l(S)$. Substitute these inequalities for $j=1,...,k-1$ in the above identity to deduce the upper bound in (4.8.1). Clearly the condition (4.8.2) implies that $\operatorname{tr} ST$ is equal to the upper bound in (4.8.1). Assume now that $\operatorname{tr} ST$ is equal to the upper bound in (4.8.1). Then $\sum_{l=1}^{i_j}\langle S\mathbf{y}_l,\mathbf{y}_l\rangle = \sum_{l=1}^{i_j}\lambda_l(S)$ for $j=1,...,k-1$. Theorem 4.4.8 yields that span $(\mathbf{y}_1,...,\mathbf{y}_{i_j})$ is spanned by some i_j eigenvectors of S corresponding to the first i_j eigenvalues of S for $j=1,...,k-1$. Let $\mathbf{x}_1,...,\mathbf{x}_{i_1}$ be an orthonormal basis of span $(\mathbf{y}_1,...,\mathbf{y}_{i_1})$ consisting of the eigenvectors of S corresponding to the eigenvalues of $\lambda_1(S),...,\lambda_{i_1}(S)$. Since any $0\ne\mathbf{x}\in$ span $(\mathbf{y}_1,...,\mathbf{y}_{i_1})$ is an eigenvector of T corresponding to the eigenvalue $\lambda_{i_1}(T)$ it follows that (4.8.2) holds for $i=1,...,i_1$. Consider span $(\mathbf{y}_1,...,\mathbf{y}_{i_2})$. The above arguments imply that this subspace contains i_2 eigenvectors of S and T corresponding to the first i_2 eigenvalues of S and T. Hence $\mathbf{U}_2$, the orthogonal complement of span $(\mathbf{x}_1,...,\mathbf{x}_{i_1})$ in span $(\mathbf{y}_1,...,\mathbf{y}_{i_2})$, spanned by $\mathbf{x}_{i_1+1},...,\mathbf{x}_{i_2}$, which are i_2-i_1 orthonormal eigenvectors of S corresponding to the eigenvalues $\lambda_{i_1+}(S),...,\lambda_{i_2}(S)$. Since any nonzero vector in $\mathbf{U}_2$ is an eigenvector of T corresponding to the eigenvalue $\lambda_{i_2}(T)$ we deduce that (4.8.2) holds for $i=1,...,i_2$. Continuing in the same manner we obtain (4.8.2).

To prove the equality case in the lower bound consider the equality in the upper bound for $\operatorname{tr} S(-T)$. □

Corollary 4.8.2 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F}=\mathbb{R},\mathbb{C}$. Assume that $S,T\in\mathbf{S}(\mathbf{V})$. Then*

$$\sum_{i=1}^{n}(\lambda_i(S)-\lambda_i(T))^2 \le \operatorname{tr}(S-T)^2. \tag{4.8.4}$$

Equality holds if and only if $ST=TS$ and $\mathbf{V}$ has an orthonormal basis $\mathbf{x}_1,...,\mathbf{x}_n$ satisfying (4.8.2).

Proof. Note

$$\sum_{i=1}^{n}(\lambda_i(S)-\lambda_i(T))^2 = \operatorname{tr} S^2 + \operatorname{tr} T^2 - 2\sum_{i=1}^{n}\lambda_i(S)\lambda_i(T).$$

□

Corollary 4.8.3 *Let $S,T\in\mathbf{H}_n$. Then the inequalities (4.8.1) and (4.8.4) hold. Equalities in the upper bounds hold if and only if there exists $U\in\mathbf{U}_n$ such that $S=U\operatorname{diag}\lambda(S)U^*, T=U\operatorname{diag}\lambda(T)U^*$. Equality in the lower bound of (4.8.1) if and only if there exists $V\in\mathbf{U}_n$ such that $S=V\operatorname{diag}\lambda(S)V^*, -T=V\operatorname{diag}\lambda(-T)V^*$.*

Problems

1. Let $\mathbf{V}$ be a n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$.

   (a) Assume that $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation. Show that for any o.n. basis $\mathbf{x}_1, ..., \mathbf{x}_n$
   $$\operatorname{tr} T = \sum_{i=1}^{n} \langle T\mathbf{x}_i, \mathbf{x}_i \rangle.$$
   Furthermore, if $\mathbb{F} = \mathbb{C}$ then $\operatorname{tr} T$ is the sum of the n eigenvalues of T.

   (b) Let $S, T \in \mathbf{S}(\mathbf{V})$. Show that $\operatorname{tr} ST = \operatorname{tr} TS \in \mathbb{R}$.

4.9 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 $A \in \mathbb{F}^{m \times n}$ as given by (1.10.2). 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

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

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

$$T^*T\mathbf{c}_i = \lambda_i(T^*T)\mathbf{c}_i, \ \langle \mathbf{c}_i, \mathbf{c}_k \rangle_{\mathbf{V}} = \delta_{ik}, \ i, k = 1, ..., n, \tag{4.9.1}$$
$$\lambda_1(T^*T) \ge ... \ge \lambda_n(T^*T) \ge 0,$$
$$TT^*\mathbf{d}_j = \lambda_j(TT^*)\mathbf{d}_j, \ \langle \mathbf{d}_j, \mathbf{d}_l \rangle_{\mathbf{U}} = \delta_{jl}, \ j, l = 1, ..., m, \tag{4.9.2}$$
$$\lambda_1(TT^*) \ge ... \ge \lambda_m(TT^*) \ge 0,$$

Proposition 4.9.1 *Let $\mathbf{U}, \mathbf{V}$, be finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $T : \mathbf{V} \to \mathbf{U}$. Then* rank T = rank T* = rank T*T = rank TT* = r. *Furthermore the selfadjoint nonnegative definite operators T^*T and TT^* have exactly r positive eigenvalues, and*

$$\lambda_i(T^*T) = \lambda_i(TT^*) > 0, \quad i = 1, ..., \text{rank T}. \tag{4.9.3}$$

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 (4.9.1) then $\tilde{\mathbf{d}}_i := \frac{T\mathbf{c}_i}{||T\mathbf{c}_i||}, i = 1, ..., r$ satisfy (4.9.2) 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

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

Thus T^*T and TT^* have exactly r positive eigenvalues. Let $i \in [1, r]$. Then $T^*T\mathbf{c}_i \neq 0$. Hence $T\mathbf{c}_i \neq 0$. (4.9.1) 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 (4.9.3) holds. Assume that $\mathbf{c}_1, ..., \mathbf{c}_r$ satisfy (4.9.1). 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 $1 \le i < j \le r$. Then

$$0 = \langle \mathbf{c}_i, \mathbf{c}_j\rangle = \lambda_i(T^*T)\langle \mathbf{c}_i, \mathbf{c}_j\rangle = \langle T^*T\mathbf{c}_i, \mathbf{c}_j\rangle = \langle T\mathbf{c}_i, T\mathbf{c}_j\rangle \Rightarrow \langle \tilde{\mathbf{d}}_i, \tilde{\mathbf{d}}_j\rangle = 0.$$

Hence $\tilde{\mathbf{d}}_1, ..., \tilde{\mathbf{d}}_r$ is an orthonormal system. □

Let

$$\begin{aligned} &\sigma_i(T) = \sqrt{\lambda_i(T^*T)} \text{ for } i = 1, ...r, \quad \sigma_i(T) = 0 \text{ for } i > r, \\ &\sigma_{(p)}(T) := (\sigma_1(T), ..., \sigma_p(T))^\top \in \mathbb{R}^p_{\searrow}, \quad p \in \mathbb{N}. \end{aligned} \tag{4.9.4}$$

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 (4.9.1). 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 span $(\mathbf{d}_1, ..., \mathbf{d}_r)$ is spanned by all eigenvectors of TT^* corresponding to nonzero eigenvalues of TT^* it follows that $\ker T^* =$ span $(\mathbf{d}_{r+1}, ..., \mathbf{d}_m)$. Hence (4.9.2) holds. In these orthonormal bases of $\mathbf{U}$ and $\mathbf{V}$ the operators T and T^* represented quite simply:

$$\begin{aligned} &T\mathbf{c}_i = \sigma_i(T)\mathbf{d}_i,\ i = 1, ..., n, \quad \text{where } \mathbf{d}_i = 0 \text{ for } i > m, \\ &T^*\mathbf{d}_j = \sigma_j(T)\mathbf{c}_j,\ j = 1, ..., m, \quad \text{where } \mathbf{c}_j = 0 \text{ for } j > n.. \end{aligned} \tag{4.9.5}$$

Let

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

In the case $m \neq n$ we call Σ a diagonal matrix with the diagonal $\sigma_1, ..., \sigma_{\min(m,n)}$. Then in the bases $[\mathbf{d}_1, ..., \mathbf{d}_m]$ and $[\mathbf{c}_1, ..., \mathbf{c}_n]$ T and T^* represented by the matrices Σ and $\Sigma^\top$ respectively.

Lemma 4.9.2 *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*

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

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

Definition 4.9.3 *(4.9.7) is called the* singular value decomposition *(SVD) of A.*

Proposition 4.9.4 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and denote by $\mathcal{R}_{m,n,k}(\mathbb{F}) \subset \mathbb{F}^{m\times n}$ the set of all matrices of rank $k \in [1, \min(m, n)]$ at most. Then $A \in \mathcal{R}_{m,n,k}(\mathbb{F})$ if and only if A can be expressed as a sum of at most k matrices of rank 1. Furthermore $\mathcal{R}_{m,n,k}(\mathbb{F})$ is a variety in $\mathbb{F}^{m\times n}$ given by the polynomial conditions: Each $(k + 1) \times (k + 1)$ minor of A is equal to zero.*

For the proof see Problem 2

Definition 4.9.5 *Let $A \in \mathbb{C}^{m\times n}$ and assume that A has the SVD given by (4.9.7), where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_m], V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$. Denote by $A_k := \sum_{i=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^* \in \mathbb{C}^{m\times n}$ for $k = 1, \ldots, \text{rank A}$. For $k > \text{rank A}$ we define $A_k := A\ (= A_{\text{rank A}})$.*

Note that for $1 \leq k < \text{rank A}$, the matrix A_k is uniquely defined if and only if $\sigma_k > \sigma_{k+1}$. (See Problem 1.)

Theorem 4.9.6 *For $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and $A = (a_{ij}) \in \mathbb{F}^{m\times n}$ the following conditions hold:*

$$(4.9.8) \qquad ||A||_F := \sqrt{\mathrm{tr}\, A^*A} = \sqrt{\mathrm{tr}\, AA^*} = \sqrt{\sum_{i=1}^{\text{rank A}} \sigma_i(A)^2}.$$

$$(4.9.9) \qquad ||A||_2 := \max_{\mathbf{x}\in\mathbb{F}^n, ||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 = \sigma_1(A).$$

(4.9.10)
$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} ||A-B||_2 = ||A-A_k|| = \sigma_{k+1}(A), k = 1, ..., \text{rank A} - 1.$$

$$\sigma_i(A) \ge \sigma_i((a_{i_pj_q})_{p=1,q=1}^{m',n'}) \ge \sigma_{i+(m-m')+(n-n')}(A),$$
(4.9.11)
$$m' \in [1,m],\ n' \in [1,n],\ 1 \le i_1 < ... < i_{m'} \le m,\ 1 \le j_1 < ... < j_{n'} \le n.$$

Proof. The proof of (4.9.8) is left a Problem 7. We now show the equality in (4.9.9). View A as an operator $A : \mathbb{C}^n \to \mathbb{C}^m$. From the definition of $||A||_2$ it follows

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

which proves (4.9.9).

We now prove (4.9.10). In the SVD decomposition of A (4.9.7) assume that $U = (\mathbf{u}_1, ..., \mathbf{u}_m)$ and $V = (\mathbf{v}_1, ..., \mathbf{v}_n)$. Then (4.9.7) is equivalent to the following representation of A:
(4.9.12)
$$A = \sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^*,\ \mathbf{u}_1, ..., \mathbf{u}_r \in \mathbb{R}^m,\ \mathbf{v}_1, ..., \mathbf{v}_r \in \mathbb{R}^n,\ \mathbf{u}_i^*\mathbf{u}_j = \mathbf{v}_i^*\mathbf{v}_j = \delta_{ij},\ i,j = 1, ..., r,$$

where $r = \text{rank A}$. Let $B = \sum_{i=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^* \in \mathcal{R}_{m,n,k}$. Then in view of (4.9.9)

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

Let $B \in \mathcal{R}_{m,n,k}$. To show (4.9.10) it is enough to show that $||A-B||_2 \ge \sigma_{k+1}$. Let

$$\mathbf{W} := \{\mathbf{x} \in \mathbb{R}^n : \quad B\mathbf{x} = 0\}.$$

Then codim $\mathbf{W} \geq k$. Furthermore

$$||A-B||_2^2 \geq \max_{||\mathbf{x}||_2=1,\mathbf{x}\in\mathbf{W}} ||(A-B)\mathbf{x}||^2 = \max_{||\mathbf{x}||_2=1,\mathbf{x}\in\mathbf{W}} \mathbf{x}^*A^*A\mathbf{x} \geq \lambda_{k+1}(A^*A) = \sigma_{k+1}^2,$$

where the last inequality follows from the min-max characterization of $\lambda_{k+1}(A^*A)$.

Let $C = (a_{ij_q})_{i,q=1}^{m,n'}$. Then C^*C is an a principal submatrix of A^*A of dimension n'. The interlacing inequalities between the eigenvalues of A^*A and C^*C yields (4.9.11) for $m' = m$. Let $D = (a_{i_pj_q})_{p,q=1}^{m',n'}$. Then DD^* is a principle submatrix of CC^*. Use the interlacing properties of the eigenvalues of CC^* and DD^* to deduce (4.9.11). □

We now restate the above results for linear operators.

Definition 4.9.7 *Let* $\mathbf{U}, \mathbf{V}$ *be finite dimensional vector spaces over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$. *For* $k \in \mathbb{Z}_+$ *denote* $\mathrm{L}_k(\mathbf{V}, \mathbf{U}) := \{T \in \mathrm{L}(\mathbf{V}, \mathbf{U}) : \ \mathrm{rank\ T} \leq \mathrm{k}\}$. *Assume furthermore that* $\mathbf{U}, \mathbf{V}$ *are IPS. Let* $T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$ *and assume that the orthonormal bases of* $[\mathbf{d}_1, \ldots, \mathbf{d}_m], [\mathbf{c}_1, \ldots, \mathbf{c}_n]$ *of* $\mathbf{U}, \mathbf{V}$ *respectively satisfy (4.9.5). Define* $T_0 := \mathbf{0}$ *and* $T_k := T$ *for an integer* $k \geq \mathrm{rank\ T}$. *Let* $k \in [1, \mathrm{rank\ T} - 1] \cap \mathbb{N}$. *Define* $T_k \in \mathrm{L}(\mathbf{V}, \mathbf{U})$ *by the equality* $T_k(\mathbf{v}) = \sum_{i=1}^k \sigma_i(T)\langle \mathbf{v}, \mathbf{c}_i\rangle \mathbf{d}_i$ *for any* $\mathbf{v} \in \mathbf{V}$.

It is straightforward to show that $T_k \in \mathrm{L}_k(\mathbf{V}, \mathbf{U})$ and T_k is unique if and only if $\sigma_k(T) > \sigma_{k+1}(T)$. See Problem 8. Theorem 4.9.6 yields:

Corollary 4.9.8 *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*

$$(4.9.13) \qquad ||T||_F := \sqrt{\mathrm{tr}\, T^*T} = \sqrt{\mathrm{tr}\, TT^*} = \sqrt{\sum_{i=1}^{\mathrm{rank\ T}} \sigma_i(T)^2}.$$

$$(4.9.14) \qquad ||T||_2 := \max_{\mathbf{x}\in\mathbf{V}, ||\mathbf{x}||_2=1} ||T\mathbf{x}||_2 = \sigma_1(T).$$

$$(4.9.15) \qquad \min_{Q\in\mathrm{L}_k(\mathbf{V},\mathbf{U})} ||T - Q||_2 = \sigma_{k+1}(T), \quad k = 1, ..., \mathrm{rank\ T} - 1.$$

Problems

1. Let $\mathbf{U}, \mathbf{V}$ be finite dimensional inner product spaces. Assume that $T \in \mathrm{L}(\mathbf{U}, \mathbf{V})$. Show that for any complex number $t \in \mathbb{C}$ $\sigma_i(tT) = |t|\sigma_i(T)$ for all i.

2. Prove Proposition 4.9.4. (Use SVD to prove the nontrivial part of the Proposition.)

3. Let $A \in \mathbb{C}^{m\times n}$ and assume that $U \in \mathbf{U}(m), V \in \mathbf{V}(m)$. Show that $\sigma_i(UAV) = \sigma_i(A)$ for all i.

4. Let $A \in \mathbf{GL}(n, \mathbb{C})$. Show that $\sigma_1(A^{-1}) = \sigma_n(A)^{-1}$.

5. Let $\mathbf{U}, \mathbf{V}$ be IPS inner product space of dimensions m and n respectively. Assume that

$$\mathbf{U} = \mathbf{U}_1 \oplus \mathbf{U}_2, \dim \mathbf{U}_1 = m_1, \dim \mathbf{U}_2 = m_2,$$
$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2, \dim \mathbf{V}_1 = n_1, \dim \mathbf{V}_2 = n_2.$$

   Assume that $T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$. Suppose furthermore that $T\mathbf{V}_1 \subseteq \mathbf{U}_1, T\mathbf{V}_2 \subseteq \mathbf{U}_2$. Let $T_i \in \mathrm{L}(\mathbf{V}_i, \mathbf{U}_i)$ be the restriction of T to $\mathbf{V}_i$ for $i = 1, 2$. Then $\mathrm{rank\ T} = \mathrm{rank\ T_1} + \mathrm{rank\ T_2}$ and $\{\sigma_1(T), \ldots, \sigma_{\mathrm{rank\ T}}(T)\} = \{\sigma_1(T_1), \ldots, \sigma_{\mathrm{rank\ T_1}}(T_1)\} \cup \{\sigma_1(T_2), \ldots, \sigma_{\mathrm{rank\ T_2}}(T_2)\}$.

6. Let the assumptions of the Definition 4.9.5 hold. Show that for $1 \le k < \mathrm{rank\ A}$ A_k is uniquely defined if and only if $\sigma_k > \sigma_{k+1}$.

7. Prove the equalities in (4.9.8).

8. Let the assumptions of Definition 4.9.7 hold. Show that for $k \in [1, \mathrm{rank\ T} - 1] \cap \mathbb{N}$ $\mathrm{rank\ T_k} = \mathrm{k}$ and T_k is unique if and only if $\sigma_k(T) > \sigma_{k+1}(T)$.

9. Let $\mathbf{V}$ be an n-dimensional IPS. Assume that $T \in \mathrm{L}(\mathbf{V})$ is a normal operator. Let $\lambda_1(T), \ldots, \lambda_n(T)$ be the eigenvalues of T arranged in the order $|\lambda_1(T)| \ge \ldots \ge |\lambda_n(T)|$. Show that $\sigma_i(T) = |\lambda_i(T)|$ for $i = 1, \ldots, n$.

4.10 Characterizations of singular values

Theorem 4.10.1 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and assume that $A \in \mathbb{F}^{m\times n}$. Define*

$$H(A) = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in \mathrm{H}_{m+n}(\mathbb{F}). \tag{4.10.1}$$

Then

$$\begin{aligned}&\lambda_i(H(A)) = \sigma_i(A),\ \lambda_{m+n+1-i}(H(A)) = -\sigma_i(A),\ i = 1, ..., \mathrm{rank\ A},\\ &\lambda_j(H(A)) = 0,\ j = \mathrm{rank\ A} + 1, ..., \mathrm{n} + \mathrm{m} - \mathrm{rank\ A}.\end{aligned} \tag{4.10.2}$$

View A as an operator $A : \mathbb{F}^n \to \mathbb{F}^m$. Choose orthonormal bases $[\mathbf{d}_1, ..., \mathbf{d}_m], [\mathbf{c}_1, ..., \mathbf{c}_n]$ in $\mathbb{F}^m, \mathbb{F}^n$ respectively satisfying (4.9.5). Then

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix} = \sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix}, \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix} = -\sigma_i(A) \begin{bmatrix} \mathbf{d}_i \\ -\mathbf{c}_i \end{bmatrix},$$

$(4.10.3)\; i = 1, ..., \text{rank A},$

$$\ker H(A) = \text{span } ((\mathbf{d}_{r+1}^*, 0)^*, ..., (\mathbf{d}_m^*, 0)^*, (0, \mathbf{c}_{r+1}^*)*, ..., (0, \mathbf{c}_n^*)^*), \; r = \text{rank A}.$$

Proof. It is straightforward to show the equalities (4.10.3). Since all the eigenvectors appearing in (4.10.3) are linearly independent we deduce (4.10.2). □

Corollary 4.10.2 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and assume that $A \in \mathbb{F}^{m \times n}$. Let $\hat{A} := A[\alpha, \beta] \in \mathbb{F}^{p \times q}$ be a submatrix of A, formed by the set of rows and columns $\alpha \in Q_{p,m}, \beta \in Q_{q,n}$ respectively. Then*

$$\sigma_i(\hat{A}) \le \sigma_i(A) \text{ for } i = 1, \ldots. \tag{4.10.4}$$

For $l \in [1, \text{rank A}] \cap \mathbb{N}$ the equalities $\sigma_i(\hat{A}) = \sigma_i(A), i = 1, \ldots, l$ hold if and only if there exists two orthonormal systems of l right and left singular vectors $\mathbf{c}_1, \ldots, \mathbf{c}_l \in \mathbb{F}^n$, $\mathbf{d}_1, \ldots, \mathbf{d}_l \in \mathbb{F}^n$ satisfying (4.10.3) for $i = 1, \ldots, l$ such that the nonzero coordinates vectors $\mathbf{c}_1, \ldots, \mathbf{c}_l$ and $\mathbf{d}_1, \ldots, \mathbf{d}_l$ are located at the indices β, α respectively.

See Problem 1.

Corollary 4.10.3 *Let $\mathbf{V}, \mathbf{U}$ be IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\mathbf{W}$ is a subspace of $\mathbf{V}$. Let $T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$ and denote by $\hat{T} \in \mathrm{L}(\mathbf{W}, \mathbf{U})$ the restriction of T to $\mathbf{W}$. Then $\sigma_i(\hat{T}) \le \sigma_i(T)$ for any $i \in \mathbb{N}$. Furthermore $\sigma_i(\hat{T}) = \sigma_i(T)$ for $i = 1, \ldots, l \le \text{rank T}$ if and only if $\mathbf{U}$ contains a subspace spanned by the first l right singular vectors of T.*

See Problem 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 4.10.4 *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 3.

Corollary 4.10.5 *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

We now translate Theorem 4.10.1 to the operator setting.

Lemma 4.10.6 *Let $\mathbf{U}, \mathbf{V}$ be finite dimensional IPS spaces with the inner products $\langle\cdot,\cdot\rangle_{\mathbf{U}}, \langle\cdot,\cdot\rangle_{\mathbf{V}}$ respectively. Define $\mathbf{W} := \mathbf{V} \oplus \mathbf{U}$ be the induced IPS with*

$$\langle(\mathbf{y}, \mathbf{x}), (\mathbf{v}, \mathbf{u})\rangle_{\mathbf{W}} := \langle\mathbf{y}, \mathbf{v}\rangle_{\mathbf{V}} + \langle\mathbf{x}, \mathbf{u}\rangle_{\mathbf{U}}.$$

Let $T : \mathbf{V} \to \mathbf{U}$ be a linear operator, and $T^* : \mathbf{U} \to \mathbf{V}$ be the adjoint of T. Define the operator

$$\hat{T} : \mathbf{W} \to \mathbf{W}, \quad \hat{T}(\mathbf{y}, \mathbf{x}) := (T^*\mathbf{x}, T\mathbf{y}). \tag{4.10.5}$$

Then $\hat{T}$ is self-adjoint operator and $\hat{T}^2 = T^*T \oplus TT^*$. Hence the spectrum of $\hat{T}$ is symmetric with respect to the origin and $\hat{T}$ has exactly 2rank T *nonzero eigenvalues. More precisely, if* $\dim \mathbf{U} = m, \dim \mathbf{V} = n$ *then:*

$$(4.10.6)\lambda_i(\hat{T}) = -\lambda_{m+n-i+1}(\hat{T}) = \sigma_i(T),\ \textit{for } i = 1, \ldots, \mathrm{rank\ T},$$
$$\lambda_j(\hat{T}) = 0,\ \textit{for } j = \mathrm{rank\ T} + 1, \ldots, \mathrm{n + m - rank\ T}.$$

Let $\{\mathbf{d}_1, \ldots, \mathbf{d}_{\min(m,n)}\} \in \mathrm{Fr}(\min(m, n), \mathbf{U}), \{\mathbf{c}_1, \ldots, \mathbf{c}_{\min(m,n)}\} \in \mathrm{Fr}(\min(m, n), \mathbf{V})$ be the set of vectors satisfying (4.9.5). Define

(4.10.7)
$$\mathbf{z}_i := \frac{1}{\sqrt{2}}(\mathbf{c}_i, \mathbf{d}_i), \mathbf{z}_{m+n-i+1} := \frac{1}{\sqrt{2}}(\mathbf{c}_i, -\mathbf{d}_i), i = 1, \ldots, \min(m, n).$$

Then $\{\mathbf{z}_1, \mathbf{z}_{m+n}, \ldots, \mathbf{z}_{\min(m,n)}, \mathbf{z}_{m+n-\min(m,n)+1}\} \in \mathrm{Fr}(2\min(m, n), \mathbf{W})$. Furthermore $\hat{T}\mathbf{z}_i = \sigma_i(T)\mathbf{z}_i, \hat{T}\mathbf{z}_{m+n-i+1} = -\sigma_i(T)\mathbf{z}_{m+n-i+1}$ for $i = 1, \ldots, \min(m, n)$.

See Problem 5.

Theorem 4.10.7 *Let $\mathbf{U}, \mathbf{V}$ be m and n-dimensional IPS over $\mathbb{C}$ respectively. Let $T : \mathbf{V} \to \mathbf{U}$ be a linear operator. Then for each $k \in [1, \min(m, n)] \cap \mathbb{Z}$*

$$(4.10.8) \sum_{i=1}^{k} \sigma_i(T) = \max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in\mathrm{Fr}(k,\mathbf{U}),\{\mathbf{g}_1,\ldots,\mathbf{g}_k\}\in\mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}} =$$

$$\max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in\mathrm{Fr}(k,\mathbf{U}),\{\mathbf{g}_1,\ldots,\mathbf{g}_k\}\in\mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} |\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}|.$$

Furthermore $\sum_{i=1}^k \sigma_i(T) = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$ *for some two k-orthonormal frames* $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\}, G_k = \{\mathbf{g}_1, ..., \mathbf{g}_k\}$ *if and only* span $((\mathbf{g}_1, \mathbf{f}_1), \ldots, (\mathbf{g}_\mathrm{k}, \mathbf{f}_\mathrm{k}))$ *is spanned by k eigenvectors of $\hat{T}$ corresponding to the first k eigenvalues of $\hat{T}$.*

Proof. Assume that $\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}), \{\mathbf{g}_1, ..., \mathbf{g}_k\} \in \mathrm{Fr}(k, \mathbf{V})$. Let $\mathbf{w}_i := \frac{1}{\sqrt{2}}(\mathbf{g}_i, \mathbf{f}_i), i = 1, \ldots, k$. Then $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \in \mathrm{Fr}(k, \mathbf{W})$. A straightforward calculation shows $\sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$. The maximal characterization of $\sum_{i=1}^k \lambda_i(\hat{T})$, (Theorem 4.4.8), and (4.10.6) yield the inequality $\sum_{i=1}^k \sigma_i(\hat{T}) \ge \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$ for $k \in [\min(m, n) \cap \mathbb{Z}$. Let $\mathbf{c}_1, \ldots, \mathbf{c}_{\min(m,n)}, \mathbf{d}_1, \ldots, \mathbf{d}_{\min(m,n)}$ satisfy (4.9.5). Then Lemma 4.10.6 yields that $\sum_{i=1}^k \sigma_i(\hat{T}) = \sum_{i=1}^k \Re\langle T\mathbf{c}_i, \mathbf{d}_i\rangle_{\mathbf{U}}$ for $k \in [\min(m, n) \cap \mathbb{Z}$. This proves the first equality of (4.10.8). The second equality of (4.10.8) is straightforward. (See Problem 6).)

Assume now that $\sum_{i=1}^k \sigma_i(T) = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$ for some two k-orthonormal frames $F_k = \{\mathbf{f}_1, ..., \mathbf{f}_k\}, G_k = \{\mathbf{g}_1, ..., \mathbf{g}_k\}$. Define $\mathbf{w}_1, \ldots, \mathbf{w}_k$ as above. The above arguments yield that $\sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \lambda_i(\hat{T})$. Theorem 4.4.8 yields that span $((\mathbf{g}_1, \mathbf{f}_1), \ldots, (\mathbf{g}_\mathrm{k}, \mathbf{f}_\mathrm{k}))$ is spanned by k eigenvectors of $\hat{T}$ corresponding to the first k eigenvalues of $\hat{T}$. Vice versa, assume that $\{\mathbf{f}_1, ..., \mathbf{f}_k\} \in \mathrm{Fr}(k, \mathbf{U}), \{\mathbf{g}_1, ..., \mathbf{g}_k\} \in \mathrm{Fr}(k, \mathbf{V})$ and span $((\mathbf{g}_1, \mathbf{f}_1), \ldots, (\mathbf{g}_\mathrm{k}, \mathbf{f}_\mathrm{k}))$ is spanned by k eigenvectors of $\hat{T}$ corresponding to the first k eigenvalues of $\hat{T}$. Define $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \in \mathrm{Fr}(\mathbf{W})$ as above. Then span $(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ contains k linearly independent eigenvectors corresponding to the the first k eigenvalues of $\hat{T}$. Theorem 4.4.8 and Lemma 4.10.6 yield that $\sigma_i(T) = \sum_{i=1}^k \langle \hat{T}\mathbf{w}_i, \mathbf{w}_i\rangle_{\mathbf{W}} = \sum_{i=1}^k \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}$. □

Theorem 4.10.8 $\mathbf{U}, \mathbf{V}$ *be m and n dimensional IPS spaces. Assume that Let $S, T : \mathbf{V} \to \mathbf{U}$ be linear operators. Then*

$$\Re\,\mathrm{tr}(S^*T) \le \sum_{i=1}^{\min(m,n)} \sigma_i(S)\sigma_i(T). \tag{4.10.9}$$

Equality holds if and only if there exists two orthonormal set $\{\mathbf{d}_1, \ldots, \mathbf{d}_{\min(m,n)}\} \in \mathrm{Fr}(\min(m, n), \mathbf{U}), \{\mathbf{c}_1, \ldots, \mathbf{c}_{\min(m,n)}\} \in \mathrm{Fr}(\min(m, n), \mathbf{V})$, *such that*

$$S\mathbf{c}_i = \sigma_i(S)\mathbf{d}_i, T\mathbf{c}_i = \sigma_i(T)\mathbf{d}_i, S^*\mathbf{d}_i = \sigma_i(S)\mathbf{c}_i, T^*\mathbf{d}_i = \sigma_i(T)\mathbf{c}_i, i = 1, \ldots, \min(m, n). \tag{4.10.10}$$

Proof. Let $A, B \in \mathbb{C}^{n\times m}$. Then $\mathrm{tr}\, B^*A = \overline{\mathrm{tr}\, AB^*}$. Hence $2\Re\,\mathrm{tr}\, AB^* = \mathrm{tr}\, H(A)H(B)$. Therefore $2\Re\,\mathrm{tr}\, S^*T = \mathrm{tr}\, \hat{S}\hat{T}$. Use Theorem 4.8.1 for $\hat{S}, \hat{T}$ and Lemma 4.10.6 to deduce (4.10.9). Equality in (4.10.9) if and only if $\mathrm{tr}\, \hat{S}\hat{T} = \sum_{i=1}^{m+n} \lambda_i(\hat{S})\lambda_i(\hat{T})$.

Clearly, the assumptions that $\{\mathbf{d}_1,\ldots,\mathbf{d}_{\min(m,n)}\}\in \mathrm{Fr}(\min(m,n),\mathbf{U})$, $\{\mathbf{c}_1,\ldots,\mathbf{c}_{\min(m,n)}\}\in \mathrm{Fr}(\min(m,n),\mathbf{V})$, and the equalities (4.10.10) imply equality in (4.10.9).

Assume equality in (4.10.9). Theorem 4.8.1 and the definitions of $\hat{S},\hat{T}$ yields the existence $\{\mathbf{d}_1,\ldots,\mathbf{d}_{\min(m,n)}\}\in \mathrm{Fr}(\min(m,n),\mathbf{U})$, $\{\mathbf{c}_1,\ldots,\mathbf{c}_{\min(m,n)}\}\in \mathrm{Fr}(\min(m,n),\mathbf{V})$, such that (4.10.10) hold. □

Theorem 4.10.9 *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*

$$(4.10.11)\qquad \min_{Q\in \mathrm{L}_k(\mathbf{V},\mathbf{U})}||T-Q||_F=\sqrt{\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T)},\quad k=1,...,\text{rank T}-1.$$

Furthermore $||T-Q||_F=\sqrt{\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T)}$ *for some* $Q\in \mathrm{L}_k(\mathbf{V},\mathbf{U}), k<$ rank T, *if and only there* $Q=T_k$, *where* T_k *is defined in Definition 4.9.7.*

Proof. Use Theorem 4.10.8 to deduce that for any $Q\in \mathrm{L}(\mathbf{V},\mathbf{U})$ one has

$$||T-Q||_F^2=\operatorname{tr}T^*T-2\Re\operatorname{tr}Q^*T+\operatorname{tr}Q^*Q\ge$$
$$\sum_{i=1}^{\text{rank T}}\sigma_i^2(T)-2\sum_{i=1}^{k}\sigma_i(T)\sigma_i(Q)+\sum_{i=1}^{k}\sigma_i^2(Q)=$$
$$\sum_{i=1}^{k}(\sigma_i(T)-\sigma_i(Q))^2+\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T)\ge\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T).$$

Clearly $||T-T_k||_F^2=\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T)$. Hence (4.10.11) holds. Vice versa if $Q\in \mathrm{L}_k(\mathbf{V},\mathbf{U})$ and $||T-Q||_F^2=\sum_{i=k+1}^{\text{rank T}}\sigma_i^2(T)$ then the equality case in Theorem 4.10.8 yields that $Q=T_k$. □

Corollary 4.10.10 *Let* $F=\mathbb{R},\mathbb{C}$ *and* $A\in\mathbb{F}^{m\times n}$. *Then*

$$(4.10.12)\qquad \min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})}||A-B||_F=\sqrt{\sum_{i=k+1}^{\text{rank A}}\sigma_i^2(A)},\quad k=1,...,\text{rank A}-1.$$

Furthermore $||A-B||_F=\sqrt{\sum_{i=k+1}^{\text{rank A}}\sigma_i^2(A)}$ *for some* $B\in\mathcal{R}_{m,n,k}(\mathbb{F}), k<$ rank A, *if and only there* $B=A_k$, *where* A_k *is defined in Definition 4.9.5.*

Theorem 4.10.11 *Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and $A \in \mathbb{F}^{m\times n}$. Then*

$$\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} \sum_{i=1}^{j} \sigma_i(A-B) = \sum_{i=k+1}^{k+j} \sigma_i(A), \tag{4.10.13}$$

$$j = 1, ..., \min(m,n) - k,\ k = 1, \dots, \min(m,n) - 1.$$

Proof. Clearly, for $B = A_k$ we have the equality $\sum_{i=1}^{j} \sigma_i(A-B) = \sum_{i=k+1}^{k+j} \sigma_i(A)$. Let $B \in \mathcal{R}_{m,n,k}(\mathbb{F})$. Let $\mathbf{X} \in \mathrm{Gr}(k, \mathbb{C}^n)$ be an subspace which contains the columns of B. Let $\mathbf{W} = \{(\mathbf{0}^\top, \mathbf{x}^\top)^\top \in \mathbb{F}^{m+n}, \mathbf{x} \in \mathbf{X}\}$. Observe that for any $\mathbf{z} \in \mathbf{W}^\perp$ one has the equality $\mathbf{z}^* H((A-B))\mathbf{z} = \mathbf{z}^* H(A)\mathbf{z}$. Combine Theorems 4.4.9 and 4.10.1 to deduce $\sum_{i=1}^{j} \sigma_i(B-A) \geq \sum_{i=k+1}^{k+j} \sigma_i(A)$. □

Theorem 4.10.12 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{C}$. Let $T : \mathbf{V} \to \mathbf{V}$ be a linear operator. Assume the n eigenvalues of T $\lambda_1(T), \dots, \lambda_n(T)$ are arranged the order $|\lambda_1(T)| \geq \dots \geq |\lambda_n(T)|$. Let $\boldsymbol{\lambda}_a(T) := (|\lambda_1(T)|, \dots, |\lambda_n(T)|)$, $\boldsymbol{\sigma}(T) := (\sigma_1(T), \dots, \sigma_n(T))$. Then $\boldsymbol{\lambda}_a(T) \preceq \boldsymbol{\sigma}(T)$. That is*

$$\sum_{i=1}^{k} |\lambda_i(T)| \leq \sum_{i=1}^{k} \sigma_i(T), \quad i = 1, \dots, n. \tag{4.10.14}$$

Furthermore, $\sum_{i=1}^{k} |\lambda_i(T)| = \sum_{i=1}^{k} \sigma_i(T)$ for some $k \in [1,n] \cap \mathbb{Z}$ if and only if the following conditions are satisfied. There exists an orthonormal basis $\mathbf{x}_1, \dots, \mathbf{x}_n$ of $\mathbf{V}$ such that:

1. *$T\mathbf{x}_i = \lambda_i(T)\mathbf{x}_i, T^*\mathbf{x}_i = \overline{\lambda_i(T)}\mathbf{x}_i$ for $i = 1, \dots, k$.*

2. *Denote by $S : \mathbf{U} \to \mathbf{U}$ the restriction of T to the invariant subspace $\mathbf{U} = \mathrm{span}\,(\mathbf{x}_{k+1}, \dots, \mathbf{x}_n)$. Then $||S||_2 \leq |\lambda_k(T)|$.*

Proof. Use Theorem 4.2.12 to choose an orthonormal basis $\mathbf{g}_1, \dots, \mathbf{g}_n$ of $\mathbf{V}$, such that T is represented by an upper diagonal matrix $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ such that $a_{ii} = \lambda_i(T), i = 1, \dots, n$. Let $\epsilon_i \in \mathbb{C}, |\epsilon_i| = 1$ such that $\bar{\epsilon}_i\lambda_i(T) = |\lambda_i(T)|$ for $i = 1, \dots, n$. Let $S \in \mathrm{L}(\mathbf{V})$ be presented in the basis $\mathbf{g}_1, \dots, \mathbf{g}_n$ by a diagonal matrix $\mathrm{diag}(\epsilon_1, \dots, \epsilon_k, 0, \dots, 0)$. Clearly, $\sigma_i(S) = 1$ for $i = 1, \dots, k$ and $\sigma_i(S) = 0$ for $i = k+1, \dots, n$. Furthermore, $\Re\,\mathrm{tr}\, S^*C = \sum_{i=1}^{k} |\lambda_i(T)|$. Hence Theorem 4.10.8 yields (4.10.14).

Assume now that $\sum_{i=1}^{k} |\lambda_i(T)| = \sum_{i=1}^{k} \sigma_i(T)$. Hence equality sign holds in (4.10.9). Hence there exists two orthonormal bases $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}, \{\mathbf{d}_1, \dots, \mathbf{d}_n\}$ in $\mathbf{V}$ such that (4.10.10) holds. It easily follows that $\{\mathbf{c}_1, \dots, \mathbf{c}_k\}, \{\mathbf{d}_1, \dots, \mathbf{d}_k\}$

are orthonormal bases of $\mathbf{W} := \mathrm{span}\,(\mathbf{g}_1, \ldots, \mathbf{g}_k)$. Hence $\mathbf{W}$ is an invariant subspace of T and T^*. Hence $A = A_1 \oplus A_2$, i.e. A is a block diagonal matrix. Thus $A_1 = (a_{ij})_{i,j=1}^k \in \mathbb{C}^{k\times k}, A_2 = (a_{ij})_{i,j=k+1}^n \in \mathbb{C}^{(n-k)\times(n-k)}$ represent the restriction of T to $\mathbf{W}, \mathbf{U} := \mathbf{W}^\perp$, denoted by T_1 and T_2 respectively. Hence $\sigma_i(T_1) = \sigma_i(T)$ for $i = 1, \ldots, k$. Note that the restriction of S to $\mathbf{W}$, denoted by S_1 is given by the diagonal matrix $D_1 := \mathrm{diag}(\epsilon_1, \ldots, \epsilon_k) \in \mathbf{U}(k)$. (4.10.10) yield that $S_1^{-1}T_1\mathbf{c}_i = \sigma_i(T)\mathbf{c}_i$ for $i = 1, \ldots, k$,i.e. $\sigma_1(T), \ldots, \sigma_k(T)$ are the eigenvalues of $S_1^{-1}T_1$. Clearly $S_1^{-1}T_1$ is presented in the basis $[\mathbf{g}_1, \ldots, \mathbf{g}_k]$ by the matrix $D_1^{-1}A_1$, which is a diagonal matrix with $|\lambda_1(T)|, \ldots, |\lambda_k(T)|$ on the main diagonal. That is $S_1^{-1}T_1$ has eigenvalues $|\lambda_1(T)|, \ldots, |\lambda_k(T)|$. Therefore $\sigma_i(T) = |\lambda_i(T)|$ for $i = 1, \ldots, k$. Theorem 4.9.6 yields that

$$\mathrm{tr}\, A_1^* A_1 = \sum_{i,j=1}^k |a_{ij}|^2 = \sum_{i=1}^k \sigma_i^2(A_1) = \sum_{i=1}^k \sigma_i^2(T_1) = \sum_{i=1}^k |\lambda_i(T)|^2.$$

As $\lambda_1(T), \ldots, \lambda_k(T)$ are the diagonal elements of A_1 is follows from the above equality that A_1 is a diagonal matrix. Hence we can choose $\mathbf{x}_i = \mathbf{g}_i$ for $i = 1, \ldots, n$ to obtain the part *1* of the equality case.

Let $T\mathbf{x} = \lambda\mathbf{x}$ where $||\mathbf{x}|| = 1$ and $\rho(T) = |\lambda|$. Recall $||T||_2 = \sigma_1(T)$, where $\sigma_1(T)^2 = \lambda_1(T^*T)$ is the maximal eigenvalue of the self-adjoint operator T^*T. The maximum characterization of $\lambda_1(T^*T)$ yields that $|\lambda|^2 = \langle T\mathbf{x}, T\mathbf{x}\rangle = \langle T^*T\mathbf{x}, \mathbf{x}\rangle \le \lambda_1(T^*T) = ||T||_2^2$. Hence $\rho(T) \le ||T||_2$.

Assume now that $\rho(T) = ||T||_2$. $\rho(T) = 0$ then $||T||_2 = 0 \Rightarrow T = 0$, 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 $\mathbf{U} = \mathrm{span}\,(\mathbf{x})^\perp$ be the orthogonal complement of $\mathrm{span}\,(\mathbf{x})$. Since $T\mathrm{span}\,(\mathbf{x}) = \mathrm{span}\,(\mathbf{x})$ it follows that $T^*\mathbf{U} \subseteq \mathbf{U}$. Similarly, since $T^*\mathrm{span}\,(\mathbf{x}) = \mathrm{span}\,(\mathbf{x})$ $T\mathbf{U} \subseteq \mathbf{U}$. Thus $\mathbf{V} = \mathrm{span}\,(\mathbf{x}) \oplus \mathbf{U}$ and $\mathrm{span}\,(\mathbf{x}), \mathbf{U}$ are invariant subspaces of T and T^*. Hence $\mathrm{span}\,(\mathbf{x}), \mathbf{U}$ 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 $||T_1||_2^2 = \lambda_1(T_1 * T_1) \ge \lambda_1(T^*T) = ||T||_2^2$. 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 4.10.13 *Let* $\mathbf{U}$ *be an* n*-dimensional IPS over* $\mathbb{C}$. *Let* $T : \mathbf{U} \to \mathbf{U}$ *be a linear operator. Denote by* $|\lambda(T)| = (|\lambda_1(T)|, ..., |\lambda_n(T)|)^T$ *the absolute eigenvalues of* T, *(counting with their multiplicities), arranged in* *a decreasing order. Then* $|\lambda(T)| = (\sigma_1(T), ..., \sigma_n(T))^T$ *if and only if* T *is a normal operator.*

Problems

1. Let the assumptions of Corollary 4.10.2 hold.

   (a) Since rank Â $\le$ rank A show that the inequalities (4.10.4) reduce to $\sigma_i(\hat{A}) = \sigma_i(A) = 0$ for $i >$ rank A.

   (b) Since $H(\hat{A})$ is a submatrix of $H(A)$ use the Cauchy interlacing principle to deduce the inequalities (4.10.4) for $i = 1, \ldots,$ rank A. Furthermore, if $p' := m - \#\alpha, q' = n - \#\beta$ then the Cauchy interlacing principle gives the complementary inequalities $\sigma_i(\hat{A}) \ge \sigma_{i+p'+q'}(A)$ for any $i \in \mathbb{N}$.

   (c) Assume that $\sigma_i(\hat{A}) = \sigma_i(A)$ for $i = 1, \ldots, l \le$ rank A. Compare the maximal characterization of the sum of the first k eigenvalues of $H(\hat{A})$ and $H(A)$ given by Theorem 4.4.8 for $k = 1, \ldots, l$ to deduce the last part of Corollary (4.10.2).

2. Prove Corollary 4.10.3 by choosing any orthonormal basis in **U**, an orthonormal basis in **V** whose first dim **W** elements span **W**, and using Problem 1.

3. Combine Theorems 4.7.15 and 4.10.1 to deduce Theorem 4.10.4.

4. (a) Prove Corollary 4.10.5

   (b) Recall the definition of a norm on a vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ 7.4.1. Show that the function f defined in Corollary 4.10.5 is a norm. For $p = \min(m, n)$ and $w_1 = ... = w_p = 1$ this norm is called the $q - Schatten$ norm.

5. Prove Lemma 4.10.6.

6. Under the assumptions of Theorem 4.10.7 show.

   (a)

$$\max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in\mathrm{Fr}(k,\mathbf{U}),\{\mathbf{g}_1,\ldots,\mathbf{g}_k\}\in\mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}} =$$
$$\max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in\mathrm{Fr}(k,\mathbf{U}),\{\mathbf{g}_1,\ldots,\mathbf{g}_k\}\in\mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} |\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}|.$$

(b) For $w_1 \geq \ldots \geq w_k \geq 0$

$$\sum_{i=1}^{k} w_i\sigma_i(T) = \max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in \mathrm{Fr}(k,\mathbf{U}),\{\mathbf{g}_1,\ldots,\mathbf{g}_k\}\in \mathrm{Fr}(k,\mathbf{V})} \sum_{i=1}^{k} w_i \Re\langle T\mathbf{g}_i, \mathbf{f}_i\rangle_{\mathbf{U}}.$$

7. Under the assumptions of Theorem 4.10.7 is it *true* that that for $k > 1$

$$\sum_{i=1}^{k} \sigma_i(T) = \max_{\{\mathbf{f}_1,\ldots,\mathbf{f}_k\}\in \mathrm{Fr}(k,\mathbf{U})} \sum_{i=1}^{k} ||T\mathbf{f}_i||_{\mathbf{V}}.$$

I doubt it.

8. Let $\mathbf{U},\mathbf{V}$ be finite dimensional IPS. Assume that $P,T \in \mathrm{L}(\mathbf{U},\mathbf{V})$. Show that $\Re\,\mathrm{tr}(P^*T) \geq -\sum_{i=1}^{\min(m,n)} \sigma_i(S)\sigma_i(T)$. Equality holds if and only if $S = -P$ and T satisfy the conditions of Theorem 4.10.8.

4.11 Moore-Penrose generalized inverse

Let $A \in \mathbb{C}^{m\times n}$. Then (4.9.12) is called the *reduced* SVD of A. It can be written as

$$A = U_r\Sigma_rV_r^*, \quad r = \mathrm{rank\ A}, \quad \Sigma_\mathrm{r} := \mathrm{diag}(\sigma_1(\mathrm{A}),\ldots,\sigma_\mathrm{r}(\mathrm{A})) \in \mathrm{S_r}(\mathbb{R}),$$
(4.11.1)
$$U_r = [\mathbf{u}_1,\ldots,\mathbf{u}_r] \in \mathbb{C}^{m\times r}, V_r = [\mathbf{v}_1,\ldots,\mathbf{v}_r] \in \mathbb{C}^{n\times r}, U_r^*U_r = V_r^*V_r = I_r,.$$

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

$$A^\dagger := V_r\Sigma_r^{-1}U_r^* \in \mathbb{C}^{n\times m} \tag{4.11.2}$$

is the *Moore-Penrose* generalized inverse of A. If $A \in \mathbb{R}^{m\times n}$ then we assume that $U \in \mathbb{R}^{m\times r}$ and $V \in R^{n\times r}$, i.e. U,V are real values matrices over the real numbers $\mathbb{R}$.

Theorem 4.11.1 *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. *$A^\dagger AA^\dagger = A^\dagger$, $AA^\dagger A = A$, $A^*AA^\dagger = A^\dagger AA^* = A^*$.*

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 $A^\dagger = A^{-1}$.*

To prove the above theorem we need the following proposition.

Proposition 4.11.2 *Let $E \in \mathbb{C}^{l\times m}, G \in \mathbb{C}^{m\times n}$. Then* $\mathrm{rank\ EG \le \min(rank\ E, rank\ G)}$*. If $l = m$ and E is invertible then* $\mathrm{rank\ EG = rank\ G}$*. If $m = n$ and G is invertible then* $\mathrm{rank\ EG = rank\ E}$*.*

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 $\mathrm{span}\,(\mathrm{E}\mathbf{g}_1, \ldots, E\mathbf{g}_n) \subseteq \mathrm{span}\,(\mathbf{e}_1, \ldots, \mathbf{e}_l)$, which implies that $\mathrm{rank\ EG \le rank\ E}$. Note that $(EG)^T = G^TE^T$. Hence $\mathrm{rank\ EG = rank\ (EG)^T \le rank\ G^T = rank\ G}$. Thus $\mathrm{rank\ EG \le \min(rank\ E, rank\ G)}$. Suppose E is invertible. Then $\mathrm{rank\ EG \le rank\ G = rank\ E^{-1}(EG) \le rank\ EG}$. Hence $\mathrm{rank\ EG = rank\ G}$. Similarly $\mathrm{rank\ EG = rank\ E}$ if G is invertible. □

Proof of Theorem 4.11.1.

1. Proposition 4.11.2 yields that $\mathrm{rank\ A^\dagger = rank\ V_r\Sigma_r^{-1}U_r^* \le rank\ \Sigma_r^{-1}U_r^* \le rank\ \Sigma_r^{-1} = r = rank\ A}$. Since $\Sigma_r = V_r^*A^\dagger U_r$ Proposition 4.11.2 yields that $\mathrm{rank\ A^\dagger \ge rank\ \Sigma_r^{-1} = r}$. Hence $\mathrm{rank\ A = rank\ A^\dagger}$.

2. $AA^\dagger = (U_r\Sigma_rV_r^*)(V_r\Sigma_r^{-1}U_r^*) = U_r\Sigma_r\Sigma_r^{-1}U_r^* = U_rU_r^*$. Hence
$$AA^\dagger A = (U_rU_r^*)(U_r\Sigma_rV_r^*) = U_r\Sigma V_r^* = A.$$
Hence $A^*AA^\dagger = (V_r\Sigma_rU_r^*)(U_rU_r^*) = A^*$. Similarly $A^\dagger A = V_rV_r^*$ and $A^\dagger AA^\dagger = A^\dagger, A^\dagger AA^* = A^*$.

3. Since $AA^\dagger = U_r U_r^*$ we deduce that $(AA^\dagger)^* = (U_r U_r^*)^* = (U_r^*)^* U_r^* = AA^\dagger$, i.e. $AA^\dagger$ is Hermitian. Next $(AA^\dagger)^2 = (U_r U_r^*)^2 = (U_r U_r^*)(U_r U_r^*) = (U_r U_r^*) = AA^\dagger$, i.e. $AA^\dagger$ is idempotent. Hence $AA^\dagger$ is nonnegative definite. As $AA^\dagger = U_r I_r U_r^*$, the arguments of part *1* yield that rank $\mathrm{AA}^\dagger = \mathrm{r}$. Similar arguments apply to $A^\dagger A = V_r V_r^*$.

4. Since $A^* A A^\dagger = A^*$ it follows that $A^* A(A^\dagger \mathbf{b}) = A^* \mathbf{b}$, i.e. $\mathbf{y} = A^\dagger \mathbf{b}$ is a least square solution. It is left to show that if $A^* A \mathbf{x} = A^* \mathbf{b}$ then $||\mathbf{x}|| \ge ||A^\dagger \mathbf{b}||$ and equality holds if and only if $\mathbf{x} = A^\dagger \mathbf{b}$.

   We now consider the system $A^* A \mathbf{x} = A^* \mathbf{b}$. To analyze this system we use the full form of SVD given in (4.9.7). It is equivalent to

   $(V\Sigma^T U^*)(U\Sigma V^*)\mathbf{x} = V\Sigma^T U^* \mathbf{b}$. 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 $\mathbf{z}^* \mathbf{z} = \mathbf{x}^* V V \mathbf{x} = \mathbf{x}^* \mathbf{x}$, 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 $\sigma_i(A)^2 z_i = \sigma_i(A) c_i$ 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 $\mathrm{A^*A}$ = rank A = n it follows that $A^* A$ is an invertible matrix. Hence the least square solution is unique and is given by $\mathbf{x} = (A^* A)^{-1} A^* \mathbf{b}$. Thus for each $\mathbf{b}$ one has $(A^* A)^{-1} A^* \mathbf{b} = A^\dagger \mathbf{b}$, hence $A^\dagger = (A^* A)^{-1} A^*$.

   If A is an $n \times n$ 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 $\mathbb{V} \subseteq \mathbb{R}^n$ be the subspace spanned by the columns of P. Show that for any $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{V}$ $||\mathbf{a} - \mathbf{b}|| \ge ||\mathbf{a} - P\mathbf{a}||$ 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 (4.9.7), 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!

4.12 Approximation by low rank matrices

We now restate Theorem 4.10.8 in matrix terms. That is we view $A, B \in \mathbb{C}^{m\times n}$ as linear operators $A, B : \mathbb{C}^n \to \mathbb{C}^m$, where $\mathbb{C}^m, \mathbb{C}^n$ are IPS equipped with the standard inner product.

Theorem 4.12.1 *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*

$$-\sum_{i=1}^{m} \sigma_i(A)\sigma_i(B) \leq \Re \operatorname{tr} AB^* \leq \sum_{i=1}^{m} \sigma_i(A)\sigma_i(B). \tag{4.12.1}$$

Equality in the right-hand side holds if and only if $\mathbb{C}^n, \mathbb{C}^m$ have two orthonormal bases $[\mathbf{c}_1, \ldots, \mathbf{c}_n], [\mathbf{d}_1, \ldots, \mathbf{d}_m]$ such that (4.10.10) is satisfied for $T = A$ and $S = B$. Equality for the left-hand side holds if and only if $\mathbb{C}^n, \mathbb{C}^m$ have two orthonormal bases $[\mathbf{c}_1, \ldots, \mathbf{c}_n], [\mathbf{d}_1, \ldots, \mathbf{d}_m]$ such that (4.10.10) is satisfied for $T = A$ and $S = -B$.

Theorem 4.10.9 yields:

Corollary 4.12.2 *For $A \in \mathbb{C}^{m\times n}$ Let A_k be defined as in Definition 4.9.5. Then $\min_{B\in\mathcal{R}_{m,n,k}(\mathbb{F})} ||A - B||_F^2 = ||A - A_k||^2 = \sum_{i=k+1}^{m} \sigma_i(A)^2$. A_k is the unique solution to this minimal problem if and only if $1 \leq k <$ rank A and $\sigma_k(A) \geq \sigma_{k+1}(A)$.*

We now give a generalization of Corollary 4.10.9. 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 (4.9.7). Let $U_A = [\mathbf{u}_1 \ \mathbf{u}_2 \ \ldots \mathbf{u}_m], V_A = [\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \mathbf{v}_n]$ be the representations of U, V in terms of their m, n columns respectively. Then

$$P_{A,\mathrm{left}} := \sum_{i=1}^{\mathrm{rank\ A}} \mathbf{u}_i \mathbf{u}_i^* \in \mathbb{C}^{m\times m}, \quad P_{A,\mathrm{right}} := \sum_{i=1}^{\mathrm{rank\ A}} \mathbf{v}_i \mathbf{v}_i^* \in \mathbb{C}^{n\times n}, \tag{4.12.2}$$

are the orthogonal projections on the range of A and A^*, respectively.

Theorem 4.12.3 *Let $A \in \mathbb{C}^{m\times n}, C \in \mathbb{C}^{m\times p}, R \in \mathbb{C}^{q\times n}$ be given. Then $X = C^\dagger(P_{C,\mathrm{left}}AP_{R,\mathrm{right}})_k R^\dagger$ is a solution to the minimal problem*

$$\min_{X\in\mathcal{R}(p,q,k)} ||A - CXR||_F, \tag{4.12.3}$$

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

Proof. Assume that $C = U_C\Sigma_C V_C^*, R = U_R\Sigma_R V_R^*$ are the SVD decomposition of C, R, respectively. Recall that the Frobenius norm is invariant under the multiplication from the left and the right by the corresponding unitary matrices. Hence $||A - BXC||_F = ||\tilde{A} - \Sigma_C\tilde{X}\Sigma_R||$, where $\tilde{A} := U_C^* A V_R, \tilde{X} := V_C^* X U_R$. Clearly, X and $\tilde{X}$ have the same rank and the same Frobenius norm. Thus it is enough to consider the minimal problem $\min_{\tilde{X}\in\mathcal{R}(p,q,k)} ||\tilde{A} - \Sigma_C\tilde{X}\Sigma_R||_F$. Let $s = \mathrm{rank\ C}, \mathrm{t = rank\ R}$. Clearly if C or R is a zero matrix, then $X = 0_{p\times q}$ is the solution to the minimal problem (4.12.3). In this case either $P_{C,\mathrm{left}}$ or $P_{R,\mathrm{right}}$ 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 $C_1 := \mathrm{diag}(\sigma_1(C), \ldots, \sigma_s(C)) \in \mathbb{C}^{s\times s}, R_1 := \mathrm{diag}(\sigma_1(R), \ldots, \sigma_t(R)) \in \mathbb{C}^{t\times t}$. 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 $Z := \Sigma_C\tilde{X}\Sigma_R = [Z_{ij}]_{i,j=1}^2$, where $Z_{11} = C_1X_{11}R_1$ 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 $X_{11} = C_1^{-1}(A_{11})_k R_1^{-1}$ and $X_{ij} = 0$ for all $(i,j) \ne (1,1)$ is a solution $\min_{\tilde{X}\in\mathcal{R}(p,q,k)} ||\tilde{A} - \Sigma_C\tilde{X}\Sigma_R||_F$ 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 $k \ge \mathrm{rank\ A_{11}}$ or $1 \le k < \mathrm{rank\ A_{11}}$ and $\sigma_k(A_{11}) > \sigma_{k+1}(A_{11})$. A straightforward calculation shows that $\hat{X} = \Sigma_C^\dagger(P_{\Sigma_C,\mathrm{left}}\tilde{A}P_{\Sigma_R,\mathrm{right}})_k\Sigma_R^\dagger$. This shows that $X = C^\dagger(P_{C,\mathrm{left}}AP_{R,\mathrm{right}})_k R^\dagger$ is a solution of (4.12.3) with the minimal Frobenius norm. This solution is unique if and only if either $k \ge \mathrm{rank}\ \mathrm{P_{C,left}AP_{R,right}}$ or $1 \le k < \mathrm{rank}\ \mathrm{P_{C,left}AP_{R,right}}$ and $\sigma_k(P_{C,\mathrm{left}}AP_{R,\mathrm{right}}) > \sigma_{k+1}(P_{C,\mathrm{left}}AP_{R,\mathrm{right}})$. □

Corollary 4.12.4 *Let the assumptions of Theorem 4.12.3 hold. Then* $X = C^\dagger A R^\dagger$ *is the unique solution to the minimal problem* $\min_{X\in\mathbb{C}^{p\times q}} ||A-CXR||_F$ *with the minimal Frobenius norm.*

Theorem 4.12.5 *Let* $\mathbf{a}_1,\ldots,\mathbf{a}_n \in \mathbb{C}^m$ *and* $k \in [1,m-1]\cap\mathbb{N}$ *be given. Let* $A = [\mathbf{a}_1\ldots\mathbf{a}_n] \in \mathbb{C}^{m\times n}$*. Denote by* $L_k \in \mathrm{Gr}(k,\mathbb{C}^m)$ *a k-dimensional subspace spanned by the first k left singular vectors of A. Then*

$$\min_{L\in\mathrm{Gr}(k,\mathbb{C}^m)} \sum_{i=1}^{n} \min_{\mathbf{b}_i\in L} ||\mathbf{a}_i-\mathbf{b}_i||_2^2 = \sum_{i=1}^{n} \min_{\mathbf{b}_i\in L_k} ||\mathbf{a}_i-\mathbf{b}_i||_2^2. \tag{4.12.4}$$

Proof. Let $L \in \mathrm{Gr}(k,\mathbb{C}^m)$ and $\mathbf{b}_1,\ldots,\mathbf{b}_n \in L$. Then $B := [\mathbf{b}_1\ldots\mathbf{b}_n] \in \mathcal{R}(m,n,k)$. Vice versa, given $B \in \mathcal{R}(m,n,k)$ then the column space of B is contained in some $L \in \mathrm{Gr}(k,\mathbb{C}^m)$. Hence $\sum_{i=1}^n ||\mathbf{a}_i-\mathbf{b}_i||_2^2 = ||A-B||_2^2$. Corollary 4.12.2 implies that the minimum stated in the left-hand side of (4.12.4) is achieved by the n columns of A_k. Clearly, the column space of A is equal to L_k. (Note that L_k is not unique. See Problem 3.) □

Problems

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 $|\alpha_1| \ge \ldots \ge |\alpha_n| \ge 0$.

   (a) Find the SVD of A.

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

2. Let k,m,n be a positive integers such that $k \le \min(m,n)$. 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 $\mathbb{R}^{m\times n}$.

3. Show that the minimal subspace for the problem (4.12.4) is unique if and only if $\sigma_k(A) > \sigma_{k+1}(A)$.

4. Prove Corollary 4.12.4.

4.13 *CUR*-approximations

Let $A = (a_{ij})_{i,j=1}^{m,n} \in \mathbb{C}^{m\times n}$, where m,n are big, e.g. $m,n \ge 10^6$. Then the low rank approximation of A given by its SVD has prohibitively high computational complexity and storage requirements. In this section we discuss a low rank approximation of A given by of the form CUR, where $C \in \mathbb{R}^{m\times p}, R \in \mathbb{R}^{q\times n}$ are obtained from A by reading p, q columns and rows of A, respectively. If one chooses U as the best least squares approximation given by Corollary 4.12.4 then $U = C^\dagger A R^\dagger$. Again, for very large m, n this U has too high computational complexity. In this section we give different ways to compute U of a relatively low computational complexity.

Let

$$I = \{1 \le \alpha_1 < \ldots < \alpha_q \le m\} \subset \langle m\rangle, \quad J = \{1 < \beta_1 < \ldots < \beta_p \le n\} \subset \langle n\rangle$$

be two nonempty sets of cardinality q, p respectively. Using the indices in I, J, we consider the submatrices

$$\begin{aligned} &A_{IJ} = (a_{\alpha_k\beta_l})_{k,l=1}^{q,p} \in \mathbb{C}^{q\times p},\\ &R = A_{I\langle n\rangle} = (a_{\alpha_k j})_{k,j=1}^{q,n} \in \mathbb{C}^{q\times n},\\ &C = A_{\langle m\rangle J} = (a_{i\beta_l})_{i,l=1}^{m,p} \in \mathbb{C}^{m\times p}.\end{aligned} \tag{4.13.1}$$

Thus, $C = A_{\langle m\rangle J}$ and $R = A_{I\langle n\rangle}$ are composed of the columns in J and the rows I of A, respectively. The read entries of A are in the index set

$$\mathcal{S} := \langle m\rangle \times \langle n\rangle \backslash ((\langle m\rangle\backslash I) \times (\langle n\rangle\backslash J)), \quad \#\mathcal{S} = mp + qn - pq. \tag{4.13.2}$$

We look for a matrix $F = CUR \in \mathbb{C}^{m\times n}$, with $U \in \mathbb{C}^{p\times q}$ still to be determined. We determine U_{opt} as a solution to the least square problem of minimizing $\sum_{(i,j)\in\mathcal{S}} |a_{ij} - (CUR)_{ij}|^2$, i.e.,

$$U_{\mathrm{opt}} = \arg\min_{U\in\mathbb{C}^{p\times q}} \sum_{(i,j)\in\mathcal{S}} |a_{ij} - (CUR)_{ij}|^2. \tag{4.13.3}$$

It is straightforward to see that the above least squares is the least squares solution of the following overdetermined system

$$(4.13.4) T\hat{U} = \hat{A},\ T = (t_{(i,j)(k,l)}) \in \mathbb{C}^{(mp+qn-pq)\times pq},\ t_{(i,j)(k,l)} = a_{ik}a_{lj},$$
$$\hat{U} = (u_{(k,l)}) \in \mathbb{C}^{pq},\ \hat{A} = (a_{(ij)}) \in \mathbb{C}^{mp+qn-pq},\ (i,j) \in \mathcal{S},\ (k,l) \in \langle p\rangle \times \langle q\rangle.$$

Here $\hat{U}, \hat{A}$ is viewed as a vector whose coordinates are the entries of U and the entries of A which are either in C or R. Note that T is a corresponding submatrix of $A \otimes A$.

Theorem 4.13.1 *Let $A \in \mathbb{C}^{m\times n}$, and let $I \subset \langle m\rangle$, $J \subset \langle n\rangle$ have cardinality q and p, respectively. Let $C = A_{\langle m\rangle J} \in \mathbb{C}^{m\times p}$, and $R = A_{I\langle n\rangle} \in$*

$\mathbb{C}^{p\times n}$ be as in (4.13.1) and suppose that A_{IJ} is invertible. Then the overdetermined system (4.13.4) has a unique solution $U = A_{IJ}^{-1}$, i.e., the rows in I and the columns in J of the matrix $CA_{IJ}^{-1}R$ are equal to the corresponding rows and columns of A, respectively.

Proof. For any $I \subset \langle m\rangle$, $J \subset \langle n\rangle$, with $\#I = q$, $\#J = p$, and $U \in \mathbb{C}^{m\times n}$ we have the identity

$$(A_{\langle m\rangle J}UA_{I\langle n\rangle})_{IJ} = A_{IJ}UA_{IJ}. \tag{4.13.5}$$

Hence the part of the system (4.13.4) corresponding to $(CUR)_{IJ} = A_{IJ}$ reduces to the equation

$$A_{IJ}UA_{IJ} = A_{IJ} \tag{4.13.6}$$

If A_{IJ} is a square matrix and invertible, then the unique solution to this matrix equation is $U = A_{IJ}^{-1}$. Furthermore

$$(A_{\langle m\rangle J}A_{IJ}^{-1}A_{I\langle n\rangle})_{I\langle n\rangle} = A_{IJ}A_{IJ}^{-1}A_{I\langle n\rangle} = A_{I\langle n\rangle},$$
$$(A_{\langle m\rangle J}A_{IJ}^{-1}A_{I\langle n\rangle})_{\langle m\rangle J} = A_{\langle m\rangle J}A_{IJ}^{-1}A_{IJ} = A_{\langle m\rangle J}.$$

□

This results extends to the general nonsquare case.

Theorem 4.13.2 *Let $A \in \mathbb{C}^{m\times n}$, and let $I \subset \langle m\rangle$, $J \subset \langle n\rangle$ have cardinality q and p, respectively. Let $C = A_{\langle m\rangle J} \in \mathbb{C}^{m\times p}$, and $R = A_{I\langle n\rangle} \in \mathbb{C}^{p\times n}$ be as in (4.13.1). Then $U = A_{IJ}^{\dagger}$ is the minimal solution (in Frobenius norm) of (4.13.3).*

Proof. Consider the SVD decomposition of A_{IJ}

$$A_{IJ} = W\Sigma V^*,\ W \in \mathbb{C}^{q\times q},\ V \in \mathbb{C}^{p\times p},\ \Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_r,0,\ldots,0) \in \mathbb{R}_+^{q\times p},$$

where W, V are unitary matrices and $\sigma_1,\ldots,\sigma_r$ are the positive singular values of A_{IJ}. In view of Theorem 4.13.1 it is enough to assume that $\max(p,q) > r$. W.l.o.g. we may assume that $I = \langle q\rangle$, $J = \langle p\rangle$. Let

$$W_1 = \begin{pmatrix} W & 0_{q\times(m-q)} \\ 0_{(m-q)\times q} & I_{m-q} \end{pmatrix} \in \mathbb{C}^{m\times m},$$
$$V_1 = \begin{pmatrix} V & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{n-p} \end{pmatrix} \in \mathbb{C}^{n\times n}.$$

Replace A by $A_1 = W_1AV_1^*$. It is easy to see that it is enough to prove the theorem for A_1. For simplicity of the notation we assume that $A_1 = A$.

That is, we assume that $A_{IJ} = \Sigma_r \oplus 0_{(q-r)\times(p-r)}$, where $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ and $r = \mathrm{rank\ A_{IJ}}$. For $U \in \mathbb{C}^{p\times q}$ denote by $U_r \in \mathbb{R}^{p\times q}$ the matrix obtained from U by replacing the last $p-r$ rows and $q-r$ columns by rows and columns of zeroes, respectively. Note that then $CUR = CU_rR$ and $\|U_r\|_F \le \|U\|_F$, and equality holds if and only if $U = U_r$. Hence, the minimal Frobenius norm least squares solution U of is given by $U = U_r$. Using the fact that the rows $r+1, \ldots, q$ and columns $r+1, \ldots, p$ of CUR are zero it follows that the minimum in (4.13.3) is reduced to the minimum on $\mathcal{S}' = \langle m\rangle \times \langle r\rangle \cup \langle r\rangle \times \langle n\rangle$. Then, by Theorem 4.13.1 the solution to the minimal Frobenius norm least square problem is given by $\Sigma^\dagger$. □

For a matrix A define the *entrywise maximal norm*

$$\|A\|_{\infty,e} := \max_{i\in\langle m\rangle, j\in\langle n\rangle} |a_{ij}|, \quad A = (a_{ij}) \in \mathbb{C}^{m\times n}. \tag{4.13.7}$$

Theorem 4.13.3 *Let $A \in \mathbb{C}^{m\times n}, p \in [1, \mathrm{rank\ A}] \cap \mathbb{N}$. Define*

$$\mu_p := \max_{I\subset\langle m\rangle, J\subset\langle n\rangle, \#I=\#J=p} |\det A_{IJ}| > 0. \tag{4.13.8}$$

Suppose that

$$|\det A_{IJ}| \ge \delta\mu_p, \delta \in (0,1], I \subset \langle m\rangle, J \subset \langle n\rangle, \#I = \#J = p. \tag{4.13.9}$$

Then for C, R defined by (4.13.1) we have

$$||A - CA_{IJ}^{-1}R||_{\infty,e} \le \frac{p+1}{\delta}\sigma_{p+1}(A). \tag{4.13.10}$$

Proof. We now estimate $|a_{ij} - (CA_{IJ}^{-1}R)_{ij}|$ from above. In the case $p = \mathrm{rank\ A}$, i.e. $\sigma_{p+1}(A) = 0$, we deduce from Problem 1 that $a_{ij} - (CA_{IJ}^{-1}R)_{ij} = 0$. Assume $\sigma_{p+1}(A) > 0$. By Theorem 4.13.1 $a_{ij} - (CA_{IJ}^{-1}R)_{ij} = 0$ if either $i \in I$ or $j \in J$. It is left to consider the case $i \in \langle m\rangle\backslash I, j \in \langle n\rangle\backslash J$. Let $K = I \cup \{i\}, L = J \cup \{j\}$. Let $B = A_{KL}$. If $\mathrm{rank\ B = p}$ then Problem 1 yields that $B = B_{KJ}A_{IJ}^{-1}B_{IK}$. Hence $a_{ij} - (CA_{IJ}^{-1}R)_{ij} = 0$. Assume that $\det B \ne 0$. We claim that

$$a_{ij} - (CA_{IJ}^{-1}R)_{ij} = \pm\frac{\det B}{\det A_{IJ}}. \tag{4.13.11}$$

It is enough to consider the case where $I = J = \langle p\rangle, i = j = p+1K = L = \langle p+1\rangle$. In view of Theorem 4.13.1 $B - B_{KJ}A_{IJ}^{-1}B_{JL} = \mathrm{diag}(0, \ldots, 0, t)$, where t is equal to the left-hand side of (4.13.11). Multiply this matrix equality from the left by $B^{-1} = (b_{st,-1})_{s,t=1}^{p+1}$. Note that the last row of $B^{-1}B_{KI}$ is zero. Hence we deduce that $b_{(p+1)(p+1),-1}t = 1$, i.e $t = b^{-1}_{(p+1)(p+1),-1}$. Use the identity $B^{-1} = (\det B)^{-1}$adj B to deduce the equality (4.13.11).

We now estimate $\sigma_1(B^{-1})$ from above. Note that each entry of $B^{-1} = (\det B)^{-1}$adj B is bounded above by $|\det B|^{-1}\mu_p$. Hence $\sigma_1(B^{-1}) \le \frac{(p+1)\mu_p}{|\det B|}$. Recall that $\sigma_1(B^{-1} = \sigma_{p+1}(B)^{-1}$. Thus

$$\frac{|\det B|}{\mu_p} \le (p+1)\sigma_{p+1}(B) \Rightarrow \frac{|\det B|}{|\det A_{IJ}|} \le \frac{(p+1)\sigma_{p+1}(B)}{\delta}.$$

Since B is a submatrix of A we deduce $\sigma_{p+1}(B) \le \sigma_{p+1}(A)$. □

Problems

1. Let $A \in \mathbb{C}^{m\times n}$, rank A = r. Assume that $I \subset \langle m\rangle, J \subset \langle n\rangle, \#I = \#J = r$. Assume that det $A_{IJ} \ne 0$. Show that $A = CA_{IJ}^{-1}R$.

4.14 Some special maximal spectral problems

Let S $\subset$ **V**. Then the convex hull of S, denoted by conv S, is the minimal convex set containing S. Thus $\Pi_n := \mathrm{conv}\{\mathbf{e}_1, ..., \mathbf{e}_n\} \subset \mathbb{R}^n$, where $\mathbf{e}_1, ..., \mathbf{e}_n$ is the standard basis in $\mathbb{R}^n$, is the set of *probability* vectors in $\mathbb{R}^n$.

Assume that $C \subset \mathbf{V}$ is convex. A point $\mathbf{e} \in C$ is called an extremal point if for any $\mathbf{x}, \mathbf{y} \in C$ such that $\mathbf{e} \in [\mathbf{x}, \mathbf{y}]$ the equality $\mathbf{x} = \mathbf{y} = \mathbf{e}$ holds. For a convex set C denote by ext C the set of the extremal points of C. It is known that ext conv S $\subset$ S [Roc70] or see Problem 1.

Definition 4.14.1 *Let* S $\subset$ **V**. *For each* $j \in \mathbb{N}$ *let*

$$\mathrm{conv}_{j-1}\mathrm{S} = \{\mathbf{z} \in \mathbf{V} :\ \mathbf{z} = \sum_{i=1}^{j} p_i\mathbf{x}_i, \text{ for all } \mathbf{p} = (p_1, ..., p_j)^\top \in \mathit{\Pi}_j,\ \mathbf{x}_1, ..., \mathbf{x}_j \in \mathrm{S}\}.$$

Let C *be a convex set in* **V**. *Assume that* ext $C \ne \emptyset$. *Then* $\mathrm{F_{j-1}(C)} := \mathrm{conv_{j-1}(ext\ C)}$ *is called* $j-1$ *dimensional face of* C *for any* $j \in \mathbb{N}$.

Suppose that S is a finite set of cardinality $N \in \mathbb{N}$. Then conv S = $\mathrm{conv_{N-1}S}$, see Problem 1, and conv S is called a *finitely generated convex set.* Note that $F_0(\mathrm{S}) = \mathrm{S}$. The following result is well known [Roc70]. (See Problem 1 for finitely generated convex sets.)

Theorem 4.14.2 *Let* $\mathbf{V}$ *be a vector space over real of finite dimension. Let* $C \subset \mathbf{V}$ *be a nonempty compact convex set. Then* $\operatorname{conv}\operatorname{ext} C = C$. *Let* $d := \dim C$. *Then* $\mathrm{F_d(ext\ C) = C}$. *More general, for any* $\mathrm{S} \subset \mathbf{V}$ *let* $d = \dim\ \operatorname{conv}\mathrm{S}$. *Then* $\mathrm{F_d(S)} = \operatorname{conv}\mathrm{S}$.

Throughout this book we assume that $\mathbf{V}$ is finite dimensional, unless stated otherwise. In many case we shall identify $\mathbf{V}$ with $\mathbb{R}^d$. Assume that the $C \subset \mathbf{V}$ is a nonempty compact convex set of dimension d. Then the following facts are known. If $d = 2$ then $\operatorname{ext} C$ is a closed set. For $d \geq 3$ there exist C such that $\operatorname{ext} C$ is not closed.

The following result is well known (see Problem 3):

Proposition 4.14.3 *Let* $\mathrm{S} \subset \mathbf{V}$ *and assume that* $f : \operatorname{conv}\mathrm{S} \to \bar{\mathbb{R}}$ *is a convex function. Then*

$$\sup_{\mathbf{x} \in \operatorname{conv}\mathrm{S}} f(\mathbf{x}) = \sup_{\mathbf{y} \in \mathrm{S}} f(\mathbf{y}).$$

If in addition S *is compact and* f *is continuous on* $\operatorname{conv}\mathrm{S}$ *then one can replace* sup *by* max.

See Problem 4 for a generalization of this proposition.

Corollary 4.14.4 *Let* $\mathbf{V}$ *be a finite dimensional IPS over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$. *Let* $\mathrm{S} \subset \mathbf{S}(\mathbf{V})$. *Let* $f : \operatorname{conv}\mathrm{S} \to \bar{\mathbb{R}}$ *be a convex function. Then*

$$\sup_{A \in \operatorname{conv}\mathrm{S}} f(A) = \sup_{B \in \mathrm{S}} f(B).$$

The aim of this section to give a generalization of this result to certain spectral nonconvex functions f.

Definition 4.14.5 *For* $\mathbf{x} = (x_1, ..., x_n)^\top, \mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$ *let* $\mathbf{x} \leq \mathbf{y} \iff x_i \leq y_i,\ i = 1, ..., n$. *Let* $D \subset \mathbb{R}^n$ *and* $f : D \to \bar{\mathbb{R}}$. f *is called a nondecreasing function on* D *if for any* $\mathbf{x}, \mathbf{y} \in D$ *one has the implication* $\mathbf{x} \leq \mathbf{y} \Rightarrow f(\mathbf{x}) \leq f(\mathbf{y})$.

Theorem 4.14.6 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS over* $\mathbb{R}$. *Let* $p \in [1, n] \cap \mathbb{N}$ *and* $D \subset \mathbb{R}^n_{\searrow}$ *be a convex Schur set. Let* D_p *be the projection of* D *on the first* p *coordinates. Let* $h : D_p \to \mathbb{R}$ *and assume that* $f : \lambda^{-1}(D) \to \mathbb{R}$ *be the spectral function given by* $A \mapsto h(\lambda_{(p)}(A))$, *where* $\lambda_{(p)}(A) = (\lambda_1(A), ..., \lambda_p(A))^\top$. *Let* $\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$. *Assume that* h *is nondecreasing on* D_p. *Then*

$$\sup_{A \in \operatorname{conv}\mathrm{S}} f(A) = \sup_{B \in \operatorname{conv}_{\binom{p+1}{2}-1}\mathrm{S}} f(B), \tag{4.14.1}$$

and this result is sharp.

Proof. Since dim $\mathrm{conv}\,\mathrm{S} \le \dim \mathbf{S}(\mathbf{V}) = \binom{n+1}{2}$ Theorem 4.14.2 implies that it is enough to prove the theorem in the case $\mathrm{S} = \mathrm{T} := \{\mathrm{A}_1, ..., \mathrm{A}_\mathrm{N}\}$, where $N \le \binom{n+1}{2} + 1$. Observe next that since D is a convex Schur set and $\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$ it follows that $\mathrm{conv}\,\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$ (Problem 6).

Let $A \in \mathrm{conv}\,\mathrm{T}$. Assume that $\mathbf{x}_1, ..., \mathbf{x}_p$ are p-orthonormal eigenvectors of A corresponding to the eigenvalues $\lambda_1(A), ..., \lambda_p(A)$. For any $B \in \mathbf{S}(\mathbf{V})$ let $B(\mathbf{x}_1, ..., \mathbf{x}_p) := (\langle B\mathbf{x}_i, \mathbf{x}_j\rangle)_{i,j=1}^p \in \mathbf{S}_p(\mathbb{R})$. We view $\mathbf{S}_p(\mathbb{R})$ as a real vector space of dimension $\binom{p+1}{2}$. Let $\mathrm{T}' := \{\mathrm{A}_1(\mathbf{x}_1, ..., \mathbf{x}_p), ..., A_N(\mathbf{x}_1, ..., \mathbf{x}_p)\} \subset \mathbf{S}_p(\mathbb{R})$. It i straightforward to show that for any $B \in \mathrm{conv}\,\mathrm{T}$ one has $B(\mathbf{x}_1, ..., \mathbf{x}_p) \in \mathrm{conv}\,\mathrm{T}'$. Let $\tilde{\mathrm{T}}$ the restriction of $\mathrm{conv}\,\mathrm{T}'$ to the line in $\mathbf{S}_p(\mathbb{R})$

$$\{X = (x_{ij}) \in \mathbf{S}_p(\mathbb{R}) : \quad x_{ij} = \lambda_i(A)\delta_{ij}, \text{ for } i + j > 2\}.$$

Clearly $A(\mathbf{x}_1, ..., \mathbf{x}_p) \in \tilde{\mathrm{T}}$. Hence $\tilde{\mathrm{T}} = [C(\mathbf{x}_1, ..., \mathbf{x}_p), D(\mathbf{x}_1, ..., \mathbf{x}_p)]$ for some $C, D \in \mathrm{conv}\,\mathrm{T}$. It is straightforward to show that $C, D \in \mathrm{conv}_{\binom{p+1}{2}-1}\,\mathrm{T}$. See Problem 1. Hence $\max_{X \in \tilde{\mathrm{T}}} x_{11} = \max(\langle C\mathbf{x}_1, \mathbf{x}_1\rangle, \langle D\mathbf{x}_1, \mathbf{x}_1\rangle)$. Without loss of generality we may assume that the above maximum is achieved for the matrix C. Hence $C(\mathbf{x}_1, ..., \mathbf{x}_p)$ is a diagonal matrix such that $\lambda_1(C(\mathbf{x}_1, ..., \mathbf{x}_p) \ge \lambda_1(A)$ and $\lambda_i(C(\mathbf{x}_1, ..., \mathbf{x}_p))) = \lambda_i(A)$ for $i = 2, ..., p$. Let $\mathbf{U} = \mathrm{span}\,(\mathbf{x}_1, ..., \mathbf{x}_p)$. Since $\mathbf{x}_1, ..., \mathbf{x}_p$ are orthonormal it follows that $\lambda_i(Q(C, \mathbf{U})) = \lambda_i(C(\mathbf{x}_1, ..., \mathbf{x}_p))$ for $i = 1, ..., p$. Corollary 4.4.7 yields that $\lambda_{(p)}(C) \ge \lambda_{(p)}(A)$. Since h is increasing on D we get $h(\lambda_{(p)}(C)) \ge h(\lambda_{(p)}(A))$. See Problem 7 which shows that (4.14.4) is sharp. □

Theorem 4.14.7 *Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{C}$. Let $p \in [1, n] \cap \mathbb{N}$ and $D \subset \mathbb{R}^n_{\searrow}$ be a convex Schur set. Let D_p be the projection of D on the first p coordinates. Let $h : D_p \to \mathbb{R}$ and assume that $f : \lambda^{-1}(D) \to \mathbb{R}$ be the spectral function given by $A \mapsto h(\lambda_{(p)}(A))$, where $\lambda_{(p)}(A) = (\lambda_1(A), ..., \lambda_p(A))^\top$. Let $\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$. Assume that h is nondecreasing on D_p. Then*

$$\sup_{A \in \mathrm{conv}\,\mathrm{S}} f(A) = \sup_{B \in \mathrm{conv}_{p^2-1}\mathrm{S}} f(B), \tag{4.14.2}$$

and this result is sharp.

See Problems 8 for the proof of the theorem.

It is possible to improve Theorems 4.14.6 and 4.14.7 in special interesting cases for $p > 1$.

Definition 4.14.8 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$ *and* $p \in [1, n] \cap \mathbb{Z}$*. Let* $A \in \mathbf{S}(\mathbf{V})$*. Then* p*-upper multiplicity of* $\lambda_p(A)$*, denoted by* $\mathrm{upmul}(A, p)$*, is a natural number in* $[1, p]$ *such that*

$$\lambda_{p-\mathrm{upmul}(A,p)}(A) > \lambda_{p-\mathrm{upmul}(A,p)+1}(A) = ... = \lambda_p(A), \quad \text{where } \lambda_0(A) = \infty.$$

For any $\mathrm{C} \subset \mathbf{S}(\mathbf{V})$ *let* $\mathrm{upmul(C, p)} := \max_{\mathrm{A} \in \mathrm{C}} \mathrm{upmul(A, p)}$.

See Problem 14 for sets satisfying $\mathrm{upmul(C, p) \leq k}$ for any $k \in \mathbb{N}$.

Theorem 4.14.9 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS over* $\mathbb{R}$*. Let* $p \in [1, n] \cap \mathbb{N}$ *and denote* $\mu(p) := \mathrm{upmul}(\mathrm{conv}\, S, p)$*. Then*

$$\sup_{A \in \mathrm{conv}\,\mathrm{S}} \lambda_p(A) = \sup_{B \in \mathrm{conv}_{\frac{\mu(p)(2p-\mu(p)+1)}{2}-1}\mathrm{S}} \lambda_p(B). \tag{4.14.3}$$

Proof. For $\mu(p) = p$ (4.14.4) follows from Theorem 4.14.6. Thus, it is enough to consider the case $p > 1$ and $\mu(p) < p$. As in the proof of Theorem 4.14.6 we may assume that $\mathrm{S} = \{\mathrm{A}_1, ..., \mathrm{A}_\mathrm{N}\}$ where $N \leq \binom{n+1}{2} + 1$. Let $\mathcal{M} := \{B \in \mathrm{conv}\,\mathrm{S} : \ \lambda_\mathrm{p}(\mathrm{B}) = \max_{\mathrm{A} \in \mathrm{conv}\,\mathrm{S}} \lambda_\mathrm{p}(\mathrm{A})\}$. Since $\lambda_p(A)$ is a continuous function on $\mathbf{S}(\mathbf{V})$ and $\mathrm{conv}\,\mathrm{S}$ is a compact set it follows that $\mathcal{M}$ is a nonempty compact set of $\mathrm{conv}\,\mathrm{S}$. Let $\nu := \frac{\mu(p)(2p-\mu(p)+1)}{2} - 1$. Assume to the contrary that the theorem does not hold, i.e. $\mathcal{M} \cap \mathrm{conv}_\nu \mathrm{S} = \emptyset$. Let $\mathcal{M}' := \{\mathbf{p} = (p_1, ..., p_N)^\top \in \mathcal{P}_N : \ \sum_{i=1}^n A_i \in \mathcal{M}\}$. Then $\mathcal{M}'$ is a nonempty compact set of $\mathcal{P}_N$ and any $\mathbf{p} \in \mathcal{M}'$ at least $\nu + 2$ positive coordinates. Introduce the following complete order on $\mathcal{P}_N$. Let $\mathbf{x} = (x_1, ..., x_N)^\top, \mathbf{y} = (y_1, ..., y_N)^\top \in \mathbb{R}^N$. As in Definition 4.6.1 let $\bar{\mathbf{x}} = (\bar{x}_1, ..., \bar{x}_N)^\top, \bar{\mathbf{y}} = (\bar{y}_1, ..., \bar{y}_N)^\top \in \mathbb{R}^N$ be the rearrangements of the coordinates of the vectors $\mathbf{x}$ and $\mathbf{y}$ in the nonincreasing order. Then $\mathbf{x} \ll \mathbf{y}$ if either $\bar{\mathbf{x}} = \bar{\mathbf{y}}$ or $\bar{\mathbf{x}}_i = \bar{\mathbf{y}}_i$ for $i = 0, ..., m-1$ and $\bar{\mathbf{x}}_m < \bar{\mathbf{y}}_m$ for some $m \in [1, n] \cap \mathbb{N}$. (We assume that $\mathbf{x}_0 = \mathbf{y}_0 = \infty$.) Since $\mathcal{M}'$ is compact there exists a maximal element $\mathbf{p} = (p_1, ..., p_N)^\top \in \mathcal{M}'$, i.e. $\mathbf{q} \in \mathcal{M}' \Rightarrow \mathbf{q} \ll \mathbf{p}$. Let $\mathcal{I} := \{i \in \langle N \rangle : \ p_i > 0\}$. Then $\#\mathcal{I} \geq \nu + 2$. Let $B = \sum_{i=1}^N p_i A_i \in \mathrm{conv}\,\mathrm{S}$ be the corresponding matrix with the maximal λ_p on $\mathrm{conv}\,\mathrm{S}$. Assume that $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbf{V}$ be an orthonormal basis of $\mathbf{V}$, consisting of the eigenvectors of B corresponding to the eigenvalues $\lambda_1(B), ..., \lambda_n(B)$ respectively. Let $m := \mathrm{upmul}(B, p) \leq \mu(p)$. Consider the following systems of $\frac{m(2p-m+1)}{2}$ equations in $\#\mathcal{I}$ unknowns $q_i \in \mathbb{R}, i \in \mathcal{I}$:

$$q_i = 0, \text{ for } i \in \langle N \rangle \backslash \mathcal{I}, \quad \sum_{i \in \mathcal{I}} q_i = 0,$$

$$\sum_{i \in \mathcal{I}} q_i \langle A_i \mathbf{x}_j, \mathbf{x}_k \rangle = 0, \ j = 1, ..., k-1, \ k = p, ..., p-m+1,$$

$$\sum_{i\in\mathcal{I}} q_i\langle A_i\mathbf{x}_j,\mathbf{x}_j\rangle = \sum_{i\in\mathcal{I}} q_i\langle A_i\mathbf{x}_p,\mathbf{x}_p\rangle \ j = p-1,...,p-m+1 \text{ if } m>1.$$

Since $\#\mathcal{I} \ge \nu + 2 = \frac{\mu(p)(2p-\mu(p)+1)}{2} + 1 > \frac{m(2p-m+1)}{2}$ it follows that there exists $0 \ne \mathbf{q} = (q_1,...,q_N)^\top \in \mathbb{R}^N$ whose coordinates satisfy the above equations. Let $B(t) := B + tC,\ C := \sum_{i=1}^N q_iA_i, t\in\mathbb{R}$. Then there exists $a>0$ such that for $t\in[-a,a]$ $\mathbf{p}(t) := \mathbf{p}+t\mathbf{q}\in\mathcal{P}_N \Rightarrow B(t)\in \operatorname{conv}\mathrm{S}$. As in the proof of Theorem 4.14.6 consider the matrix $B(t)(\mathbf{x}_1,...,\mathbf{x}_p)\in\mathbf{S}_p(\mathbb{R})$. Since $B(0)(\mathbf{x}_1,...,\mathbf{x}_p) = B(\mathbf{x}_1,...,\mathbf{x}_p)$ is the diagonal matrix $\operatorname{diag}(\lambda_1(B),...,\lambda_p(B)$ the conditions on the coordinates of $\mathbf{q}$ imply that $B(t)(\mathbf{x}_1,...,\mathbf{x}_p)$ is of the form $(\operatorname{diag}(\lambda_1(B),...,\lambda_{p-m}(B)) + tC_1)\oplus(\lambda_p+tb)I_m$ for a corresponding $C_1\in\mathbf{S}_{p-m}(\mathbb{R})$. Since $\lambda_{p-m}(B) > \lambda_p(B)$ it follows that there exists $a'\in(0,a]$ such that

$$\lambda_{p-m}(B(t)(\mathbf{x}_1,...,\mathbf{x}_p)) = \lambda_{p-m}(\operatorname{diag}(\lambda_1(B),...,\lambda_{p-m}(B)) + tC_1) > \\ \lambda_p(B)+|tb|,\ \lambda_p(B(t)) = \lambda_p(B)+tb, \quad \text{for } |t|\le a'.$$

Hence $\lambda_p(B(t)) \ge \lambda_p(B)+tb$ for $|t|\le a'$. As $B(t)\in\operatorname{conv}\mathrm{S}$ for $|t|\le a'$ and $\lambda_p(B)\ge\lambda_p(B(t))$ for $|t|\le a'$ it follows that $b=0$ and $\lambda_p(B(t)) = \lambda_p(B)$ for $|t|\le a'$. Hence $\mathbf{p}+t\mathbf{q}\in\mathcal{M}'$ for $|t|\le a'$. Since $\mathbf{q}\ne 0$, it is impossible to have the inequalities $\mathbf{p}-a'\mathbf{q}\ll\mathbf{p}$ and $\mathbf{p}+a'\mathbf{q}\ll\mathbf{p}$. This contradiction proves the theorem. □

It is possible to show that the above theorem is sharp in the case $\mu(p) = 1$ Problem 13 (d2). Similarly one can show, see Problem 10.

Theorem 4.14.10 *Let* $\mathbf{V}$ *be an n-dimensional IPS over* $\mathbb{C}$. *Let* $p\in[1,n]\cap\mathbb{N}$ *and denote* $\mu(p) := \operatorname{upmul}(\operatorname{conv} S, p)$. *Then*

$$\sup_{A\in\operatorname{conv}\mathrm{S}} \lambda_p(A) = \sup_{B\in\operatorname{conv}_{\mu(p)(2p-\mu(p))-1}\mathrm{S}} \lambda_p(B). \tag{4.14.4}$$

Problems

1. Let $\mathrm{S}\subset\mathbb{R}^{\mathrm{n}}$ be a nonempty finite set. Show

   a. Let $\mathrm{S} = \{\mathbf{x}_1,...,\mathbf{x}_N\}$. Then $\operatorname{conv}\mathrm{S} = \operatorname{conv}_{\mathrm{N}-1}(\mathrm{S})$.

   b. Any finitely generated convex set is compact.

   c. $\mathrm{S}\subset\operatorname{ext}\ \operatorname{conv}\mathrm{S}$.

   d. Let $f_1,...,f_m:\mathbb{R}^n\to\mathbb{R}$ be linear functions and $a_1,...,a_m\in\mathbb{R}^m$. Denote by A the affine space $\{\mathbf{x}\in\mathbb{R}^n:\ f_i(\mathbf{x}) = a_i,\ i=1,...,m\}$. Assume that $C := \operatorname{conv}\mathrm{S}\cap\mathrm{A}\ne\emptyset$. Then C is a finitely generate convex set such that $\mathrm{ext}\ C \subset \mathrm{conv}_m \mathrm{S}$. (*Hint*: Describe C by $m+1$ equations with $\#\mathrm{S}$ variables as in part a. Use the fact that any homogenous system in $m+1$ equations and $l > m+1$ variables has a nontrivial solution.)

e. Prove Theorem 4.14.2 for a finitely generated convex set C and a finite S.

2. Let C be a convex set of dimension d with a nonempty $\mathrm{ext}\ C$. Let $C' = \mathrm{conv}\,\mathrm{ext}\ C$. Show

   (a) $\mathrm{ext}\ C' = \mathrm{ext}\ C$.

   (b) Let $d' = \dim C'$. Then $d' \le d$ and the equality holds if and only if $C = C'$.

   (c) $F_j(C) \subset F_{j+1}(C)$ and equality holds if and only if $j \ge d'$.

3. Prove Proposition 2.

4. Let $C \subset \mathbf{V}$ be a convex set. A function $f : C \to \bar{\mathbb{R}}$ is called concave if $-f$ is a convex function on C. Assume the assumptions of Proposition 4.14.3. Assume in addition $f(C) \subset [0,\infty]$ and $g : C \to (0,\infty)$ is concave. Then
$$\sup_{\mathbf{x}\in \mathrm{conv}\, S} \frac{f(\mathbf{x})}{g(\mathbf{x})} = \sup_{\mathbf{y}\in S} \frac{f(\mathbf{y})}{g(\mathbf{y})}.$$
   If in addition S is compact and f, g are continuous on $\mathrm{conv}\, S$ then one can replace sup by max.

5. a. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Show the implication $\mathbf{x} \le \mathbf{y} \Rightarrow \mathbf{x} \preceq \mathbf{y}$.

   b. Let $D \subset \mathbb{R}^n_{\searrow}$ and assume that $f : D \to \mathbb{R}$ is strong Schur's order preserving. Show that f is nondecreasing on D.

   c. Let $i \in [2,n] \cap \mathbb{N}$ and f be the following function on $\mathbb{R}^n$: $(x_1, ..., x_n)^\top \mapsto x_i$. Show that f is nondecreasing on $\mathbb{R}^n$ but not Schur's order preserving on $\mathbb{R}^n_{\searrow}$.

6. Let $D \subset \mathbb{R}^n_{\searrow}$ be a convex Schur set. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $\mathrm{S} \subset \mathbf{S}(\mathbf{V})$ be a finite set such that $\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$. Show that $\mathrm{conv}\,\mathrm{S} \subset \lambda^{-1}(\mathrm{D})$.

7. a. Let $A \in \mathbf{H}_n$ and assume that $\mathrm{tr}\, A = 1$. Show that $\lambda_n(A) \le \frac{1}{n}$ and equality holds if and only if $A = \frac{1}{n} I_n$.

   b. Let $E_{kl} := (\frac{\delta_{(k,l)(i,j)}+\delta_{(l,k)(i,j)}}{2})_{i,j=1}^p \in \mathbf{S}_p(\mathbb{R})$ for $1 \le k \le l \le p$ be the symmetric matrices which have at most two nonzero equal entries at the locations (k,l) and $(l.k)$ which sum to 1. Let $Q_1,...,Q_{\binom{p+1}{2}} \in \mathbf{S}_p(\mathbb{R})$ be defined as follows:

$$\begin{aligned}&Q_1 := E_{11}+E_{12},\ Q_2 := E_{11}-E_{12}+E_{13},...,Q_p := E_{11}-E_{1p}+E_{23},...,\\ &Q_{2p-3} := E_{11}-E_{2(p-1)}+E_{2p},...,Q_{\binom{p}{2}} := E_{11}-E_{(p-2)p}+E_{(p-1)p},\\ &Q_{\binom{p}{2}+1} = E_{11}-E_{(p-1)p},\quad Q_{\binom{p}{2}+i} = E_{ii},\ \text{for } i=2,...,p.\end{aligned}$$

Let $\mathrm{S}=\{\mathrm{Q}_1,...,\mathrm{Q}_{\binom{p+1}{2}}\}$ Show that $\frac{1}{p}I_p \in \operatorname{conv}\mathrm{S} = \operatorname{conv}_{\binom{p+1}{2}-1}\mathrm{S}$ and $\frac{1}{p}I_p \notin \operatorname{conv}_{\binom{p+1}{2}-2}\mathrm{S}$.

c. Let $\mathrm{S}\subset \mathbf{S}_p(\mathbb{R})$ be defined as in b. Show that $\operatorname{tr} A = 1$ for each $A\in \operatorname{conv}\mathrm{S}$. Hence

$$\max_{A\in\operatorname{conv}\mathrm{S}} \lambda_p(A) = \lambda_p(\frac{1}{p}I_p) = \frac{1}{p} > \max_{B\in\operatorname{conv}_{\binom{p+1}{2}-2}\mathrm{S}} \lambda_p(B).$$

d. Assume that $n>p$ and let $R_i := Q_i\oplus 0 \in \mathbf{S}_n(\mathbb{R})$, where Q_i is defined in b, for $i=1,...,\binom{p+1}{2}$. Let $\mathrm{S}=\{\mathrm{R}_1,...,\mathrm{R}_{\binom{p+1}{2}}\}$. Show that

$$\max_{A\in\operatorname{conv}\mathrm{S}} \lambda_p(A) = \lambda_p(\frac{1}{p}I_p\oplus 0) = \frac{1}{p} > \max_{B\in\operatorname{conv}_{\binom{p+1}{2}-2}\mathrm{S}} \lambda_p(B).$$

8. a. Prove Theorem 4.14.7 repeating the arguments of Theorem 4.14.6. (*Hint*: Note that the condition $\langle B\mathbf{x}_i,\mathbf{x}_j\rangle = o$ for two distinct orthonormal vectors $\mathbf{x}_i,\mathbf{x}_j\in\mathbf{V}$ is equivalent to two real conditions, while the condition $\langle B\mathbf{x}_i,\mathbf{x}_i\rangle = \lambda_i(A)$ is one real conditions for $B\in\mathbf{S}(\mathbf{V})$.)

   b. Modify the example in Problem 7 to show that Theorem 4.14.7 is sharp.

9. Let $C = A+\sqrt{-1}B \in \mathrm{M_n}(\mathbb{C}), \mathrm{A},\mathrm{B}\in\mathrm{M_n}(\mathbb{R})$.

   a. Show $C\in\mathbf{H}_n$ if and only if A is symmetric and B is antisymmetric: $B^\top = -B$.

   b. Assume that $C\in\mathbf{H}_n$ and let $\hat{C}\in\mathrm{M_{2n}}(\mathbb{R})$ be defined as in Problem 6. Show that $\hat{C}\in\mathbf{S}_{2n}(\mathbb{R})$ and $\lambda_{2i-1}(\hat{C}) = \lambda_{2i}(\hat{C}) = \lambda_i(C)$ for $i=1,...,n$.

   c. Use the results of b to obtain a weaker version Theorem 4.14.7 directly from Theorem 4.14.6.

10. Prove Theorem 4.14.10.

11. Let $\mathbb{F}$ be a field and $k \in \mathbb{Z}_+$. $A = (a_{ij}) \in \mathrm{M_n}(\mathbb{F})$ is called a $2k+1$-diagonal matrix if $a_{ij} = 0$ if $|i-j| > k$. (1-diagonal are diagonal and 3-diagonal are called tridiagonal.) Then the entries $a_{1(k+1)}, ..., a_{(n-k)n}$ are called the k-upper diagonal.

    a. Assume that $A \in \mathrm{M_n}(\mathbb{F})$, $n > k$ and A is $2k+1$-diagonal. Suppose furthermore that k-upper diagonal of A does not have zero elements. Show that $\mathrm{rank\ A \geq n - k}$.

    b. Suppose in addition to a. $A \in \mathbf{H}_n$. Then $\mathrm{upmul}(A, p) \leq k$ for any $p \in \langle n \rangle$.

12. Let $\mathbf{V}$ be an n-dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $\mathrm{S} \subset \mathbf{S}(\mathbf{V})$ and $p \in \langle n \rangle$. Define the *weak* p-upper multiplicity denoted by $\mathrm{wupmul(conv\,S, p)}$ as follows. It is the smallest positive integer $m \leq p$ such that for any $N = \binom{n+1}{2} + 1$ operators $A_1, ..., A_N \in \mathrm{S}$ there exists a sequence $A_{j,k} \in \mathbf{S}(\mathbf{V}), j \in \langle N \rangle, k \in \mathbb{N}$, such that $\lim_{k\to\infty} A_{j,k} = A_j, j \in \langle N \rangle$ and $\mathrm{upmul}(\mathrm{conv}\{A_{1,k}, ..., A_{N,k}\}, p) \leq m$ for $k \in \mathbb{N}$.

    a. Show that $\mathrm{wupmul(conv\,S, p) \leq upmul(conv\,S, p)}$.

    b. Show that in Theorems 4.14.9 and 4.14.10 one can replace $\mathrm{upmul(conv\,S, p)}$ by $\mathrm{wupmul(conv\,S, p)}$.

13. a. Show that for any set $\mathrm{S} \subset \mathbf{D}(n, \mathbb{R})$ and $p \in \langle n \rangle$ $\mathrm{wupmul(conv\,S, p)} = 1$. (*Hint*: Use Problem 11.)

    b. Let $D_i = \mathrm{diag}(\delta_{i1}, ..., \delta_{in}), i = 1, ..., n$. Let $\mathrm{S := \{D_1, ..., D_n\}}$. Show that for $p \in [2, n] \cap \mathbb{Z}$

$$\max_{D \in \mathrm{conv\,S}} \lambda_p(D) = \max_{D \in \mathrm{conv}_{p-1}\mathrm{S}} \lambda_p(D) = \frac{1}{p} > \max_{D \in \mathrm{conv}_{p-2}\mathrm{S}} \lambda_p(D) = 0.$$

    c. Show that the variation of Theorem 4.14.9 as in Problem 12b for $\mathrm{wupmul(conv\,S, p)} = 1$ is sharp.

    d. Let $A \in \mathbf{S}_n(\mathbb{R})$ be a tridiagonal matrix with nonzero elements on the first upper diagonal as in 11b. Let $t \in \mathbb{R}$ and define $D_i(t) = D_i + tA$, where D_i is defined as in b, for $i = 1, ..., n$. Let $\mathrm{S(t) = \{D_1(t), ..., D_n(t)\}}$. Show

    d1. For $t \neq 0$ $\mathrm{upmul(conv\,S(t), p)} = 1$ for $p \in [2, n] \cap \mathbb{Z}$.

    d2. There exists $\epsilon > 0$ such that for any $|t| \leq \epsilon$

$$\max_{A \in \mathrm{conv\,S(t)}} \lambda_p(A) = \max_{B \in \mathrm{conv}_{p-1}\mathrm{S(t)}} \lambda_p(B) > \max_{C \in \mathrm{conv}_{p-2}\mathrm{S(t)}} \lambda_p(C).$$

    Hence Theorem 4.14.9 is sharp in the case $\mathrm{upmul(conv\,S, p)} = 1$.

14. a. Let $\mathrm{S} \subset \mathrm{H_n}$ be a set of $2k+1$-diagonal matrices. Assume that either each k-upper diagonal of any $A \in \mathrm{S}$ consists of positive elements, or each k-upper diagonal of any A is fixed and consists of nonzero elements. Show that $\mathrm{upmul}(\mathrm{conv}\,\mathrm{S}, \mathrm{p}) \le \mathrm{k}$.

    b. Let $\mathrm{S} \subset \mathrm{H_n}$ be a set of $2k+1$-diagonal matrices. Show that $\mathrm{wupmul}(\mathrm{conv}\,\mathrm{S}, \mathrm{p}) \le \mathrm{k}+1$.

4.15 Multiplicity index of a subspace of $\mathbf{S}(\mathbf{V})$

Definition 4.15.1 *Let $\mathbf{V}$ be a finite dimensional IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $\mathbf{U}$ be a nontrivial subspace of $\mathbf{S}(\mathbf{V})$. Then the multiplicity index of $\mathbf{U}$ is defined*

$$\mathrm{mulind}\,\mathbf{U} := \{\max p \in \mathbb{N} : \quad \exists A \in \mathbf{U}\backslash\{0\} \text{ such that } \lambda_1(A) = ... = \lambda_p(A)\}.$$

Clearly for any nontrivial $\mathbf{U}$ $\mathrm{mulind}\,\mathbf{U} \in [1, \dim \mathbf{V}]$. Also $\mathrm{mulind}\,\mathbf{U} = \dim \mathbf{V} \iff I \in \mathbf{V}$. The aim of this section to prove the following theorem.

Theorem 4.15.2 *Let $\mathbf{V}$ be an IPS over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ of dimension $n \ge 3$. For $r \in [2, n-1]$ let $\kappa(r) := \frac{(r-1)(2n-r+2)}{2}$ if $\mathbb{F} = \mathbb{R}$ and $\kappa(r) := (r-1)(2n-r+2)$ if $\mathbb{F} = \mathbb{C}$. Let $\mathbf{U}$ be a subspace of $\mathbf{S}(\mathbf{V})$. Then* $\mathrm{mulind}\,\mathbf{U} \ge r$ *if* $\dim \mathbf{U} \ge \kappa(r)$ *and this result is sharp.*

Proof. Assume first that $\mathbb{F} = \mathbb{R}$. Assume $\dim \mathbf{U} \ge \kappa(r)$. Suppose to the contrary that $\mathrm{index}\,\mathbf{U} = p < r$. Let $A \in \mathbf{U}$ such that $\lambda_1(A) = ... = \lambda_p(A) > \lambda_{p+1}(A)$. Assume that

$$A\mathbf{x}_i = \lambda_i(A)\mathbf{x}_i,\ \mathbf{x}_i \in \mathbf{V}, \quad \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{ij},\ i, j = 1, ..., n.$$

By representing $\mathbf{S}(\mathbf{V})$ is $S_n(\mathbb{R})$ with respect to the orthonormal basis $\mathbf{x}_1, ..., \mathbf{x}_n$ we may assume that $\mathbf{U}$ is a subspace of $\mathbf{S}_n(\mathbb{R})$ and $A = \mathrm{diag}(\lambda_1(A), ..., \lambda_n(A))$.

4.16 Analytic functions of hermitian matrices

Denote by $\mathbf{H}_n$ the set of all $n \times n$ hermitian matrices. For $A \in \mathbf{H}_n$ denote by $\mathrm{spec\,A} \subset \mathbb{R}$ the spectrum of A. $A(z) := A_0 + zA_1, A_0, A_1 \in \mathbf{H}_n$ is called a *hermitian pencil.* It is known that it is possible to rename the eigenvalues of $A(z)$ as $\alpha_1(z), \ldots, \alpha_n(z)$ such that each $\alpha_i(z)$ is analytic in some neighborhood $\mathcal{N}$ of the real axis $\mathbb{R}$. Furthermore the eigenprojection

$P_i(z)$ on each $\alpha_i(z)$ is analytic in z in $\mathcal{N}$. (Note that if $\alpha_i(z)$ has multiplicity m_i for all but a finite number of points on $\mathbb{R}$, then $P_i(z)$ has rank m_i on $\mathcal{N}$.) Furthermore, the corresponding eigenvectors $\mathbf{x}_1(z), \ldots, \mathbf{x}_n(z)$ can be chosen to be analytic in $z \in \mathcal{N}$, such that for $\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t)$ are orthonormal for $t \in \mathbb{R}$. See [Kat80, II.6.1-II.6.2].

Let $t \in \mathbb{R}$ be fixed. An eigenvalue $\alpha(z) := \alpha_i(z)$ is called regular at t if the multiplicity of $\alpha(z)$ is fixed for $|z-t| < r$ for some $r = r(t) > 0$. So

$$\alpha(z) = \sum_{j=0}^{\infty} a_j (z-t)^j, \quad a_j \in \mathbb{R}, j \in \mathbb{Z}_+. \tag{4.16.1}$$

Furthermore, given a normalized eigenvector $A(t)\mathbf{x}_0 = a_0\mathbf{x}_0, \mathbf{x}_0^*\mathbf{x}_0 = 1$, there exists an analytic eigenvector $\mathbf{x}(z)$ corresponding to $\alpha(z)$ satisfying the conditions:

$$\mathbf{x}(z) = \sum_{j=0}^{\infty} (z-t)^j \mathbf{x}_j, \quad \mathbf{x}_j \in \mathbb{C}^n, j \in \mathbb{Z}_+, \tag{4.16.2}$$

$$A(z)\mathbf{x}(z) = \alpha(z)\mathbf{x}(z), \quad \mathbf{x}(s)^*\mathbf{x}(s) = 1 \text{ for } s \in \mathbb{R}.$$

Let $\mathbf{S}_n$ denote the space of all $n \times n$ real symmetric matrices. Suppose that $A_0, A_1 \in \mathbf{S}_n$. Then $A(z)$ is called a symmetric pencil. In that case the projections induced by $A(s), s \in \mathbb{R}$ on each $\alpha_i(s)$ must be a real orthogonal projection. Hence in the expansion (4.16.2) each $\mathbf{x}_j$ can be chosen to be a real vector.

One can find the formulas for $a_j, \mathbf{x}_j, j \in \mathbb{Z}_+$ in terms of A_0, A_1, in particular for a_1, a_2, in [Kat80, II.2.4]. In this note we give slightly different formulas for a_1, a_2, a_3.

Let $B \in \mathbf{H}_n$. Denote by $B^\dagger \in \mathbf{H}_n$ the Moore-Penrose inverse of B. That is B is uniquely characterized by the condition that $B^\dagger B = BB^\dagger$ is the projection on the subspace spanned by all eigenvectors of B corresponding to the nonzero eigenvalues of B.

Theorem 4.16.1 *Let $A_0,\ A_1 \in \mathbf{H}_n, A(z) = A_0 + zA_1$. Let $t \in \mathbb{R}, A = A_0 + tA_1$ and assume that the eigenvalue $a_0 \in$ spec A is simple for the pencil $A(z)$ at $z = t$. Suppose that $A\mathbf{x}_0 = a_0\mathbf{x}_0,\ \mathbf{x}_0^*\mathbf{x}_0 = 1$. Let (4.16.1-4.16.2) be the Taylor expansion of the eigenvalue $\alpha(z) \in$ spec A(z) and a local Taylor expanison of the corresponding eigenvector $\mathbf{x}(z)$ satisfying $\alpha(t) = a_0,\ \mathbf{x}(t) = \mathbf{x}_0$. Then*

$$
\begin{aligned}
&a_1 = \mathbf{x}_0^* A_1 \mathbf{x}_0,\\
&\mathbf{x}_1 = (a_0 I - A)^\dagger A_1 \mathbf{x}_0,\\
&a_2 = \mathbf{x}_0^* A_1 \mathbf{x}_1 = \mathbf{x}_0^* A_1 (a_0 I - A)^\dagger A_1 \mathbf{x}_0,\\
&\mathbf{x}_2 = (a_0 I - A_0)^\dagger (A_1 - a_1 I)\mathbf{x}_1 - (\frac{1}{2}\mathbf{x}_1^* \mathbf{x}_1)\mathbf{x}_0\\
&a_3 = \mathbf{x}_0^*((A_1 - a_1 I)(a_0 I - A)^\dagger)^2 A_1 \mathbf{x}_0,
\end{aligned}
\tag{4.16.3}
$$

Proof. Without loss of generality we assume that $t = 0$, hence $A = A_0$. Next we consider first the case where A_0, A_1 are real symmetric. Furthermore, by replacing A_0, A_1 with $Q^\top A_0 Q, Q^\top A_1 Q$, where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix we may assume that A_0 is a diagonal matrix $\mathrm{diag}(d_1, \dots, d_n)$, where $d_1 = a_0$ and $d_i \neq a_0$ for $i > 1$. Moreover, we can assume that $\mathbf{x}_0 = (1, 0, \dots, 0)^\top$. Note that $(a_0 I - A_0)^\dagger = \mathrm{diag}(0, (a_0 - d_1)^{-1}, \dots, (a_0 - d_n)^{-1})$. (We are not going to use explicitly these assumptions, but the reader can see more transparently our arguments using these assumptions.)

Recall that we may assume that $\mathbf{x}(s) \in \mathbb{R}^n$. The orthogonality condition $\mathbf{x}(s)^\top \mathbf{x}(s) = 1$ yields that

$$
\sum_{j=0}^{k} \mathbf{x}_j^\top \mathbf{x}_{k-j} = 0, \quad k \in \mathbb{N}. \tag{4.16.4}
$$

The equality $A(z)\mathbf{x}(z) = \alpha(z)\mathbf{x}(z)$ yields

$$
A_0 \mathbf{x}_k + A_1 \mathbf{x}_{k-1} = \sum_{j=0}^{k} a_j \mathbf{x}_{k-j}, \quad k \in \mathbb{N}. \tag{4.16.5}
$$

Since $\alpha(s)$ is real for a real s we deduce that $a_j \in \mathbb{R}$. Consider the equality (4.16.5) for $k = 1$. Multiply it by $\mathbf{x}_0^\top$ and use the equality $\mathbf{x}_0^\top A_0 = a_0 \mathbf{x}_0^\top$ to deduce the well known equality $a_1 = \mathbf{x}_0^\top A_0 \mathbf{x}_0$, which is the first equality of (4.16.3). The equality (4.16.5) for $k = 1$ is equivalent to $(a_0 I - A_0)\mathbf{x}_1 = A_1 \mathbf{x}_0 - a_1 \mathbf{x}_0$. Hence $\mathbf{x}_1$ is of the form

$$
\mathbf{x}_1 = (a_0 I - A)^\dagger (A_1 \mathbf{x}_0 - a_1 \mathbf{x}_0) + b_1 \mathbf{x}_0 = (a_0 I - A)^\dagger A_1 \mathbf{x}_0 + b_1 \mathbf{x}_0, \tag{4.16.6}
$$

for some b_1. The orthogonality condition $\mathbf{x}_0^\top \mathbf{x}_1$ implies that $b_1 = 0$. Hence the second equality of (4.16.3) holds.

Multiply the equality (4.16.5) for $k = 2$ by $\mathbf{x}_0^\top$ and use $\mathbf{x}_0^\top A_0 = a_0 \mathbf{x}_0^\top, \mathbf{x}_0^\top \mathbf{x}_1 = 0$ to obtain $a_2 = \mathbf{x}_0^\top A_1 \mathbf{x}_1$. This establishes the third equality of (4.16.3).

Rewrite the equality (4.16.5) for $k = 2$ as $(a_0I - A_0)\mathbf{x}_2 = A_1\mathbf{x}_1 - a_1\mathbf{x}_1 - a_2\mathbf{x}_0$. Hence

$$\mathbf{x}_2 = (a_0I - A_0)^\dagger(A_1 - a_1I)\mathbf{x}_1 + b_2\mathbf{x}_0. \tag{4.16.7}$$

Multiply the above equality by $\mathbf{x}_0^\top$ to deduce that $\mathbf{x}_0^\top\mathbf{x}_2 = b_2$. (4.16.4) for $k = 2$ yields $b_2 = -\frac{1}{2}\mathbf{x}_1^\top\mathbf{x}_1$. This establishes the fourth equality of (4.16.3).

Multiply the equality (4.16.5) for $k = 3$ by $\mathbf{x}_0^\top$ to deduce

$$a_3 = \mathbf{x}_0^\top A_1\mathbf{x}_2 - a_1\mathbf{x}_0^\top\mathbf{x}_2 = \mathbf{x}_0^\top(A_1 - a_1I)\mathbf{x}_2.$$

Observe next that from the first equality in (4.16.3) $\mathbf{x}_0^T(A_1 - a_1I)\mathbf{x}_0 = 0$. This establishes the last equality of (4.16.3).

We now show that the same formulas hold when A_0, A_1 are hermitian. Observe that if we have a local Taylor expansion of $\mathbf{x}(z)$ then we can replace $\mathbf{x}(z)$ by $e^{\phi(z)}\mathbf{x}(z)$, where $\phi(z) = \sum_{j=1}^\infty \phi_j(z-t)^j$ is locally analytic at $z = t$ and each ϕ_j is purely imaginary. Now we repeat the proof of (4.16.3). The first formula of (4.16.3) holds as in the symmetric case. The equality (4.16.6) also holds. We now can only deduce the equality $\Re b_1 = 0$. Now choose ϕ_1 such that $b_1 = 0$. Hence the second equality of (4.16.3) holds. Now deduce the third equality of (4.16.3). Next we deduce (4.16.7) and the equality $\Re b_2 = 0$. Now use the corresponding choice of ϕ_2 to obtain that $b_2 = 0$. Hence the fourth equality of (4.16.3) holds. Now deduce the last equality of (4.16.3). □

Note that for the real symmetric case the formulas of $\mathbf{x}_1, \mathbf{x}_2$ in (4.16.3) are global formulas.

Theorem 4.16.2 *Let $A_0 \in \mathbf{H}_n, n \geq 2$. Assume that a_0 is a simple eigenvalue of A_0, with the corresponding eigenvector $A\mathbf{x}_0 = a_0\mathbf{x}_0, \mathbf{x}_0^*\mathbf{x}_0 = 1$. Suppose furthermore that $|\lambda - a_0| \geq r > 0$ for any other eigenvalue λ of A_0. Let $A_1 \in \mathbf{H}_n, A_1 \neq 0$, and denote by $||A_1||$ the l_2 norm of A_1, i.e. the maximal absolute value of the eigenvalues of A_1. Let $\alpha(z)$ be the eigenvalue of $A(z) = A_0 + zA_1$, which is analytic in the neighborhood of $\mathbb{R}$ and satisfying the condition $\alpha(0) = a_0$. Let a_1, a_2 will be given by (4.16.3), where $A = A_0$. Fix $0 < c < \frac{r}{2||A_1||}$. Then*

$$|\alpha(s) - (a_0 + a_1s + a_2s^2)| \leq \frac{4||A_1||^3|s|^3}{(r - 2c||A_1||)^2} \text{ for all } s \in [-c, c]. \tag{4.16.8}$$

Proof. Let $\lambda_1(s) \geq \ldots \geq \lambda_n(s)$ be the eigenvalues of $A(s), s \in \mathbb{R}$. Note that $\lambda_1(0), \ldots, \lambda_n(0)$ are the eigenvalues of A_0. Assume that $\lambda_i(0) = a_1$. Let $\rho(s) = \min(\lambda_{i-1}(s) - \lambda_i(s), \lambda_i(s) - \lambda_{i+1}(s))$, where $\lambda_0(s) = \infty, \lambda_{n+1}(s) =$

$-\infty$. Thus $r \le \rho(0)$. Let $\beta_1 \ge \ldots \ge \beta_n$ be the eigenvalues of A_1. Then $||A_1|| = \max(|\beta_1|, |\beta_n|)$. Apply Lidskii's theorem to $C = A(s) - A_0$ [Kat80, III.Thm 6.10] to deduce that

$$|\lambda_j(s) - \lambda_j(0)| \le |s|\ ||A_1||, \quad j = 1, \ldots, n.$$

Hence

$$\rho(s) \ge r(0) - 2|s|||A_1|| > 0 \text{ for } s \in (-\frac{\rho(0)}{2||A_1||}, \frac{\rho(0)}{2||A_1||}). \tag{4.16.9}$$

In particular, $\lambda_i(s)$ is a simple eigenvalue of $A(s)$ in the above interval. Assume that s is in the interval given in (4.16.9). It is straightforward to show that

$$||(\alpha(s)I - A(s))^\dagger|| = \frac{1}{\rho(s)} \le \frac{1}{\rho(0) - 2|s|\ ||A_1||}. \tag{4.16.10}$$

(One can assume that $A(s)$ is a diagonal matrix.)

Use the Taylor theorem with remainder to obtain the equality

$$\alpha(s) - (a_0 + sa_1 + s^2a_2) = \frac{1}{6}\alpha^{(3)}(t)s^3 \text{ for some } t,\ |t| < |s|. \tag{4.16.11}$$

Use Theorem 4.16.1 to deduce that

$$\frac{1}{6}\alpha^{(3)}(t) = \mathbf{x}_i(t)^*((A_1 - \alpha_i'(t)I)(\alpha_i(t)I - A(t))^\dagger)^2 A_1\mathbf{x}_i(t),$$

where $\mathbf{x}_i(s)$ is an eigenvector of $A(s)$ of length one corresponding to $\alpha_i(s)$. As $\alpha'(t) = \mathbf{x}_i(t)^* A_1 \mathbf{x}_i(t)$ we deduce that $|\alpha_i'(t)| \le ||A_1||$. Hence $||A_1 - \alpha_i'(t)I|| \le 2||A_1||$. Therefore

$$|\frac{1}{6}\alpha^{(3)}(t)| \le ||((A_1 - \alpha_i'(t)I)(\alpha_i(t)I - A(t))^\dagger)^2 A_1|| \le \frac{4||A_1||^3}{\rho(t)^2}$$

Use the inequality (4.16.10) and the inequality $r \le \rho(0)$ to deduce the theorem. □

4.17 Eigenvalues of sum of hermitian matrices

Put here

- Lidskii's theorem which is equivalent to $\lambda(A) - \lambda(B) \prec \lambda(A - B)$.
- Weyl's inequality.
- Kato's inequality [Kat80, II.5.Thm 6.11] If $f : \mathbb{R} \to \mathbb{R}$ is convex then

$$\sum_{i=1}^{n} f(\lambda_i(A) - \lambda_i(B)) \le \sum_{i=1}^{n} f(\lambda_i(A - B).$$

In particular use $f(x) = |x|^p, p \ge 1$.

Chapter 5

Elements of Multilinear Algebra

5.1 Tensor product of two free modules

Let $\mathbb{D}$ be a domain. Recall that $\mathbf{N}$ is called a *free* finite dimensional module if $\mathbf{N}$ has a finite basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, i.e. $\dim \mathbf{N} = n$. Then $\mathbf{N}' := \mathrm{Hom}\,(\mathbf{N}, \mathbb{D})$ is a free n-dimensional module. Furthermore we can identify $\mathrm{Hom}\,(\mathbf{N}', \mathbb{D})$ with $\mathbf{N}$. (See Problem 1.)

Definition 5.1.1 *Let* $\mathbf{M}, \mathbf{N}$ *be two free finite dimensional modules over an integral domain* $\mathbb{D}$. *Then the tensor product* $\mathbf{M} \otimes \mathbf{N}$ *is identified with* $\mathrm{Hom}\,(\mathbf{N}', \mathbf{M})$. *Moreover, for each* $\mathbf{m} \in \mathbf{M}, \mathbf{n} \in \mathbf{N}$ *we identify* $\mathbf{m} \otimes \mathbf{n} \in \mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ *with the linear transformation* $\overline{\mathbf{m} \otimes \mathbf{n}} : \mathbf{N}' \to \mathbf{M}$ *given by* $\mathbf{f} \mapsto \mathbf{f}(\mathbf{n})\mathbf{m}$ *for any* $\mathbf{f} \in \mathbf{N}'$.

Proposition 5.1.2 *Let* $\mathbf{M}, \mathbf{N}$ *be free modules over a domain* $\mathbb{D}$ *with bases* $[\mathbf{d}_1, \ldots, \mathbf{d}_m]$, $[\mathbf{e}_1, \ldots, \mathbf{e}_n]$ *respectively. Then* $\mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ *is a free module with the basis* $\mathbf{d}_i \otimes \mathbf{e}_j, i = 1, \ldots, m, j = 1, \ldots, n$. *In particular*

$$\dim \mathbf{M} \otimes \mathbf{N} = \dim \mathbf{M} \dim \mathbf{N}. \tag{5.1.1}$$

(See Problem 3.) For an abstract definition of $\mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ for any two $\mathbb{D}$-modules see Problem 16.

Intuitively, one views $\mathbf{M} \otimes \mathbf{N}$ as a linear span of all elements of the form $\mathbf{m} \otimes \mathbf{n}$, where $\mathbf{m} \in \mathbf{M}, \mathbf{n} \in \mathbf{N}$ satisfying the following *natural* properties:

- $a(\mathbf{m} \otimes \mathbf{n}) = (a\mathbf{m}) \otimes \mathbf{n} = \mathbf{m} \otimes (a\mathbf{n})$ for all $a \in \mathbb{D}$.

- $(a_1\mathbf{m}_1 + a_2\mathbf{m}_2) \otimes \mathbf{n} = a_1(\mathbf{m}_1 \otimes \mathbf{n}) + a_2(\mathbf{m}_2 \otimes \mathbf{n})$ for all $a_1, a_2 \in \mathbb{D}$. (Linearity in the first variable.)
- $\mathbf{m} \otimes (a_1\mathbf{n}_1 + a_2\mathbf{n}_2) = a_1(\mathbf{m} \otimes \mathbf{n}_1) + a_2(\mathbf{u} \otimes \mathbf{n}_2)$ for all $a_1, a_2 \in \mathbb{D}$. (Linearity in the second variable.)

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

Proposition 5.1.3 *Let $\mathbf{M}, \mathbf{N}$ be free modules over a domain $\mathbb{D}$ with bases*
$[\mathbf{d}_1, \ldots, \mathbf{d}_m]$, $[\mathbf{e}_1, \ldots, \mathbf{e}_n]$ respectively. Then any $\tau \in \mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ is given by

$$\tau = \sum_{i=j=1}^{i=m,j=n} a_{ij}\mathbf{d}_i \otimes \mathbf{e}_j, \quad A = [a_{ij}] \in \mathbb{D}^{m\times n}. \tag{5.1.2}$$

Let $[\mathbf{u}_1, \ldots, \mathbf{u}_m], [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be another bases of $\mathbf{M}, \mathbf{N}$ respectively. Assume that $\tau = \sum_{i,j=1}^{m,n} b_{ij}\mathbf{u}_i \otimes \mathbf{v}_j$ and let $B = [b_{ij}] \in \mathbb{D}^{m\times n}$. Then $B = PAQ^T$, where P and Q are the transition matrices from the bases $[\mathbf{d}_1, \ldots, \mathbf{d}_m]$ to $[\mathbf{u}_1, \ldots \mathbf{u}_m]$ and $[\mathbf{e}_1, \ldots, \mathbf{e}_n]$ to $[\mathbf{v}_1, \ldots, \mathbf{v}_n]$.
($[\mathbf{d}_1, \ldots, \mathbf{d}_m] = [\mathbf{u}_1, \ldots \mathbf{u}_m]P$, $[\mathbf{e}_1, \ldots, \mathbf{e}_n] = [\mathbf{v}_1, \ldots, \mathbf{v}_n]Q$.)

See Problem 6.

Definition 5.1.4 *Let $\mathbf{M}, \mathbf{N}$ be free finite dimensional modules over a domain $\mathbb{D}$. Let $\tau \in \mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ be given by (5.1.2). The rank of τ, denoted by* $\operatorname{rank} \tau$*, is the rank of the representation matrix A, i.e.* $\operatorname{rank} \tau = \operatorname{rank} \mathrm{A}$. *The tensor rank of τ, denoted by* $\operatorname{Rank} \tau$*, is the minimal k such that $\tau = \sum_{l=1}^{k} \mathbf{m}_l \otimes \mathbf{n}_l$ for some $\mathbf{m}_l \in \mathbf{M}, \mathbf{n}_l \in \mathbf{N}, l = 1, \ldots, k$.*

$\operatorname{rank} \tau$ is independent of the choice of bases in $\mathbf{M}$ and $\mathbf{N}$. (Problem 7.) Since $\mathbf{M}\otimes_{\mathbb{D}}\mathbf{N}$ has a basis consisting of decomposable tensors it follows that

$$\operatorname{Rank} \tau \le \min(\dim \mathbf{M}, \dim \mathbf{N}) \text{ for any } \tau \in \mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}. \tag{5.1.3}$$

See Problem 8.

Proposition 5.1.5 *Let $\mathbf{M}, \mathbf{N}$ be free finite dimensional modules over a domain $\mathbb{D}$. Let $\tau \in \mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$. Then* $\operatorname{rank} \tau \le \operatorname{Rank} \tau$. *If $\mathbb{D}$ is a Bezout domain then* $\operatorname{rank} \tau = \operatorname{Rank} \tau$

Proof. Assume that $\mathbf{M}, \mathbf{N}$ have bases as in Proposition 5.1.3. Suppose that (5.1.2) holds. Let $\tau = \sum_{l=1}^{k} \mathbf{m}_l \otimes \mathbf{n}_l$. Clearly, each $\mathbf{m}_l \otimes \mathbf{n}_l = \sum_{i,j=1}^{m,n} a_{ij,l}\mathbf{d}_i \otimes \mathbf{e}_j$, where $A_l := [a_{ij,l}]_{i,j=1}^{m,n} \in \mathbb{D}^{m\times n}$ is rank one matrix.

Then $A = \sum_{l=1}^{k} A_l$. It is straightforward to show that rank A $\le$ k. This shows that rank $\tau \le$ Rank τ.

Assume that $\mathbb{D}$ is BD. Let $P \in \mathbf{GL}(m, \mathbb{D})$ such that $PA = [b_{ij}] \in \mathbb{D}^{m \times n}$ is a Hermite normal form of A. In particular, the first $r :=$ rank A rows of B are nonzero rows, and all other rows of B are zero rows. Let $[\mathbf{u}_1, \ldots, \mathbf{u}_m] := [\mathbf{d}_1, \ldots, \mathbf{d}_m]P^{-1}$ be a basis in $\mathbf{M}$. Proposition 5.1.3 yields that $\tau = \sum_{i,j=1}^{m,n} b_{ij} \mathbf{u}_i \otimes \mathbf{e}_j$. Define $\mathbf{n}_l = \sum_{j=1}^{n} b_{lj} \mathbf{e}_j, l = 1, \ldots, r$. Then $\tau = \sum_{l=1}^{r} \mathbf{u}_l \otimes \mathbf{n}_l$. Hence $r \ge$ Rank τ, which implies that rank $\tau =$ Rank τ. □

Proposition 5.1.6 *Let $\mathbf{M}_i, \mathbf{N}_i$ be free finite dimensional modules over $\mathbb{D}$. Let $T_i : \mathbf{M}_i \to \mathbf{N}_i$ be homomorphisms. Then there exists a unique homomorphism on $T : \mathbf{M}_1 \otimes \mathbf{M}_2 \to \mathbf{N}_1 \otimes \mathbf{N}_2$ such that $T(\mathbf{m}_1 \otimes \mathbf{m}_2) = (T_1\mathbf{m}_1) \otimes (T_2\mathbf{m}_2)$ for all $\mathbf{m}_1 \in \mathbf{M}_1, \mathbf{m}_2 \in \mathbf{M}_2$. This homomorphism is denoted by $T_1 \otimes T_2$.*

Suppose furthermore that $\mathbf{W}_1, \mathbf{W}_2$ are free finite dimensional $\mathbb{D}$-modules, and $P_i : \mathbf{N}_i \to \mathbf{W}_i, i = 1, 2$ are homomorphisms. Then $(P_1 \otimes P_2)(T_1 \otimes T_2) = (P_1T_1) \otimes (P_2T_2)$.

See Problem 9.

Since each homomorphism $T_i : \mathbf{M}_i \to \mathbf{N}_i, i = 1, 2$ 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 $A_1 \in \mathbb{D}^{n_1 \times m_1}, A_2 \in \mathbb{D}^{n_2 \times m_2}$. This tensor product is called the *Kronecker* product.

Definition 5.1.7 *Let $A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{D}^{m \times n}, B = [b_{ij}]_{i,j=1}^{p,q} \in \mathbb{D}^{p \times q}$. Then $A \otimes B \in \mathbb{D}^{mp \times nq}$ is the following block matrix:*

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ a_{21}B & a_{22}B & \ldots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m_1 2}B & \ldots & a_{mn}B \end{bmatrix} \tag{5.1.4}$$

In the rest of the section we discuss the symmetric and skew symmetric tensor products of $\mathbf{M} \otimes \mathbf{M}$.

Definition 5.1.8 *Let $\mathbf{M}$ be a free finite dimensional module over $\mathbb{D}$. Denote $\mathbf{M}^{\otimes 2} := \mathbf{M} \otimes \mathbf{M}$. The submodule* $\mathrm{Sym}^2\mathbf{M} \subset \mathbf{M}^{\otimes 2}$*, called a* 2-symmetric power *of $\mathbf{M}$, is spanned by tensors of the form* $\mathrm{sym}^2(\mathbf{m}, \mathbf{n}) :=$

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

Since $\mathbf{M}^{\otimes 2}$ can be identified with $\mathbb{D}^{m\times m}$ it follows that $\mathrm{Sym}^2(\mathbf{M})$ and $\bigwedge^2\mathbf{M}$ can be identified with the submodules of symmetric and skew symmetric matrices respectively. See Problem 12. Observe next that $2\mathbf{m}\otimes\mathbf{n}=\mathrm{sym}^2(\mathbf{m},\mathbf{n})+\mathbf{m}\wedge\mathbf{n}$. Assume that 2 is a unit in $\mathbb{D}$. Then $\mathbf{M}^{\otimes 2}=\mathrm{Sym}^2(\mathbf{M})\oplus\bigwedge^2\mathbf{M}$. Hence any tensor $\tau\in\mathbf{M}^{\otimes 2}$ can be decomposed uniquely to a sum $\tau=\tau_s+\tau_a$ where $\tau_s,\tau_a\in\mathbf{M}^{\otimes 2}$ are symmetric and skew symmetric tensors respectively. (See Problem 12.)

Proposition 5.1.9 *Let* $\mathbf{M},\mathbf{N}$ *be a finite dimensional module over* $\mathbb{D}$*. Let* $T:\mathrm{Hom}\,(\mathbf{M},\mathbf{N})$*. Then*

$$T\otimes T:\mathrm{Sym}^2\mathbf{M}\to\mathrm{Sym}^2\mathbf{N},\quad T\otimes T:\bigwedge^2\mathbf{M}\to\bigwedge^2\mathbf{N}.$$

See Problem 13.

Definition 5.1.10 *Let* $\mathbf{M},\mathbf{N}$ *be finite dimensional modules over* $\mathbb{D}$*. Let* $T:\mathrm{Hom}\,(\mathbf{M},\mathbf{N})$*. Then* $T\wedge T\in\mathrm{Hom}\,(\bigwedge^2\mathbf{M},\bigwedge^2\mathbf{N})$ *is defined as the restriction of* $T\otimes T$ *to* $\bigwedge^2\mathbf{M}$*.*

Proposition 5.1.11 *Let* $\mathbf{M},\mathbf{N}$ *be a finite dimensional module over* $\mathbb{D}$*. Let* $T:\mathrm{Hom}\,(\mathbf{M},\mathbf{N})$*. Then*

1. *Assume that* $[\mathbf{d}_1,\ldots,\mathbf{d}_m]$*.*
2. *Assume that* $S:\mathrm{Hom}\,(\mathbf{L},\mathbf{M})$*. Show that* $ST\wedge ST=(S\wedge S)(T\wedge T)$*.*

Problems

1. Let $\mathbf{N}$ be a free module with a basis $[\mathbf{e}_1,\ldots,\mathbf{e}_n]$. Show
   - $\mathbf{N}':=\mathrm{Hom}\,(\mathbf{N},\mathbb{D})$ is a free module with a basis $[\mathbf{f}_1,\ldots,\mathbf{f}_n]$, where $\mathbf{f}_i(\mathbf{e}_j)=\delta_{ij}, i,j=1,\ldots,n$.

- Show that $(\mathbf{N}')'$ can be identified with $\mathbf{N}$ as follows. To each $\mathbf{n} \in \mathbf{N}$ associate the following functional $\widehat{\mathbf{n}} : \mathbf{N}' \to \mathbb{D}$ defined by $\widehat{\mathbf{n}}(\mathbf{f}) = \mathbf{f}(\mathbf{n})$ for each $\mathbf{f} \in \mathbf{N}'$. Show that $\widehat{\mathbf{n}}$ is a linear functional on $\mathbf{N}'$ and any $\tau \in (\mathbf{N}')'$ is equal to a unique $\widehat{\mathbf{n}}$.

2. . Let $\mathbb{F}$ be a field and $\mathbf{V}$ and n-dimensional subspace of $\mathbf{V}$. Then $\mathbf{V}' :=$ Hom $(\mathbf{V}, \mathbb{F})$ is called the *dual space* of $\mathbf{V}$. Show

   (a) $(\mathbf{V}')'$ can be identified with $\mathbf{V}$. I.e. for each $\mathbf{v} \in \mathbf{V}$ let $\hat{\mathbf{v}} : \mathbf{V}' \to \mathbb{F}$ be the linear functional given by $\hat{\mathbf{v}}(\mathbf{f}) = \mathbf{f}(\mathbf{v})$. Then any $\psi \in (\mathbf{V}')'$ is of the form $\hat{\mathbf{v}}$ for some $\mathbf{v} \in \mathbf{V}$

   (b) For $X \subseteq \mathbf{V}$, $F \subseteq \mathbf{V}'$ denote by $X^\perp := \{\mathbf{f} \in \mathbf{V}' : \mathbf{f}(\mathbf{x}) = 0, \forall \mathbf{x} \in X\}, F^\perp := \{\mathbf{v} \in \mathbf{V} : \mathbf{f}(\mathbf{v}) = 0, \forall \mathbf{f} \in F\}$. Then $X^\perp, F^\perp$ are subspaces of $\mathbf{V}', \mathbf{V}$ respectively satisfying

   $$(X^\perp)^\perp = \text{span (X)}, \quad \text{dim X}^\perp = \text{n} - \text{dim span (X)},$$
   $$(F^\perp)^\perp = \text{span (F)}, \quad \text{dim F}^\perp = \text{n} - \text{dim span (F)}.$$

   (c) Let $\mathbf{U}_1, \ldots, \mathbf{U}_k$ be k-subspaces of either $\mathbf{V}$ or $\mathbf{V}'$. Then

   $$(\cap_{i=1}^k \mathbf{U}_i)^\perp = \sum_{i=1} \mathbf{U}_i^\perp, \quad (\sum_{i=1}^{k} \mathbf{U}_i)^\perp = \cap_{i=1} \mathbf{U}_i^\perp,$$

   (d) For each bases $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}, \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ in $\mathbf{V}, \mathbf{V}'$ respectively there exists unique dual bases $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n\}, \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ in $\mathbf{V}', \mathbf{V}$ respectively such that $\mathbf{g}_i(\mathbf{v}_j) = \mathbf{f}_i(\mathbf{u}_j) = \delta_{ij},\ i, j = 1, \ldots, n$.

   (e) Let $\mathbf{U} \subset \mathbf{V}, \mathbf{W} \subset \mathbf{V}'$ two m-dimensional subspaces. TFAE

      i. $\mathbf{U} \cap \mathbf{W}^\perp = \{\mathbf{0}\}$.

      ii. $\mathbf{U}^\perp \cap \mathbf{W} = \{\mathbf{0}\}$.

      iii. There exists bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}, \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ in $\mathbf{U}, \mathbf{W}$ respectively such that $\mathbf{f}_j(\mathbf{u}_i) = \delta_{ij}, i, j = 1, \ldots, m$.

3. Show Proposition 5.1.2.

4. 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 are basic tools for solving PDE (partial differential equations), using *separation of variables*, i.e. Fourier series.)

5. Let $\mathbf{M} = \mathbb{D}^m, \mathbf{N} = \mathbb{D}^n$. Show

   - $\mathbf{M}\otimes\mathbf{N}$ can be identified with the space of $m \times n$ matrices $\mathbb{D}^{m\times n}$. More precisely each $A \in \mathbb{D}^{m\times n}$ is viewed as a homomorphism $A : \mathbb{D}^n \to \mathbb{D}^m$, where $\mathbb{D}^m$ is identified with $\mathbf{M}'$.
   - The decomposable tensor $\mathbf{m} \otimes \mathbf{n}$ is identified with $\mathbf{m}\mathbf{n}^T$. (Note $\mathbf{m}\mathbf{n}^T$ is indeed rank one matrix.)

6. Prove Proposition 5.1.3.

7. Show that rank τ defined in Definition 5.1.4 is independent of choices of bases in $\mathbf{M}$ and $\mathbf{N}$.

8. Let the assumptions of Proposition 5.1.3 holds. Show that the equalities
$$\tau = \sum_{i=1}^{m} \mathbf{d}_i \otimes (\sum_{j=1}^{n} b_{ij}\mathbf{e}_j) = \sum_{j=1}^{n}(\sum_{i=1}^{m} b_{ij}\mathbf{d}_i) \otimes \mathbf{e}_j$$
   yield (5.1.3).

9. Prove Proposition 5.1.6.

10. Let the assumptions of Proposition 5.1.2 hold. Arrange the basis of $\mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ is the lexicographical order: $\mathbf{d}_1 \otimes \mathbf{e}_1, \ldots, \mathbf{d}_1 \otimes \mathbf{e}_n, \mathbf{d}_2 \otimes \mathbf{e}_1, \ldots, \mathbf{d}_2 \otimes \mathbf{e}_n, \ldots, \mathbf{d}_m \otimes \mathbf{e}_1, \ldots, \mathbf{d}_m \otimes \mathbf{e}_n$. We denote this basis by $[\mathbf{d}_1, \ldots, \mathbf{d}_m] \otimes [\mathbf{e}_1, \ldots, \mathbf{e}_n]$.

    Let $\mathbf{M}_l, \mathbf{N}_l$ be free modules with the bases $[\mathbf{d}_{1,l}, \ldots, \mathbf{d}_{m_l,l}], [\mathbf{e}_{1,l}, \ldots, \mathbf{e}_{n_l,l}]$ for $l = 1, 2$. Let $T_l : \mathbf{M}_l \to \mathbf{N}_l$ be a homomorphism represented by $A_l \in \mathbb{D}^{n_l\times m_l}$ in the above bases for $l = 1, 2$. Show that $T_1 \otimes T_2$ is represented by the matrices $A_1 \otimes A_2$ with respect to the bases $[\mathbf{d}_{1,1}, \ldots, \mathbf{d}_{m_1,1}]\otimes[\mathbf{e}_{1,1}, \ldots, \mathbf{e}_{n_1,1}]$ and $[\mathbf{d}_{1,2}, \ldots, \mathbf{d}_{m_2,2}]\otimes[\mathbf{e}_{1,2}, \ldots, \mathbf{e}_{n_2,2}]$.

11. Let $A \in \mathbb{D}^{m\times n}, B \in \mathbb{D}^{p\times q}$. Show

    - If $m = n$ and A is an upper triangular than $A\otimes B$ is block upper triangular.
    - If $m = n, p = q$ and A and B are upper triangular then $A \otimes B$ is upper triangular.
    - If A and B are diagonal matrices then $A\otimes B$ is a diagonal matrix. In particular $I_m \otimes I_p = I_{mp}$.
    - Let $C \in \mathbb{D}^{l\times m}, D \in \mathbb{D}^{r\times p}$. Then $(C\otimes D)(A\times B) = (CA)\otimes(DB)$.
    - $A \in \mathbf{GL}(m, \mathbb{D}), B \in \mathbf{GL}(p, \mathbb{D})$ then $A \otimes B \in \mathbf{GL}(mp, \mathbb{D})$ and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

- rank A⊗B = rank A rank B. (Use the fact that over the quotient field $\mathbb{F}$ of $\mathbb{D}$, A and B are equivalent to diagonal matrices.)
- Let $m = n, p = q$. Show that det $A \otimes B =$ det A det B.

12. Let $\mathbf{M}$ be a free module with a basis $[\mathbf{d}_1, \ldots, \mathbf{d}_m]$. Identify $\mathbf{M}^{\otimes 2}$ with $\mathbb{D}^{m\times m}$. Show that $\mathrm{Sym}^2\mathbf{M}$ is identified with $\mathrm{S_m}(\mathbb{D}) \subset \mathbb{D}^{\mathrm{m\times m}}$, the module of $m\times m$ symmetric matrices: $A^T = A$, and $\bigwedge^2 \mathbf{M}$ is identified with $\mathbf{AS}(m, \mathbb{D})$, the module of $m\times m$ skew symmetric matrices: $A^T = -A$.

    Assume that 2 is a unit in $\mathbb{D}$. Show the decomposition $\tau \in \mathbf{M}^{\otimes 2}$ as sum of symmetric and skew symmetric tensor is equivalent to the following fact: Any matrix $A \in \mathbb{D}^{m\times m}$ is of the form $A = 2^{-1}(A + A^T) + 2^{-1}(A - A^T)$, which is the unique decomposition to a sum of symmetric and skew symmetric matrices.

13. - Prove Proposition 5.1.9.
    - Show that $(\mathrm{Sym}^2\mathbf{M}, \mathrm{Sym}^2\mathbf{N})$ and $(\bigwedge^2 \mathbf{M}, \bigwedge^2 \mathbf{N})$ are the only invariant pairs of submodules of $T^{\otimes 2}$ for all choices of $T \in$ Hom $(\mathbf{M}, \mathbf{N})$.

14. Let $\mathbf{M}$ be a module over the domain $\mathbb{D}$. Let $X \subseteq \mathbf{M}$ be a subset of $\mathbf{M}$. Then span X is the set of all finite linear combinations of the elements from X.

    - Show that span X is a submodule of $\mathbf{M}$.
    - span X is called the submodule generated by X.

15. Let X be a nonempty set. For a given domain $\mathbb{D}$ denote by $\mathbf{M}_{\mathbb{D}}(X)$ the free $\mathbb{D}$-module generated by X. That is $\mathbf{M}_{\mathbb{D}}(X)$ has a set of elements $\mathbf{e}(x), x \in X$ with the following properties:

    - For each finite nonempty subset $Y \subseteq X$, the set of vectors $e(y), y \in Y$ are linearly independent.
    - $M_{\mathbb{D}}(X)$ is generated by $\{\mathbf{e}(x), x \in X\}$.

16. Let $\mathbf{M}, \mathbf{N}$ be two modules over an integral domain $\mathbb{D}$. Let $\mathbf{P}$ be the free module generated by $\mathbf{M} \times \mathbf{N} := \{(\mathbf{m}, \mathbf{n}) : \mathbf{m} \in \mathbf{M}, \mathbf{n} \in \mathbf{N}\}$. (See Problem 15.) Let $\mathbf{Q} \subseteq \mathbf{P}$ generated by the elements of the form

    $$\mathbf{e}((a\mathbf{m}_1{+}b\mathbf{m}_2, c\mathbf{n}_1{+}d\mathbf{n}_2)){-}ac\mathbf{e}((\mathbf{m}_1, \mathbf{n}_1)){-}ad\mathbf{e}((\mathbf{m}_1, \mathbf{n}_2)){-}bc\mathbf{e}((\mathbf{m}_2, \mathbf{n}_1){-}bd\mathbf{e}(\mathbf{m}_2, \mathbf{n}_2)),$$

    for all $a, b, c, d \in \mathbb{D}$ and $\mathbf{m}_1, \mathbf{m}_2 \in \mathbf{M}, \mathbf{n}_1, \mathbf{n}_2 \in \mathbf{N}$ Then $\mathbf{M} \otimes_{\mathbb{D}} \mathbf{N} := \mathbf{P}/\mathbf{Q}$ is called the tensor product of $\mathbf{M}$ and $\mathbf{N}$ over $\mathbb{D}$.

    Show that if $\mathbf{M}, \mathbf{N}$ are two free finite dimensional modules then the above definition of $\mathbf{M} \otimes_{\mathbb{D}} \mathbf{N}$ is isomorphic to Definition 5.1.1.

5.2 Tensor product of several free modules

Definition 5.2.1 *Let $\mathbf{M}_i$ be free finite dimensional modules over a domain $\mathbb{D}$ for $i = 1,\ldots,k$, where $k \ge 2$. Then $\mathbf{M} := \otimes_{i=1}^{k}\mathbf{M}_i = \mathbf{M}_1 \otimes \mathbf{M}_2 \otimes \ldots \otimes \mathbf{M}_k$ is the* tensor product space *of $\mathbf{M}_1,\ldots,\mathbf{M}_k$ is defined as follows. For $k = 2$ $\mathbf{M}_1 \otimes \mathbf{M}_2$ is defined in Definition 5.1.1. For $k \ge 3$ $\otimes_{i=1}^{k}\mathbf{M}_i$ is defined recursively as $(\otimes_{i=1}^{k-1}\mathbf{M}_i) \otimes \mathbf{M}_k$.*

Note that from now on we suppress in our notation the dependence on $\mathbb{D}$. When we need to emphasize $\mathbb{D}$ we use the notation $\mathbf{M}_1 \otimes_{\mathbb{D}} \ldots \otimes_{\mathbb{D}} \mathbf{M}_k$.
$\mathbf{M}$ is spanned by the *decomposable tensors*

$$\otimes_{i=1}^{k}\mathbf{m}_i := \mathbf{m}_1 \otimes \mathbf{m}_2 \otimes \ldots \otimes \mathbf{m}_k, \quad \mathbf{m}_i \in \mathbf{M}_i, i = 1,\ldots,k,$$

called also *rank one tensors*. One have the basic identity:

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

Furthermore, the above decomposable tensor is multilinear in each variable. Clearly

(5.2.1) $\otimes_{i=1}^{k}\mathbf{m}_{j_i,i},\ j_i = 1,\ldots,m_i, i = 1,\ldots,k$ is a basis of $\otimes_{i=1}^{k}\mathbf{M}_i$ if $\mathbf{m}_{1,i},\ldots,\mathbf{m}_{m_i,i}$ is a basis of $\mathbf{M}_i$ for $i = 1,\ldots,k$.

Hence

$$\dim\ \otimes_{i=1}^{k}\mathbf{M}_i = \prod_{i=1}^{k}\dim\mathbf{M}_i. \tag{5.2.2}$$

Thus

$$\alpha = \sum_{j_1=j_2=\cdots=j_k=1}^{m_1,m_2,\ldots,m_k} a_{j_1j_2\ldots j_k} \otimes_{i=1}^{k}\mathbf{m}_{j_i,i}, \text{ for any } \alpha \in \otimes_{i=1}^{k}\mathbf{M}_i. \tag{5.2.3}$$

Denote

$$\mathbb{D}^{m_1\times\ldots\times m_k} := \otimes_{i=1}^{k}\mathbb{D}^{m_i}, \text{ for } k \in \mathbb{N} \text{ and } m_i \in \mathbb{N}, i = 1,\ldots,k. \tag{5.2.4}$$

$\mathcal{A} \in \mathbb{D}^{m_1\times\ldots\times m_k}$ is given as $\mathcal{A} := [a_{j_1\ldots j_k}]_{j_1=\ldots=j_k=1}^{m_1,\ldots,m_k}$, where $a_{j_1\ldots j_k} \in \mathbb{D}, j_i = 1,\ldots,m_i, i = 1,\ldots,k$. $\mathcal{A}$ is called a $k-$ *tensor*. So 1-tensor is a vector and 2-tensor is a matrix.

In particular $\otimes_{i=1}^{k}\mathbf{M}_i$ is isomorphic to $\mathbb{D}^{m_1\times\ldots\times m_k}$. Furthermore, after choosing a basis of $\otimes_{i=1}^{k}\mathbf{M}_i$ of the form (5.2.1) we correspond to each $\tau \in \otimes_{i=1}^{k}\mathbf{M}_i$ of the form (5.2.3) the tensor $\mathcal{A} = [a_{j_1\ldots j_k}]_{j_1=\ldots=j_k=1}^{m_1,\ldots,m_k} \in \mathbb{D}^{m_1\times\ldots\times m_k}$.

Proposition 5.2.2 *Let $\mathbf{M}_i, \mathbf{N}_i, i = 1, \ldots, k$ be free finite dimensional modules over $\mathbb{D}$. Let $T_i : \mathbf{M}_i \to \mathbf{N}_i, i = 1, \ldots, k$ be homomorphisms. Then there exists a unique homomorphism on $T : \otimes_{i=1}^k \mathbf{M}_i \to \otimes_{i=1}^k \mathbf{N}_i$ such that $T(\otimes_{i=1}^k \mathbf{m}_i) = \otimes_{i=1}^k (T_i \mathbf{m}_i)$ for all $\mathbf{m}_i \in \mathbf{M}_i, i = 1, \ldots, k$. This homomorphism is denoted by $\otimes_{i=1}^k T_i$.*

Suppose furthermore that $\mathbf{W}_i, i = 1, \ldots, k$ are free finite dimensional $\mathbb{D}$-modules, and $P_i : \mathbf{N}_i \to \mathbf{W}_i, i = 1, \ldots, k$ are homomorphisms. Then $(\otimes_{i=1}^k P_i)(\otimes_{i=1}^k T_i) = \otimes_{i=1}^k (P_i T_i)$.

See Problem 2.

Since each homomorphism $T_i : \mathbf{M}_i \to \mathbf{N}_i, i = 1, \ldots, k$ is represented by a matrix, one can reduce the definition of $\otimes_{i=1}^k T_i$ to the notion of *tensor* product of k matrices.

Definition 5.2.3 *Let $A_i = [a_{lj,i}]_{l,j=1}^{m_i,n_i} \in \mathbb{D}^{m_i \times n_i}, i = 1, \ldots, k$. Then the Kronecker product $A := \otimes_{i=1}^k A_i \in \mathbb{D}^{m_1 \ldots m_k \times n_1 \ldots n_k}$ is the matrix with the entries*

$$A = [a_{(l_1,\ldots,l_k)(j_1,\ldots,j_k)}], \; a_{(l_1,\ldots,l_k)(j_1,\ldots,j_k)} := \prod_{i=1}^k a_{l_i j_i, i},$$
$$\textit{for } l_i = 1, \ldots, m_i, \; j_i = 1, \ldots, n_i, \; i = 1, \ldots, k.$$

where the indices $(l_1, \ldots, l_k), l_i = 1, \ldots, m_i, i = 1, \ldots, k$, and the indices $(j_1, \ldots, j_k)$, $j_i = 1, \ldots, n_i, i = 1, \ldots, k$ are arranged in the lexicographical order.

It is straightforward to show that the above tensor product of matrices can be recursively defined by the Kronecker product of two matrices as defined in Definition 5.1.7. See Problem 3. The tensor products of k matrices have similar properties as in the case $k = 2$. See Problem 4.

We now consider the k-symmetric and k-exterior products of a free finite dimensional module $\mathbf{M}$. In view of the previous section we may assume that $k \ge 3$. Denote by $\mathcal{S}_k$ the permutation group of k elements of $\{1, \ldots, k\}$, i.e. $\mathcal{S}_k$ is the group of injections $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$. Recall that $\text{sgn}(\sigma) \in \{1, -1\}$ is the sign of the permutation. That is $\text{sgn}(\sigma) = 1$ is σ is an even permutation, and $\text{sgn}(\sigma) = -1$ is σ is an odd permutation.

Definition 5.2.4 *Let $\mathbf{M}$ be a free finite dimensional module over $\mathbb{D}$ and $2 \le k \in \mathbf{M}$. Denote $\mathbf{M}^{\otimes k} := \otimes_{i=1}^k \mathbf{M}_i$, where $\mathbf{M}_i = \mathbf{M}$ for $i = 1, \ldots, k$. The submodule $\mathrm{Sym}^k \mathbf{M} \subset \mathbf{M}^{\otimes k}$, called a k-*symmetric power *of $\mathbf{M}$, is spanned by tensors of the form*

$$\text{sym}^k(\mathbf{m}_1, \ldots, \mathbf{m}_k) := \sum_{\sigma \in \mathcal{S}_k} \otimes_{i=1}^k \mathbf{m}_{\sigma(i)}, \tag{5.2.5}$$

for all $\mathbf{m}_i \in \mathbf{M}, i = 1, \ldots, k$. $\mathrm{sym}^k(\mathbf{m}_1, \ldots, \mathbf{m}_k)$ *is called a k-symmetric product of* $\mathbf{m}_1, \ldots, \mathbf{m}_k$*, or simply a symmetric product. Any tensor* $\tau \in \mathrm{Sym}^{\mathrm{k}}\mathbf{M}$ *is a called a k-symmetric tensor, or simply a symmetric tensor. The subspace* $\bigwedge^k \mathbf{M} \subset \mathbf{M}^{\otimes k}$*, called k-exterior power of* $\mathbf{M}$*, is spanned by all tensors of the form*

$$\wedge_{i=1}^k \mathbf{m}_i = \mathbf{m}_1 \wedge \ldots \wedge \mathbf{m}_k := \sum_{\sigma \in \mathcal{S}_k} \mathrm{sgn}(\sigma) \otimes_{i=1}^k \mathbf{m}_{\sigma(i)} \tag{5.2.6}$$

for all $\mathbf{m}_i \in \mathbf{M}, i = 1, \ldots, k$. $\wedge_{i=1}^k \mathbf{m}_i$ *is called k- wedge product of* $\mathbf{m}_1, \ldots, \mathbf{m}_k$. *Any vector* $\tau \in \bigwedge^k \mathbf{M}$ *is a called a k-skew symmetric tensor, or simply a skew symmetric tensor.*

Proposition 5.2.5 *Let* $\mathbf{M}, \mathbf{N}$ *be free finite dimensional module over* $\mathbb{D}$. *Let* $T : \mathrm{Hom}\,(\mathbf{M}, \mathbf{N})$. *For* $k \in \mathbb{N}$ *let* $T^{\otimes k} : \mathbf{M}^{\otimes k} \to \mathbf{N}^{\otimes k}$ *be* $\underbrace{T \otimes \ldots \otimes T}_{k}$. *Then*

$$T^{\otimes k} : \mathrm{Sym}^{\mathrm{k}}\mathbf{M} \to \mathrm{Sym}^{\mathrm{k}}\mathbf{N}, \quad T^{\otimes k} : \bigwedge^{k} \mathbf{M} \to \bigwedge^{k} \mathbf{N}.$$

See Problem 5.

Definition 5.2.6 *Let* $\mathbf{M}, \mathbf{N}$ *be free finite dimensional modules over* $\mathbb{D}$. *Let* $T : \mathrm{Hom}\,(\mathbf{M}, \mathbf{N})$. *Then* $\wedge^k T \in \mathrm{Hom}\,(\bigwedge^k \mathbf{M}, \bigwedge^k \mathbf{N})$ *is defined as the restriction of* $T^{\otimes k}$ *to* $\bigwedge^k \mathbf{M}$.

Proposition 5.2.7 *Let* $\mathbf{M}, \mathbf{N}$ *be free finite dimensional modules over* $\mathbb{D}$. *Let* $T : \mathrm{Hom}\,(\mathbf{M}, \mathbf{N})$. *Then*

1. *Let* $[\mathbf{d}_1, \ldots, \mathbf{d}_m], [\mathbf{e}_1, \ldots, \mathbf{e}_n]$ *be bases in* $\mathbf{M}, \mathbf{N}$ *respectively. Assume that T is represented by the matrix* $A = [a_{ij}] \in \mathbb{D}^{n \times m}$ *in these bases. Then* $\wedge^k T$ *represented in the bases*

   $$\wedge_{i=1}^k \mathbf{d}_{j_i},\ 1 \le j_1 < \ldots < j_k \le m,\ \wedge_{i=1}^k \mathbf{e}_{l_i},\ 1 \le l_1 < \ldots < l_k \le n.$$

   by the matrix $\wedge^k A \in \mathbb{D}^{\binom{n}{k} \times \binom{m}{k}}$*, where the entry* $((l_1, \ldots, l_k), (j_1, \ldots, j_k))$ *of* $\wedge^k A$ *is the the* $k \times k$ *minor based of A on the* $(l_1, \ldots, l_k)$ *rows and* $(j_1, \ldots, j_k)$ *columns of A.*

2. *Let* $\mathbf{L}$ *be a free finite dimensional module and assume that* $S : \mathrm{Hom}\,(\mathbf{L}, \mathbf{M})$. *Then* $\wedge^k(TS) = (\wedge^k T)(\wedge^k S)$.

See Problem 5.

Remark 5.2.8 *In the classical matrix books as [Gan59] and [MaM64] the matrix $\wedge^k A$ is called the kth compound matrix or kth adjugate of A.*

Proposition 5.2.9 *Let $\mathbf{M}_1,\ldots,\mathbf{M}_k, \mathbf{M} := \otimes_{i=1}^k \mathbf{M}_i$ be free finite dimensional modules over $\mathbb{D}$ with bases given in (5.2.1). Let $[\mathbf{n}_{1,i},\ldots,\mathbf{n}_{m_i,i}] = [\mathbf{m}_{1,i},\ldots,\mathbf{m}_{m_i,i}]T_i^{-1}$, $T_i = [t_{lj,i}] \in \mathbf{GL}(m_i,\mathbb{D})$ be another basis of $\mathbf{M}_i$ for $i=1,\ldots,m_i$. Let $\alpha \in \mathbf{M}$ be given by (5.2.3). Then*

$$(5.2.7)\alpha = \sum_{l_1=\cdots=l_k=1}^{m_1,\ldots,m_k} b_{l_1\ldots l_k} \otimes_{i=1}^k \mathbf{n}_{l_i,i}, \text{ where}$$

$$b_{l_1,\ldots,l_k} = \sum_{j_1,\ldots,j_k=1}^{m_1,\ldots,m_k} (\prod_{i=1}^k t_{l_ij_i,i}) a_{j_1\ldots j_k} \text{ for } l_i = 1,\ldots,m_i, i=1,\ldots,k.$$

That is if $\mathcal{A} := [a_{j_1\ldots j_k}], \mathcal{B} := [b_{l_1\ldots l_k}]$ then $\mathcal{B} = (\otimes_{i=1}^k T_i)\mathcal{A}$.

Definition 5.2.10 *Let $\mathbf{M}_1,\ldots,\mathbf{M}_k$ be free finite dimensional modules over a domain $\mathbb{D}$. Let $\tau \in \otimes_{i=1}^k \mathbf{M}_i$. The tensor rank of τ, denoted by* $\mathrm{Rank}\,\tau$, *is the minimal R such that $\tau = \sum_{l=1}^R \otimes_{i=1}^k \mathbf{m}_{l,i}$ for some $\mathbf{m}_{l,i} \in \mathbf{M}_i, l = 1,\ldots,R, i=1,\ldots,k$.*

We shall see that for $k \ge 3$ it is hard to determine the tensor rank a general k-tensor even in the case $\mathbb{D} = \mathbb{C}$.

Let $\mathbf{M}$ be a $\mathbb{D}$-module, and let $\mathbf{M}' = \mathrm{Hom}(\mathbf{M},\mathbb{D})$ the *dual* module of $\mathbf{M}$. For $\mathbf{m} \in \mathbf{M}, \mathbf{g} \in \mathbf{M}'$ we denote $\langle \mathbf{m},\mathbf{g}\rangle := \mathbf{g}(\mathbf{m})$. Let

$$\mathbf{m}_1,\ldots,\mathbf{m}_k \in \mathbf{M}, \quad \mathbf{g}_1,\ldots,\mathbf{g}_k \in \mathbf{M}'.$$

It is straightforward to show

$$(5.2.8)\langle \mathbf{m}_1 \wedge \ldots \wedge \mathbf{m}_k, \mathbf{g}_1 \wedge \ldots \wedge \mathbf{g}_k\rangle = k!\langle \otimes_{i=1}^k, \mathbf{g}_1 \wedge \ldots \wedge \mathbf{g}_k\rangle =$$

$$k!\det \begin{bmatrix} \langle \mathbf{m}_1,\mathbf{g}_1\rangle & \ldots & \langle \mathbf{m}_1,\mathbf{g}_k\rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{m}_k,\mathbf{g}_1\rangle & \ldots & \langle \mathbf{m}_k,\mathbf{g}_k\rangle \end{bmatrix}$$

See Problem 8b.

Assume that $\mathbf{M}$ is an m-dimensional free module over $\mathbb{D}$, with the basis $\mathbf{d}_1,\ldots,\mathbf{d}_m$. Recall that $\mathbf{M}'$ is an m-dimensional free module with the *dual basis* $\mathbf{f}_1,\ldots,\mathbf{f}_m$:

$$(5.2.9) \qquad \langle \mathbf{d}_i,\mathbf{f}_j\rangle = \mathbf{f}_j(\mathbf{d}_i) = \delta_{ij}, \quad i,j = 1,\ldots,m.$$

Let $\mathbf{M}_1, \ldots, \mathbf{M}_k, \mathbf{M} := \otimes_{i=1}^k \mathbf{M}_i$ be free finite dimensional modules over $\mathbb{D}$ with bases given in (5.2.1). Let $\mathbf{f}_{1,i}, \ldots, \mathbf{f}_{m_i,i}$ be the dual basis of $\mathbf{M}_i'$ for $i = 1, \ldots, k$. Then $\mathbf{M}'$ is isomorphic to $\otimes_{i=1}^k \mathbf{M}_i'$, where we assume that (5.2.10)

$$\langle \otimes_{i=1}^k \mathbf{m}_i, \otimes_{i=1}^k \mathbf{g}_i \rangle := \prod_{i=1}^k \langle \mathbf{m}_i, \mathbf{g}_i \rangle, \ \mathbf{m}_i \in \mathbf{M}_i, \mathbf{g}_i \in \mathbf{M}', \ i = 1, \ldots, k.$$

In particular, $\mathbf{M}'$ has the dual basis $\otimes_{i=1}^k \mathbf{f}_{j_i,i}, j_i = 1, \ldots, m_i, i = 1, \ldots, k$.

Assume that $\mathbf{d}_1, \ldots, \mathbf{d}_m$ is a basis of $\mathbf{M}$ and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ is the dual basis of $\mathbf{M}'$. Note that $\bigwedge^k \mathbf{M}'$ is a submodule of $(\bigwedge^k \mathbf{M})'$. See Problem 8c. Note that if $\mathbb{Q} \subseteq \mathbb{D}$ then $\bigwedge^k \mathbf{M}' = (\bigwedge^k \mathbf{M})'$.

Let $\mathbf{N}$ be a module over $\mathbb{D}$ of dimension n, as defined in Problem 1.6.1. Assume that $\mathbf{M} \subseteq \mathbf{N}$ is a submodule of dimension $m \le n$. For any $k \in \mathbb{N}$ we view $\bigwedge^k \mathbf{M}$ as a submodule of $\bigwedge^k \mathbf{N}$. $\bigwedge^0 \mathbf{M} := 1$, $\bigwedge^m \mathbf{M}$ is a one dimensional module, while for $k > m$ it is agreed that $\bigwedge^k \mathbf{M}$ is a trivial subspace consisting of zero vector. (See Problem 10.)

Let $\mathbf{O} \subseteq \mathbf{N}$ be another submodule of $\mathbf{N}$. Then $(\bigwedge^p \mathbf{M}) \bigwedge (\bigwedge^q \mathbf{O})$ is a submodule of $\bigwedge^{p+q}(\mathbf{M} + \mathbf{O})$ of $\bigwedge^{p+q} \mathbf{N}$, spanned by $(\mathbf{m}_1 \wedge \ldots \mathbf{m}_p) \wedge (\mathbf{o}_1 \wedge \ldots \wedge \mathbf{o}_q)$, where $\mathbf{m}_1, \ldots, \mathbf{m}_p \in \mathbf{U}, \mathbf{o}_1, \ldots, \mathbf{o}_q \in \mathbf{O}$ for $p, q \ge 1$. If $p = 0$ or $q = 0$ then $(\bigwedge^p \mathbf{M}) \bigwedge (\bigwedge^q \mathbf{O})$ is equal to $\bigwedge^q \mathbf{O}$ or $\bigwedge^p \mathbf{M}$ respectively.

In in the next sections we need the following lemma

Lemma 5.2.11 *Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$. Assume that $0 \le p_1, p_2, \ 1 \le q_1, q_2, k := p_1 + q_1 = p_2 + q_2 \le n$. Suppose that $\mathbf{U}_1, \mathbf{U}_2, \mathbf{W}_1, \mathbf{W}_2$ are subspaces of $\mathbf{V}$ such that $\dim \mathbf{U}_i = p_i, \dim \mathbf{W}_i \ge q_i$ for $i = 1, 2$ and $\mathbf{U}_1 \cap \mathbf{W}_1 = \mathbf{U}_2 \cap \mathbf{W}_2 = \{\mathbf{0}\}$. Then*

$$(\overset{p_1}{\bigwedge} \mathbf{U}_1) \bigwedge (\overset{q_1}{\bigwedge} \mathbf{W}_1) \cap (\overset{p_2}{\bigwedge} \mathbf{U}_2) \bigwedge (\overset{q_2}{\bigwedge} \mathbf{W}_1) \ne \{\mathbf{0}\} \tag{5.2.11}$$

if and only if the following condition holds. There exists a subspace $\mathbf{V}_1 \subseteq \mathbf{V}$ of dimension k at such that

$$\mathbf{U}_1 \subset \mathbf{V}_1, \ \mathbf{U}_2 \subset \mathbf{V}_1, \ \mathbf{V}_1 \subseteq (\mathbf{U}_1 + \mathbf{W}_1), \ \mathbf{V}_1 \subseteq (\mathbf{U}_2 + \mathbf{W}_2). \tag{5.2.12}$$

Proof. Assume first that (5.2.11) holds. Note that

$$(\overset{p_1}{\bigwedge} \mathbf{U}_1) \bigwedge (\overset{q_1}{\bigwedge} \mathbf{W}_1) \subseteq \overset{k}{\bigwedge} (\mathbf{U}_1 + \mathbf{W}_1), \ (\overset{p_2}{\bigwedge} \mathbf{U}_2) \bigwedge (\overset{q_2}{\bigwedge} \mathbf{W}_2) \subseteq \overset{k}{\bigwedge} (\mathbf{U}_2 + \mathbf{W}_2).$$

Let $\mathbf{V}_2 := (\mathbf{U}_1 + \mathbf{W}_1) \cap (\mathbf{U}_2 + \mathbf{W}_2)$. Problem 10a yields that

$$(\overset{p_1}{\bigwedge} \mathbf{U}_1) \bigwedge (\overset{q_1}{\bigwedge} \mathbf{W}_1) \cap (\overset{p_2}{\bigwedge} \mathbf{U}_2) \bigwedge (\overset{q_2}{\bigwedge} \mathbf{W}_1) \subseteq \overset{k}{\bigwedge} \mathbf{V}_2. \tag{5.2.13}$$

The assumption (5.2.11) implies that dim $\mathbf{V}_2 \geq k$. We now show that $\mathbf{U}_1 \subset \mathbf{V}_2$. Assume to the contrary that dim $\mathbf{U}_1 \cap \mathbf{V}_2 = i < p_1$. Choose a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ such that in $\mathbf{V}$ such that $\mathbf{v}_1, \ldots, \mathbf{v}_{p_1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{p_1+1}, \ldots, \mathbf{v}_r$ is a basis of $\mathbf{U}_1$ and $\mathbf{V}_2$ respectively. Observe that the span of vectors $\mathbf{v}_1 \wedge \ldots \wedge \mathbf{v}_{p_1} \wedge \mathbf{v}_{i_1} \ldots \mathbf{v}_{i_{q_1}}$ for $p_1 < i_1 < \ldots < i_{q_1} \leq n$ contain the subspace $(\bigwedge^{p_1} \mathbf{U}_1) \bigwedge (\bigwedge^{q_1} \mathbf{W}_1)$. On the other hand the subspace $\bigwedge^k \mathbf{V}_2$ is has a basis formed by the exterior products of k vectors out of $\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{v}_{p_1+1}, \ldots, \mathbf{v}_r$. Hence $((\bigwedge^{p_1} \mathbf{U}_1) \bigwedge (\bigwedge^{q_1} \mathbf{W}_1)) \cap \bigwedge^k \mathbf{V}_2 = \{\mathbf{0}\}$, which contradicts (5.2.11-5.2.13). So $\mathbf{U}_1 \subset \mathbf{V}_2$. Similarly $\mathbf{U}_2 \subset \mathbf{V}_2$.

Next we claim that dim $(\mathbf{U}_1 + \mathbf{U}_2) \leq k$. Assume to the contrary that dim $(\mathbf{U}_1 + \mathbf{U}_2) = j > k$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is a basis of $\mathbf{V}$, such that

$$\mathbf{u}_1, \ldots, \mathbf{u}_{p_1} \text{ and } \mathbf{u}_1, \ldots, \mathbf{u}_{p_1+p_2-j}, \mathbf{u}_{p_1+1}, \ldots, \mathbf{u}_j$$

are bases of $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively. Then $(\bigwedge^{p_1} \mathbf{U}_1) \bigwedge (\bigwedge^{q_1} \mathbf{W}_1)$ is spanned by $\binom{n-p_1}{q_1}$ linearly independent vectors $\mathbf{u}_{i_1} \wedge \ldots \mathbf{u}_{i_k}$, where $1 \leq i_1 < \ldots < i_k \leq n$ and $\{1, \ldots, p_1\} \subset \{i_1, \ldots, i_k\}$. Similarly, $(\bigwedge^{p_2} \mathbf{U}_2) \bigwedge (\bigwedge^{q_2} \mathbf{W}_2)$ is spanned by $\binom{n-p_2}{q_2}$ linearly independent vectors $\mathbf{u}_{j_1} \wedge \ldots \mathbf{u}_{j_k}$, where $1 \leq j_1 < \ldots < j_k \leq n$ and $\{1, \ldots, p_1 + p_2 - j, p_1 + 1, \ldots, j\} \subset \{i_1, \ldots, i_k\}$. Since $j > k$ it follows that these two subset of vectors of the full set of the basis of $\bigwedge^k V$ do not have any common vector, which contradicts (5.2.11). So dim $(\mathbf{U}_1 + \mathbf{U}_2) \leq k$. Choose $\mathbf{V}_1$ any k dimensional subspace of $\mathbf{V}_2$ which contains $\mathbf{U}_1 + \mathbf{U}_2$.

Vice versa, suppose that $\mathbf{V}_1$ is a k-dimensional subspace of $\mathbf{V}$ satisfying (5.2.12). So $\bigwedge^k \mathbf{V}_1$ is a one dimensional subspace which is contained in $(\bigwedge^{p_i} \mathbf{U}_i) \bigwedge (\bigwedge^{q_i} \mathbf{W}_i)$ for $i = 1, 2$. Hence (5.2.11) holds.

□

Problems

1. Let $\mathbf{M}_1, \ldots, \mathbf{M}_k$ be free finite dimensional modules over $\mathbb{D}$. Show that for any $\sigma \in \mathcal{S}_k$ $\otimes_{i=1}^k \mathbf{M}_{\sigma(i)}$ is isomorphic to $\otimes_{i=1}^k \mathbf{M}_i$.

2. Prove Proposition 5.2.2.

3. Show

   - Let $A \in \mathbb{D}^{m \times n}$ and $B \in \mathbb{D}^{p \times q}$. Then the definitions of $A \otimes B$ given by Definitions 5.1.7 and 5.2.3 coincide.

- Let the assumptions of Definition 5.2.3 hold. Assume that $k \geq 3$. Then the recursive definition of $\otimes_{i=1}^k A_i := (\otimes_{i=1}^{k-1} A_i) \otimes A_k$ coincides with the definition of $\otimes_{i=1}^k A_i$ given in Definition 5.2.3.

4. Let $A_i \in \mathbb{D}^{m_i \times n_i}, i = 1, \ldots, k \geq 3$. Show

   - $\otimes_{i=1}^k (a_i A_i) = (\prod_{i=1}^k a_i) \otimes_{i=1}^k A_i$.
   - $(\otimes_{i=1}^k A_i)^T = \otimes_{i=1}^k A_i^T$.
   - If $m_i = n_i$ and A_i is an upper triangular for $i = 1, \ldots, k$ then $\otimes_{i=1}^k A_i$ is upper triangular.
   - If $A_1, \ldots, A_k$ are diagonal matrices then $\otimes_{i=1}^k A_i$ is a diagonal matrix. In particular $\otimes_{i=1}^k I_{m_i} = I_{m_1 \ldots m_k}$.
   - Let $B_i \in \mathbb{D}^{l_i \times m_i}, i = 1, \ldots, k$. Then $(\otimes_{i=1}^k B_i)(\otimes_{i=1}^k A_i) = \otimes_{i=1}^k (B_i A_i)$.
   - $A_i \in \mathbf{GL}(m_i, \mathbb{D}), i = 1, \ldots, k$ then $\otimes_{i=1}^k A_i \in \mathbf{GL}(m_1 \ldots m_k, \mathbb{D})$ and $(\otimes_{i=1}^k A_i)^{-1} = \otimes_{i=1}^k A_i^{-1}$.
   - $\mathrm{rank}\ \otimes_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{A} = \prod_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{rank\ A_i}$.
   - For $m_i = n_i, i = 1, \ldots, k$, det $\otimes_{i=1}^k A_i = \prod_{i=1}^k (\det A_i)^{\frac{\prod_{j=1}^k m_j}{m_i}}$.

5. Prove Proposition 5.2.7.

6. (a) Let $A \in \mathbb{D}^{m \times n}, B \in \mathbb{D}^{n \times p}$. Show that $\wedge^k AB = \wedge^k A \wedge^k B$ for any $k \in [1, \min(m, n, p)] \cap \mathbb{N}$.

   (b) Let $A \in \mathbb{D}^{n \times n}$. Then $\wedge^k A$ is upper triangular, lower triangular, diagonal if A is upper triangular, lower triangular, diagonal respectively.

   (c) $\wedge^k I_n = I_{\binom{n}{k}}$.

   (d) If $A \in \mathbf{GL}(n, \mathbb{D})$ then $\wedge^k A \in \mathbf{GL}(\binom{n}{k}, \mathbb{D})$ and $(\wedge^k A)^{-1} = \wedge^k A^{-1}$.

7. Let $\mathbb{F}$ be an algebraically closed field. Recall that over an algebraically closed $A \in \mathbb{F}^{n \times n}$ is similar to an upper triangular matrix.

   (a) Let $A_i \in \mathbb{F}^{n_i \times n_i}$ for $i = 1, \ldots, k$. Show that there exists $T_i \in \mathbf{GL}(n_i, \mathbb{F})$ such that $(\otimes_{i=1}^k T_i)(\otimes_{i=1}^k A_i)(\otimes_{i=1}^k T_i)^{-1}$ is an un upper triangular matrix. Furthermore, let $\lambda_{1,i}, \ldots, \lambda_{n_i,i}$ be the eigenvalues of A_i, counted with their multiplicities. Then $\prod_{i=1}^k \lambda_{j_i,i}$ for $j_i = 1, \ldots, n_i, i = 1, \ldots, k$ are the eigenvalues of $\otimes_{i=1}^k A_i$ counted with their multiplicities.

(b) Let $A \in \mathbb{F}^{n\times n}$ and assume that $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A counted with their multiplicities. Show that $\prod_{i=1}^{k} \lambda_{j_i}$ for $1 \le j_1 < \ldots < j_k \le n$ are all the eigenvalues of $\wedge^k A$ counted with their multiplicites.

8. Let $\mathbf{M}$ be a finitely generated module over $\mathbb{D}$.

   (a) Let $\mathbf{m}_1, \ldots, \mathbf{m}_k \in \mathbf{M}$. Show that for any $\sigma \in \mathcal{S}_k$ $\mathbf{m}_{\sigma(1)} \wedge \ldots \wedge \mathbf{m}_{\sigma(k)} = \mathrm{sgn}(\sigma)\mathbf{m}_1 \wedge \ldots \wedge \mathbf{m}_k$. In particular, if $\mathbf{m}_i = \sum_{j\ne i} a_j \mathbf{m}_j$ then $\mathbf{m}_1 \wedge \ldots \wedge \mathbf{m}_k = \mathbf{0}$.

   (b) Prove the equality (5.2.8).

   (c) Assume that $\mathbf{d}_1, \ldots, \mathbf{d}_m$ is a basis of $\mathbf{M}$ and $\mathbf{f}_1, \ldots, \mathbf{f}_m$ is a dual basis of $\mathbf{M}'$. Show that $\frac{1}{k!}\mathbf{f}_{i_1} \wedge \ldots \wedge \mathbf{f}_{i_k}, 1 \le i_1 < \ldots < i_k \le m$ can be viewed as a basis for $(\bigwedge^k \mathbf{M})'$ for $k \in [1, m]$.

9. Let $\mathbf{M}$ be an m-dimensional module over $\mathbb{D}$ as defined in Problem 1.6.1. Show

   - $\bigwedge^m \mathbf{M}$ is a 1-dimensional module over $\mathbb{D}$.
   - $\bigwedge^k \mathbf{V}$ is a zero module over $\mathbb{D}$ for $k > m$.

10. (a) Let $\mathbf{V}$ be an finite dimensional vector space over $\mathbb{F}$ and assume that $\mathbf{U}, \mathbf{W}$ are subspaces of $\mathbf{V}$. Show that $\bigwedge^k \mathbf{U} \cap \bigwedge^k \mathbf{W} = \bigwedge^k (\mathbf{U} \cap \mathbf{W})$.

    Hint: Choose a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in $\mathbf{V}$ satisfying the following property. $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\mathbf{v}_1, \ldots, \mathbf{v}_l, \mathbf{v}_{m+1}, \ldots \mathbf{v}_{m+p-l}$ are bases for $\mathbf{U}$ and $\mathbf{W}$ respectively. Recall that $\mathbf{v}_{i_1} \wedge \ldots \wedge \mathbf{v}_{i_k}, 1 \le i_1 < \ldots < i_k \le n$ form a basis in $\bigwedge^k \mathbf{V}$. Observe next that bases of $\mathbf{U}$ and $\mathbf{W}$ are of the form of exterior, (wedge), product of k vectors from $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\mathbf{v}_1, \ldots, \mathbf{v}_l, \mathbf{v}_{m+1}, \ldots \mathbf{v}_{m+p-l}$ respectively.

    (b) Assume that $\mathbf{V}$ is a an n-dimensional module of $\mathbb{D}_b$. Suppose furthermore that $\mathbf{U}, \mathbf{W}$ are finitely generated submodules of $\mathbf{V}$. Show that $\bigwedge^k \mathbf{U} \cap \bigwedge^k \mathbf{W} = \bigwedge^k (\mathbf{U} \cap \mathbf{W})$.

11. Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$ and $\mathbf{U} \subset \mathbf{V}, \mathbf{W} \subset \mathbf{V}'$ be m-dimensional subspaces. Show

    (a) Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}, \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ be bases of $\mathbf{U}, \mathbf{W}$ respectively. Then vanishing of the determinant det $[\langle \mathbf{u}_i, \mathbf{f}_j \rangle]_{i,j=1}^m$ is independent of the choice of bases in $\mathbf{U}, \mathbf{W}$.

    (b) Let $\mathbb{F}$ be a field of infinite characteristic. TFAE

i. $\dim \mathbf{U}^\perp \cap \mathbf{W} > 0$.

ii. $\dim \mathbf{U} \cap \mathbf{W}^\perp > 0$.

iii. $\bigwedge^{m-1} \mathbf{U} \subset (\bigwedge^{m-1} \mathbf{W})^\perp$.

iv. $\bigwedge^{m-1} \mathbf{W} \subset (\bigwedge^{m-1} \mathbf{U})^\perp$.

v. For any bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}, \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ of $\mathbf{U}, \mathbf{W}$ respectively $\langle \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_m, \mathbf{f}_1 \wedge \ldots \wedge \mathbf{f}_m \rangle = 0$.

Hint: If $\dim \mathbf{U}^\perp \cap \mathbf{W} = 0$ use Problem 2(e). If $\dim \mathbf{U}^\perp \cap \mathbf{W} > 0$ choose at least one vector of a basis in $\mathbf{W}$ to be in $\mathbf{U}^\perp \cap \mathbf{W}$ and use (5.2.8).

5.3 Sparse bases of subspaces

Definition 5.3.1 *1. For* $\mathbf{0} \neq \mathbf{x} \in \mathbb{F}^n$ *denote* $\mathrm{span}\,(\mathbf{x})^* := \mathrm{span}\,(\mathbf{x}) \backslash \{\mathbf{0}\}$.

2. The support of $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{F}^n$ *is defined as* $\mathrm{supp}\,(\mathbf{x}) = \{i \in \{1, \ldots, n\} : \ x_i \neq 0\}$.

3. For a nonzero subspace $\mathbf{U} \subseteq \mathbb{F}^n$*, a nonzero vector* $\mathbf{x} \in \mathbf{U}$ *is called elementary if for every* $\mathbf{0} \neq \mathbf{y} \in \mathbf{U}$ *the condition* $\mathrm{supp}\,(\mathbf{y}) \subseteq \mathrm{supp}\,(\mathbf{x})$ *implies* $\mathrm{supp}\,(\mathbf{y}) = \mathrm{supp}\,(\mathbf{x})$. $\mathrm{span}\,(\mathbf{x})^*$ *is called an elementary class, in* $\mathbf{U}$*, if* $\mathbf{x} \in \mathbf{U}$ *is elementary.*

4. Denote by $\mathcal{E}(\mathbf{U})$ *the union of all elementary classes in* $\mathbf{U}$.

5. A basis in $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ *in* $\mathbf{U}$ *is called sparse if* $\mathbf{u}_1, \ldots, \mathbf{u}_m$ *are elementary.*

Proposition 5.3.2 *Let* $\mathbf{U}$ *be a a subspace of* $\mathbb{F}^n$ *of dimension* $m \in [1, n]$. *Then*

1. $\mathbf{x} \in \mathbf{U}$ *is elementary if and only if for each* $\mathbf{0} \neq \mathbf{y} \in \mathbf{U}$ *the condition* $\mathrm{supp}\,(\mathbf{x}) \subseteq \mathrm{supp}\,(\mathbf{x})$ *implies that* $\mathbf{y} \in \mathrm{span}\,(\mathbf{x})^*$.

2. $\mathcal{E}(\mathbf{U})$ *consists of a finite number of elementary classes.*

3. $\mathrm{span}\,(\mathcal{E}(\mathbf{U})) = \mathbf{U}$.

4. For each subset I *of* $\{1, \ldots, n\}$ *of cardinality* $m - 1$ *there exists an elementary* $\mathbf{x} \in \mathbf{U}$ *such that* $\mathrm{supp}\,(\mathbf{x})^c := \{1, \ldots, n\} \backslash \mathrm{supp}\,(\mathbf{x})$ *contains* I.

See Problem 1 for proof.

Definition 5.3.3 *Let* $\mathbb{F}$ *be a field of* 0 *characteristic.*

1. *$A = (a_{ij}) \in \mathbb{F}^{k\times n}$ is called generic if all the entries of A are algebraically independent over $\mathbb{Q}$, i.e. there is no nontrivial polynomial p in kn variable with integer coefficients such that $p(a_{11}, \ldots, a_{kn}) = 0$.*
2. *A is called nondegenerate if all $\min(k, n)$ minors of A are nonzero.*
3. *An $1 \le m$-dimensional subspace $\mathbf{U} \subseteq \mathbb{F}^n$ is called nondegenerate if for $J \subset \{1, \ldots, n\}$ of cardinality $n - m + 1$ there exists a unique elementary set $\mathrm{span}\, \mathbf{x}^*$ such that $J = \mathrm{supp}\,(\mathbf{x})$.*

Lemma 5.3.4 *Let $A \in \mathbb{F}^{k\times n}, 1 \le k < n$ be of rank k. TFAE:*

1. *A is nondegenerate.*
2. *The row space of A, (viewed as a column space of $A^\top$), is nondegenerate.*
3. *The null space of A is nondegenerate.*

Proof. Consider first the column space of $A^\top$ denoted by $\mathbf{U} \subseteq \mathbb{F}^n$. Recall that any vector in $\mathbf{U}$ is of the form $\mathbf{x} = A^\top \mathbf{y}$ for some $\mathbf{y} \in \mathbb{F}^k$. Let $I \subset \{1, \ldots, n\}$ be a set of cardinality $k-1$. Let $B = (A^\top)[I, :] \in \mathbb{F}^{k-1\times k}$ be submatrix of $A^\top$ with the rows indexed by the set I. The condition that $\mathrm{supp}\,(\mathbf{x}) \subseteq I^c$ is equivalent to the condition $B\mathbf{y} = \mathbf{0}$. Since $\mathrm{rank\ B \le k - 1}$ there exists $\mathbf{0} \ne \mathbf{x} \in \mathbf{U}$ such that $\mathrm{supp}\,(\mathbf{x}) \subseteq I^c$. Let $\mathbf{d}$ be defined as in Problem 3.

Assume that $\mathrm{rank\ B < k - 1}$. Then $\mathbf{d} = \mathbf{0}$, see Problem 3(b). Furthermore, it is straightforward to show that for each each $j \in I^c$ there exists a nonzero $\mathbf{x} \in \mathbf{U}$ such that $\mathrm{supp}\,(\mathbf{x}) \subseteq (I \cup \{j\})^c$. So $\det A[; , I \cup \{j\}] = 0$ and A is not degenerate.

Suppose that $\mathrm{rank\ B = k - 1}$, i.e. $\mathbf{d} \ne \mathbf{0}$. Then any nonzero $\mathbf{x} \in \mathbf{U}, \mathrm{supp}\,(\mathbf{x}) \subset I^c$ is in $\mathrm{span}\,(\mathrm{A}^\top \mathbf{d})^*$. Let $j \in I^c$. Expand $\det A[:, I \cup \{j\}$ by the column j to deduce that and $(A^\top \mathbf{d})_j = \pm \det A[:, I \cup \{j\}]$. Thus $\mathrm{supp}\,(\mathbf{x}) = I^c$ if and only $\det A[:, I \cup \{j\}] \ne 0$ for each $j \in I^c$. These arguments show the equivalence of *1* and *2*.

The equivalence of *1* and *3* are shown in a similar way and are discussed in Problem 4.

□

Definition 5.3.5 *Let $\mathcal{J} = \{J_1, \ldots, J_t\}$ be t subsets of $\langle n\rangle$ each of cardinality $m - 1$. Then $\mathcal{J}$ satisfies the m-intersection property provided that*

$$\# \cap_{i\in P} J_i \le m - \#P \textit{ for all } \emptyset \ne P \subseteq \langle t\rangle. \tag{5.3.1}$$

It is known that given a set $\mathcal{J}$ of t satisfying the above assumptions, one can check effectively, i.e. in polynomial time, if $\mathcal{J}$ satisfies the m-intersection property. See Problems 5 - 7.

The aim of this section to prove the following theorem.

Theorem 5.3.6 *Let $\mathbb{F}$ be a field of 0 characteristic and assume that $A \in \mathbb{F}^{k\times n}$ is generic over $\mathbb{Q}$.*

1. *Let $\mathcal{I} = \{I_1, \dots, I_s\}$ denote the collection of $s \le k$ subsets of $\langle n\rangle$ each of cardinality $n-k+1$. Then the elementary vectors $\mathbf{x}(I_1), \dots, \mathbf{x}(I_s)$ in the row space of A with supports $I_1, \dots, I_s$ are linearly independent if and only if $\mathcal{I}' := \{I_1^c, \dots, I_s^c\}$, consisting of the complements of the supports, have k intersection property.*

2. *Let $\mathcal{J} = \{J_1, \dots, J_t\}$ denote the collection of $t \le n-k$ subsets of $\langle n\rangle$ each of cardinality $k+1$. Then the elementary vectors $\mathbf{y}(J_1), \dots, \mathbf{y}(J_t)$ in the null space of A with supports $J_1, \dots, J_t$ are linearly independent if and only if $\mathcal{J}' := \{J_1^c, \dots, J_t^c\}$, consisting of the complements of the supports, have $n-k$ intersection property.*

The proof of this theorem needs a number of auxiliary results.

Lemma 5.3.7 *Let $A \in \mathbb{F}^{k\times n}$ be nondegenerate.*

1. *Let $\mathcal{I} = \{I_1, \dots, I_s\}$ denote the collection of $s \le k$ subsets of $\langle n\rangle$ each of cardinality $n-k+1$. Then the elementary vectors $\mathbf{x}(I_1), \dots, \mathbf{x}(I_s)$ in the row space of A with supports $I_1, \dots, I_s$ are linearly independent if and only if the $k \times s$ submatrix of $\wedge^{k-1}A$ determined by its columns indexed by $I_1^c, \dots, I_s^c$ has rank s.*

2. *Let $\mathbf{b}_1, \dots, \mathbf{b}_{n-k} \in \mathbb{R}^n$ be a basis in the null space of A and denote by $B^\top \in \mathbb{F}^{n\times(n-k)}$ the matrix whose columns are $\mathbf{b}_1, \dots, \mathbf{b}_{n-k}$. Let $\mathcal{J} = \{J_1, \dots, J_t\}$ denote the collection of $t \le n-k$ subsets of $\langle n\rangle$ each of cardinality $k+1$. Then the elementary vectors $\mathbf{y}(J_1), \dots, \mathbf{y}(J_t)$ in the null space of A with supports $J_1, \dots, J_t$ are linearly independent if and only if the $(n-k-1)\times t$ matrix $\wedge^{n-k-1}B$ determined by its columns indexed by $J_1^c, \dots, J_t^c$ has rank t.*

See Problems 8-9 for the proof of the lemma.

Corollary 5.3.8 *et $A \in \mathbb{F}^{k\times n}$ be nondegenerate.*

1. *Let $\mathcal{I} = \{I_1, \dots, I_k\}$ denote the collection of k subsets of $\langle n\rangle$ each of cardinality $n-k+1$. Then the elementary vectors $\mathbf{x}(I_1), \dots, \mathbf{x}(I_s)$ in the row space of A with supports $I_1, \dots, I_s$ not linearly independent if*

and only if the determinant of the full row $k \times k$ submatrix of $\wedge^{k-1}X$ determined by its columns indexed by $I_1^c, \ldots, I_k^c$ is identically zero for any $X \in \mathbb{F}^{k\times n}$

2. *Let $\mathbf{b}_1, \ldots, \mathbf{b}_{n-k} \in \mathbb{R}^n$ be a basis in the null space of A and denote by $B^\top \in \mathbb{F}^{n\times(n-k)}$ the matrix whose columns are $\mathbf{b}_1, \ldots, \mathbf{b}_{n-k}$. Let $\mathcal{J} = \{J_1, \ldots, J_t\}$ denote the collection of $t \le n-k$ subsets of $\langle n\rangle$ each of cardinality $k+1$. Then the elementary vectors $\mathbf{y}(J_1), \ldots, \mathbf{y}(J_{n-k})$ in the null space of A with supports $J_1, \ldots, J_{n-k}$ are linearly independent if and only if the determinant of the full row $(n-k-1)\times(n-k-1)$ submatrix $\wedge^{n-k-1}Y$ determined by its columns indexed by $J_1^c, \ldots, J_{n-k}^c$ is identically zero for any $Y \in \mathbb{F}^{(n-k)\times n}$.*

(One may use Problem 10 to show part *2* of the above Corollary.)

Definition 5.3.9 *Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$. Let $\mathbf{U}_1, \ldots, \mathbf{U}_t \subset \mathbf{V}$ be t subspaces of dimension $m-1$.. Then $\{\mathbf{U}_1, \ldots, \mathbf{U}_t\}$ satisfies the dimension m-intersection property provided that*

$$\text{dim } \cap_{i\in P} \mathbf{U}_i \le m - \#P \textit{ for all } \emptyset \ne P \subseteq \langle t\rangle. \tag{5.3.2}$$

Theorem 5.3.6 follows from the following theorem.

Theorem 5.3.10 *Let $\mathbf{V}$ be an n-dimensional vector space over a field $\mathbb{F}$ of 0 characteristic and $n \ge 2$. Let $2 \le m \in \langle n\rangle$ and assume that $\mathbf{U}_1, \ldots, \mathbf{U}_m \in \mathrm{Gr}_{m-1}(\mathbf{V}')$. Let $\mathcal{W}_m(\mathbf{U}_1, \ldots, \mathbf{U}_m) \subseteq \mathrm{Gr}_m(\mathbf{V})$ be the variety of all subspaces $\mathbf{X} \in \mathrm{Gr}_m(\mathbf{V})$ such that the one dimensional subspace $\mathbf{Y} := \bigwedge^m(\bigwedge^{m-1}\mathbf{X}) \subset \otimes^{m(m-1)}\mathbf{V}$ is orthogonal on the subspace $\mathbf{W} := (\bigwedge^{m-1}\mathbf{U}_1)\bigwedge(\bigwedge^{m-1}\mathbf{U}_2)\bigwedge\cdots\bigwedge(\bigwedge^{m-1}\mathbf{U}_m) \subset \otimes^{m(m-1)}\mathbf{V}'$ of dimension one at most. Then $\mathcal{W}_m(\mathbf{U}_1, \ldots, \mathbf{U}_m)$ is a strict subvariety of $\mathrm{Gr}_m(\mathbf{V})$ if and only if $\mathbf{U}_1, \ldots, \mathbf{U}_m$ satisfy the dimension m-intersection property.*

Proof. Since each $\mathbf{U}_i$ is $m-1$ dimensional we assume that $\bigwedge^{m-1}\mathbf{U}_i = \text{span}\,(\mathbf{w}_i)$ for some $\mathbf{w}_i \in \bigwedge^{m-1}\mathbf{U}_i$ for $i = 1, \ldots, m$. Then $\mathbf{W} = \text{span}\,(\mathbf{w}_1 \wedge \ldots \wedge \mathbf{w}_m)$. Choose a basis $\mathbf{x}_1, \ldots, \mathbf{x}_m$ in $\mathbf{X}$. Let $\mathbf{y}_i$ be the wedge product of $m-1$ vectors from $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}\backslash\{\mathbf{x}_i\}$ for $i = 1, \ldots, m$. Then $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are linearly independent and $\mathbf{Y} = \text{span}\,(\mathbf{y}_1 \wedge \ldots \wedge \mathbf{y}_n)$. The condition that $\mathbf{Y} \perp \mathbf{W}$, i.e. $\mathbf{Y}^\perp \cap \mathbf{W}$ is a nontrivial subspace, is equivalent to the condition

$$\langle \mathbf{y}_1 \wedge \ldots \wedge \mathbf{y}_m, \mathbf{w}_1 \wedge \ldots \wedge \mathbf{w}_m\rangle = m!\text{det}\,(\langle \mathbf{y}_i, \mathbf{w}_j\rangle)_{i,j=1}^m = 0. \tag{5.3.3}$$

See Problem 5.2.11. Since $\mathbb{F}$ has 0 characteristic, the condition (5.3.3) is equivalent to the vanishing of the determinant in the above formula. We will use the formula (5.2.8) for each $\langle \mathbf{y}_i, \mathbf{w}_j \rangle$.

Assume first that $\mathbf{U}_1, \ldots, \mathbf{U}_m$ do not satisfy the dimension intersection property. By interchanging the order of $\mathbf{U}_1, \ldots, \mathbf{U}_m$ if necessary, we may assume that there exists $2 \le p \le m$ such that $\mathbf{Z} := \cap_{j=1}^p \mathbf{U}_j$ has dimension $m-p+1$ at least. Let $\mathbf{Z}_1 \subseteq \mathbf{Z}$ be a subspace of dimension $m-p+1$. Then $\dim \mathbf{X} \cap \mathbf{Z}_1^\perp \ge m-(m-p+1) = p-1$. Let $\mathbf{F} \subseteq \mathbf{X} \cap \mathbf{Z}_1^\perp$ be a subspace of dimension $p-1$. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_m$ is a basis of $\mathbf{X}$ such that $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$ is a basis of $\mathbf{F}$. So $\mathbf{X}_i \subset \mathbf{F}$ for $i = p, \ldots, m$. Hence

$$\mathbf{X}_i \cap \mathbf{U}_j^\perp \supseteq \mathbf{X}_i \cap \mathbf{Z}^\perp \supseteq \mathbf{X}_i \cap \mathbf{Z}_1^\perp \supseteq \mathbf{F} \cap \mathbf{Z}_1^\perp \ne \{\mathbf{0}\} \text{ for } i = p, \ldots, m,\ j = 1, \ldots, p.$$

Thus $\langle \mathbf{y}_i, \mathbf{w}_j \rangle = 0$ for $i = p, \ldots, m,\ j = 1, \ldots, p$. See Problem 5.2.11. Hence any $p \times p$ submatrix $[\langle \mathbf{y}_i, \mathbf{w}_j \rangle]_{i,j=1}^m$, with the set of columns $\langle p \rangle$, must have a a zero row. Expand $\det\, [\langle \mathbf{y}_i, \mathbf{w}_j \rangle]_{i,j=1}^m$ by the columns $\langle p \rangle$ to deduce that this determinant is zero. Hence $\mathcal{W}_m(\mathbf{U}_1, \ldots, \mathbf{U}_m) = \mathrm{Gr}_m(\mathbf{V})$.

We now show by induction on m that if $\mathbf{U}_1, \ldots, \mathbf{U}_m \in \mathrm{Gr}_{m-1}(\mathbf{V}')$ satisfy the dimension m-intersection property then there exists $\mathbf{X} \in \mathrm{Gr}_m(\mathbf{V})$ such that $\dim \mathbf{Y}^\perp \cap \mathbf{W} = 0$, for each $n = m, m+1, \ldots$. Assume that $m = 2$. As $\dim\, (\mathbf{U}_1 \cap \mathbf{U}_2) = 0$ we deduce that $\dim\, (\mathbf{U}_1 + \mathbf{U}_2) = 2$. Let $\mathbf{U}_i = \mathrm{span}\, (\mathbf{u}_i), i = 1, 2$. Then $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a basis in $\mathbf{Z} = \mathrm{span}\, (\mathbf{u}_1, \mathbf{u}_2)$. Hence $\mathbf{Z}^\perp$ is a subspace of $\mathbf{V}$ of dimension $n-2$. Thus there exists a subspace $\mathbf{X} \in \mathrm{Gr}_2(\mathbf{V})$ such that $\dim \mathbf{X} \cap \mathbf{Z}^\perp = 0$. Note that $\bigwedge^{m-1} \mathbf{X} = \mathbf{X}, \bigwedge^{m-1} \mathbf{U}_i = \mathbf{U}_i, i = 1, 2$. Let $\mathbf{x}_1, \mathbf{x}_2$ be a basis in $\mathbf{X}$. The negation of the condition (5.3.3) is equivalent to $\langle \mathbf{x}_1 \wedge \mathbf{x}_2, \mathbf{u}_1 \wedge \mathbf{u}_2 \rangle \ne 0$. Use Problems 5.1.2(e) and 5.2.11 to deduce this negation.

Assume the induction hypothesis that for $2 \le l < n$ and any l dimensional subspaces $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_l \subset \mathbf{V}'$ satisfying the l-dimensional intersection property there exists $\hat{\mathbf{X}} \in \mathrm{Gr}_l(\mathbf{V})$ such that $\dim \hat{\mathbf{Y}}^\perp \cap \hat{\mathbf{W}} = 0$. Let $m = l+1$ and assume that $\mathbf{U}_1, \ldots, \mathbf{U}_m$ satisfy the m-dimensional intersection property. Let $\mathcal{P} := \{P \subseteq \langle m-1 \rangle :\ \dim \cap_{i \in P} \mathbf{U}_i = m - \#P\}$. Note that $\{i\} \in \mathcal{P}$ for each $i \in \langle m-1 \rangle$. The m-intersection property yields that $\mathbf{U}_m \cap (\cap_{i \in P} \mathbf{U}_i)$ is a strict subspace of $\cap_{i \in P} \mathbf{U}_i$ for each $P \in \mathcal{P}$. I.e. $\cap_{i \in P} \mathbf{U}_i \subsetneq \mathbf{U}_m$ for each $P \in \mathcal{P}$. Equivalently $(\cap_{i \in P} \mathbf{U}_i)^\perp \supsetneq \mathbf{U}_m^\perp$. Problem 12(d) yields that $\mathbf{U}_m^\perp \setminus \cup_{P \in \mathcal{P}} (\cap_{i \in P} \mathbf{U}_i)^\perp \ne \emptyset$. Let $\mathbf{x}_m \in \mathbf{U}_m^\perp \setminus \cup_{P \in \mathcal{P}} (\cap_{i \in P} \mathbf{U}_i)^\perp$. Define $\hat{\mathbf{U}}_i := \mathbf{U}_i \cap \{\mathbf{x}_m\}^\perp, i = 1, \ldots, l$. For $i \in \langle l \rangle$ we have that $\{i\} \in \mathcal{P}$, hence $\mathbf{x}_m \notin \mathbf{U}_i^\perp$. Thus $\hat{\mathbf{U}}_i \in \mathrm{Gr}_{l-1}(\mathbf{V}), i = 1, \ldots, l$. We claim that $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_l$ satisfy the l-dimensional intersection property.

Assume to the contrary that the l-dimensional intersection property is violated. By renaming the indices in $\langle l \rangle$ we may assume that there is $2 \le$

$k \in \langle l\rangle$ such that dim $\cap_{i\in\langle k\rangle}\hat{\mathbf{U}}_i > l-k = m-k-1$. Since $\hat{\mathbf{U}}_i \subset \mathbf{U}_i, i\in\langle l\rangle$ we deduce that dim $\cap_{i\in\langle k\rangle}\mathbf{U}_i > m-k-1$. The assumption that $\mathbf{U}_1,\ldots,\mathbf{U}_m$ satisfy the m-dimensional intersection property yields dim $\cap_{i\in\langle k\rangle}\mathbf{U}_i = m-k$, i.e. $\langle k\rangle\in\mathcal{P}$. Since $\mathbf{x}_m \notin (\cap_{i\in\langle k\rangle}\mathbf{U}_i)^\perp$ we deduce that dim $(\cap_{i\in\langle k\rangle}\mathbf{U}_i)\cap\{\mathbf{x}_m\}^\perp =$ dim $\cap_{i\in\langle k\rangle}\hat{\mathbf{U}}_i = m-k-1$, contradicting our assumption. Hence $\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_l$ satisfy the l-dimensional intersection property.

Let $\mathbf{v}_1,\ldots,\mathbf{v}_{n-1},\mathbf{x}_m$ be a basis in $\mathbf{V}$. Let $\mathbf{f}_1,\ldots,\mathbf{f}_n$ be the dual basis in $\mathbf{V}'$. (See Problem 5.1.2(d).) Note that $\hat{\mathbf{U}}_i\subset$ span $(\mathbf{f}_1,\ldots,\mathbf{f}_{n-1})$. Let $\mathbf{V}_1 =$ span $(\mathbf{v}_1,\ldots,\mathbf{v}_{n-1})$. Then we can identify span $(\mathbf{f}_1,\ldots,\mathbf{f}_{n-1})$ with $\mathbf{V}_1'$. The induction hypothesis yields the existence of $\hat{\mathbf{X}}\in \mathrm{Gr}_l(\mathbf{V}_1)$ such that dim $\hat{\mathbf{Y}}^\perp\cap\hat{\mathbf{W}} = 0$. Assume that $\hat{\mathbf{X}}$ is the columns space of the matrix $X = [x_{ij}]\in\mathbb{F}^{n\times l}$. The existence of the above $\mathbf{X}$ is equivalent to the statement that the polynomial $p_{\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_l}(x_{11},\ldots,x_{nl})$, defined in as in the Problem 15, is not identically zero. Recall that $\mathbf{U}_m\in$ span $(\mathbf{f}_1,\ldots,\mathbf{f}_{n-1})$. Problem 13 yields the existence of a nontrivial polynomial $p_{\mathbf{U}}(x_{11},\ldots,x_{nl})$ such that $\hat{\mathbf{X}}\in\mathrm{Gr}_l(\mathbf{V}_1)$, equal to the column space of $X=[x_{ij}]\in\mathbb{F}^{n\times l}$, satisfies the condition dim $\hat{\mathbf{X}}\cap\mathbf{U}_m^\perp = 0 \iff p_{\mathbf{U}}(x_{11},\ldots,x_{nl})\neq 0$. As $p_{\mathbf{U}_m}p_{\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_l}$ is a nonzero polynomial we deduce the existence of $\hat{\mathbf{X}}\in\mathrm{Gr}(\mathbf{V}_1)$ such that dim $\hat{\mathbf{X}}\cap\mathbf{U}_m^\perp = 0$ and $\hat{\mathbf{X}}\notin\mathcal{W}_m(\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_l)$.

Assume that $\mathbf{x}_1,\ldots,\mathbf{x}_{m-1}$ is a basis of $\hat{\mathbf{X}}$. Let $\mathbf{X} :=$ span $(\mathbf{x}_1,\ldots,\mathbf{x}_m)$. We claim that $\mathbf{X}\notin\mathcal{W}_m(\mathbf{U}_1,\ldots,\mathbf{U}_m)$. Let $\mathbf{X}_i$ be the $m-1$ dimensional subspace spanned by $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}\backslash\{\mathbf{x}_i\}$ for $i = 1,\ldots,m$. Then $\bigwedge^{m-1}\mathbf{X}_i =$ span $(\mathbf{y}_i), i = 1,\ldots,m$ and $\bigwedge^{m-1}\mathbf{X} =$ span $(\mathbf{y}_1,\ldots,\mathbf{y}_m)$. Let $\bigwedge^{m-1}\mathbf{U}_i =$ span $(\mathbf{w}_i), i = 1,\ldots,m$. Note that $\mathbf{x}_m\in\mathbf{X}_i\cap\mathbf{U}_m^\perp$ for $i = 1,\ldots,m-1$. Problem 13 yields that $\langle\mathbf{y}_i,\mathbf{w}_m\rangle = 0$ for $i = 1,\ldots,m-1$. Hence det $[\langle\mathbf{y}_i,\mathbf{w}_j\rangle]_{i,j=1}^m = \langle\mathbf{y}_m,\mathbf{w}_m\rangle$det $[\langle\mathbf{y}_i,\mathbf{w}_j\rangle]_{i,j=1}^{m-1}$. Since $\mathbf{X}_m = \hat{\mathbf{X}}$ we obtain that dim $\mathbf{X}_m\cap\mathbf{U}_m^\perp = 0$. Hence $\langle\mathbf{y}_m,\mathbf{w}_m\rangle\neq 0$. It is left to show that det $[\langle\mathbf{y}_i,\mathbf{w}_j\rangle]_{i,j=1}^{m-1}\neq 0$. Let $\hat{\mathbf{X}}_i\subset\mathbf{X}_i$ be the subspace of dimension $l-1 = m-2$ spanned by $\{\mathbf{x}_1,\ldots,\mathbf{x}_{m-1}\}\backslash\{\mathbf{x}_i\}$ for $i = 1,\ldots,m-1$. Note that $\hat{\mathbf{X}}_i\subset\mathbf{X}$. So $\bigwedge^{l-1}\hat{\mathbf{X}}_i =$ span $(\hat{\mathbf{y}}_\mathrm{i})$ and we can assume that $\mathbf{y}_i = \hat{\mathbf{y}}_i\wedge\mathbf{x}_m$ for $i = 1,\ldots,m-1$. Recall that $\hat{\mathbf{U}}_i = \{\mathbf{x}_m\}^\perp\cap\mathbf{U}_i$. As dim $\hat{\mathbf{U}}_i =$ dim $\mathbf{U}_i - 1$ we deduce that there exists $\mathbf{u}_i\in\mathbf{U}_i$ such that $\langle\mathbf{x}_m,\mathbf{u}_i\rangle = 1$ for $i = 1,\ldots,m-1$. So $\mathbf{U}_i = \hat{\mathbf{U}}_i\oplus$ span $(\mathbf{u}_i)$. Let $\bigwedge^{l-1}\hat{\mathbf{U}}_i =$ span $(\hat{\mathbf{w}}_\mathrm{i})$. We can assume that $\mathbf{w}_i = \hat{\mathbf{w}}_i\wedge\mathbf{u}_i$ for $i = 1,\ldots,m-1$. Problem 14 yields that $\langle\hat{\mathbf{y}}_i\wedge\mathbf{x}_m,\hat{\mathbf{w}}_j\wedge\mathbf{u}_j\rangle = l\langle\hat{\mathbf{y}}_i,\hat{\mathbf{w}}_j\rangle$ for $i,j = 1,\ldots,m-1$. Hence det $[\langle\mathbf{y}_i,\mathbf{w}_j\rangle]_{i,j=1}^{m-1} = l^{m-1}$det $[\langle\hat{\mathbf{y}}_i,\hat{\mathbf{w}}_j\rangle]_{i,j=1}^{m-1}$. Since $\hat{\mathbf{X}}\notin\mathcal{W}_m(\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_l)$ we deduce that det $[\langle\hat{\mathbf{y}}_i,\hat{\mathbf{w}}_j\rangle]_{i,j=1}^{m-1}\neq 0$, i.e. $\mathbf{X}\notin\mathcal{W}_m(\mathbf{U}_1,\ldots,\mathbf{U}_m)$. □

Lemma 5.3.11 *Let $J_1,\ldots,J_t$ be $t < m\le n$ subsets of $\langle n\rangle$ each of car-*

dinality $m-1$. Assume that $J_1,\ldots,J_t$ satisfy the m-intersection property. Then there exists $m-t$ subsets $J_{t+1},\ldots,J_m$ of $\langle n\rangle$ of cardinality m such that the sets $J_1,\ldots,J_m$ satisfy the m-intersection property.

Proof. It suffices to show that there is a subset $J_{t+1}\subset\langle n\rangle$ of cardinality $m-1$ such that $J_1,\ldots,J_{t+1}$ that satisfies the m-intersection property. If $t=1$ then choose $J_2\neq J_1$. Assume that $t\geq 2$. Let $\mathcal{P}=\{P\subset\langle t\rangle : \#\cap_{i\in P}J_i=m-\#P\}$. Note that $\{i\}\in\mathcal{P}$ for $i\in\langle t\rangle$.

Let $P,Q\in\mathcal{P}$ and assume that $P\cap Q\neq\emptyset$. We claim that $P\cup Q\in\mathcal{P}$. Let $X:=\cap_{i\in P}J_i, Y:=\cap_{j\in Q}J_j$. Then $\#X=m-\#P,\#Y=m-\#Q$. Furthermore $\#(X\cap Y)=\#X+\#Y-\#(X\cup Y)$. Observe next $X\cup Y\subset\cap_{k\in P\cap Q}J_k$. Hence the m-intersection property of $J_1,\ldots,J_t$ yields $\#(X\cup Y)\leq m-\#(P\cap Q)$. Combine the m-intersection property with all the above facts to deduce

$$m-\#(P\cup Q)\geq\#\cap_{l\in P\cup Q}J_l=\#(X\cap Y)=m-\#P+m-\#Q-\#(X\cup Y)\geq$$
$$m-\#P+m-\#Q-(m-\#(P\cap Q))=m-\#(P\cup Q).$$

It follows that there exists a partition $\{P_1,\ldots,P_l\}$ of $\langle t\rangle$ into l sets that equality in (5.3.1) holds for each P_i, and each $P\subset\langle t\rangle$ satisfying equality in (5.3.1) is a subset of some P_i.

As $\#\cap_{i\in P_1}J_i=m-\#P_1\geq m-t\geq 1$, we let $x\in\cap_{i\in P_1}J_i$. Choose J_{t+1} be any subset of cardinality $m-1$ such that $J_{t+1}\cap(\cap_{i\in P_1}J_i)=(\cap_{i\in P_1}J_i)\backslash\{x\}$. Since $\#J_{t+1}=m-1$ it follows that J_{t+1} contains exactly $\#P_1$ elements not in $\cap_{i\in P_1}J_i$.

We now show that $J_1,\ldots,J_{t+1}$ satisfy the m-intersection property. Let $Q\subseteq\langle t\rangle$ and $P:=Q\cup\{t+1\}$. If $Q\notin\mathcal{P}$ then $\#\cap_{i\in P}J_i\leq m-\#Q-1=m-\#P$. Assume that $Q\in\mathcal{P}$. To show (5.3.1) we need to show that $\cap_{i\in Q}J_i\not\subseteq J_{t+1}$. Suppose first that $Q\subseteq P_1$. Then $x\in\cap_{i\in Q}J_i$ and $x\notin J_{t+1}$. Assume that $Q\subset P_j, j>1$. So $P_1\cap Q=\emptyset$ and $P_1\cup Q\notin\mathcal{P}$. Hence

$$q:=\#((\cap_{i\in P_1}J_i)\cap(\cap_{j\in Q}J_i))=\#\cap_{k\in P_1\cup Q}J_k\leq m-(\#P_1+\#Q)-1.$$

Thus $\#((\cap_{j\in Q}J_i)\backslash(\cap_{i\in P_1}J_i))=m-\#Q-q\geq\#P_1+1$. We showed above that $\#(J_t\backslash(\cap_{i\in P_1}J_i))=\#P_1$. Therefore $\cap_{i\in Q}J_i\not\subseteq J_{t+1}$. □

Proof of Theorem 5.3.6.
1. Suppose first that $J_1:=I_1^c,\ldots,J_s:=I_s^c$ do not satisfy the intersection k intersection property. Let $P\subset\langle s\rangle$ for which $q:=\#(\cap_{i\in P}J_i)\geq k-\#P+1$. Note that $\#P\geq 2$. We can assume that $P=\langle p\rangle$ for some $2\leq p\leq s$. We claim that $\mathbf{y}_i:=\wedge^{k-1}A[;,J_i], i=1,\ldots,p$ are linearly dependent. Let $J=\cap_{i=1}^p J_i$. By renaming the indices if necessary we may assume that $J=\langle q\rangle$. Suppose first that the columns $i=1,\ldots,q$ are linearly dependent.

Hence any $k-1$ columns in J_i are linearly dependent for $i=1,\ldots,p$. Thus $\mathbf{y}_i=\mathbf{0}$ for $i=1,\ldots,p$ and $\mathbf{y}_1,\ldots,\mathbf{y}_p$ are linearly dependent.

Assume now that the columns in $i=1,\ldots,q$ are linearly independent. Let $C\in\mathbb{F}^{k\times k}$ be an invertible matrix. Then $\wedge^{k-1}C$ is also invertible. Thus $\mathbf{y}_1,\ldots,\mathbf{y}_p$ are linearly dependent if and only if $(\wedge^{k-1}C)\mathbf{y}_1,\ldots,(\wedge^{k-1}C)\mathbf{y}_p$ are linearly dependent. Thus we may replace A by $A_1:=CA$. Choose C such that $A_1=\begin{bmatrix} I_q & X\\ O & F\end{bmatrix}$, where $O\in\mathbb{F}^{(k-q)\times q}$ is the zero matrix and $F\in\mathbb{F}^{(k-q)\times(n-q)}$.

Consider a $k-1$ minor $A_1[\{i\}^c,K]$ for some $K\subset\langle n\rangle$ of cardinality $k-1$ containing set J. Expanding this minor by the first q columns we deduce that it is equal to zero, unless, unless $i=q+1,\ldots,k$. Let $J_i':=J_i\backslash J, i=1,\ldots,p$. Observe next that the submatrix of $\wedge^{k-1}A_1$ based on the rows $\{q+1\}^c,\ldots,\{k\}^c$ and columns $J_1,\ldots,J_p$ is equal to the matrix $\wedge^{k-q-1}F[;,\{J_1',\ldots,J_p'\}]$. Hence $\mathrm{rank}\ \wedge^{\mathrm{k-1}}\mathrm{A_1[;,\{J_1,\ldots,J_p\}]} =$ $\mathrm{rank}\ \wedge^{\mathrm{k-q-1}}\mathrm{F[;,\{J_1',\ldots,J_p'\}]}$. Since $q\ge k-p+1$ it follows that F has at most $k-(k-p+1)=p-1$ rows, which implies that $\wedge^{k-q-1}F$ has at most $p-1$ rows. Hence $\mathrm{rank}\ \wedge^{\mathrm{k-q-1}}\mathrm{F[;,\{J_1',\ldots,J_p'\}]\le rank}\ \wedge^{\mathrm{k-q-1}}\mathrm{F\le p-1}$. Lemma 5.3.7 implies that $\mathbf{x}(I_1),\ldots,\mathbf{x}(I_p)$ are linearly dependent, which yield that $\mathbf{x}(I_1),\ldots,\mathbf{x}(I_s)$ are linearly dependent.

Assume now that $J_1:=I_1^c,\ldots,J_s:=I_s^c$ satisfy the intersection k intersection property. By Lemma 5.3.11 we can extend these s sets to k subsets $J_1,\ldots,J_k\subset\langle n\rangle$ of cardinality $k-1$ which satisfy the intersection k intersection property. Let $\mathbf{V}:=\mathbb{F}^n$ and identify $\mathbf{V}':=\mathbb{F}^n$, where $\langle\mathbf{v},\mathbf{f}\rangle=\mathbf{f}^\top\mathbf{v}$. Let $\{\mathbf{f}_1,\ldots,\mathbf{f}_n\}$ be the standard basis in $\mathbb{F}^n$. Let Let $\mathbf{U}_i=\oplus_{j\in J_i}\mathrm{span}\,(\mathbf{f}_j), j=1,\ldots,k$. Then $\mathbf{U}_1,\ldots,\mathbf{U}_k$ have the k-dimensional intersection property. (See Problem 16). Theorem 5.3.10 yields that there exists a subspace $\mathbf{X}\in\mathrm{Gr}_k(\mathbf{V})$ such that $\mathbf{X}\notin\mathcal{W}_k(\mathbf{U}_1,\ldots,\mathbf{U}_k)$. Assume that $\mathbf{X}$ is the column space of $B^\top\in\mathbb{F}^{n\times k}$. Assume that the columns of $B^\top$ are $\mathbf{b}_1,\ldots,\mathbf{b}_k$. As in the proof of Theorem 5.3.10 let $\mathbf{y}_i=\wedge_{j\in\langle k\rangle\backslash\{i\}}\mathbf{b}_j$, $\mathbf{w}_i=\wedge_{j\in J_i}\mathbf{f}_j, i=1,\ldots,k$. Note that $\mathbf{y}_i$ is i-th column of $\wedge^{k-1}B^\top$. Furthermore $\langle\mathbf{y}_i,\mathbf{w}_j\rangle=\wedge^{k-1}B[\{i\}^c,J_j]$. The choice of B is equivalent to the condition $\det\,[\langle\mathbf{y}_i,\mathbf{w}_j\rangle]_{i,j=1}^k\neq 0$. This is equivalent to the condition that the minor of $k\times k$ submatrix of $\wedge^{k-1}B$ based on the columns $J_1,\ldots,J_k$ is not equal to zero. Since A is generic, the corresponding minor of $\wedge^{k-1}A\neq 0$. (Otherwise the entries of A will satisfy some nontrivial polynomial equations with integer coefficients.) Hence the k columns of $\wedge^{k-1}A$ corresponding to $J_1,\ldots,J_k$ are linearly independent. In particular the s columns of $\wedge^{k-1}A$ corresponding to $J_1,\ldots,J_s$ are linearly independent. Lemma 5.3.7 implies that $\mathbf{x}(I_1),\ldots,\mathbf{x}(I_s)$ are linearly independent.

2. For a generic A let $B \in \mathbb{F}^{(n-k)\times n}$ such that the columns of $B^\top$ span the null space of A. So $AB^\top = 0$ and $\mathrm{rank\, B = n - k}$. According to Problem 4(e) B is generic. Note that for any $J \subset \langle n \rangle$ of cardinality $k+1$ $\mathbf{x}(J, B) = \mathbf{y}(J, A)$.

Assume that $J_1^c, \ldots, J_t^c$ do not satisfy the the $n-k$ intersection property. The above arguments and *1* imply that $\mathbf{y}(J_1, A), \ldots, \mathbf{y}(J_t, A)$ are linearly dependent.

Suppose now that $J_1^c, \ldots, J_t^c$ satisfy the the $n-k$ intersection property. Extend this set to the set $J_1, \ldots, J_{n-k}$, each set of cardinality $k+1$, such that $J_1^c, \ldots, J_{n-k}^c$ satisfy the $n-k$ intersection property. Let $B \in \mathbb{F}^{(n-k)\times n}$ be generic. *1* implies the $n-k$ vectors $\mathbf{x}(J_1, B), \ldots, \mathbf{x}(J_{n-k}, B)$ in the row space of B are linearly independent. Let $A \in \mathbb{F}^{k\times n}$ such that the columns of $A^\top$ span the null space of B. So $\mathrm{rank\, A = k}$ and $BA^\top = 0$. According to Problem 4(e) A is generic. Hence it follows that $\mathbf{x}(J_i, B) = \mathbf{y}(J_i, A), i = 1, \ldots, n-k$ are linearly independent vectors. Problems 3- 4 yield that we can express the coordinate of each vector elementary vector in the null space of A in terms of corresponding $k \times k$ minors of A. Form the matrix $C = [\mathbf{y}(J_1, A)\ \ldots\ \mathbf{y}(J_{n-k}, A)] \in \mathbb{F}^{n\times(n-k)}$. Since $\mathrm{rank\, C = n - k}$ it follows that some $(n-k)$ minor of C is different form zero. Hence for any generic A the corresponding minor of C is different from zero. I.e. the vectors $\mathbf{y}(J_1, A), \ldots, \mathbf{y}(J_{n-k}, A)$ are always linearly independent for a generic A. In particular, the vectors $\mathbf{y}(J_1, A), \ldots, \mathbf{y}(J_t, A)$ are linearly independent. □

Problems

1. Prove Proposition 5.3.2.

2. Let $A \in \mathbb{F}^{k\times n}$ be generic. Show

   (a) All entries of A are nonzero.

   (b) A is nondegenerate.

3. Let $\mathbb{D}$ be a domain and $B \in \mathbb{D}^{(k-1)\times k}$, where $k \geq 1$. Let $B_i \in \mathbb{D}^{(k-1)\times(k-1)}$ be the matrix obtained by deleting the column i for $i = 1, \ldots, k$. Denote $\mathbf{d} = (d_1, -d_2, \ldots, (-1)^{k-1}d_k)^\top$. Show

   (a) $\mathbf{d} = \mathbf{0}$ if and only if $\mathrm{rank\, B < k - 1}$.

   (b) $B\mathbf{d} = \mathbf{0}$.

(c) Assume that $\mathbf{x} \in \ker B$. If rank B = k − 1 then $\mathbf{x} = b\mathbf{d}$ for some b in the division field of $\mathbb{D}$.

4. Let $A \in \mathbb{F}^{k\times n}, 1 \le k \le n$. Assume that $1 \le$ rank A = l $\le$ k. Show

   (a) For any $I \subset \{1, \ldots, n\}$ of cardinality $n - l - 1$ there exist $\mathbf{0} \neq \mathbf{x} \in \text{nul } A$ such that $\text{supp}\,(\mathbf{x}) \subseteq I^c$.

   (b) Let $I \subset \{1, \ldots, n\}$ be of cardinality $n - k - 1$ and denote $B := A[:, I^c] \in \mathbb{R}^{k\times(k+1)}$. Then $\dim \{\mathbf{x} \in \text{nul } A : \text{ supp}\,(\mathbf{x}) \subseteq I^c\} = 1$ if and only if rank B = k.

   (c) Let $I \subset \{1, \ldots, n\}$ be of cardinality $n - k - 1$ and denote $B := A[:, I^c] \in \mathbb{R}^{k\times(k+1)}$. Then there exists an elementary vector $\mathbf{x} \in \text{nul } A$ with $\text{supp}\,(\mathbf{x}) = I^c$ if and only if for each $j \in I^c$ $\det A[:, I^c\backslash\{j\}] \neq 0$.

   (d) The conditions *1* and *3* of Lemma 5.3.4 are equivalent.

   (e) Let rank A = l < k. The nul A is nondegenerate if all minors of A of order l are nonzero.

5. Let $\mathcal{J}$ be defined in Definition 5.3.5. Show

   (a) The condition (5.3.1) is equivalent to

   $$\#(\cup_{i\in P} J_i^c) \ge n - m + \#P \text{ for all } \emptyset \neq P \subseteq \langle t\rangle.$$

   (b) Assume that $\mathcal{J}$ satisfies (5.3.1). Let J_{t+1} be a subset of $\langle n\rangle$ of cardinality $m - 1$. Then $\mathcal{J}' := \mathcal{J} \cup \{J_{t+1}\}$ satisfies the m-intersection property if and only if

   $$\# \cup_{i\in P} (J_i^c \cap J_{t+1}) \ge \#P \text{ for all } \emptyset \neq P \subseteq \langle t\rangle.$$

   In particular, if $\mathcal{J}'$ satisfies m-intersection property then each $J_i^c \cap J_{t+1}$ is nonempty for $i = 1, \ldots, t$. **Hint**: Observe that $J_{t+1}^c \cup (\cup_{i\in P} J_i^c)$ decomposes to union of two disjoint sets J_{t+1}^c and $\cup_{i\in P}(J_i^c \cap J_{t+1})$.

6. Let $S_1, \ldots, S_t$ be t nonempty subsets of a finite nonempty set $\mathcal{S}$ of cardinality t at least. $S_1, \ldots, S_t$ is said to have *a set of distinct representatives* if there exists a subset $\{s_1, \ldots, s_t\} \subseteq \mathcal{S}$ of cardinality t such that $s_i \in S_i$ for $i = 1, \ldots, t$. Show that if $S_1, \ldots, S_t$ has a set of distinct representatives then

   $$\# \cup_{i\in P} S_i \ge \#P \text{ for all } \emptyset \neq P \subseteq \langle t\rangle.$$

   Hall's Theorem states that the above conditions are necessary and sufficient for existence of a set of distinct representatives [Hal35].

7. Let the assumptions of Problem 6 hold. Let G be a bipartite graph on a set of vertices $V = \langle t\rangle \cup \mathcal{S}$ and the set of edges $E \subseteq \langle t\rangle \times \mathcal{S}$ as follows. $(i, s) \in \langle t\rangle \times \mathcal{S}$ if and only if $s \in S_i$. Show that $S_1, \ldots, S_t$ has a set of distinct representatives if and only if G has a match $M \subset E$, i.e. no two distinct edges in M have a common vertex, of cardinality t.

   Remark: There exist effective algorithms in bipartite graphs $G = (V_1 \cup V_2, E), E \subseteq V_1 \times V_2$ to find a match of size $\min(\#V_1, \#V_2)$.

8. Let $A \in \mathbb{F}^{k\times n}$ be nondegenerate. Show

   (a) Let $I \subseteq \langle n\rangle$ be of cardinality $n-k+1$. Then there exists $\mathbf{x}(I) = (x_1, \ldots, x_n)$ in the row space of A, with supp $(\mathbf{x}(I)) = I$, whose nonzero coordinates are given by $x_j = (-1)^{p_j+1}\det A[:, I^c \cup \{j\}]$ for any $j \in I$, where p_j is the number of integers in I^c less than j.

   (b) Let I and $\mathbf{x}(I)$ be defined as in (a). Show that there exists a unique $\mathbf{z}(I) = (z_1, \ldots, z_k) \in \mathbb{F}^k$ such that $\mathbf{x}(I) = \mathbf{z}(I)A$. Use the fact that $(\mathbf{z}A)_j = 0$ for any $j \in I^c$ and Cramer's rule to show that $z_i = (-1)^i\det A[\{i\}^c, I^c]$ for $i = 1, \ldots, k$.

   (c) Let $I_1, \ldots, I_s \subset \langle n\rangle$ be sets of cardinality $n-k+1$. Let $\mathbf{x}(I_1), \mathbf{z}(I_1), \ldots, \mathbf{x}(I_s), \mathbf{z}(I_s)$ be defined as above. Then

      i. $\mathbf{x}(I_1), \ldots, \mathbf{x}(I_s)$ are linearly independent if and only if $\mathbf{z}(I_1), \ldots, \mathbf{z}(I_s)$ are linearly independent.

      ii. Let $D = \mathrm{diag}(-1, 1, -1, \ldots) \in \mathbb{F}^{k\times k}$. Then the matrix $D[\mathbf{z}(I_1)^\top\ \mathbf{z}(I_2)\ \ldots\ \mathbf{z}(I_s)] \in \mathbb{F}^{k\times s}$ is the submatrix $\wedge^{k-1}A[;, \{I_1^c, \ldots, I_s^c\}]$. Hence $\mathbf{z}(I_1), \ldots, \mathbf{z}(I_s)$ are linearly independent if and only if the submatrix $\wedge^{k-1}A[;, \{I_1^c, \ldots, I_s^c\}]$ has rank s.

      iii. The submatrix $\wedge^{k-1}A[;, \{I_1^c, \ldots, I_s^c\}]$ has rank s if and only if not all the determinants $\det \wedge^{k-1}A[\{i_1\}^c, \ldots, \{i_s\}^c\}, \{\{I_1^c, \ldots, I_s^c\}]$ for $1 \le i_1 < i_2 < \ldots < i_s \le k$ are equal to zero.

      iv. $\mathbf{x}(I_1), \ldots, \mathbf{x}(I_k)$ is a basis in the row space of A if and only if the determinant of the full row submatrix of $\wedge^{k-1}A$ corresponding to the columns determined by $I_1^c, \ldots, I_k^c$ is not equal to zero.

9. Let $A \in \mathbb{F}^{k\times n}$ be nondegenerate. Let $\mathbf{b}_1, \ldots, \mathbf{b}_{n-k} \in \mathbb{R}^n$ be a basis in the null space of A and denote by $B^\top \in \mathbb{F}^{n\times(n-k)}$ the matrix whose columns are $\mathbf{b}_1, \ldots, \mathbf{b}_{n-k}$. Show

   (a) Let $J \subseteq \langle n\rangle$ be of cardinality $k+1$. Then there exists $\mathbf{y}(J) = (y_1, \ldots, y_n)^\top$ in the column space of $B^\top$, with supp $(\mathbf{y}(J)) = J$, whose nonzero coordinates are given by $y_j = (-1)^{p_j+1}\det B[:, J^c \cup \{j\}]$ for any $j \in J$, where p_j is the number of integers in J^c less than j.

(b) Let J and $\mathbf{y}(J)$ be defined as in (a). Show that there exists a unique $\mathbf{u}(J) = (u_1, \ldots, u_{n-k})^\top \in \mathbb{F}^{n-k}$ such that $\mathbf{y}(J) = B^\top \mathbf{u}(J)$. Use the fact that $(B^\top \mathbf{u})_j = 0$ for any $j \in J^c$ and Cramer's rule to show that $u_i = (-1)^i \det B[\{i\}^c, J^c]$ for $i = 1, \ldots, n-k$.

(c) Let $J_1, \ldots, J_t \subset \langle n \rangle$ be sets of cardinality $k+1$. Let $\mathbf{y}(J_1), \mathbf{u}(J_1), \ldots, \mathbf{y}(J_t), \mathbf{u}(J_t)$ be defined as above. Then

   i. $\mathbf{y}(J_1), \ldots, \mathbf{y}(J_t)$ are linearly independent if and only if $\mathbf{u}(J_1), \ldots, \mathbf{u}(J_t)$ are linearly independent.

   ii. Let $D = \mathrm{diag}(-1, 1, -1, \ldots) \in \mathbb{F}^{n-k \times n-k}$. Then the matrix $D[\mathbf{u}(J_1)\ \mathbf{u}(J_2)\ \ldots\ \mathbf{u}(J_t)] \in \mathbb{F}^{(n-k)\times t}$ is the submatrix $\wedge^{n-k-1}B[;, \{J_1^c, \ldots, J_t^c\}]$. Hence $\mathbf{u}(J_1), \ldots, \mathbf{u}(J_t)$ are linearly independent if and only if the submatrix $\wedge^{n-k-1}B[;, \{J_1^c, \ldots, J_t^c\}]$ has rank t.

   iii. The submatrix $\wedge^{n-k-1}B[;, \{J_1^c, \ldots, J_t^c\}]$ has rank t if and only if not all the determinants $\det \wedge^{n-k-1}B[\{i_1\}^c, \ldots, \{i_t\}^c\}, \{\{J_1^c, \ldots, J_t^c\}]$ for $1 \le i_1 < i_2 < \ldots < i_t \le n-k$ are equal to zero.

   iv. $\mathbf{y}(J_1), \ldots, \mathbf{x}(J_{n-k})$ is a basis in the null space of A if and only if the determinant of the full row submatrix of $\wedge^{n-k-1}B$ corresponding to the columns determined by $J_1^c, \ldots, J_{n-k-1}^c$ is not equal to zero.

10. Let $C \in \mathbb{F}^{n\times(n-k)}$ be a matrix of rank $n-k$. Show that there exists $A \in \mathbb{F}^{k\times n}$ of rank k such that $AC = \mathbf{0}$.

11. Let $\mathbb{F}$ be a field of 0 characteristic. Let $p(x_1, \ldots, x_n) \in \mathbb{F}[x_1, \ldots, x_n]$. Show that $p(x_1, \ldots, x_n) = 0$ for all $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{F}^n$ if and only if p is the zero polynomial. **Hint**: Use induction.

12. Let $\mathbb{F}$ be a field of 0 characteristic. Assume that $\mathbf{V} = \mathbb{F}^n$. Identify $\mathbf{V}'$ with $\mathbb{F}^n$. So for $\mathbf{u} \in \mathbf{V}, \mathbf{f} \in \mathbf{V}'$ $\langle \mathbf{u}, \mathbf{f} \rangle = \mathbf{f}^\top \mathbf{v}$. Show

    (a) $\mathbf{U} \subset \mathbf{V}$ is a subspace of dimension of $n-1$ if and only if there exists a nontrivial linear polynomial $l(\mathbf{x}) = a_1x_1 + \ldots + a_nx_n$ such that $\mathbf{U}$ is the zero set of $l(\mathbf{x})$, i.e. $\mathbf{U} = Z(l)$.

    (b) Let $\mathbf{U}_1, \ldots, \mathbf{U}_k$ be k subspaces of $\mathbf{V}$ of dimension $n-1$. Show that there exists a nontrivial polynomial $p = \prod_{i=1}^k l_i \in \mathbb{F}[x_1, \ldots, x_n]$, where each l_i is a nonzero linear polynomial, such that $\cup_{i=1}\mathbf{U}_i$ is $Z(p)$.

(c) Show that if $\mathbf{U}_1, \ldots, \mathbf{U}_k$ are k strict subspaces of $\mathbf{V}$ then $\cup_{i=1}^k \mathbf{U}_i$ is a strict subset of $\mathbf{V}$. **Hint**: One can assume that dim $\mathbf{U}_i = n - 1, i = 1, \ldots, k$ and the use Problem 11.

(d) Let $\mathbf{U}, \mathbf{U}_1, \ldots, \mathbf{U}_k$ be subspaces of $\mathbf{V}$. Assume that $\mathbf{U} \subseteq \cup_{i=1}^k \mathbf{U}_i$. Show that there exists a subspace $\mathbf{U}_i$ which contains $\mathbf{U}$. **Hint**: Observe $\mathbf{U} = \cup_{i=1}^k (\mathbf{U}_i \cap \mathbf{U})$.

13. Let the assumptions of Problem 12 hold. Let $X = [x_{ij}], U = [u_{ij}] \in \mathbb{F}^{n\times l}$. View the matrices $\wedge^l X, \wedge^l U$ as column vectors in $\mathbb{F}^{\binom{n}{l}}$. Let $p_U(x_{11}, \ldots, x_{nl}) :=$ det $(X^\top U) = (\wedge^l X)^\top \wedge^l U$. View p_U as a polynomial in nl variables with coefficients in $\mathbb{F}$. Show

(a) p_U a homogeneous multilinear polynomial of degree l.

(b) p_U is a trivial polynomial if and only if $\mathrm{rank\ U} < \mathrm{l}$.

(c) Let $\mathbf{X} \in \mathrm{Gr}_l(\mathbf{V}), \mathbf{U} \in \mathrm{Gr}_l(\mathbf{V}')$ and assume that the column space of $X = [x_{ij}] = [\mathbf{x}_1, \ldots, \mathbf{x}_l], U = [u_{ij}] = [\mathbf{u}_1, \ldots, \mathbf{u}_l] \in \mathbb{F}^{n\times l}$ are $\mathbf{X}, \mathbf{U}$ respectively. Then

$$p_U(x_{11}, \ldots, x_{nl}) = \det\, [\mathbf{u}_j^\top \mathbf{x}_i]_{i,j=1}^l,\ \langle \mathbf{x}_1 \wedge \ldots \wedge \mathbf{x}_l, \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_l \rangle = l! p_U(x_{11}, \ldots, x_{nl}).$$

In particular, dim $\mathbf{X} \cap \mathbf{U}^\perp = 0 \iff p_U(x_{11}, \ldots, x_{nl}) \neq 0$.

14. Let $\mathbb{F}$ be a field of 0 characteristic. Assume that $\mathbf{V}$ is an n-dimensional vector space with $n \geq 2$. Let $\mathbf{X} \subset \mathbf{V}, \mathbf{U} \subset \mathbf{V}'$ be $m \geq 2$ dimensional subspaces. Assume that $\mathbf{X} \subsetneq \mathbf{U}^\perp$. Let $\mathbf{x}_m \in \mathbf{X} \backslash \mathbf{U}^\perp$. Let $\hat{\mathbf{U}} = \{\mathbf{x}_m\}^\perp \cap \mathbf{U}$. Let $\hat{\mathbf{X}}$ be any $m-1$ dimensional subspace of $\mathbf{X}$ which does not contain $\mathbf{x}_m$. Show

(a) dim $\hat{\mathbf{U}} = m - 1, \mathbf{X} = \hat{\mathbf{X}} \oplus$ span $(\mathbf{x}_m)$.

(b) There exists $\mathbf{u}_m \in \mathbf{U}$ such that $\langle \mathbf{x}_m, \mathbf{u}_m \rangle = 1$. Furthermore $\mathbf{U} = \hat{\mathbf{U}} \oplus$ span $(\mathbf{u}_m)$.

(c) Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}\}, \{\mathbf{u}_1, \ldots, \mathbf{u}_{m-1}\}$ be bases of $\hat{\mathbf{X}}, \hat{\mathbf{Y}}$ respectively. Then

$$\langle \mathbf{x}_1 \wedge \ldots \wedge \mathbf{x}_m, \mathbf{u}_1 \wedge \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_m \rangle = m \langle \mathbf{x}_1 \wedge \ldots \wedge \mathbf{x}_{m-1}, \mathbf{u}_1 \wedge \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_{m-1} \rangle,$$

where $\mathbf{u}_m$ is defined in (b). **Hint**: Use (5.2.8) and expand the determinant by the last row.

(d) Assume that $\bigwedge^{m-1} \hat{\mathbf{X}} =$ span $(\hat{\mathbf{y}}), \bigwedge^{\mathrm{m}-1} \hat{\mathbf{U}} =$ span $(\hat{\mathbf{w}})$. Then $\langle \hat{\mathbf{y}} \wedge \mathbf{x}_m, \hat{\mathbf{w}} \wedge \mathbf{u}_n \rangle = m \langle \hat{\mathbf{y}}, \hat{\mathbf{w}} \rangle$.

15. Let the assumptions of Theorem 5.3.10 hold. View $\mathbf{V}$ and $\mathbf{V}'$ as $\mathbb{F}^n$. So for $\mathbf{u} \in \mathbf{V}, \mathbf{f} \in \mathbf{V}'$ $\langle \mathbf{u}, \mathbf{f}\rangle = \mathbf{f}^\top \mathbf{v}$. Let $\mathbf{X} \in \mathrm{Gr}_m(\mathbf{V})$ be the column space of $X = [x_{ij}] \in \mathbb{F}^{n\times m}$. Show that there exists a homogeneous polynomial $p_{\mathbf{U}_1,\ldots,\mathbf{U}_m}(x_{11},\ldots,x_{nm})$ of degree $m(m-1)$ such that $\mathbf{X} \in \mathcal{W}_m(\mathbf{U}_1,\ldots,\mathbf{U}_m)$ if and only if $p_{\mathbf{U}_1,\ldots,\mathbf{U}_m}(x_{11},\ldots,x_{nm}) = 0$. **Hint**: Choose a basis of $\mathbf{X}$ to be the columns of X. Then use the first paragraph of Proof of Theorem 5.3.10 and Problem 13.

16. Let $\mathbb{F}$ be a field and $\mathbf{V}$ an n-dimensional subspace over $\mathbb{F}$. Let $\mathbf{v}_1,\ldots,\mathbf{v}_n$ be a basis of $\mathbf{V}$. For $\emptyset \ne K \subset \langle n\rangle$ let $\mathbf{U}_K = \oplus_{i\in K}\mathrm{span}\,(\mathbf{v}_i)$. Let $t \le m$ and assume that $K_1,\ldots,K_t \subset \langle n\rangle$ be sets of cardinality $m-1$ for any $2 \le m \in \langle n\rangle$. Show that $\mathbf{U}_{K_1},\ldots,\mathbf{U}_{K_t}$ satisfy the m-dimensional intersection property if and only if $K_1,\ldots,K_t$ satisfy the m-intersection property.

17. Let $\mathbf{V}$ be an n-dimensional vector space over $\mathbb{F}$ of characteristic 0. Show

    (a) Let $2 \le m \le n$. If $\mathbf{U}_1,\ldots,\mathbf{U}_m$ are $m-1$ dimensional vector spaces satisfying the m-dimensional intersection property then $\dim \sum_{i=1}^m \bigwedge^{m-1}\mathbf{U}_i = m$.

    (b) For $m = 3, n = 4$ there exist $\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3$, which do not satisfy the 3-dimensional intersection property such that $dim \sum_{i=1}^3 \bigwedge^2 \mathbf{U}_i = 3$. **Hint**: Choose a basis in $\mathbf{V}$ and assume that each $\mathbf{U}_i$ is spanned by by some two vectors in the basis.

    (c) Show that for $2 \le t \le m = n$ $\mathbf{U}_1,\ldots,\mathbf{U}_t$ satisfy the n-intersection property if and only if $\dim \sum_{i=1}^t \bigwedge^{n-1}\mathbf{U}_i = t$.

5.4 Tensor products of inner product spaces

Let $\mathbb{F} = \mathbb{R}, \mathbb{C}$ and assume that $\mathbf{V}_i$ is a n_i-dimensional vector space with the inner product $\langle\cdot,\cdot\rangle_i$ for $i = 1,\ldots,k$. Then $\mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i$ has a unique inner product $\langle\cdot,\cdot\rangle$ satisfying the property

$$\langle \otimes_{i=1}^k \mathbf{x}_i, \otimes_{i=1}^k \mathbf{y}_j\rangle = \prod_{i=1}^k \langle \mathbf{x}_i, \mathbf{y}_i\rangle_i, \text{ for all } \mathbf{x}_i, \mathbf{y}_i \in \mathbf{V}_i,\ i = 1,\ldots,k. \tag{5.4.1}$$

(See Problem 1.) We will assume that $\mathbf{Y}$ has the above *canonical* inner product, unless stated otherwise.

Proposition 5.4.1 *Let $\mathbf{U}_i, \mathbf{V}_i$ be a finite dimensional IPS over $\mathbb{F} := \mathbb{R}, \mathbb{C}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathbf{U}_i}, \langle\cdot,\cdot\rangle_{\mathbf{V}_i}$ for $i = 1,\ldots,k$ respectively.*

Let $\mathbf{X} := \otimes_{i=1}^k \mathbf{U}_i, \mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i$ be IPS with the canonical inner products $\langle \cdot, \cdot \rangle_{\mathbf{X}}, \langle \cdot, \cdot \rangle_{\mathbf{Y}}$ respectively. Then the following claims hold.

1. *Assume that $T_i \in \mathrm{L}(\mathbf{V}_i, \mathbf{U}_i)$ for $i = 1, \ldots, k$. Then $\otimes_{i=1}^k T_i \in \mathrm{L}(\mathbf{Y}, \mathbf{X})$ and $(\otimes_{i=1}^k T_i)^* = \otimes_{i=1}^k T_i^* \in \mathrm{L}(\mathbf{X}, \mathbf{Y})$.*

2. *Assume that $T_i \in \mathrm{L}(\mathbf{V}_i)$ is normal for $i = 1, \ldots, k$. Then $\otimes_{i=1}^k T_i \in \mathrm{L}(\mathbf{Y})$ is normal. Moreover, $\otimes_{i=1}^k T_i$ is hermitian or unitary, if each T_i is hermitian or unitary respectively.*

3. *Assume that $T_i \in \mathrm{L}(\mathbf{V}_i, \mathbf{U}_i)$ for $i = 1, \ldots, k$. Let $\sigma_1(T_i) \ge \ldots \ge \sigma_{\mathrm{rank\ T_i}}(T_i) > 0, \sigma_j(T_i) = 0, j > \mathrm{rank\ T_i}$ be the singular values of T_i. Let $\mathbf{c}_{1,i}, \ldots, \mathbf{c}_{n_i,i}$ and $\mathbf{d}_{1,i}, \ldots, \mathbf{d}_{m_i,i}$ be orthonormal bases of $\mathbf{V}_i$ and $\mathbf{U}_i$ consisting of right and left singular vectors of T_i as described in (4.9.5):*

$$T_i \mathbf{c}_{j_i,i} = \sigma_{j_i}(T_i) \mathbf{d}_{j_i,i}, \quad j_i = 1, \ldots, \ i = 1, \ldots, k.$$

Then

$$(\otimes_{i=1}^k T_i) \otimes_{i=1}^k \mathbf{c}_{j_i,i} = \Big(\prod_{i=1}^k \sigma_{j_i}(T_i) \Big) \otimes_{i=1}^k \mathbf{d}_{j_i,i}, \ j_i = 1, \ldots, \ i = 1, \ldots, k.$$

In particular

$$(5.4.2) \quad || \otimes_{i=1}^k T_i || = \sigma_1(\otimes_{i=1}^k T_i) = \prod_{i=1}^k ||T_i|| = \prod_{i=1}^k \sigma_1(T_i),$$

$$\sigma_{\prod_{i=1}^k \mathrm{rank\ T_i}}(\otimes_{i=1}^k T_i) = \prod_{i=1}^k \sigma_{\mathrm{rank\ T_i}}(T_i).$$

We consider a fixed IPS vector space $\mathbf{V}$ of dimension n and its exterior products $\bigwedge^k \mathbf{V}$ for $k = 1, \ldots, n$. Since $\bigwedge^k \mathbf{V}$ is a subspace of $\mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i, \mathbf{V}_1 = \ldots = \mathbf{V}_k = \mathbf{V}$, it follows that $\bigwedge^k \mathbf{V}$ has a canonical inner product induced by $\langle \cdot, \cdot \rangle_{\mathbf{Y}}$. See Problem 3a.

Proposition 5.4.2 *Let $\mathbf{V}, \mathbf{U}$ be IPS of dimension n and m respectively. Assume that $T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$. Suppose that $\mathbf{c}_1, \ldots, \mathbf{c}_n$ and $\mathbf{d}_1, \ldots, \mathbf{d}_m$ be orthonormal bases of $\mathbf{V}$ and $\mathbf{U}$ composed of the right and left singular eigenvectors of T respectively, as given in (4.9.5). Let $k \in \mathbb{N} \cap [1, \min(m, n)]$. Then the orthonormal bases*

$$\frac{1}{\sqrt{k!}} \mathbf{c}_{i_1} \wedge \ldots \wedge \mathbf{c}_{i_k} \in \bigwedge^k \mathbf{V}, \ 1 \le i_1 < \ldots < i_k \le n,$$

$$\frac{1}{\sqrt{k!}} \mathbf{d}_{j_1} \wedge \ldots \wedge \mathbf{d}_{j_k} \in \bigwedge^k \mathbf{U}, \ 1 \le j_1 < \ldots < j_k \le m,$$

are the right and the left singular vectors of $\wedge^k T \in \mathrm{L}(\bigwedge^k \mathbf{V}, \bigwedge^k \mathbf{U})$*, with the corresponding singular values* $\prod_{l=1}^k \sigma_{i_l}(T)$ *and* $\prod_{l=1}^k \sigma_{j_l}(T)$ *respectively. In particular*

$$\wedge^k T\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k = || \wedge^k T||\mathbf{d}_1 \wedge \ldots \wedge \mathbf{d}_k, \quad || \wedge^k T|| = \sigma_1(\wedge^k T) = \prod_{l=1}^k \sigma_i(T),$$

$$\wedge^k T\mathbf{c}_{\mathrm{rank\ T-k+1}} \wedge \ldots \wedge \mathbf{c}_{\mathrm{rank\ T}} = \prod_{l=1}^k \sigma_{\mathrm{rank\ T-k+l}}(T)\mathbf{d}_{\mathrm{rank\ T-k+1}} \wedge \ldots \wedge \mathbf{d}_{\mathrm{rank\ T}}$$

are the biggest and the smallest positive singular value of $\wedge^k T$ *for* $k \leq$ rank T.

Corollary 5.4.3 *Suppose that* $\mathbf{V}$ *is an IPS of dimension* n*. Assume that* $T \in \mathrm{S}_+(\mathbf{V})$*. Let* $\lambda_1(T) \geq \ldots \geq \lambda_n(T) \geq 0$ *be the eigenvalues of* T *with the corresponding orthonormal eigenbasis* $\mathbf{c}_1, \ldots, \mathbf{c}_n$ *of* $\mathbf{V}$*. Then* $\wedge^k T \in \mathrm{S}_+(\bigwedge^k \mathbf{V})$*. Let* $k \in \mathbb{N} \cap [1, n]$*. Then the orthonormal base* $\frac{1}{\sqrt{k!}}\mathbf{c}_{i_1} \wedge \ldots \wedge \mathbf{c}_{i_k}$*,* $1 \leq i_1 < \ldots < i_k \leq n$ *of* $\bigwedge^k \mathbf{V}$ *is an eigensystem of* $\wedge^k T$*, with with the corresponding eigenvalues* $\prod_{l=1}^k \lambda_{i_l}(T)$*. In particular*

$$\wedge^k T\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k = || \wedge^k T||\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k, \quad || \wedge^k T|| = \lambda_1(\wedge^k T) = \prod_{l=1}^k \lambda_i(T),$$

$$\wedge^k T\mathbf{c}_{\mathrm{rank\ T-k+1}} \wedge \ldots \wedge \mathbf{c}_{\mathrm{rank\ T}} = \prod_{l=1}^k \lambda_{\mathrm{rank\ T-k+l}}(T)\mathbf{d}_{\mathrm{rank\ T-k+1}} \wedge \ldots \wedge \mathbf{d}_{\mathrm{rank\ T}}$$

are the biggest and the smallest positive eigenvalue of $\wedge^k T$ *for* $k \leq$ rank T.

See Problem 4.

Theorem 5.4.4 *Let* $\mathbf{U}, \mathbf{V}, \mathbf{W}$ *be finite dimensional IPS. Assume that* $P \in \mathrm{L}(\mathbf{U}, \mathbf{W}), T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$*. Then*

$$\prod_{i=1}^k \sigma_i(PT) \leq \prod_{i=1}^k \sigma_i(P)\sigma_i(T), \quad k = 1, \ldots \tag{5.4.3}$$

For $k \leq \min(\mathrm{rank\ P}, \mathrm{rank\ T})$ *equality in (5.4.3) holds if and only if the following condition is satisfied. There exists a* k*-dimensional subspace* $\mathbf{V}_k$ *of* $\mathbf{V}$ *which spanned by the first* k*-orthonormal right singular vectors of* T*, such that* $T\mathbf{V}_k$ *is a* k*-dimensional subspace of* $\mathbf{U}$ *which is spanned the first* k*-orthonormal right singular vectors of* P*.*

Proof. Suppose first that $k = 1$. Then $||PT|| = ||PT\mathbf{v}||$, where $\mathbf{v} \in \mathbf{V}, ||\mathbf{v}|| = 1$ is the right singular vector of PT. Clearly, $||PT\mathbf{v}|| = ||P(T\mathbf{v})|| \le ||P||\ ||T\mathbf{v}|| \le ||P||\ ||T||$, which implies the inequality (5.4.3) for $k = 1$. Assume that $||P||\ ||T|| > 0$. For the equality $||PT|| = ||P||\ ||T||$ we must have that $T\mathbf{v}$ is the right singular vector corresponding to P and $\mathbf{v}$ the the right singular vector corresponding to T. This shows the equality case in the theorem for $k = 1$.

Assume that $k > 1$. If the right-hand side of (5.4.3) is zero then $\mathrm{rank\ PT} \le \min(\mathrm{rank\ P}, \mathrm{rank\ Q}) < \mathrm{k}$ and $\sigma_k(PT) = 0$. Hence (5.4.3) trivially holds. Assume that $k \le \min(\mathrm{rank\ P}, \mathrm{rank\ Q})$. Then the right-hand side of (5.4.3) is positive. Clearly $\min(\mathrm{rank\ P}, \mathrm{rank\ T}) \le \min(\dim \mathbf{U}, \dim \mathbf{V}, \dim \mathbf{W})$. Observe that $\wedge^k T \in \mathrm{L}(\bigwedge^k \mathbf{V}, \bigwedge^k \mathbf{U}), \wedge^k P \in \mathrm{L}(\bigwedge^k \mathbf{U}, \bigwedge^k \mathbf{W})$. Hence (5.4.3) for $k = 1$ applied to $\wedge^k PT = \wedge^k P \wedge^k T$ yields $\sigma_1(\wedge^k PT) \le \sigma_1(\wedge^k P)\sigma_1(\wedge^k T)$. Use Proposition 5.4.2 to deduce (5.4.3). In order to have $\sigma_1(\wedge^k PT) = \sigma_1(\wedge^k P)\sigma_1(\wedge^k T)$ the operator $\wedge^k T$ has a right first singular vector $\mathbf{x} \in \bigwedge^k \mathbf{V}$, such that $\mathbf{0} \ne \wedge^k T\mathbf{x}$ is a right singular vector of $\wedge^k P$ corresponding to $\sigma_1(\wedge^k P)$. It is left to show to show that $\mathbf{x}$ can be chosen as $\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k$, where $\mathbf{c}_1, \ldots, \mathbf{c}_k$ are the right singular vectors of T corresponding to the first k-singular values of T.

Suppose that

$$\sigma_1(T) = \ldots = \sigma_{l_1}(T) > \sigma_{l_1+1}(T) = \ldots = \sigma_{l_2}(T) > \ldots > 0 = \sigma_j(T) \text{ for } j > l_p. \tag{5.4.4}$$

Assume first that $k = l_i$ for some $i \le p$. Then $\sigma_1(\wedge^k T) > \sigma_2(\wedge^k T)$ and $\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k$ is the right singular vector of $\wedge^k T$ corresponding to $\sigma_1(\wedge^k T)$. Then $\sigma_1(\wedge^k P \wedge^k T) = \sigma_1(\wedge^k P)\sigma_1(\wedge^k T)$ if and only if $(\wedge^k T)\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k = T\mathbf{c}_1 \wedge \ldots \wedge T\mathbf{c}_k$ is the right singular vector of $\wedge^k P$ corresponding to $\sigma_1(\wedge^k P)$.

Assume that

$$\sigma_1(P) = \ldots = \sigma_{m_1}(P) > \sigma_{m_1+1}(P) = \ldots = \sigma_{m_2}(p) > \ldots > 0 = \sigma_j(P) \text{ for } j > m_q.$$

Suppose that $k = m_{j-1} + r$, where $1 \le r \le m_j - m_{j-1}$. (We assume here that $m_0 = 0$.) Let $\mathbf{U}_1$ be the subspace spanned by the m_{j-1} right singular vectors of P corresponding to the first m_{j-1} singular values of P and $\mathbf{W}_1$ be the subspace spanned by $m_j - m_{j-1}$ right singular vectors of P corresponding the $\sigma_{m_{j-1}+1}(P), \ldots, \sigma_{m_j}(P)$. Then any right singular vector of $\wedge^k P$ corresponding to $\sigma_1(\wedge^k P)$ is in the subspace $\big(\bigwedge^{m_{j-1}} \mathbf{U}_1\big) \bigwedge \big(\bigwedge^r \mathbf{W}_1\big)$. Let $\mathbf{V}_k = \mathrm{span}\ (\mathbf{c}_1, \ldots, \mathbf{c}_k)$. So $\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k$ is a nonzero vector in $\bigwedge^k \mathbf{V}_k$ and $(\wedge^k T)\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k$ is a nonzero vector in $\bigwedge^k \mathbf{W}_2$, where $\mathbf{W}_2 := T\mathbf{V}_k$ and $\mathbf{U}_2 = \{\mathbf{0}\}$. The equality in (5.4.3) yields that $\Big(\bigwedge^{m_{j-1}} \mathbf{U}_1\Big) \bigwedge \Big(\bigwedge^r \mathbf{W}_1\Big) \cap$

$\left(\bigwedge^{0}\mathbf{U}_2\right)\bigwedge\left(\bigwedge^{k}\mathbf{W}_2\right) \neq \{\mathbf{0}\}$. Lemma 5.2.11 yields that $\mathbf{U}_1 \subset T\mathbf{V}_k$ and $T\mathbf{V}_k \subseteq \mathbf{U}_1 + \mathbf{W}_1$. So $T\mathbf{V}_k$ is spanned by the first k right singular vectors of P.

Assume now that $k = l_{i-1} + s, 1 \le s < l_i - l_{i-1}$. Then the subspace spanned by all right singular vectors of $\wedge^k T$ corresponding to $\sigma_1(\wedge^k T)$ is equal to
$\left(\bigwedge^{l_{i-1}}\mathbf{U}_2\right)\bigwedge\left(\bigwedge^{s}\mathbf{W}_2\right)$, where $\mathbf{U}_3$ and $\mathbf{W}_3$ are the subspaces spanned the right singular vectors of T corresponding to the first l_{i-1} and the next $l_i - l_{i-1}$ singular values of T respectively. Let $\mathbf{U}_2 := T\mathbf{U}_3, \mathbf{W}_2 := T\mathbf{W}_3$. The equality in (5.4.3) yields that $\left(\bigwedge^{m_{j-1}}\mathbf{U}_1\right)\bigwedge\left(\bigwedge^{r}\mathbf{W}_1\right) \cap \left(\bigwedge^{l_{i-1}}\mathbf{U}_2\right)\bigwedge\left(\bigwedge^{s}\mathbf{W}_2\right)$ contains a right singular vector of $\wedge^k P$ corresponding to $\sigma_1(\wedge^k P)$. Lemma 5.2.11 yields that there exists a k dimensional subspace $\mathbf{V}_1'$ such that $\mathbf{V}_1' \supset \mathbf{U}_1 + \mathbf{U}_2$ and $\mathbf{V}_1' \subset (\mathbf{U}_1 + \mathbf{W}_1) \cap (\mathbf{U}_2 + \mathbf{W}_2)$. Hence there exists a k-dimensional subspace $\mathbf{V}_k$ of $\mathbf{U}_3 + \mathbf{W}_3$ containing $\mathbf{U}_3$ such that $\mathbf{V}_1' = T\mathbf{V}_k$ contains $\mathbf{U}_1$ and is contained in $\mathbf{U}_1 + \mathbf{W}_1$. Hence $T\mathbf{V}_k$ is spanned by the first k right singular vectors of P.

Assume now that $\prod_{i=1}^{l} \sigma_i(P)\sigma_i(T) > 0$. Then $0 < \sigma_i(P), 0 < \sigma_i(T)$ for $i = 1, \ldots, l$. Assume that for $k = 1, \ldots, l$ equality holds in (5.4.3). We prove the existence of orthonormal sets $\mathbf{c}_1, \ldots, \mathbf{c}_l$, $\mathbf{d}_1, \ldots, \mathbf{d}_l$ of right singular vectors of T and P respectively such that $\frac{1}{\sigma_k(T)}\mathbf{c}_k = \mathbf{d}_k, k = 1, \ldots, l$ by induction on l. For $l = 1$ the result is trivial. Assume that the result holds for $l = m$ and let $l = m + 1$. The equality in (5.4.3) for $k = m + 1$ yields the existence of $m + 1$ dimensional subspace $\mathbf{X} \subseteq \mathbf{U}$ such that $\mathbf{X}$ is spanned by the first $m + 1$ right singular vectors of T and $T\mathbf{X}$ is spanned by the first $m + 1$ right singular vectors of P. □

Theorem 5.4.5 *Let the assumptions of Theorem 5.4.4 hold. Then equalities in (5.4.3) hold for* $k = 1, \ldots, l \le \min(\mathrm{rank\,P}, \mathrm{rank\,T})$ *if and only if there exits first l orthonormal right singular vectors* $\mathbf{c}_1, \ldots, \mathbf{c}_l$ *of T, such that* $\frac{1}{\sigma_1(T)}T\mathbf{c}_1, \ldots, \frac{1}{\sigma_l(T)}T\mathbf{c}_l$ *are first l orthonormal right singular vectors of P.*

Proof. We prove the theorem by induction on l. For $l = 1$ the theorem follows from Theorem 5.4.4. Suppose that the theorem holds for $l = j$. Let $l = j + 1$. Since we assumed that equality holds in (5.4.3) for $k = l$ Theorem 5.4.4 yields that there exists an subspace l-dimensional subspace $\mathbf{V}_l$ of $\mathbf{V}$ which is spanned by the first l right singular vectors of T, and $T\mathbf{V}_l$ is spanned by the first l right singular vectors of P. Let $\hat{T} \in \mathrm{L}(\mathbf{V}_l, T\mathbf{V}_l), \hat{P} \in \mathrm{L}(T\mathbf{V}_l, PT\mathbf{V}_l)$ be the restrictions of T and P to the subspaces $\mathbf{V}_l, T\mathbf{V}_l$ respectively. Clearly

$$\sigma_i(T) = \sigma_i(\hat{T}) > 0,\ \sigma_i(P) = \sigma_i(\hat{P}) > 0,\ \text{for } i = 1,\ldots,l. \tag{5.4.5}$$

The equalities in (5.4.3) for $k = 1,\ldots,l$ imply that $\sigma_i(PT) = \sigma_i(P)\sigma_i(T)$ for $i = 1,\ldots,l$. Let $\hat{Q} := \hat{P}\hat{T} \in \mathrm{L}(\mathbf{V}_l, PT\mathbf{V}_l)$. Clearly $\hat{Q}$ is the restriction of $Q := PT$ to $\mathbf{V}_l$. Corollary 4.10.3 yields that $\sigma_i(\hat{Q}) \le \sigma_i(Q)$ for $i = 1,\ldots$. Since $\det \hat{Q} = \det \hat{P}\ \det \hat{T}$ we deduce that $\prod_{i=1}^{l} \sigma_i(\hat{Q}) = \prod_{i=1}^{l} \sigma_i(\hat{P})\ \prod_{i=1}^{l} \sigma_i(\hat{T})$. The above arguments show that $\prod_{i=1}^{l} \sigma_i(\hat{Q}) = \prod_{i=1}^{l} \sigma_i(Q) > 0$. Corollary 4.10.3 yields that $\sigma_i(\hat{Q}) = \sigma_i(Q)$. Hence we have equalities $\prod_{i=1}^{k} \sigma_i(\hat{P}\hat{Q}) = \prod_{i=1}^{k} \sigma_i(\hat{P})\sigma_i(\hat{T})$ for $i = 1,\ldots,l$. The induction hypothesis yields that there exist first $l-1$ orthonormal right singular vectors of $\hat{T}$ $\mathbf{c}_1,\ldots,\mathbf{c}_{l-1}$, such that $\frac{1}{\sigma_1(\hat{T})}\hat{T}\mathbf{c}_1,\ldots,\frac{1}{\sigma_l(\hat{T})}\hat{T}\mathbf{c}_{l-1}$ are first l orthonormal right singular vectors of $\hat{P}$. Complete $\mathbf{c}_1,\ldots,\mathbf{c}_{l-1}$ to an orthonormal basis $\mathbf{c}_1,\ldots,\mathbf{c}_l$ of $\mathbf{V}_l$. Then $\mathbf{c}_l$ is a right singular vector of $\hat{T}$ corresponding $\sigma_l(\hat{T})$. Since $\hat{T}\mathbf{c}_l$ is orthogonal to $\hat{T}\mathbf{c}_1,\ldots,\hat{T}\mathbf{c}_{l-1}$, which are right singular vectors of $\hat{P}$ it follows that $\frac{1}{\sigma_l(\hat{T})}\hat{T}\mathbf{c}_l$ is a right singular vector of $\hat{P}$ corresponding to $\sigma_l(\hat{T})$. Use (5.4.5) and the fact that $\hat{T}$ and $\hat{P}$ are the corresponding restrictions of T and P respectively to deduce the theorem. □

In what follows we need to consider the half closed infinite interval $[-\infty,\infty)$. We assume that

$$-\infty < a,\ a - \infty = -\infty + a = -\infty - \infty = -\infty \text{ for any } a \in [-\infty,\infty).$$

Denote by $[-\infty,\infty)^n_{\searrow} \subset [-\infty,\infty)^n$ the set of $\mathbf{x} = (x_1,\ldots,x_n)$ where $x_1 \ge \ldots \ge x_n \ge -\infty$.

We now extend the notions of majorizatrions, Schur set, Schur order preserving function to subsets of $[-\infty,\infty)^n_{\searrow}$. Let $\mathbf{x} = (x_1,\ldots,x_n), \mathbf{y} = (y_1,\ldots,y_n) \in [-\infty,\infty)^n_{\searrow}$. Then $\mathbf{x} \preceq \mathbf{y}$, i.e. $\mathbf{x}$ is weakly majorized by $\mathbf{y}$, if the inequalities $\sum_{i=1}^{k} x_i \le \sum_{i=1}^{k} y_i$ hold for $i = 1,\ldots,n$. $\mathbf{x} \prec \mathbf{y}$, i.e. $\mathbf{x}$ majorized by $\mathbf{y}$, if $\mathbf{x} \preceq \mathbf{y}$ and $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. A set $D \subseteq [-\infty,\infty)^n_{\searrow}$ is called *Schur set* if for any $\mathbf{y} \in D$ and any $\mathbf{x} \prec \mathbf{y}$ $\mathbf{x} \in D$.

Let $I \subseteq [-\infty,\infty)$ be interval, which may be open, closed or half closed. Denote by I_o the interior of I. $f : I \to \mathbb{R}$ is called continuous if $f|I_0$ and continuous. If $a \in [-\infty,\infty)$ is an end point of I then f is continuous from the at a from the left or right respectively. Suppose that $-\infty \in I$. Then $f : I \to \mathbb{R}$ is called convex on I if f is continuous on I and a nondecreasing convex function on I_0. (See Problem 6.) f is called strictly convex on I if f continuous on I and strictly convex on I_0. If $-\infty \in I$ then f is continuous on I, and is an increasing strictly convex function on I_0. Note that the function e^x is a strictly convex function on $[-\infty, \infty)$.

Let $D \subseteq [-\infty, \infty)^n$. Then $f : D \to \mathbb{R}$ is continuous, if for any $\mathbf{x} \in D$ and any sequence of points $\mathbf{x}_k \in D, k \in \mathbb{N}$ the equality $\lim_{k\to\infty} f(\mathbf{x}_k) = f(\mathbf{x})$ holds if $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}$. D is convex if for any $\mathbf{x}, \mathbf{y} \in D$ the point $t\mathbf{x} + (1 - t)\mathbf{y} \in D$ for any $t \in (0, 1)$. For a convex D, $f : D \to \mathbb{R}$ is convex if f is continuous and $f(t\mathbf{x} + (1-t)\mathbf{y}) \le tf(\mathbf{x}) + (1-t)f(\mathbf{y})$ for any $t \in (0, 1)$. For a Schur set $D \subseteq [-\infty, \infty)^n$ $f : D \to \mathbb{R}$ is called if Schur order preserving, strict Schur order preserving, strong Schur order preserving, strict strong Schur order preserving if f is a continuous function satisfying the properties described in the beginning of §4.7. It is straightforward to generalize the results on Schur order preserving functions established in §4.7 using Problem 7.

Let the assumptions of Theorem 5.4.4 hold. For any $k \in \mathbb{N}$ let

$$\boldsymbol{\sigma}_k(T) := (\sigma_1(T), \ldots, \sigma_k(T)) \in \mathbb{R}^k_{+,\searrow}, \ \log \boldsymbol{\sigma}_k := (\log \sigma_1(T), \ldots, \log \sigma_k(T)) \in [-\infty, \infty)^k. \tag{5.4.6}$$

Theorem 5.4.4 yields

$$\begin{aligned}
&\log \boldsymbol{\sigma}_k(PT) \preceq \log \boldsymbol{\sigma}_k(P) + \log \boldsymbol{\sigma}_k(T) \text{ for any } k \in [1, \max(\text{rank P}, \text{rank T})]\\
&\log \boldsymbol{\sigma}_k(PT) \prec \log \boldsymbol{\sigma}_k(P) + \log \boldsymbol{\sigma}_k(T) \text{ for } k > \max(\text{rank P}, \text{rank T}),\\
&\log \boldsymbol{\sigma}_k(PT) \prec \log \boldsymbol{\sigma}_k(P) + \log \boldsymbol{\sigma}_k(T) \text{ if } k = \text{rank P} = \text{rank T} = \text{rank PT}.
\end{aligned} \tag{5.4.7}$$

See Problem 8.

Theorem 5.4.6 *Let* $\mathbf{U}, \mathbf{V}, \mathbf{W})$ *be IPS. Assume that* $T \in \mathrm{L}(\mathbf{V}, \mathbf{U}), P \in \mathrm{L}(\mathbf{U}, \mathbf{W})$ *and* $l \in \mathbb{N}$.

1. *Assume that* $D \subset [-\infty, \infty)^l_{\searrow}$ *be a strong Schur set containing* $\log \boldsymbol{\sigma}_l(PT), \log \boldsymbol{\sigma}_l(P) + \log \boldsymbol{\sigma}_l(T)$. *Let* $h : D \to \mathbb{R}$ *be a strong Schur order preserving function. Then* $h(\log \boldsymbol{\sigma}_l(PT)) \le h(\log \boldsymbol{\sigma}_l(PT) + \log \boldsymbol{\sigma}_l(PT))$. *Suppose furthermore that h is a strict strong Schur order preserving. Then equality holds in the above inequality if and only if equality holds in (5.4.3) for* $k = 1, \ldots, l$.

2. *Assume that* $\log \boldsymbol{\sigma}_l(PT) \prec \log \boldsymbol{\sigma}_l(P) + \log \boldsymbol{\sigma}_l(T)$, *and* $D \subset [-\infty, \infty)^l_{\searrow}$ *be a Schur set containing* $\log \boldsymbol{\sigma}_l(PT), \log \boldsymbol{\sigma}_l(P) + \log \boldsymbol{\sigma}_l(T)$. *Let* $h : D \to \mathbb{R}$ *be a Schur order preserving function. Then* $h(\log \boldsymbol{\sigma}_l(PT)) \le h(\log \boldsymbol{\sigma}_l(PT) + \log \boldsymbol{\sigma}_l(PT))$. *Suppose furthermore that h is a strict Schur order preserving. Then equality holds in the above inequality if and only if equality holds in (5.4.3) for* $k = 1, \ldots, l - 1$.

See Problem 9.

Corollary 5.4.7 *Let* $\mathbf{U},\mathbf{V},\mathbf{W})$ *be IPS. Assume that* $T\in \mathrm{L}(\mathbf{V},\mathbf{U}), P\in \mathrm{L}(\mathbf{U},\mathbf{W})$ *and* $l\in\mathbb{N}$. *Assume that* $\log\boldsymbol{\sigma}_l(PT)\preceq\log\boldsymbol{\sigma}_l(P)+\log\boldsymbol{\sigma}_l(T)$, *and* $I\subset[-\infty,\infty)$ *is an interval set containing* $\log\boldsymbol{\sigma}_1(P)+\log\boldsymbol{\sigma}_1(T),\log\sigma_l(PT)$. *Let* $h: I\to\mathbb{R}$ *be a convex function. Then*

$$\sum_{i=1}^{l} h(\log\sigma_i(PT))\le\sum_{i=1}^{l} h(\log\sigma_i(P))+h(\log\sigma_i(T)). \tag{5.4.8}$$

Corollary 5.4.8 *Let* $\mathbf{U},\mathbf{V},\mathbf{W})$ *be IPS. Assume that* $T\in \mathrm{L}(\mathbf{V},\mathbf{U}), P\in \mathrm{L}(\mathbf{U},\mathbf{W})$ *and* $l\in\mathbb{N}$. *Then for any* $t>0$

$$\sum_{i=1}^{l}\sigma_i(PT)^t\ge\sum_{i=1}^{l}\sigma_i(P)^t\sigma_i(T)^t. \tag{5.4.9}$$

equality holds if and only if one has equality sign in (5.4.3) for $k=1,\ldots,l$.

Proof. Observe that the function $h:[-\infty,\infty)^l_{\searrow}\to\mathbb{R}$ given by $h((x_1,\ldots,x_l))=\sum_{i=1}^l e^{tx_i}$ is a strictly strongly Schur order preserving for any $t>0$. □

The following theorem improves the results of Theorem 4.10.12.

Theorem 5.4.9 *Let* $\mathbf{V}$ *be an* n*-dimensional IPS vector space over* $\mathbb{C}$ *and assume that* $T\in\mathrm{L}(\mathbf{V})$. *Let* $\lambda_1(T),\ldots,\lambda_n(T)\in\mathbb{C}$ *be the eigenvalues of* T *counted with their multiplicities and arranged in order* $|\lambda_1(T)\ge\ldots\ge|\lambda_n(T)|$. *Let* $\boldsymbol{\lambda}_a(T):=(|\lambda_1(T)|,\ldots,|\lambda_n(T)|)$ *and* $\boldsymbol{\lambda}_{a,k}(T):=(|\lambda_1(T)|,\ldots,|\lambda_k(T)|)$ *for* $k=1,\ldots,n$. *Then*

$$\prod_{i=1}^{l}|\lambda_i(T)|\le\prod_{i=1}^{l}\sigma_i(T)\ \textit{for}\ l=1,\ldots,n-1,\ \textit{and}\ \prod_{i=1}^{n}|\lambda_i(T)|=\prod_{i=1}^{n}\sigma_i(T). \tag{5.4.10}$$

For $l=1,\ldots,k\le n$ *equalities hold in the above inequalities if and only if the conditions 1 and 2 of Theorem 4.10.12 hold.*

In particular $\log\boldsymbol{\lambda}_{a,k}(T)\preceq\log\boldsymbol{\sigma}_k(T)$ *for* $k=1,\ldots,n-1$ *and* $\log\boldsymbol{\lambda}_a(T)\prec\log\boldsymbol{\sigma}(T)$.

Proof. By Theorem 4.10.12 $|\lambda_1(\wedge^l T)|\le\sigma_1(\wedge^l T)$. Use Problem 7 and Proposition 5.4.2 to deduce the inequalities in (5.4.10). The equality $\prod_{i=1}^n|\lambda_i(T)|=\prod_{i=1}^l\sigma_i(T)$ is equivalent to the identity $|\det T|^2=\det TT^*$.

Suppose that for $l=1,\ldots,k\le n$ equalities hold in (5.4.10). Then $|\lambda_i(T)|=\sigma_i(T)$ for $i=1,\ldots,k$. Hence equality holds in (4.10.14). Theorem 4.10.12 implies that conditions *1,2* hold. Vice versa, assume that the conditions *1,2* of Theorem 4.10.12 hold. Then from the proof of Theorem 4.10.12 it follows that $|\lambda_i(T)| = \sigma_i(T)$ for $i = 1, \ldots, k$. Hence for $l = 1, \ldots, k$ equalities hold in (5.4.10). □

Corollary 5.4.10 *Let* $\mathbf{V}$ *be an* n *dimensional IPS. Assume that* $T \in \mathrm{L}(\mathbf{V})$.

1. *Assume that* $k \in [1, n-1] \cap \mathbb{N}$ *and* $D \subset [-\infty, \infty)^k_{\searrow}$ *be a strong Schur set containing* $\log \boldsymbol{\sigma}_k(T)$. *Let* $h : D \to \mathbb{R}$ *be a strong Schur order preserving function. Then* $h(\log \boldsymbol{\lambda}_k(T)) \le h(\log \boldsymbol{\sigma}_k(T))$. *Suppose furthermore that* h *is a strict strong Schur order preserving. Then equality holds in the above inequality if and only if equality holds in (5.4.10) for* $l = 1, \ldots, k$.

2. *Let* $I \subset [-\infty, \infty)$ *be an interval containing* $\log \sigma_1(T), \log \sigma_k(T), \log |\lambda_k(T)|$. *Assume that* $f : I \to \mathbb{R}$ *is a convex nondecreasing function. Then* $\sum_{i=1}^k f(\log |\lambda_i(T)|) \le \sum_{i=1}^k f(\log |\sigma_i(T)|)$. *If* f *is a strictly convex increasing function on* I *then equality holds if and only if equality holds in (5.4.10) for* $l = 1, \ldots, k$. *In particular for any* $t > 0$

$$\sum_{i=1}^k |\lambda_i(T)|^t \le \sum_{i=1}^k \sigma_i(T)^t. \tag{5.4.11}$$

Equality holds if and only if equality holds in (5.4.10) for $l = 1, \ldots, k$.

3. *Assume that* $D \subset [-\infty, \infty)^n_{\searrow}$ *is a Schur set containing* $\log \boldsymbol{\sigma}_n(T)$. *Let* $h : D \to \mathbb{R}$ *be a Schur order preserving function. Then* $h(\log \boldsymbol{\lambda}_a(T)) \le h(\log \boldsymbol{\sigma}_n(T))$. *Suppose furthermore that* h *is a strict Schur order preserving. Then equality holds in the above inequality if and only if equality holds* T *is a normal operator.*

Problems

1. Let $\mathbf{V}_i$ be an n_i-dimensional vector space with the inner product $\langle \cdot, \cdot \rangle_i$ for $i = 1, \ldots, k$.

   (a) Let $\mathbf{e}_{1,i}, \ldots, \mathbf{e}_{n_i,i}$ be an orthonormal basis of $\mathbf{V}_i$ with respect $\langle \cdot, \cdot \rangle_i$ for $i = 1, \ldots, k$. Let $\langle \cdot, \cdot \rangle$ be the inner product in $\mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i$ such that $\otimes_{i=1}^k \mathbf{e}_{j_i,i}$ where $j_i = 1, \ldots, n_i, i = 1, \ldots, k$ is an orthonormal basis of $\mathbf{Y}$. Show that (5.4.1) holds.

(b) Prove that there exists a unique inner product on $\mathbf{Y}$ satisfying (5.4.1).

2. Prove Proposition 5.4.1.

3. Let $\mathbf{V}$ be an n-dimensional IPS with an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Let $\mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i, \mathbf{V}_1 = \ldots = \mathbf{V}_k = \mathbf{V}$ be an IPS with the canonical inner product $\langle \cdot, \cdot \rangle_{\mathbf{Y}}$. Show

   (a) Let $k \in \mathbb{N} \cap [1, n]$. Then the subspace $\bigwedge^k \mathbf{V}$ of $\mathbf{Y}$ has an orthonormal basis

   $$\frac{1}{\sqrt{k!}} \mathbf{e}_{i_1} \wedge \ldots \wedge \mathbf{e}_{i_k}, 1 \le i_1 < i_2 < \ldots < i_k \le n.$$

   (b) Let $k \in \mathbb{N}$. Then the subspace $\mathrm{Sym}^{\mathrm{k}}\mathbf{V}$ of $\mathbf{Y}$ has an orthonomal basis $\alpha(i_1, \ldots, i_k)\mathrm{sym}^k(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}), 1 \le i_1 \le \ldots \le i_k \le n$. The coefficient $\alpha(i_1, \ldots, i_k)$ is given as follows. Assume that $i_1 = \ldots = i_{l_1} < i_{l_1+1} = \ldots = i_{l_1+l_2} < \ldots < i_{l_1+\ldots+l_{r-1}+1} = \ldots = i_{l_1+\ldots+l_r}$, where $l_1 + \ldots + l_r = k$. Then $\alpha(i_1, \ldots, i_k) = \frac{1}{\sqrt{k! l_1! \ldots l_r!}}$.

4. (a) Prove Proposition 5.4.2.

   (b) Prove Corollary 5.4.3.

5. Let $\mathbf{U}, \mathbf{V}$ be IPS of dimensions n and m respectively. Let $T \in \mathrm{L}(\mathbf{V}, \mathbf{U})$ and assume that we chose orthonormal base $[\mathbf{c}_1, \ldots, \mathbf{c}_n], [\mathbf{d}_1, \ldots, \mathbf{d}_m]$ of $\mathbf{V}, \mathbf{U}$ respectively satisfying (4.9.5). Suppose furthermore that

   $$\sigma_1(T) = \ldots = \sigma_{l_1}(T) > \sigma_{l_1+1}(T) = \ldots = \sigma_{l_2}(T) > \ldots >$$
   (5.4.12)
   $$\sigma_{l_{p-1}+1}(T) = \ldots = \sigma_{l_p}(T) > 0, \ 1 \le l_1 < \ldots < l_p = \text{rank T}.$$

   Let

   $$\mathbf{V}_i := \text{span}\,(\mathbf{c}_{l_{i-1}+1}, \ldots, \mathbf{c}_{l_i}), \quad i = 1, \ldots, p, \ l_0 := 0. \tag{5.4.13}$$

   (a) Let $k = l_i$ for some $i \in [1, p]$. Show that $\sigma_1(\wedge^k T) > \sigma_2(\wedge^k T)$. Furthermore the vector $\mathbf{c}_1 \wedge \ldots \wedge \mathbf{c}_k$ is a unique right singular vector, (up to a multiplication by scalar), of $\wedge^k T$ corresponding to $\sigma_1(\wedge^k T)$. Equivalently, the one dimensional subspace spanned by the the right singular vectors of $\wedge^k T$ is given by $\bigwedge^k \oplus_{j=1}^i \mathbf{V}_i$.

(b) Assume that $l_i - l_{i-1} \geq 2$ and $l_{i-1} < k < l_i$ for some $1 \leq i \leq p$. Show that

$$(5.4.14) \qquad \sigma_1(\wedge^k T) = \ldots = \sigma_{\binom{l_i-l_{i-1}}{k-li-1)}}(T) > \sigma_{\binom{l_i-l_{i-1}}{k-l_{i-1}}+1}(T).$$

The subspace spanned by all right singular vectors of $\wedge^k T$ corresponding to $\sigma_1(\wedge^k T)$ is given by the subspace:

$$(\bigwedge^{l_{i-1}} \oplus_{j=1}^{i-1} \mathbf{V}_j) \bigwedge (\bigwedge^{k-l_{i-1}} \mathbf{V}_i).$$

6. Let $I := [-\infty, a), a \in \mathbb{R}$ and assume that $f : I \to \mathbb{R}$ is continuous. f is called convex on I if $f(tb + (1-t)c) \leq tf(b) + (1-t)f(c)$ for any $b, c \in I$ and $t \in (0,1)$. We assume that $t(-\infty) = -\infty$ for any $t > 0$. Show that if f is convex on I if and only f is a convex nondecreasing bounded below function on I_o.

7. $D \subset [-\infty, \infty)^n_{\searrow}$ such that $D' := D \cap \mathbb{R}^n$ is nonempty. Assume that $f : D \to \mathbb{R}$ is continuous. Show

   (a) D is a Schur set if and only if D' is a Schur set.

   (b) f is Schur order preserving if and only if $f|D'$ is Schur order preserving.

   (c) f is strict Schur order preserving if and only if $f|D'$ is string Schur order preserving.

   (d) f is strong Schur order preserving if and only if $f|D'$ is strong Schur order preserving.

   (e) f is strict strong Schur order preserving if and only if $f|D'$ is strict strong Schur order preserving.

8. Let the assumptions of Theorem 5.4.4 hold. Assume that rank P = rank T = rank PT. Let $k =$ rank P. Show that the arguments of the proof of Theorem 5.4.5 implies that $\prod_{i=1}^k \sigma_i(PT) = \prod_{i=1}^k \sigma_i(P)\sigma_i(T)$. Hence $\log \boldsymbol{\sigma}_k(PT) \prec \log \boldsymbol{\sigma}_k(P) + \log \boldsymbol{\sigma}_k(T)$.

9. Prove Theorem 5.4.6 using the results of Section 4.7.

10. Show that under the assumptions of Theorem 4.10.8 one has the inequality $\sum_{i=1}^l \sigma_i(S^*T)^t \leq \sum_{i=1}^l \sigma_i(S)^t \sigma_i(T)^t$ for any $l \in \mathbb{N}$ and $t > 0$.

11. (a) Let the assumptions of Theorem 5.4.9 hold. Show that (5.4.10) imply that $\lambda_i(T) = 0$ for $i >$ rank T.

(b) Let $\mathbf{V}$ be a finite dimensional vector field over the algebraically closed field $\mathbb{F}$. Let $T \in \mathrm{L}(\mathbf{V})$. Show that the number of nonzero eigenvalues, counted with their multiplicities, does not exceed rank T. (*Hint*: Use the Jordan canonical form of T.)

5.5 Tensor products of exponents

Proposition 5.5.1 *(The Lie-Trotter formula.) Let $a > 0$, $A(t) : (-a, a) \to \mathbb{C}^{n\times n}$, assume that*

$$\lim_{t\to 0} \frac{A(t)}{t} = B. \tag{5.5.1}$$

Then for any $s \in \mathbb{R}$

$$\lim_{t\to 0}(I + A(t))^{\frac{s}{t}} = e^{sB} \tag{5.5.2}$$

Proof. The assumption (5.5.1) yields that $B(t) := \frac{1}{t}A(t), t \neq 0, B(0) := B$ is continuous at $t = 0$. Hence, there exists $\delta > 0$ such that for $|t| \le \delta$ $||B(t)||_2 = \sigma_1(B(t)) \le c$. Without loss of generality we can assume that $c\delta \le \frac{1}{2}$. Hence, for $|t| \le \delta$ all the eigenvalues of $A(t)$ are in the disk $|z| \le \frac{1}{2}$. Consider the analytic function $\log(1 + z)$ in the unit disk $|z| < 1$ with $\log 1 = 0$. The results of §3.1 that for $|t| < \delta$

$$\frac{s}{t}\log(I + A(t)) = s\sum_{i=1}^{\infty}\frac{(-t)^{i-1}B(t)}{i} = sB(t) + \sum_{i=2}^{\infty}\frac{s(-t)^{i-1}B(t)^i}{i}.$$

Recall that

$$||\sum_{i=2}^{\infty}\frac{s(-t)^{i-1}B(t)^i}{i}||_2 \le \sum_{i=2}^{\infty}||\frac{st^{i-1}B(t)^i}{i}||_2 \le$$
$$\sum_{i=2}^{\infty}\frac{|s||t|^{i-1}||B(t)||_2^i}{i} \le \frac{|s|(-|t|c - \log(1-|t|c)}{|t|}.$$

Hence

$$\lim_{t\to 0}(I + A(t))^{\frac{s}{t}} = \exp(\lim_{t\to 0}\frac{s}{t}\log(I + A(t)) = e^{sB}.$$

Proposition 5.5.2 *Let $\mathbf{V}_i$ is a n_i-dimensional vector space for $i = 1, \ldots, k$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Let $\mathbf{Y} := \otimes_{i=1}^k \mathbf{V}_i$. Assume that $A_i \in \mathrm{L}(\mathbf{V}_i), i =$*

$1, \ldots$. *Then*

$$\otimes_{i=1}^k e^{A_i t} = e^{(A_1,\ldots,A_k)_\otimes\, t} \in \mathrm{L}(\otimes_{i=1}^k \mathbf{V}_i), \text{ where}$$

(5.5.3)

$$(A_1,\ldots,A_k)_\otimes := \sum_{i=1}^k I_{n_1} \otimes \ldots \otimes I_{n_{i-1}} \otimes A_{n_i} \otimes I_{n_{i+1}} \otimes \ldots \otimes I_{n_k} \in \mathrm{L}(\otimes_{i=1}^k \mathbf{V}_i),$$

for any $t \in \mathbb{F}$.

See Problem 1.

Definition 5.5.3 *Let* $\mathbf{V}$ *be a* n*-dimensional vector space over* $\mathbb{F}$. *Assume that* $A \in \mathrm{L}(\mathbf{V})$. *Denote by* $A_{\wedge^k}$ *the restriction of* $(\underbrace{A,\ldots,A}_{k})_\otimes$ *to* $\wedge^k \mathbf{V}$.

Corollary 5.5.4 *Let the assumptions of Definition 5.5.3 hold for* $\mathbb{F} = \mathbb{R}, \mathbb{C}$. *Then* $\wedge^k e^{At} = e^{A_{\wedge^k} t}$ *for any* $t \in \mathbb{F}$.

Definition 5.5.5 *A subspace* $\mathbf{U} \subset \mathbf{H}_n$ *is called a commuting subspace if any two matrices* $A, B \in \mathbf{U}$ *commute.*

Recall that if $A, B \in \mathbf{H}_n$ then each eigenvalue of $e^A e^B$ are positive. (See Problem 4.)

Theorem 5.5.6 *Let* $\mathbf{U}, \mathbf{V} \subset \mathbf{H}_n$ *be two computing subspaces. Then the functions*

$$(5.5.4) \qquad f_k : \mathbf{U} \times \mathbf{V} \to \mathbb{R}, \quad f_k(A,B) := \sum_{i=1}^k \log \lambda_i(e^A e^B), \ k = 1, \ldots, n,$$

are convex functions on $\mathbf{U} \times \mathbf{V}$. *(The eigenvalues of* $e^A e^B$ *are arranged in a decreasing order.)*

Proof. Since $e^A e^B$ has positive eigenvalues for all pairs $A, B \in \mathbf{H}_n$ it follows that each f_k is a continuous function on $\mathbf{U} \times \mathbf{V}$. Hence it is enough to show that

(5.5.5)

$$f_k(\frac{1}{2}(A_1+A_2), \frac{1}{2}(B_1+B_2)) \le \frac{1}{2}(f_k(A_1,B_1)+f_k(A_2,B_2)), \quad k = 1, \ldots, n,$$

for any $A_1, A_2 \in \mathbf{U}, B_1, B_2 \in \mathbf{V}$. (See Problem 5.)

We first consider the case $k = 1$. Since $A_1A_2 = A_2A_1, B_1B_2 = B_2B_1$ it follows that

$$e^{\frac{1}{2}(A_1+A_2)}e^{\frac{1}{2}(B_1+B_2)} = e^{\frac{1}{2}A_2}e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2}.$$

Observe next that

$$e^{-\frac{1}{2}A_2}(e^{\frac{1}{2}A_2}e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2})e^{\frac{1}{2}A_2} = e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2} \Rightarrow$$
$$f_1(\frac{1}{2}(A_1+A_2), \frac{1}{2}(B_1+B_2)) = \lambda_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2}).$$

Hence

$$(5.5.6) \qquad \lambda_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2}) \le \sigma_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2}) \le$$
$$\sigma_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1})\sigma_1(e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2}) = \lambda_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}A_1})^{\frac{1}{2}}\lambda_1(e^{\frac{1}{2}A_2}e^{\frac{1}{2}B_2}e^{\frac{1}{2}B_2}e^{\frac{1}{2}A_2})^{\frac{1}{2}} =$$
$$\lambda_1(e^{\frac{1}{2}A_1}e^{\frac{1}{2}A_1}e^{\frac{1}{2}B_1}e^{\frac{1}{2}B_1})^{\frac{1}{2}}\lambda_1(e^{\frac{1}{2}A_2}e^{\frac{1}{2}A_2}e^{\frac{1}{2}B_2}e^{\frac{1}{2}B_2})^{\frac{1}{2}} = \lambda_1(e^{A_1}e^{B_1})^{\frac{1}{2}}\lambda_1(e^{A_2}e^{B_2})^{\frac{1}{2}}.$$

This proves the convexity of f_1.

We now show the convexity of f_k. Use Problem 6 to deduce that we may assume that $\mathbf{U} = U\mathbf{D}_n(\mathbb{R})U^*, \mathbf{V} = V\mathbf{D}_n(\mathbb{R})V^*$ for some $U, V \in \mathbf{U}(n)$. Let $\mathbf{U}_k, \mathbf{V}_k \subset \mathbf{H}_{\binom{n}{k}}$ be two commuting subspaces defined in Problem 6(d). The above result imply that $g : \mathbf{U}_k \times \mathbf{V}_k \to \mathbb{R}$ given by $g(C, D) = \log(e^Ce^D)$ is convex. Hence

$$g(\frac{1}{2}((A_1)_{\wedge^k}+(A_2)_{\wedge^k}), \frac{1}{2}((B_1)_{\wedge^k}+(B_2)_{\wedge^k}) \le \frac{1}{2}(g((A_1)_{\wedge^k},(B_1)_{\wedge^k})+g((A_1)_{\wedge^k},(B_1)_{\wedge^k})).$$

The definitions of $\underbrace{(A, \ldots, A)}_{k}{}_{\otimes}$ and $A_{\wedge^k}$ yield the equality $\frac{1}{2}(A_{\wedge^k} + B_{\wedge^k}) = (\frac{1}{2}(A+B))_{\wedge^k}$. Use Problem 3 to deduce that the convexity of g implies the convexity of f_k. □

Theorem 5.5.7 *Let $A, B \in \mathbf{H}_n, k \in [1, n] \cap \mathbb{N}$. Assume that $f_k(tA, tB), t \in \mathbb{R}$ is defined as in (5.5.4). Then the function $\frac{f_k(tA,tB)}{t}$ increases on $(0, \infty)$. In particular*

$$(5.5.7) \qquad \sum_{i=1}^{k} \lambda_i(A+B) \le \sum_{i=1}^{k} \log\lambda_i(e^Ae^B), \quad k = 1, \ldots, n,$$
$$(5.5.8) \qquad \operatorname{tr} e^{A+B} \le \operatorname{tr}(e^Ae^B)..$$

Equality in (5.5.8) holds if and only if $AB = BA$.

Proof. Theorem 5.5.6 yields that $g_k(t) := f_k(tA, tB)$ is convex on $\mathbb{R}$. (Assume $\mathbf{U} = \text{span (A)}, \mathbf{V} = \text{span (B)}$). Note that $g_k(0) = 0$. Problem 8 implies that $\frac{g_k(t)}{t}$ nondecreasing on $(0,\infty)$. Problem 9 implies that $\lim_{t\searrow 0} \frac{g_k(t)}{t} = \sum_{i=1}^k \lambda_i(A+B)$. Hence (5.5.7) holds. Use Problem 7 to deduce that in (5.5.7) equality holds for $k = n$. Hence (5.5.7) is equivalent to

$$\boldsymbol{\lambda}(A+B) \prec \log \boldsymbol{\lambda}(e^A e^B). \tag{5.5.9}$$

Apply the convex function e^x to this relation to deduce (5.5.8).

We now show that equality holds in (5.5.8) if and only if $AB = BA$. Clearly if $AB = BA$ then $e^{At}e^{Bt} = e^{(A+B)t}$, hence we have equality in (5.5.8).

It is left to show the claim that equality in (5.5.8) implies that A and B commutes. We prove this claim by induction on n. For $n = 1$ this claim trivially holds. Assume that this claim holds for $n = m - 1$. Let $n = m$. Since e^x is strictly convex it follows that equality in (5.5.8) yields the equality

$AB = BA$ hence we have equality in (5.5.8) $\boldsymbol{\lambda}(A+B) = \log \boldsymbol{\lambda}(e^A e^B)$ in particular $\lambda(A+B) = \log \lambda_1(e^A e^B)$. Hence $\frac{g_1(t)}{t}$ is a constant function on $(0,1]$, i.e. $g_1(t) = Kt$ for $t \in [0,1]$. Use the inequalities (5.5.6) for $k = 1, A_1 = At, A_2 = A, B_1 = Bt, B_2 = B, t \in (0,1)$ to conclude that we must have equalities in all inequalities in (5.5.6). In particular we first must have the equalities $\lambda_1(e^A e^B) = \sigma_1(e^A e^B)$. Similarly we conclude that

$$\prod_{i=1}^k \lambda_i(e^A e^B) = \prod_{i=1}^k \sigma_i(e^A e^B),\ i = 1,\ldots,n \Rightarrow \lambda_i(e^A e^B) = \sigma_i(e^A e^B),\ i = 1,\ldots,n.$$

Theorem 4.10.12 yields that $e^A e^B$ is a normal matrix.

Assume first that all the eigenvalues of $e^A e^B$ are equal. Hence $e^A e^B = cI \Rightarrow e^A e^B = e^B e^A \Rightarrow AB = BA$. Assume now that $e^A e^B$ has l-distinct eigenvalues $\gamma_1 > \ldots > \gamma_l > 0$. Let $\mathbf{W}_i$ be the eigenspace of $e^A e^B$ corresponding to γ_i. Clearly $e^B \mathbf{W}_i$ is the eigenspace of $e^B e^A$ corresponding γ_i. Hence $e^B \mathbf{W}_i = \mathbf{W}_i \Rightarrow B\mathbf{W}_i \subseteq \mathbf{W}_i$. Similarly $e^A \mathbf{W}_i = \mathbf{W}_i \Rightarrow A\mathbf{W}_i \subseteq \mathbf{W}_i$. Since $e^A e^B|_{\mathbf{W}_i} = \gamma_i I_{\mathbf{W}_i}$ it follows that $AB|_{\mathbf{W}_i} = BA|_{\mathbf{W}_i}$ for $i = 1,\ldots,k$. Hence $e^A e^B = e^B e^A \Rightarrow AB = BA$. □

Let

$$C(t) := \frac{1}{t} \log e^{\frac{1}{2}At} e^{Bt} e^{\frac{1}{2}At} \in \mathbf{H}_n, \quad t \in \mathbb{R}\backslash\{\text{o}\}. \tag{5.5.10}$$

$tC(t)$ is the unique hermitian logarithm of a positive definite hermitian matrix $e^{\frac{1}{2}At}e^{Bt}e^{\frac{1}{2}At}$, which is similar to $e^{At}e^{Bt}$. Proposition 5.5.1 yields

$$\lim_{t\to 0} C(t) = C(0) := A + B. \tag{5.5.11}$$

(See Problem 11.) In what follows we give a complementary formula to (5.5.11).

Theorem 5.5.8 *Let $A, B \in \mathbf{H}_n$ and assume that $C(t)$ is be the hermitian matrix defined as (5.5.10). Then $\sum_{i=1}^k \lambda_i(C(t))$ are nondecreasing functions on $[0, \infty)$ for $k = 1, \ldots, n$ satisfying*

$$\boldsymbol{\lambda}(C(t)) \prec \boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B). \tag{5.5.12}$$

Moreover there exists $C \in \mathbf{H}_n$ such that

$$\lim_{t\to\infty} C(t) = C, \tag{5.5.13}$$

and C commutes with A. Furthermore there exist two permutations ϕ, ψ on $\{1, \ldots, n\}$ such that

$$\lambda_i(C) = \lambda_{\phi(i)}(A) + \lambda_{\psi(i)}(B), \quad i = 1, \ldots, n. \tag{5.5.14}$$

Proof. Assume that $t > 0$ and let $\lambda_i(t) = e^{t\lambda_i(C(t))}, i = 1, \ldots, n$ be the eigenvalues of $G(t) := e^{\frac{1}{2}At}e^{Bt}e^{\frac{1}{2}At}$. Clearly

$$\lambda_1(t) = ||e^{\frac{1}{2}At}e^{Bt}e^{\frac{1}{2}At}||_2 \le ||e^{\frac{1}{2}At}||_2^2 ||e^{Bt}||_2^2 = e^{(\lambda_1(A)+\lambda_1(B))t}.$$

By considering $\wedge^k G(t)$ we deduce

$$\prod_{i=1}^k \lambda_i(t) \le e^{t\sum_{i=1}^k \lambda_i(A)+\lambda_i(B)}, \quad k = 1, \ldots, n, \quad t > 0.$$

Note that for $k = n$ equality holds. (See Problem 7.) Hence (5.5.12) holds. Let $g_k(t)$ be defined as in the proof of Theorem 5.5.7. Clearly $\frac{g_k(t)}{t} = \sum_{i=1}^k \lambda_i(C(t))$. Since $\frac{g_k(t)}{t}$ is nondecreasing we deduce that $\sum_{i=1}^k \lambda_i(C(t))$ is nondecreasing on $[0, \infty)$. Furthermore (5.5.12) shows that $\frac{g_k(t)}{t}$ is bounded. Hence $\lim_{t\to\infty} \frac{g_k(t)}{t}$ exists for each $k = 1, \ldots, n$, which is equivalent to

$$\lim_{t\to\infty} \lambda_i(C(t)) = \omega_i, \quad i = 1, \ldots, n. \tag{5.5.15}$$

Let

$$\omega_1 = \ldots = \omega_{n_1} > \omega_{n_1+1} = \ldots = \omega_{n_2} > \ldots > \omega_{n_{l-1}+1} = \ldots \omega_{n_r}, \qquad n_0 = 0 < n_1 < \ldots < n_r = n. \tag{5.5.16}$$

Let $\omega_0 := \omega_1 + 1, \omega_{n+1} = \omega_n - 1$. Hence for $t > T$ the open interval $(\frac{\omega_{i-1}+\omega_i}{2}, \frac{\omega_i+\omega_{i+1}}{2})$ contains exactly $n_i - n_{i-1}$ eigenvalues of $C(t)$ for $i = 1, \ldots, r$. In what follows we assume that $t > T$. Let $P_i(t) \in \mathbf{H}_n$ be the orthogonal projection on the eigenspace of $C(t)$ corresponding to the eigenvalues $\lambda_{n_{i-1}+1}(C(t)), \ldots, \lambda_{n_i}(C(t))$ for $i = 1, \ldots, r$. Observe that $P_i(t)$ is the orthogonal projection on the eigenspace of $G(t)$ the eigenspace corresponding to the eigenvalues $\lambda_{n_{i-1}+1}(t), \ldots, \lambda_{n_i}(t)$ for $i = 1, \ldots, r$. The equality (5.5.13) is equivalent to

$$\lim_{t\to\infty} P_i(t) = P_i, \quad i = 1, \ldots, r. \tag{5.5.17}$$

The claim that $CA = AC$ is equivalent to the claim that $AP_i = P_iA$ for $i = 1, \ldots, r$.

We first show these claims for $i = 1$. Assume that the eigenvalues of A and B are of the form

$$\lambda_1(A) = \ldots = \lambda_{l_1}(A) = \alpha_1 > \lambda_{l_1+1}(A) = \ldots = \lambda_{l_2}(A) = \alpha_2 > \ldots > \lambda_{l_{p-1}+1}(A) = \ldots = \lambda_{l_p}(A) = \alpha_p,$$
$$\lambda_1(B) = \ldots = \lambda_{m_1}(B) = \beta_1 > \lambda_{m_1+1}(B) = \ldots = \lambda_{m_2}(B) = \beta_2 > \ldots > \lambda_{m_{q-1}+1}(B) = \ldots = \lambda_{m_q}(B) = \beta_q,$$
$$l_0 = 0 < l_1 < \ldots < l_p = n, \quad m_0 = 0 < m_1 < \ldots < m_q = n. \tag{5.5.18}$$

Note that if either $p = 1$ or $q = 1$, i.e. either A or B is of the form aI, then the theorem trivially holds. Assume that $p, q > 1$. Let Q_i, R_j be the orthogonal projections on the eigenspaces of A and B corresponding to the eigenvalues α_i and β_j respectively, for $i = 1, \ldots, p, j = 1, \ldots, q$. So

$$e^{\frac{1}{2}At} = \sum_{i=1}^{p} e^{\frac{1}{2}\alpha_i t} Q_i, \; e^{Bt} = \sum_{j=1}^{q} e^{\beta_j t} R_j, \; G(t) = \sum_{i_1=i_2=j=1}^{p,p,q} e^{\frac{1}{2}(\alpha_{i_1}+\alpha_{i_2}+\beta_j)t} Q_{i_1} R_j Q_{i_2}.$$

Observe next that

$$\mathrm{rank}\ \mathrm{Q_iR_j} = \mathrm{rank}\ (\mathrm{Q_iR_j})^* = \mathrm{rank}\ \mathrm{R_jQ_i} = \mathrm{rank}\ (\mathrm{Q_iR_j})(\mathrm{Q_iR_j})^* = \mathrm{rank}\ \mathrm{Q_iR_j^2Q_i} =$$
$$Q_{i_1}R_jQ_{i_2} = (Q_{i_1}R_j)(R_jQ_{i_2}) \neq 0 \Rightarrow Q_{i_1}R_jQ_{i_1} \neq 0, Q_{i_2}R_jQ_{i_2} \neq 0,$$
$$\mathcal{K} := \{(i,j) \in \{1, \ldots, p\} \times \{1, \ldots, q\},\ Q_iR_j \neq 0\},$$
$$I = (\sum_{i=1}^{p} Q_i)(\sum_{j=1}^{q} R_j) = \sum_{i,j=1}^{p,q} Q_iR_j = \sum_{(i,j)\in\mathcal{K}} Q_iR_j.$$

(See Problem 14.) Let

$$(5.5.19)\gamma_1 := \max_{(i,j)\in\mathcal{K}} \alpha_i + \beta_j, \quad \mathcal{K}_1 := \{(i,j)\in\mathcal{K}, \alpha_i+\beta_j = \gamma_1\},$$
$$n_1' = \sum_{(i,j)\in\mathcal{K}_1} \mathrm{rank}\ \mathrm{Q_iR_jQ_i}, \quad \gamma_1' = \max_{\mathrm{(i,j)}\in\mathcal{K}\setminus\mathcal{K}_1} \alpha_\mathrm{i} + \beta_\mathrm{j},.$$

From the above equalities we deduce that $\mathcal{K}_1 \neq \emptyset$. Assume that $(i,j),(i',j') \in \mathcal{K}_1$ are distinct pairs. From the maximality of γ_1 and the definition of $\mathcal{K}_1$ it follows that $i\neq i', j\neq j'$. Hence $Q_iR_jQ_i(Q_{i'}R_{j'}Q_{i'}) = 0$. Furthermore γ_1' is well defined and $\gamma_1' < \gamma_1$. Let

$$D_1(t) := \sum_{(i,j)\in\mathcal{K}\setminus\mathcal{K}_1} e^{(\alpha_i+\beta_j-\gamma_1)t}Q_iR_jQ_i +$$
$$\sum_{(i_1,j),(i_2,j)\in\mathcal{K},i_1\neq i_2} e^{\frac{1}{2}(\alpha_{i_1}+\alpha_{i_2}+2\beta_j-2\gamma_1)t}Q_{i_1}R_jQ_{i_2}.$$

$$(5.5.20) \qquad D = \sum_{(i,j)\in\mathcal{K}_1} Q_iR_jQ_i, \quad D(t) = D + D_1(t).$$

Then $n_1' = \mathrm{rank\ D}$. (See Problem 15b). We claim that

$$(5.5.21) \qquad \omega_1 = \gamma_1, \quad n_1 = n_1'.$$

From the above equalities and definitions we deduce $G(t) = e^{\gamma_1 t}D(t)$. Hence $\lambda_i(t) = e^{\gamma_1 t}\lambda_i(D(t))$. As each term $e^{\frac{1}{2}(\alpha_{i_1}+\alpha_{i_2}+2\beta_j-2\gamma_1)t}$ appearing in $D_1(t)$ is bounded above by $e^{-\frac{1}{2}(\gamma_1-\gamma_1')t}$ we deduce $D_1(t) = e^{\frac{1}{2}(\gamma_1'-\gamma_1)t}D_2(t)$ and $\leq ||D_2(t)||_2 \leq K$. Hence $\lim_{t\to\infty} D(t) = D$. Since $\mathrm{rank\ D = n_1'}$ we deduce that we have $0.5\lambda_i(D) < \lambda_i(D(t)) \leq 2\lambda_i(D)$ for $i = 1,\dots,n_1'$. If $n_1' < n$ then from Theorem 4.4.6 we obtain that

$$\lambda_i(D(t)) = \lambda_i(D+D_1(t)) \leq \lambda_i(D)+\lambda_1(D_1(t)) = \lambda_1(D_1(t)) \leq e^{\frac{1}{2}(\gamma_1'-\gamma_1)t}K,$$

for $i = n_1+1,\dots,n$. Hence

$$\omega_i = \gamma_1,\ i = 1,\dots,n_1', \quad \omega_i \leq \frac{1}{2}(\gamma+\gamma_1'),\ i = n_1'+1,\dots,n,$$

which shows (5.5.21). Furthermore $\lim_{t\to\infty} P_1(t) = P_1$, where P_1 is the projection on $D\mathbb{C}^n$. Since $Q_iQ_{i'} = \delta_{ii'}Q_i$ it follows that $Q_{i'}D = DQ_{i'}$ for $i' = 1,\dots,p$. Hence $AD = DA \Rightarrow AP_1 = P_1A$. Furthermore $P_1\mathbb{C}^n$ is a direct sum of the orthogonal subspaces $Q_iP_jQ_j\mathbb{C}^n, (i,j)\in\mathcal{K}_1$, which are the eigen-subspaces of A corresponding to $\lambda_i(A)$ for $(i,j)\in\mathcal{K}_1$.

We now define partially the permutations ϕ, ψ. Assume that $\mathcal{K}_1 = \{(i_1, j_1), \ldots, (i_o, j_o)\}$. Then $\omega_1 = \gamma_1 = \alpha_{i_k} + \beta_{j_k}$ for $k = 1, \ldots, o$. Let $e_0 = 0$ and $e_k = e_{k-1} + \mathrm{rank}\ \mathrm{Q_{i_k} R_{j_k} Q_{i_k}}$ for $k = 1, \ldots, o$. Note that $e_o = n_1'$. Define
(5.5.22)
$\phi(s) = l_{i_k - 1} + s - e_{k-1},\ \psi(s) = m_{j_k - 1} + s - e_{k-1},$ for $s = e_{k-1}+1, \ldots, e_k,\ k = 1, \ldots, o.$

Then $\omega_1 = \omega_s = \lambda_{\phi(s)}(A) + \lambda_{\psi(s)}(B)$ for $s = 1, \ldots, n_1$.

Next we consider the matrix

$$G_2 = \wedge^{n_1+1} G(t) = \wedge^{n_1+1} e^{\frac{1}{2}At} e^{Bt} e^{\frac{1}{2}At} = e^{\frac{1}{2}A_{\wedge^{n_1+1}}t} e^{B_{\wedge^{n_1+1}}t} e^{\frac{1}{2}A\wedge^{n_1+1}t}.$$

So $\lambda_1(G_2(t)) = \prod_{i=1}^{n_1+1} \lambda_i(t)$ and more generally all the eigenvalues of $G_2(t)$ are for the form

$$\prod_{i=1}^{n_1+1} \lambda_{j_1}(t) \ldots \lambda_{j_{n_1+1}}(t),\ 1 \le j_1 < j_2 < \ldots < j_{n_1+1} \le n.$$

Since we already showed that $\lim_{t\to\infty} \frac{\log \lambda_i(t)}{t} = \omega_i$ for $i = 1, \ldots n$ we deduce that

$$\lim_{t\to\infty} \big(\prod_{i=1}^{n_1+1} \lambda_{j_1}(t) \ldots \lambda_{j_{n_1+1}}(t)\big)^{\frac{1}{t}} = e^{\sum_{i=1}^{n_1+1} \omega_{j_i}}.$$

Hence all the eigenvalues of $G_2(t)^{\frac{1}{t}}$ converge to the above values for all choices of $1 \le j_1 < j_2 < \ldots < j_{n_1+1} \le n$. The limit of the maximal eigenvalue of $G_2(t)^{\frac{1}{t}}$ is equal to $e^{\omega_1 + \ldots + \omega_{n_1} + \omega_i}$ for $i = n_1 + 1, \ldots, n_2$, which is of multiplicity $n_2 - n_1$. Let $P_{2,1}(t)$ be the projection on the eigenspace of $G_2(t)$ spanned by the first $n_2 - n_1$ eigenvalues of $G_2(t)$. Our results for $G(t)$ yield that $\lim_{t\to\infty} P_{2,1}(t) = P_{2,1}$, where $P_{2,1}$ is the projection on a direct sum of eigen-subspaces of $A_{\wedge^{n_1+1}}$. Let $\mathbf{W}_2(t) = P_1(t)\mathbb{C}^n + P_2(t)\mathbb{C}^n$ be a subspace of dimension n_2 spanned by the eigenvectors of $G(t)$ corresponding to $\lambda_1(t), \ldots, \lambda_{n_2}(t)$. Then $P_{2,1}(t) \wedge^{n_1+1} \mathbb{C}^n$ is the the subspace of the form $(\bigwedge^{n_1} P_1(t)\mathbb{C}^n) \bigwedge (P_2(t)\mathbb{C}^n)$. Since $\lim_{t\to\infty} P_1(t)\mathbb{C}^n = P_1\mathbb{C}^n$ and $\lim_{t\to\infty} P_{2,1}(t) \wedge^{n_1+1} \mathbb{C}^n = P_{2,1} \wedge^{n_1+1} \mathbb{C}^n$ we deduce that $\lim_{t\to\infty} P_2(t)\mathbb{C}^n = \mathbf{W}_2$ for some subspace of dimension $n_2 - n_1$ which is orthogonal to $P_1\mathbb{C}^n$. Let P_2 be the orthogonal projection on $\mathbf{W}_2$. Hence $\lim_{t\to\infty} P_2(t) = P_2$. (See for details Problem 12.)

We now show that that there exists two permutations ϕ, ψ on $\{1, \ldots, n\}$ satisfying (5.5.22) such that $\omega_i = \alpha_{\phi(i)} + \beta_{\psi(i)}$ for $i = n_1 + 1, \ldots, n_2$. Furthermore $AP_2 = P_2A$. To do that we need to apply carefully our results for $\omega_1, \ldots, \omega_{n_1}$. The logarithm of the first $n_2 - n_1$ limit eigenvalues of $G_2(t)^{\frac{1}{t}}$

must be of the form $\lambda_a(A_{\wedge^{n_1+1}})) + \lambda_b(B_{\wedge^{n_1+1}})$. The values of indices a and b can be identified as follows. Recall that the indices $\phi(i), i = 1, \ldots, n_1$ in $\omega_i = \lambda_{\phi(i)}(A) + \lambda_{\psi(i)}(B)$ can be determined from the projection P_1, where P_1 is viewed as the sum of the projections on the orthogonal eigen-subspaces $Q_iR_j\mathbb{C}^n, (i,j) \in \mathcal{K}_1$. Recall that $P_{2,1} \wedge^{n_1+1} \mathbb{C}^n$ is of the from $\wedge^{n_1}(P_1\mathbb{C}^n) \wedge (P_2\mathbb{C}^n)$. Since $P_2\mathbb{C}^n$ is orthogonal to $P_1\mathbb{C}^n$ and $\wedge^{n_1}(P_1\mathbb{C}^n) \wedge (P_2\mathbb{C}^n)$ is an invariant subspace of $A_{\wedge^{n_1+1}}$ it follows that $P_2\mathbb{C}^n$ is an invariant subspace of A. It is spanned by eigenvectors of A, which are orthogonal to $P_1\mathbb{C}^n$ spanned by the eigenvectors corresponding $\lambda_{\phi(i)}(A), i = 1, \ldots, n_1$. Hence the eigenvalues of the eigenvectors spanning $P_2\mathbb{C}^n$ are of the form $\lambda_k(A)$ for $k \in \mathcal{I}_2$, where $\mathcal{I}_2 \subset \{1, \ldots, n\} \backslash \{\phi(1), \ldots, \phi(n_1)\}$ is a set of cardinality $n_2 - n_1$. Hence $P_2Q_i = Q_iP_2, i = 1, \ldots, p$, which implies that $P_2A = AP_2$.

Note that $\lambda_a((A_{\wedge^{n_1+1}})) = \sum_{j=1}^{n_1} \lambda_{\phi(j)}(A) + \lambda_k(A)$ for $k \in \mathcal{I}_2$. Since $G_1(t) = e^{\frac{1}{2}At}e^{Bt}e^{\frac{1}{2}At}$ is similar to the matrix $H_1(t) := e^{\frac{1}{2}Bt}e^{At}e^{\frac{1}{2}At}$ we can apply the same arguments of $H_2(t) := \wedge^{n_1+1}H_1(t)$. We conclude that that there exists a set $\mathcal{J}_2 \subset \{1, \ldots, n\} \backslash \{\psi(1), \ldots, \psi(n_1)\}$ is a set of cardinality $n_2 - n_1$ such that $\lambda_b((B_{\wedge^{n_1+1}})) = \sum_{j=1}^{n_1} \lambda_{\psi(j)}(B) + \lambda_{k'}(B)$ for $k' \in \mathcal{J}_2$. Hence the logarithm of the limit value of the largest eigenvalue of $G_2(t)^{\frac{1}{t}}$ which is equal to $\omega_1 + \ldots + \omega_{n_1} + \omega_{n_1+1}$ is given by $n_2 - n_1$ the sum of the pairs $\lambda_a(A_{\wedge^{n_1+1}})) + \lambda_b(B_{\wedge^{n_1+1}})$. The pairing (a, b) induces the pairing (k, k') in $\mathcal{I}_2 \times \mathcal{J}_2$. Choose any permutation ϕ such that $\phi(1), \ldots, \phi(n_1)$ defined as above and $\{\phi(n_1 + 1), \ldots \phi(n_2)\} = \mathcal{I}_2$. We deduce the existence of a permutation ψ, where $\psi(1), \ldots, \psi(n_1)$ be defined as above, $\{\psi(n_1 + 1), \ldots, \psi(n_2)\} = \mathcal{J}_2$, and $(\phi(i), \psi(i))$ is the pairing (k, k') for $i = n_1 + 1, \ldots, n_2$. This shows that $\omega_i = \lambda_{\phi(i)}(A) + \lambda_{\psi(i)}(B)$ for $i = n_1 + 1, \ldots, n_2$. By considering the matrices $\wedge^{n_i+1}G(t)$ for $i = 2, \ldots, r$ we deduce the theorem. □

Problems

1. Prove Proposition 5.5.2. (**Hint**: Show that the left-hand side of (5.5.3) is one parameter group in t with the generator $(A_1, \ldots, A_k)_\otimes$.)

2. Let the assumptions of Proposition 5.5.2 hold. Assume that

   $\boldsymbol{\lambda}(A_i) = (\lambda_1(A_i), \ldots, \lambda_{n_i}(A_i))$ for $i = 1, \ldots, k$. Show that the $n_1 \ldots n_k$ eigenvalues of $(A_1, \ldots, A_k)_\otimes$ are of the form $\sum_{j=1}^k \lambda_{i_j}(A_i)$, where $j_i = 1, \ldots, n_i, i = 1, \ldots k$.

   (**Hint**: Recall that the eigenvalues of $\otimes_{i=1}^k e^{A_it}$ are of the form $\prod_{i=1}^k e^{\lambda_{j_i}(A_i)t}$.)

3. Let the assumptions of Definition 5.5.3 hold for $\mathbb{F} = \mathbb{C}$. Assume that $\boldsymbol{\lambda}(A) = (\lambda_1, \ldots, \lambda_n)$. Show that the $\binom{n}{k}$ eigenvalues of $A_{\wedge^k}$ are $\lambda_{i_1} + \ldots + \lambda_{i_k}$ for all $1 \le i_1 < \ldots < i_k \le n$.

4. Let $A, B \in \mathbb{C}^{n\times n}$.

   (a) Show that $e^A e^B$ is similar $e^{\frac{1}{2}A} e^B e^{\frac{1}{2}A}$.

   (b) Show that if $A, B \in \mathbf{H}_n$ then the eigenvalues all the eigenvalues of $e^A e^B$ are real and positive.

5. Let $g \in \mathrm{C[a, b]}$. Show that the following are equivalent

   (a) $g(\frac{1}{2}(x_1 + x_2)) \le \frac{1}{2}(g(x_1) + g(x_2))$ for all $x_1, x_2 \in [a, b]$.

   (b) $g(t_1x_1 + t_2x_2)) \le t_1g(x_1) + t_2g(x_2)$ for all $t_1, t_2 \in [0, 1], t_1 + t_2 = 1$ and $x_1, x_2 \in [a, b]$.

   Hint: Fix $x_1, x_2 \in [a, b]$. First show that (a)⇒(b) for any $t_1, t_2 \in [0, 1]$ which have finite binary expansions. Use the continuity to deduce that (a)⇒(b).

6. (a) Let $\mathbf{D}_n(\mathbb{R}) \subset \mathbf{H}_n$ be the subspace of diagonal matrices. Show that $\mathbf{D}_n(\mathbb{R})$ is a maximal commuting subspace.

   (b) Let $\mathbf{U} \subset \mathbf{H}_n$ be a commuting subspace. Show that there exists a unitary matrix $U \in \mathbf{U}(n)$ such that $\mathbf{U}$ is a subspace of of a maximal commuting subspace $U\mathbf{D}_n(\mathbb{R})U^*$.

   (c) Show that a commuting subspace $\mathbf{U} \subset \mathbf{H}_n$ is maximal if and only if $\mathbf{U}$ contains A with n distinct eigenvalues.

   (d) Let $\mathbf{U} \subset \mathbf{H}_n$ be a commuting subspace. Show that for each $k \in [1, n] \cap \mathbb{N}$ the subspace $\mathbf{U}_k := \mathrm{span}\,(\mathrm{A}_{\wedge^k} :\ \mathrm{A} \in \mathbf{U})$ is a commuting subspace of $\mathbf{H}_{\binom{n}{k}}$.

7. Let $A, B \in \mathbf{H}_n$ and assume that $f_n(A, B)$ is defined as in (5.5.4). Show that $f_n(A, B) = \mathrm{tr}(A + B)$. Hence $f_n(A, B)$ is convex on $\mathbf{H}_n \times \mathbf{H}_n$.

   Remark. We suspect that $f_k : \mathbf{H}_n \times \mathbf{H}_n \to \mathbb{R}$ defined as (5.5.4) is not a convex function for $k = 1, \ldots, n - 1$.

8. Let $g : [0, \infty) \to \mathbb{R}$ be a continuous convex function. Show that if $g(0) = 0$ the the function $\frac{g(t)}{t}$ nondecreasing on $(0, \infty)$. (**Hint**: Observe that $g(x) \le \frac{x}{y}g(y) + (1 - \frac{x}{y})g(0)$ for any $0 < x < y$.)

9. Let $A, B \in \mathbb{C}^{n\times n}$.

(a) Show $\lim_{t\to\infty} \frac{1}{t}(e^{At}e^{Bt} - I) = A + B$.

(b) View $(e^{At}e^{Bt})^{\frac{1}{t}}$ as $(I + (e^{At}e^{Bt} - I))^{\frac{1}{t}}$. Show $\lim_{t\to 0}(e^{At}e^{Bt})^{\frac{1}{t}} = e^{A+B}$.

(c) Show that if $A, B \in \mathbf{H}_n$ then $\lim_{t\searrow 0} \frac{1}{t} \log \lambda_i(e^{At}e^{Bt}) = \lambda_i(A+B)$ for $i = 1, \ldots, n$.

10. Let $A, B \in \mathbf{H}_n$.

(a) Assume that there exists a vector of length $\mathbf{x}_1$ such that $A\mathbf{x}_1 = \lambda_i(A)\mathbf{x}_1, B\mathbf{x}_1 = \lambda_j(B)\mathbf{x}_1$. Then

$$(A + B)\mathbf{x}_1 = (\lambda_i(A) + \lambda_j(B))\mathbf{x}_1, \quad e^A e^B \mathbf{x}_1 = e^{\lambda_i(A)+\lambda_j(B)}\mathbf{x}_1.$$

The convexity of $\lambda_1(\cdot)$ on $\mathbf{H}_n$ implies that $\lambda_1(A+B) = \lambda_1(A) + \lambda_1(B)$. The inequalities

$$e^{\lambda_1(A)+\lambda_1(B)} \le \lambda_1(e^A e^B) \le \sigma_1(e^A e^B) \le \sigma_1(e^A)\sigma_1(e^B) = e^{\lambda_1(A)+\lambda_1(B)}$$

imply that we have equalities in the above inequalities. This show the equality holds for $k = 1$ in (5.5.7) if A and B have a common eigenvector corresponding to the first eigenvalue of A and B.

11. Let $C(t)$ be defined by (5.5.10).

(a) Show that $C(-t) = C(t)$ for any $t \neq 0$.

(b) Show the equality (5.5.11).

12. Let $\mathbf{V}$ be an n-dimensional inner product space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, with the inner product $\langle \cdot, \cdot \rangle$. Let

$$\text{(5.5.23)} \qquad \mathrm{S}(\mathbf{U}) := \{\mathbf{u} \in \mathbf{U},\ \langle \mathbf{u}, \mathbf{u} \rangle = 1\}, \quad \mathbf{U} \text{ is a subspace of } \mathbf{V}$$

be the unit sphere in $\mathbf{U}$. ($\mathrm{S}(\{\mathbf{0}\}) = \emptyset$.) For two subspaces of $\mathbf{U}, \mathbf{W} \subseteq \mathbf{V}$ the distance $\mathrm{dist}(\mathbf{U}, \mathbf{V})$ is defined to be the Hausdorff distance between the unit spheres in $\mathbf{U}, \mathbf{W}$:

$$\mathrm{dist}(\mathbf{U}, \mathbf{V}) := \max\big(\max_{\mathbf{u}\in \mathrm{S}(\mathbf{U})} \min_{\mathbf{v}\in \mathrm{S}(\mathbf{V})} ||\mathbf{u} - \mathbf{v}||, \max_{\mathbf{v}\in \mathrm{S}(\mathbf{V})} \min_{\mathbf{u}\in \mathrm{S}(\mathbf{U})} ||\mathbf{v} - \mathbf{u}||\big),$$

(5.5.24)

$$\mathrm{dist}(\{\mathbf{0}\}, \{\mathbf{0}\}) = 0, \mathrm{dist}(\{\mathbf{0}\}, \mathbf{W}) = \mathrm{dist}(\mathbf{W}, \{\mathbf{0}\}) = 1 \text{ if } \dim \mathbf{W} \ge 1.$$

(a) Let $\dim \mathbf{U}, \dim \mathbf{V} \ge 1$. Show that $\mathrm{dist}(\mathbf{U}, \mathbf{V}) \le 2$. Equality holds if either $\mathbf{U} \cap (\mathbf{V})^{\perp}$ or $\mathbf{U}^{\perp} \cap \mathbf{V}$ are nontrivial subspaces. In particular $\mathrm{dist}(\mathbf{U}, \mathbf{V}) = 2$ if $\dim \mathbf{U} \neq \dim \mathbf{V}$.

(b) Show that dist is a metric on $\cup_{m=0}^{n} \mathrm{Gr}_m(\mathbf{V})$.

(c) Show that $\mathrm{Gr}_m(\mathbf{V})$ is a compact connected space with respect to the metric $\mathrm{dist}(\cdot,\cdot)$ for $m = 0, 1 \ldots, n$. (I.e. for each sequence of m-dimensional subspaces $\mathbf{U}_i, i \in \mathbb{N}$ one can choose a subsequence $i_j, j \in \mathbb{N}$ such that $\mathbf{U}_{i_j}, j \in \mathbb{N}$ converges in the metric dist to $\mathbf{U} \in \mathrm{Gr}_m(\mathbf{V})$. **Hint**: Choose an orthonormal basis in each $\mathbf{U}_i$.)

(d) Show that $\cup_{m=0}^{n} \mathrm{Gr}_m(\mathbf{V})$ is a compact space in the metric dist.

(e) Let $\mathbf{U}, \mathbf{U}_i \in \mathrm{Gr}_m(\mathbf{V}), i \in \mathbb{N}, 1 \le m < n$. Let $P, P_i \in \mathbf{S}(\mathbf{V})$ be the orthogonal projection on $\mathbf{U}, \mathbf{U}_i$ respectively. Show that $\lim_{i\to\infty} \mathrm{dist}(\mathbf{U}_i, \mathbf{U}) = 0$ if and only if $\lim_{i\to\infty} P_i = P$.

(f) Let $\mathbf{U}_i \in \mathrm{Gr}_m(\mathbf{V}), \mathbf{W}_i \in \mathrm{Gr}_l(\mathbf{V}), 1 \le m, n$ and $\mathbf{U}_i \perp \mathbf{W}_i$ for $i \in \mathbb{N}$. Assume that $\lim_{i\to\infty} \mathrm{dist}(\mathbf{U}_i, \mathbf{U}) = 0$ and $\mathrm{dist}((\bigwedge^{\mathrm{m}} \mathbf{U}_i) \bigwedge \mathbf{W}_i, \mathbf{X}) = 0$ for some subspaces $\mathbf{U} \in \mathrm{Gr}_m(\mathbf{V}), \mathbf{X} \in \mathrm{Gr}_l(\bigwedge^{m+1} \mathbf{V})$. Show that there exists $\mathbf{W} \in \mathrm{Gr}(\mathbf{V})$ orthogonal to $\mathbf{U}$ such that $\lim_{i\to\infty} \mathrm{dist}(\mathbf{W}_i, \mathbf{W}) = 0$.

13. Let $\mathbf{V}$ be an n-dimensional inner product space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, with the inner product $\langle\cdot,\cdot\rangle$. Let $m \in [1, n-1] \cap \mathbb{N}$ and assume that $\mathbf{U}, \mathbf{W} \in \mathrm{Gr}_m(\mathbf{V})$. Choose orthonormal bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}, \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ in $\mathbf{U}, \mathbf{W}$ respectively. Show

(a) $\det(\langle \mathbf{u}_i, \mathbf{w}_j\rangle)_{i,j=1}^m = \overline{\det(\langle \mathbf{w}_j, \mathbf{u}_i\rangle)_{i,j=1}^m}$.

(b) Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ another orthonormal basis in $\mathbf{U}$, i.e. $x_i = \sum_{k=1}^m q_{ki}\mathbf{u}_k, i = 1, \ldots, m$ where $Q = (q_{ki}) \in \mathbb{F}^{m\times m}$ is orthogonal for $\mathbb{F} = \mathbb{R}$ and unitary for $\mathbb{F} = \mathbb{C}$. Then $\det(\langle \mathbf{x}_i, \mathbf{w}_j\rangle)_{i,j=1}^m = \det Q \det(\langle \mathbf{u}_k, \mathbf{w}_j\rangle)_{k,j=1}^m$.

(c) $[\mathbf{U}, \mathbf{W}] := |\det(\langle \mathbf{u}_i, \mathbf{w}_j\rangle)_{i,j=1}^m|$ is independent of the choices of orthonormal bases in $\mathbf{U}, \mathbf{W}$. Furthermore $[\mathbf{U}, \mathbf{W}] = [\mathbf{W}, \mathbf{U}]$.

(d) Fix an orthonormal basis in $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ in $\mathbf{W}$. Then there exists an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ in $\mathbf{U}$ such that the matrix $(\langle \mathbf{u}_i, \mathbf{w}_j\rangle)_{i,j=1}^m$ is upper triangular. **Hint**: Let $\mathbf{W}_i = \mathrm{span}(\mathbf{w}_{i+1}, \ldots, \mathbf{w}_n)$ for $i = 1, \ldots, m-1$. Consider $\mathrm{span}(\mathbf{w}_1)^\perp \cap \mathbf{U}$ which has dimension $m-1$ at least. Let $\mathbf{U}_1$ be an $m-1$ dimensional subspace of $\mathrm{span}(\mathbf{w}_1)^\perp \cap \mathbf{U}$. Let $\mathbf{u}_1 \in \mathrm{S}(\mathbf{U}) \cap \mathbf{U}_1^\perp$. Use $\mathbf{U}_1, \mathbf{V}_1$ to define an $m-2$ dimensional subspace $\mathbf{U}_2 \subset \mathbf{U}_1$ and $\mathbf{u}_2 \in \mathrm{S}(\mathbf{U}_1) \cap \mathbf{U}_2^\perp$ as above. Continue in this manner to find an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$.

(e) $[\mathbf{U}, \mathbf{W}] \le 1$. ($[\mathbf{U}, \mathbf{W}]$ *is called the cosine of the angle between* $\mathbf{U}, \mathbf{W}$.)

(f) $[\mathbf{U}, \mathbf{V}] = 0 \iff \mathbf{U}^\perp \cap \mathbf{V} \ne \{\mathbf{0}\} \iff \mathbf{U} \cap \mathbf{V}^\perp \ne \{\mathbf{0}\}$. **Hint**: Use (d) and induction.

14. Let $\mathbf{V}$ be an n-dimensional inner product space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, with the inner product $\langle \cdot, \cdot \rangle$. Let $l, m \in [1, n] \cap \mathbb{N}$ and assume that $\mathbf{U} \in \mathrm{Gr}_l(\mathbf{V}), \mathbf{W} \in \mathrm{Gr}_m(\mathbf{V})$. Let $P, Q \in \mathrm{S}(\mathbf{V})$ be the orthogonal projections on $\mathbf{U}, \mathbf{W}$ respectively. Show

    (a) $\mathbf{U} + \mathbf{W} = \mathbf{U} \cap \mathbf{W} \oplus \mathbf{U} \cap (\mathbf{U} \cap \mathbf{W})^\perp \oplus \mathbf{W} \cap (\mathbf{U} \cap \mathbf{W})^\perp$.

    (b) $\mathrm{rank\ PQ = rank\ QP = rank\ PQP = rank\ QPQ.}$

    (c) $\mathrm{rank\ PQ} = \dim \mathbf{W} - \dim \mathbf{W} \cap \mathbf{U}^\perp, \mathrm{rank\ QP} = \dim \mathbf{U} - \dim \mathbf{U} \cap \mathbf{W}^\perp$.

15. Let $\mathbf{V}$ be an n-dimensional inner product space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$, with the inner product $\langle \cdot, \cdot \rangle$. Assume that $\mathbf{V} = \oplus_{i=1}^{l} \mathbf{U}_i = \oplus_{j=1}^{m} \mathbf{W}_i$ be two decompositions of $\mathbf{V}$ to nontirivial orthogonal subspaces:

$$\dim \mathbf{U}_i = l_i - l_{i-1},\ i = 1, \ldots, p, \quad \dim \mathbf{W}_j = m_j - m_{j-1},\ j = 1, \ldots, q,$$
$$0 = l_0 < l_1 < \ldots < l_p = n, \quad 0 = m_0 < m_1 < \ldots < m_q = n.$$

Let $Q_i, R_j \in \mathrm{S}(\mathbf{V})$ be the orthogonal projections on $\mathbf{U}_i, \mathbf{W}_j$ respectively for $i = 1, \ldots, p, j = 1, \ldots, q$. Let $n_{ij} := \mathrm{rank\ Q_iR_j, i = 1, \ldots, p, j = 1, \ldots, q}$.

Denote $\mathcal{K} := \{(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, q\} : \ Q_i R_j \neq 0\}$. For $i \in \{1, \ldots, p\}, j \in \{1, \ldots, q\}$ let

$$\mathcal{J}_i := \{j' \in \{1, \ldots, q\},\ (i, j') \in \mathcal{K}\}, \quad \mathcal{I}_j := \{i' \in \{1, \ldots, p\}, (i', j) \in \mathcal{K}\}.$$

Show

(a) $Q_i = \sum_{j \in \mathcal{J}_i} Q_i R_j,\ i = 1, \ldots, p$.

(b) Let $(i_1, j_1), \ldots, (i_s, j_s) \in \mathcal{K}$ and assume that $i_a \neq i_b, j_a \neq j_b$. Then

$$\mathrm{rank} \sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{Q_{i_a} R_{j_a}} = \mathrm{rank}\ (\sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{Q_{i_a} R_{j_a}})(\sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{Q_{i_a} R_{j_a}})^* =$$
$$\mathrm{rank} \sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{Q_{i_a} R_{j_a} Q_{i_a}} = \sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{rank\ Q_{i_a} R_{j_a} Q_{i_a}} = \sum_{\mathrm{a=1}}^{\mathrm{s}} \mathrm{rank\ Q_{i_a} R_{j_a}}.$$

(c) $\mathrm{rank\ P_i} \leq \sum_{\mathrm{j} \in \mathcal{J}_\mathrm{i}} \mathrm{n_{ij}}$, where strict inequality may hold

(d) $\mathbf{U}_i = \sum_{j \in \mathcal{J}_i} \mathbf{U}_{ij}$, where $\mathbf{U}_{ij} := P_i \mathbf{W}_j, \dim \mathbf{U}_{ij} = n_{ij}$ for $i = 1, \ldots, p,\ j = 1, \ldots, q$.

(e) $Q_j = \sum_{i \in \mathcal{I}_j} P_i Q_j,\ j = 1, \ldots, q$.

(f) $\mathrm{rank\ Q_j} \leq \sum_{\mathrm{i} \in \mathcal{I}_\mathrm{j}} \mathrm{n_{ij}}$, where strict inequality may hold.

(g) $\mathbf{W}_j = \sum_{i \in \mathcal{I}_j} \mathbf{W}_{ji}$, where $\mathbf{W}_{ji} = Q_j \mathbf{U}_i, \dim \mathbf{W}_{ji} = n_{ij}$ for $j = 1, \ldots, q,\ i = 1, \ldots, p$.

Chapter 6

Nonnegative matrices

6.1 Graphs

6.1.1 Undirected graphs

An *undirected graph* is denoted by $G = (V, E)$. It consists of *vertices* $v \in V$, and edges which are *unordered* set of pairs (u, v), where $u, v \in V$, and $u \neq v$, which are called *edges* of G. The set of edges in G is denoted by E. Let $n = \#V$ be the cardinality of V, i.e. V has n vertices. Then it is useful to identify V with $\langle n \rangle = \{1, \ldots, n\}$. For example, the graph $G = (\langle 4 \rangle, \{(1,2), (1,4), (2,3), (2,4), (3,4)\})$ has 4 vertices and 5 edges.

In what follows we assume that $G = (V, E)$ unless stated otherwise. A graph $H = (W, F)$ is called a subgraph of $G = (V, E)$ if W is a subset of V and any edge in F is an edge in E. Given a subset W of V then $E(W) = \{(u, v) \in E, u, v \in W\}$ is the set of edges in G *induced* by W. The graph $G(W) := (W, E(W))$ is call the subgraph *induced* by W. Given a subset F of E, then $V(F)$ is the set of vertices which consist of all vertices participating in the edges in F. The graph $G(F) = (V(F), F)$ is called the subgraph *induced* by F.

The *degree* of v, denoted by deg v is the number of edges that has v as its vertex. Since each edge has two different vertices

$$\sum_{v \in V} \deg v = 2\#E, \tag{6.1.1}$$

where $\#E$ is the number of edges in E. $v \in V$ is called an *isolated* vertex if $\deg v = 0$. Note that $V(E)$ is the set of nonisolated vertices in G, and $G(E) = (V(E), E)$ the subgraph of G obtained by deleting isolated vertices in G.

The complete graph on n vertices is the graph with all possible edges. It is denoted by $K_n = (\langle n\rangle, \mathcal{E}_n)$, where $\mathcal{E}_n = \{(1,2),\ldots,(1,n),(2,3),\ldots,(n-1,n)\}$. For example, K_3 is called a *triangle*. Note that for any graph on n vertices $G = (\langle n\rangle, E)$ is a a subgraph of K_n, obtained by erasing some of edges in K_n, but not the vertices! I.e. $E \subset \mathcal{E}_n$.

$G = (V,E)$ is called *biparite* if V is a union of two disjoint sets of vertices $V_1 \cup V_2$ so that each edge in E connects some vertex in V_1 to some vertex in E_2. Thus $E \subset V_1 \times V_2 := \{(v,w), v \in V_1, w \in V_2\}$. So any bipartite graph $D = (V_1 \cup V_2, E)$ is a subgraph of the complete bipartite graph $K_{V_1,V_2} := (V_1 \cup V_2, V_1 \times V_2)$. For positive integers l,m the complete bipartite graph on l,m vertices is denoted by $K_{l,m} := (\langle l\rangle \cup \langle m\rangle, \langle l\rangle \times \langle m\rangle)$. Note that $K_{l,m}$ has $l+m$ vertices and lm edges.

6.1.2 Directed graphs

A *directed graph* is denoted by $D = (V,E)$. V is the set of vertices and E is the set of directed edges in G. So E is a subset of $V \times V = \{(v,w), v,w \in V$. Thus $(v,w) \in E$ is a directed edge from v to W. For example, the graph $D = (\langle 4\rangle, \{(1,2),(2,1),(2,3),(2,4),(3,3),(3,4),(4,1)\})$ has 4 vertices and 7 (directed) edges.

The directed edge $(v,v) \in E$ is called a *loop*, or selfloop.

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

the number of edges to v and out of v in D. $\deg_{\,in}, \deg_{\,out}$ are called the *in* or *out* degrees. Clearly we have the analog of (6.1.1)

$$\sum_{v\in V} \deg_{\,in} v = \sum_{v \in V} \deg_{\,out} v = \#E, \tag{6.1.2}$$

A subgraph $H = (W,F)$ of $D = (V,E)$ is defined, and the induced subgraphs $D(W) = (W, E(W)), D(F) = (V(F), F)$ are defined as in §6.1.1. $v \in V$ is called isolated if $\deg_{\,in}(v) = \deg_{\,out}(v) = 0$.

6.1.3 Multi graphs

A *multigraph* is graph where multiple edges, in particular and multiple loops are allowed. So undirected multigraph $G = (V,E)$ has undirected edges, which may be multiple, and may have multiple loops. A directed multigraph $D = (V,E)$ may have multiple edges.

Each directed multigraph $D = (V,E)$ induces an undirected multigraph $G(D) = (V,E')$, where each directed edge $(u,v) \in E$ is viewed as undirected edge $(u,v) \in E'$. (Each loop $(u,u) \in E$ will appear twice in E'.) Vice versa, an undirected multigraph $G = (V, E')$ induces a directed multigraph $D(G) = (V, E)$, where each undirected edge (u, v) is (u, v) and (v, u), when $u \neq v$. The loop (u, u) appears p times in $D(G)$ if it appears p times in G.

Most of the following notions are the same for directed or undirected graphs or multigraphs, unless stated otherwise. We state them for directed multigraphs $D = (V, E)$.

Definition 6.1.1

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

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

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

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

5. *D is called a* diforest *if D does not have cycles. (An undirected multigraph with no cycles is called* forest*.)*

6. *Let $D = (V, E)$ be a diforest. Then the height of $v \in V$, denoted by* height(v) *is the length of the longest path ending at v.*

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

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

9. *A multigraph $G = (V, E)$ is called connected if either $n = 1$, or $n > 1$ and any two vertices in G are connected. (*Note *that a simple graph on one vertex $G = (\langle 1 \rangle, \emptyset)$ is considered connected. The induced directed graph $D(G) = G$ is not strongly connected.)*

10. *Assume that a multidigraph $D = (V, E)$ is strongly connected. Then D is called* primitive *if there exists $k \geq 1$ such that for any two vertices $u, v \in V$ there exists a walk of length k which connects u and v. For a primitive multidigraph D, the minimal such k is called the*

index of primitivity*, and denoted by* indprim(D)*. A strongly connected multidigraph which is not primitive is called* imprimitive*.*

11. *For $W \subset V$, the multidirected subgraph $D(W) = (W, E(W)$ is called a* strongly connected component *of D if $D(W)$ is strongly connected, and for any $W \subsetneqq U \subset V$ the induced subgraph $D(U) = (U, E(U))$ is not strongly connected.*

12. *For $W \subset V$, the undirected subgraph $G(W) = (W, E(W)$, of undirected multigraph $G = (V, E)$, is called a* connected component *of G if $G(W)$ is connected, and for any $W \subsetneqq U \subset V$ the induced subgraph $G(U) = (U, E(U))$ is not connected.*

13. *An undirected forest $G = (V, E)$ is called a tree if it is connected.*

14. *A diforest $D = (V, E)$ is called a ditree if the induced undirected multigraph $G(D)$ is a tree.*

15. *Let $D = (V, E)$ be a multidigraph. The* reduced *(simple) digraph $D_r = (V_r, E_r)$ is defined as follows. Let $D(V_i), i = 1, \ldots, k$ be all strongly connected components of D. Let $V_0 = V \backslash (\cup_{i=1}^k V_i$ be all vertices in D which do not belong to any of strongly connected components of D. (It is possible that either V_0 is an empty set or $k = 0$, i.e D does not have connected components, and the two conditions are mutually exclusive.) Then $V_r = (\cup_{v \in V_0}\{v\}) \cup_{i=1}^k \{V_i\}$, i.e. V_r is the set of all vertices in V which do not belong to any connected component and the new k vertices named $\{V_1\}, \ldots, \{V_k\}$. A vertex $u' \in V_r$ is viewed as either a set consisting of one vertex $v \in V_0$ or the set V_i for some $i = 1, \ldots, k$. Then E_r does not contain loops. Furthermore $(s, t) \in E_r$, if there exists an edge from $(u, v) \in E$, where u and v are in the set of vertices represented by s and t in V, respectively.*

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

Proposition 6.1.2 *Let $G = (V, E)$ be a multigraph. Then G is a disjoint union of its connected components. That is, there is a unique decomposition of V to $\cup_{i=1}^k V_i$, up to relabeling of $V_1, \ldots, V_k$, such that the following conditions hold:*

1. *$V_1, \ldots, V_k$ are nonempty and mutually disjoint.*

2. *Each $G(V_i) = (V_i, E(V_i))$ is a connected component of G.*

3. *$E = \cup_{i=1}^{k} V_i$.*

Proof. We introduce the following relation $\sim$ on V. First, we assume that $v \sim v$ for each $v \in V$. Second, for $v, w \in V, v \neq w$ we say that $v \sim w$ if v is connected to w. It is straightforward to show that $\sim$ is an *equivalence* relation. Let $V_1, \ldots, V_k$ be the equivalence classes in V. That is $v, w \in V_i$ if and only if v and w are connected. The rest of the proposition follows straightforward. $\square$

Proposition 6.1.3 *Let $D = (E, V)$ be a multidigraph. Then the reduced digraph D_r is a diforest.*

See Problem 6.1.5.4 for proof.

Proposition 6.1.4 *Let $D = (V, E)$ be a multidigraph. Then D is diforest if and only if it is isomorphic to a digraph $D_1 = (\langle n \rangle, E_1)$ such that if $(i, j) \in E_1$ then $i < j$.*

Proof. Clearly, the graph in D_1 can not have a cycle. So if D is isomorphic to D_1 then D is a diforest. Assume now that $D = (V, E)$ is a diforest. Let V_i be all vertices in V having height i for $i = 0, \ldots, k \geq 0$, where k is the maximal height of all vertices in D. Observe that from the definition of height it follows that if $(u, v) \in D$, where $u \in V_i, w \in V_j$ then $i < j$. Rename the vertices of V such that $V_i = \{n_i + 1, \ldots, n_{i+1}\}$ where $0 = n_0 < n_1 < \ldots < n_{k+1} = n := \#V$. Then one obtains the isomorphic graph $D_1 = (\langle n \rangle, E_1$, such that if $(i, j) \in E_1$ then $i < j$. $\square$

Theorem 6.1.5 *Let $D = (V, E)$ be as strongly connected multidigraph. Let ℓ be the g.c.d, (the greatest common divisor), of lengths of all cycles in D. Then exactly one of the following conditions hold.*

1. *$\ell = 1$. Then D is primitive. Let s be the length of the shortest cycle in D. Then* $\mathrm{indprim(D)} \leq \mathrm{\#V + s(\#V - 2)}$.

2. *$\ell > 1$. Then D is imprimitive. Furthermore, it is possible to divide V to ℓ disjoint nonempty subsets $V_1, \ldots, V_\ell$ such $E \subset \cup_{i=1}^{\ell} V_i \times V_{i+1}$, where $V_{\ell+1} := V_1$.*

   Define $D_i = (V_i, E_i)$ to be the following digraph. $(v, w) \in E_i$ if there is a path or cycle of length ℓ from v to w in D, for $i = 1, \ldots, l$. Then each D_i is strongly connected and irreducible.

The proof of this theorem will be given later using the Perron-Frobenius theorem. (Obviously, one can give a pure graph theoretical proof of this theorem.) If D is a strongly connected imprimitive multidigraph, then $\ell > 1$ given in (2) is called the *index of imprimitivity* of D.

6.1.4 Matrices and graphs

For a set $\mathcal{S}$ denote by $\mathcal{S}^{m\times n}$ the set of all $m \times n$ matrices $A = [a_{ij}]_{i=j=1}^{m,n}$ where each entry a_{ij} is in $\mathcal{S}$. Then $A^\top \in \mathcal{S}^{n\times m}$ the transposed matrix of A. Denote by $\mathbb{F}$ any field, and by $\mathbb{R}, \mathbb{C}$ the field of real and complex numbers respectively. By $\mathrm{S_n}(\mathcal{S}) \subset \mathcal{S}^{\mathrm{n\times n}}$ denote the set of all symmetric matrices $A = [a_{ij}], a_{ij} = a_{ji}$ with entries in $\mathcal{S}$. Assume that $0 \in \mathcal{S}$. Then by $\mathrm{S_{n,0}}(\mathcal{S}) \subset \mathrm{S_n}(\mathcal{S})$ the subset of all symmetric matrices with entries in $\mathcal{S}$ and zero diagonal. Denote by $\mathcal{P}_n \subset \{0,1\}^{n\times n}$ the group of permutation matrices. I.e. each $P \in \mathcal{P}_n$ has one 1 in each row and column, and all other $n^2 - n$ entries are zero. Denote by $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ the vector of length n whose all coordinates are 1. For $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ we denote by $\operatorname{tr} A := \sum_{i=1}^n a_{ii}$ the *trace* of A. For any $t \in \mathbb{R}$, we let $\mathrm{sign\,t} = 0$ if $t = 0$ and $\mathrm{sign\,t} = \frac{\mathrm{t}}{|\mathrm{t}|}$ if $t \neq 0$. For $A, B \in \mathbb{R}^{m\times n}$ we denote $B - A \geq 0, B - A \gneq 0, B - A > 0$ if $B - A$ is a nonnegative matrix, a nonnegative nonzero matrix, and a positive matrix respectively.

Let $D = (V, E)$ be a digraph. Assume that $\#V = n$ and label the vertices of V as $1, \ldots, n$. So we have a bijection $\phi_1 : V \to \langle n\rangle$. This bijection induces an isomorphic graph $D_1 = (\langle n\rangle, E_1)$. With D_1 we associate the following matrix $A(D_1) = [a_{ij}]_{i,j=1}^n \in \mathbb{Z}_+^{n\times n}$. So a_{ij} is the number of directed edges from the vertex i t the vertex j. (If $a_{ij} = 0$ then there no diedges from i to j.) When no confusion arises we let $A(D) := A(D_1)$, and we call $A(D)$ a *representation matrix* of D. Note that a different bijection $\phi_2 : V \to \langle n\rangle$ gives rise to a different $A(D_2)$, where $A(D_2) = P^\top A(D_1)P$ for some permutation matrix $P \in \mathcal{P}_n$. See Problem 7.

If D is a simple digraph then $A(D) \in \{0,1\}^{n\times n}$. If G is a multigraph, then $A(G) = A(D)$ where D is is the induced digraph by G. Hence $A(G) \in \mathrm{S_n}(\mathbb{Z}_+)$. If G is a graph then $A(G) \in \mathrm{S_{n,0}}(\{0,1\})$.

Proposition 6.1.6 *Let $D = (V, E)$ be a multidigraph on n vertices. Let $A(D)$ be a representation matrix of D. For an integer $k \geq 1$ let $A(D)^k = [a_{ij}^{(k)}] \in \mathbb{Z}_+^{n\times n}$. Then $a_{ij}^{(k)}$ is the number of walks of length k from the vertex i to the vertex j. In particular, $\mathbf{1}^\top A\mathbf{1}$ and $\operatorname{tr} A$ are the total number of walks and the total number of closed walks of length k in D.*

Proof. For $k = 1$ the proposition is obvious. Assume that $k > 1$.

Recall that

$$a_{ij}^{(k)} = \sum_{i_1,\ldots,i_{k-1}\in\langle n\rangle} a_{ii_1}a_{i_1i_2}\ldots a_{i_{k-1}j}. \tag{6.1.3}$$

The summand $a_{ii_1}a_{i_1i_2}\ldots a_{i_{k-1}j}$ gives the number of walks of the form $i_0i_1i_2\ldots i_{k-1}i_k$, where $i_0 = i, i_k = j$. Indeed if one of the terms in this product is zero, i.e. the is no diedge (i_p, i_{p+1}) then the product is zero. Otherwise each positive integer $a_{i_pi_{p+1}}$ counts the number of diedges (i_p, i_{p+1}). Hence $a_{ii_1}a_{i_1i_2}\ldots a_{i_{k-1}j}$ is the number of walks of the form $i_0i_1i_2\ldots i_{k-1}i_k$. The total number of walks from $i = i_0$ to $j = i_k$ of length k is the sum given by (6.1.3). To find out the total number of walks in D of length k is $\sum_{i=j=1}^n a_{ij}^{(k)} = \mathbf{1}^\top A\mathbf{1}$. The total number of closed walks in D of length k is $\sum_{i=1}^k a_{ii}^{(k)} = \operatorname{tr} A(D)^k$. □

With a multibipartite graph $G = (V_1 \cup V_2, E)$, where $\#V_1 = m, \#V_2 = n$, we associate a representation matrix $B(G) = [b_{ij}]_{i=j=1}^{m,n}$ as follows. Let $\psi_1 : V_1 \to \langle m\rangle, \phi_1 : V_2 \to \langle m\rangle$ be bijections. This bijection induces an isomorphic graph $D_1 = (\langle m\rangle \cup \langle n\rangle, E_1)$. Then b_{ij} is the number of edges connecting $i \in \langle m\rangle$ to $j \in \langle n\rangle$ in D_1.

A nonnegative matrix $A = [a_{ij}]_{i=j=1}^n \in \mathbb{R}_+^{n\times n}$ induces the following digraph $D(A) = (\langle n\rangle, E)$. The diedge (i, j) is in E if and only if $a_{ij} > 0$. Note that of $A(D(A)) = [\mathrm{sign\ a_{ij}}] \in \{0, 1\}^{\mathrm{n\times n}}$. We have the following definitions.

Definition 6.1.7

1. *$A = [a_{ij}] \in \mathbb{R}^{n\times n}$ is* combinatorially symmetric *if* $\mathrm{sign\ a_{ij}} = \mathrm{sign\ a_{ji}}$ *for $i, j = 1, \ldots, n$.*

2. *$A \in \mathbb{R}_+^{n\times n}$ is irreducible, if $D(A)$ is strongly connected.*

3. *$A \in \mathbb{R}_+^{n\times n}$ is primitive if A^k is a positive matrix for some integer $k \geq 1$.*

4. *Assume that $A \in \mathbb{R}_+^{n\times n}$ is primitive. Then the smallest positive integer k such that $A^k > 0$ is called the index of primitivity of A, and is denoted by* $\mathrm{indprim(A)}$.

5. *$A \in \mathbb{R}_+^{n\times n}$ is imprimitive if A is irreducible but not primitive.*

Proposition 6.1.8 *Let $D = (\langle n\rangle, E)$ be a multidigraph. Then D is strongly connected if and only if $(I + A(D))^{n-1} > 0$. in particular, a nonnegative matrix $A \in \mathbb{R}_+^{n\times n}$ is irreducible if and only if $(I + A)^{n-1} > 0$.*

Proof. Apply the Newton binomial theorem for $(1+t)^{n-1}$ to the matrix $(I + A(D))^{n-1}$

$$(I + A(D))^{n-1} = \sum_{p=0}^{n-1} \binom{n-1}{p} A(D)^p.$$

Recall that all the binomial coefficients $\binom{n-1}{p}$ are positive for $p = 0, \ldots, n-1$. Assume first that $(I + A(D))^{n-1} > 0$. That is for any $i, j \in \langle n \rangle$ the (i, j) entry of $(I + A(D))^{n-1}$ is positive. Hence the (i, j) entry of $A(D)^p$ is positive for some $p = p(i, j)$. Let $i \neq j$. Since $A(D)^0 = I$, we deduce that $p(i, j) > 0$. Use Proposition 6.1.6 to deduce that there is a walk of length p from the vertex i to the vertex j.

Suppose that D is strongly connected. Then for each $i \neq j$ we must have a path of length $p \in [1, n-1]$ which connects i and j, see Problem 1. Hence all off-diagonal entries of $(I + A(D))^{n-1}$ are positive. Clearly, $(I + A(D))^{n-1} \geq I$. Hence $(I + A(D))^{n-1} > 0$.

Let $A \in \mathbb{R}_+^{n \times n}$. Then the (i, j) entry of $(I + A)^{n-1}$ is positive if and only if the (i, j) entry of $(I + A(D(A)))^{n-1}$ is positive. Hence A is irreducible if and only if $(I + A)^{n-1} > 0$. □

6.1.5 Problems

1. Assume $v_1 \ldots v_p$ is a walk in $D = (V, E)$. Show that it is possible to subdivide this walk to walks $v_{n_{i-1}+1} \ldots v_{n_i}, i = 1, \ldots, q$, where $n_0 = 0 < n_1 < \ldots < n_q = p$, and each walk is either a cycle, or a maximal path.

   Erase all cycles in $v_1 \ldots v_p$ and apply the above statement to the new walk. Conclude that a walk can be "decomposed" to a union of cycles and at most one path. item Let D be a digraph. Assume that there exists a walk from v to w. Show that

   - if $v \neq w$ then there exists a path from u to v of length $\#V - 1$ at most;
   - if $v = w$ there exists a cycle which which contains v, of lenght $\#V$ at most.

2. Let $G = (V, E)$ be a multigraph. Show that the following are equivalent.

   - G is bipartite;
   - all cycles in G have even length;

- G is imprimitive.

3. Let $D = (V, E)$ be a directed multigraph. Assume that the reduced graph D_r of D has two vertices. List all all possible D_r up to the isomorphism, and describe the structure of all possible corresponding D.

4. Prove Proposition 6.1.3.

5. Let $A(D) \in \mathbb{Z}_+^{n\times n}$ be the representation matrix of the multidigraph $D = (\langle n\rangle, E)$. Show that $A(D) + A(D)^\top$ is the representation matrix of the undirected multigraph $G(D) = (\langle n\rangle, E')$ induced by D.

6. Let $G = (\langle n\rangle, E')$ be an undirected multigraph, with the representation matrix $A(G) \in \mathrm{S_n}(\mathbb{Z}_+)$. Show that $A(G)$ is the representation matrix of the induced directed multigraph $D(G)$. In particular, if G is (simple) graph, then $D(G)$ is a (simple) graph with no loops.

7. Let $D = (V, E), D_1 = (V_1, E_1)$ be two multidigraphs with the same number of vertices. Show that D and D_1 are isomorphic if and only if $A(D_1) = P^\top A(D)P$ for some permutation matrix.

8. Let $G = (V_1 \cup V_2, E)$ be a bipartite multigraph. Assume that $\#V_1 = m, \#V_2 = n$ and $B(G) \in \mathbb{Z}_+^{m\times n}$ is a representation matrix of G. Show that a full representation matrix of G is of the form $A(G) = \begin{bmatrix} 0_{m\times m} & B(G) \\ B(G)^\top & 0_{n\times n} \end{bmatrix}$.

6.2 Perron-Frobenius theorem

The aim of this section to prove the Perron-Frobenius theorem.

Theorem 6.2.1 *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Then*

1. *The spectral radius of A, $\rho(A)$, is a positive eigenvalue of A.*

2. *$\rho(A)$ is an algebraically simple eigenvalue of A.*

3. *To $\rho(A)$ corresponds a positive eigenvector $\mathbf{0} < \mathbf{u} \in \mathbb{R}^n$, i.e. $A\mathbf{u} = \rho(A)\mathbf{u}$.*

4. *All other eigenvalues of λ of A satisfy the inequality $|\lambda| < \rho(A)$ if and only if A is primitive, i.e. $A^k > 0$ for some integer $k \geq 1$.*

5. *Assume that A is imprimitive, i.e. not primitive. If $n = 1$ then $A = 0_{1\times 1}$. Assume that $n > 1$. Then there exists exactly $h - 1 \geq 1$ distinct eigenvalues $\lambda_1, \ldots, \lambda_{h-1}$ different from $\rho(A)$ and satisfying $|\lambda_i| = \rho(A)$. Furthermore, the following conditions hold.*

   (a) *λ_i is an algebraically simple eigenvalue of A for $i = 1, \ldots, h-1$.*

   (b) *The complex numbers $\frac{\lambda_i}{\rho(A)}, i = 1, \ldots, h-1$ and 1 are all h roots of unity, i.e. $\lambda_i = \rho(A)e^{\frac{2\pi\sqrt{-1}i}{h}}$ for $i = 1, \ldots, h-1$. Furthermore, if $A\mathbf{z}_i = \lambda_i \mathbf{z}_i, \mathbf{z}_i \neq \mathbf{0}$ then $|\mathbf{z}_i| = \mathbf{u} > 0$, the Perron-Frobenius eigenvector $\mathbf{u}$ given in 3.*

   (c) *Let ζ be any h-root of 1, i.e. $\zeta^h = 1$. Then the matrix ζA is similar to A. Hence, if λ is an eigenvalue of A then $\zeta\lambda$ is an eigenvalue of A having the same algebraic and geometric multiplicity as λ.*

   (d) *There exists a permutation matrix $P \in \mathcal{P}_n$ such that $P^\top AP = B$ has a block h-circulant form*

$$B = \begin{bmatrix} 0 & B_{12} & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & B_{23} & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & 0 & B_{(h-1)h} \\ B_{h1} & 0 & 0 & 0 & \vdots & 0 & 0 \end{bmatrix},$$

$B_{i(i+1)} \in \mathbb{R}^{n_i \times n_{i+1}}, i = 1, \ldots, h, B_{h(h+1)} = B_{h1}, n_{h+1} = n_1, n_1 + \ldots + n_h = n.$

Furthermore, the diagonal blocks of B^h are all irreducible primitive matrices, i.e.
(6.2.1)
$$C_i := B_{i(i+1)} \ldots B_{(h-1)h} B_{h1} \ldots B_{(i-1)i} \in \mathbb{R}_+^{n_i \times n_i},\ i = 1, \ldots, h,$$

are irreducible and primitive.

Our proof follows closely the proof of H. Wielandt [Wie50]. For a nonnegative matrix $A = [a_{ij}] \in \mathbb{R}_+^{n\times n}$ define

$$r(\mathbf{x}) := \min_{i, x_i > 0} \frac{(A\mathbf{x})_i}{x_i}, \text{ where } \mathbf{x} = (x_1, \ldots, x_n)^\top \gneq \mathbf{0}. \tag{6.2.2}$$

It is straightforward to show, e.g. Problem 1, that

$$r(\mathbf{x}) = \max\{s \geq 0, s\mathbf{x} \leq A\mathbf{x}\}. \tag{6.2.3}$$

Theorem 6.2.2 *(Wielandt's characterization) Let $A = [a_{ij}] \in \mathbb{R}_+^{n\times n}$ be irreducible. Then*

$$\max_{\mathbf{x}\gneq\mathbf{0}} r(\mathbf{x}) = \max_{\mathbf{x}=(x_1,\ldots,x_n)^\top\gneq\mathbf{0}} \min_{x_i>0} \frac{(A\mathbf{x})_i}{x_i} = \rho(A) > 0. \tag{6.2.4}$$

The maximum in the above characterization is achieved exactly for all $\mathbf{x} > \mathbf{0}$ of the form $\mathbf{x} = a\mathbf{u}$, where $a > 0$ and $\mathbf{u} = (u_1, \ldots, u_n)^\top > \mathbf{0}$ is the unique positive probability vector, i.e. $\sum_{i=1}^n u_i = 1$, satisfying $A\mathbf{u} = \rho(A)\mathbf{u}$. Moreover, $\rho(A)$ is a geometrically simple eigenvalue.

Proof. Let $r(A) := \sup_{\mathbf{x}\gneq\mathbf{0}} r(\mathbf{x})$. So $r(A) \geq r(\mathbf{1}) = \min_i \sum_{j=1} a_{ij}$. Since an irreducible A can not have a zero row, e.g. Problem 2, it follows that $r(A) \geq r(\mathbf{1}) > 0$.

Denote by

$$\Pi_n := \{(x_1, \ldots, x_n)^\top \gneq \mathbf{0}, \sum_{i=1}^n x_i = 1\}, \tag{6.2.5}$$

the convex set of probability vectors in $\mathbb{R}_+^n$. Note that Π_n is a compact set in $\mathbb{R}_+^n$, i.e. from any sequence $\mathbf{x}_j, j = 1, \ldots,$ we can find a subsequence $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \ldots$ which converges to $\mathbf{x} \in \varPi_n$.

Clearly, for any $\mathbf{x} \gneq \mathbf{0}$ and $a > 0$ we have $r(a\mathbf{x}) = r(\mathbf{x})$. Hence

$$r(A) = \sup_{\mathbf{x}\gneq\mathbf{0}} r(\mathbf{x}) = \sup_{\mathbf{x}\in\varPi_n} r(\mathbf{x}). \tag{6.2.6}$$

Since A is irreducible, $(I + A)^{n-1} > 0$. Hence for any $\mathbf{x} \in \varPi_n$ $\mathbf{y} = (I + A)^{n-1}\mathbf{x} > \mathbf{0}$. (See Problem 3a.) As $r(\mathbf{y})$ is a continuous function on $(I + A)^{n-1}\Pi_n$, it follows that $r(\mathbf{y})$ achieves its maximum on $(I + A)^{n-1}\Pi_n$

$$r_1(A) := \max_{\mathbf{y}\in(I+A)^{n-1}\varPi_n} = r(\mathbf{v}), \text{ for some } \mathbf{v} \text{ in } (I + A)^{n-1}\varPi_n.$$

$r(A)$ is defined as the supremum of $r(\mathbf{x})$ on the set of all $\mathbf{x} \gneq \mathbf{0}$ it follows that $r(A) \geq r_1(A)$. We now show the reversed inequality $r(A) \leq r_1(A)$ which is equivalent to $r(\mathbf{x}) \leq r_1(A)$ for any $\mathbf{x} \gneq \mathbf{0}$.

One has the basic inequality

$$r(\mathbf{x}) \leq r((I + A)^{n-1}\mathbf{x}), \mathbf{x} \gneq \mathbf{0}, \text{ with equality iff } A\mathbf{x} = r(\mathbf{x})\mathbf{x}, \tag{6.2.7}$$

see Problem 3d. For $\mathbf{x} \in \varPi_n$ we have $r(\mathbf{x}) \leq r((I + A)^{n-1}\mathbf{x}) \leq r_1(A)$. In view of (6.2.6) we have $r(A) \leq r_1(A)$. Hence $r(A) = r_1(A)$.

Suppose that $r(\mathbf{x}) = r(A), \mathbf{x} \gneq \mathbf{0}$. Then the definition of $r(A)$ (6.2.6) and (6.2.7) yields that $r(\mathbf{x}) = r((I + A)^{n-1}\mathbf{x})$. The equality case in (6.2.7) yields that $A\mathbf{x} = r(A)\mathbf{x}$. Hence $(1 + r(A))^{n-1}\mathbf{x} = (I + A)^{n-1}\mathbf{x} > \mathbf{0}$, which yields that $\mathbf{x}$ is a positive eigenvector corresponding to the eigenvalue $r(A)$. So $\mathbf{x} = a\mathbf{u}, a > 0$ for the corresponding probability eigenvector $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $A\mathbf{u} = r(A)\mathbf{u}$.

Suppose that $r(\mathbf{z}) = r(A)$ for some vector $\mathbf{z} = (z_1, \ldots, z_n)^\top \gneq \mathbf{0}$. So $\mathbf{z} > 0$ and $A\mathbf{z} = r(A)\mathbf{z}$. Let $b = \min_i \frac{z_i}{u_i}$. We claim that $\mathbf{z} = b\mathbf{u}$. Otherwise $\mathbf{w} := \mathbf{z} - b\mathbf{u} \gneq \mathbf{0}$, $\mathbf{w}$ has at least one coordinate equal to zero, and $A\mathbf{w} = r(A)\mathbf{w}$. So $r(\mathbf{w}) = r(A)$. This is impossible since we showed above that $\mathbf{w} > \mathbf{0}$! Hence $\mathbf{z} = b\mathbf{u}$. Assume now that $\mathbf{y} \in \mathbb{R}^n$ is an eigenvector of A corresponding to $r(A)$. So $A\mathbf{y} = r(A)\mathbf{y}$. There exists a big positive number c such that $\mathbf{z} = \mathbf{y}+c\mathbf{u} > \mathbf{0}$. Clearly $A\mathbf{z} = r(A)\mathbf{z}$. Hence $r(\mathbf{z}) = r(A)$ and we showed above that $\mathbf{z} = b\mathbf{u}$. So $\mathbf{y} = (b - c)\mathbf{u}$. Hence $r(A)$ is a geometrically simple eigenvalue of A.

We now show that $r(A) = \rho(A)$. Let $\lambda \neq r(A)$ be another eigenvalue of A, which may be complex valued. Then

$$(\lambda\mathbf{z})_i = \lambda\mathbf{z}_i = (A\mathbf{z})_i = \sum_{j=1}^{n} a_{ij}z_j, \ i = 1, \ldots, n,$$

where $\mathbf{0} \neq \mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{C}^n$ is the corresponding eigenvector of A. Take the absolute values in the above equality, and use the triangle inequality, and the fact that A is nonnegative matrix to obtain

$$|\lambda|\ |z_i| \leq \sum_{j=1}^{n} a_{ij}|z_j|, \ i = 1, \ldots, n.$$

Let $|\mathbf{z}| := (|z_1|, \ldots, |z_n|)^\top \gneq \mathbf{0}$. Then the above inequality is equivalent to $|\lambda|\ |\mathbf{z}| \leq A|\mathbf{z}|$. Use (6.2.3) to deduce that $|\lambda| \leq r(|\mathbf{z}|)$. Since $r(|\mathbf{z}|) \leq r(A)$ we deduce that $|\lambda| \leq r(A)$. Hence $\rho(A) = r(A)$, which yields (6.2.4). □

Lemma 6.2.3 *Let $A \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Then $\rho(A)$ is an algebraically simple eigenvalue.*

Proof. (For all the notions and results used here see §**??**.) Theorem 6.2.2 implies that $\rho(A)$ is geometrically simple, i.e. nul $(\rho(A)I - A) = 1$. Hence rank $(\rho(\mathrm{A})\mathrm{I} - \mathrm{A}) = \mathrm{n} - 1$. Hence adj $(\rho(\mathrm{A})\mathrm{I} - \mathrm{A}) = \mathrm{t}\mathbf{u}\mathbf{v}^\top$, where $A\mathbf{u} = \rho(A)\mathbf{u}, A^\top\mathbf{v} = \rho(A)\mathbf{v}, \mathbf{u}, \mathbf{v} > \mathbf{0}$ and $0 \neq t \in \mathbb{R}$. Note that $\mathbf{u}\mathbf{v}^\top$ is a positive matrix, hence tr $\mathbf{u}\mathbf{v}^\top = \mathbf{v}^\top\mathbf{u} > 0$. Since

$$(\det\,(\lambda I - A))'(\lambda = \rho(A)) = \mathrm{tr\,adj}\ (\rho(\mathrm{A})\mathrm{I} - \mathrm{A}) = \mathrm{t}(\mathbf{v}^\top\mathbf{u}) \neq 0,$$

we deduce that $\rho(A)$ is a simple root of the characteristic polynomial of A. □

As usual, denote by $\mathrm{S}^1 := \{\mathrm{z} \in \mathbb{C}, |\mathrm{z}| = 1\}$ the unit circle in the complex plane

Lemma 6.2.4 *Let $A \in \mathbb{R}_+^{n\times n}$ be irreducible, $C \in \mathbb{C}^{n\times n}$. Assume that $|C| \le A$. Then $\rho(C) \le \rho(A)$. Equality holds, i.e. there exists $\lambda \in \mathrm{spec}\,\mathrm{C}$, such that $\lambda = \zeta\rho(A)$ for some $\zeta \in \mathrm{S}^1$, if and only if there exists a complex diagonal matrix $D \in \mathbb{C}^{n\times n}$, whose diagonal entries are equal to 1, such that $C = \zeta DAD^{-1}$. The matrix D is unique up to a multiplication by $t \in \mathrm{S}^1$.*

Proof. Assume that $A = [a_{ij}], C = [c_{ij}]$. Let $\mathbf{z} = (z_1, \ldots, z_n)^\top \neq \mathbf{0}$ be an eigenvector of C corresponding to an eigenvalue λ, i.e. $\lambda\mathbf{z} = C\mathbf{z}$. The arguments of the proof of Theorem 6.2.2 yield that $|\lambda|\ |\mathbf{z}| \le |C|\ |\mathbf{z}|$. Hence $|\lambda|\ |\mathbf{z}| \le |A|\ |\mathbf{z}|$, which implies that $|\lambda| \le r(|\mathbf{z}|) \le r(A) = \rho(A)$.

Suppose that $\rho(C) = \rho(A)$. So there exists $\lambda \in \mathrm{spec}\,\mathrm{C}$, such that $|\lambda| = \rho(A)$. So $\lambda = \zeta\rho(A)$ for some $\zeta \in \mathrm{S}^1$. Furthermore, for the corresponding eigenvector $\mathbf{z}$ we have the equalities

$$|\lambda|\ |\mathbf{z}| = |C\mathbf{z}| = |C|\ |\mathbf{z}| = A|\mathbf{z}| = r(A)|\mathbf{z}|.$$

Theorem 6.2.2 yields that $|\mathbf{z}|$ is a positive vector. Let $z_i = d_i|\mathbf{z}_i|, |d_i| = 1$ for $i = 1, \ldots, n$. The equality $|C\mathbf{z}| = |C|\ |\mathbf{z}| = A|\mathbf{z}|$ combined with the triangle inequality and $|C| \le A$, yields first that $|C| = A$. Furthermore for each fixed i the nonzero complex numbers $c_{i1}z_1, \ldots, c_{in}z_n$ have the same argument, i.e. $c_{ij} = \zeta_i a_{ij}\bar{d}_j$ for $j = 1, \ldots, n$ and some complex number ζ_j, where $|\zeta_i| = 1$. Recall that $\lambda z_i = (C\mathbf{z})_i$. Hence $\zeta_i = \zeta d_i$ for $i = 1, \ldots, n$. Thus $C = \zeta DAD^{-1}$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$. It is straightforward to see that D is unique up tD for any $t \in \mathrm{S}^1$.

Suppose now that for $D = \mathrm{diag}(d_1, \ldots, d_n)$, where $|d_1| = \ldots = |d_n| = 1$ and $|\zeta| = 1$ we have that $C = \zeta DAD^{-1}$. Then $\boldsymbol{\lambda}(C) = \zeta\boldsymbol{\lambda}(A)$, see Fact (**??**.**??**). So $\rho(C) = \rho(A)$. Furthermore $c_{ij} = \zeta d_i c_{ij}\bar{d}_j, i, j = 1, \ldots, n$. So $|C| = A$. □

Lemma 6.2.5 *Let $\zeta_1, \ldots, \zeta_h \in \mathrm{S}^1$ be h distinct complex numbers which form a multiplicative semi-group, i.e. for any integers $i, j \in [1, h]$ $\zeta_i\zeta_j \in \{\zeta_1, \ldots, \zeta_h\}$. Then the set $\{\zeta_1, \ldots, \zeta_h\}$ is the set, (the group), of all h roots of 1: $e^{\frac{2\pi i\sqrt{-1}}{h}}, i = 1, \ldots, h$.*

Proof. Let $\zeta \in \mathcal{T} := \{\zeta_1, \ldots \zeta_h\}$. Consider the sequence $\zeta^i, i = 1, \ldots$. Since $\zeta^{i+1} = \zeta\zeta^i$ for $i = 1, \ldots,$ and $\mathcal{T}$ is a semigroup, it follows that each ζ^i is in $\mathcal{T}$. Since $\mathcal{T}$ is a finite set, we must have two positive integers such that $\zeta^k = \zeta^l$ for $k < l$. Assume that k and l are the smallest possible positive integers. So $\zeta^p = 1$, where $p = l - k \ge 1$, and $\mathcal{T}_p := \{\zeta, \zeta^2, \ldots, \zeta^{p-1}, \zeta^p = 1\}$

are all p roots of 1. ζ is called a p-primitive root of 1. I.e. $\zeta = e^{\frac{2\pi p_1\sqrt{-1}}{p}}$ where p_1 is an positive integer less than p. Furthermore p_1 and p are *coprime*, which is denoted by $(p_1, p) = 1$. Note that $\zeta^i \in \mathcal{T}$ for any integer i.

Next we choose $\zeta \in \mathcal{T}$, such that ζ is a primitive p-root of 1 of the maximal possible order. We claim that $p = h$, which is equivalent to the equality $\mathcal{T} = \mathcal{T}_p$. Assume to the contrary that $\mathcal{T}_p \subsetneq \mathcal{T}$. Let $\eta \in \mathcal{T}\backslash\mathcal{T}_p$. The previous arguments show that $\eta =$ is a q-primitive root of 1. So $\mathcal{T}_q \subset \mathcal{T}$, and $\mathcal{T}_q \subsetneq \mathcal{T}_p$. So q can not divide p. Also the maximality of p yields that $q \le p$. Let $(p, q) = r$ be the g.c.d., the greatest common divisor of p and q. So $1 \le r < q$. Recall that Euclid algorithm, which is applied to the division of p by q with a residue, yields that there exists two integers i, j such that $ip + jq = r$. Let $l := \frac{pq}{r} > p$ be the least common multiplier of p and q. Observe that $\zeta' = e^{\frac{2\pi\sqrt{-1}}{p}} \in \mathcal{T}_p, \eta' = e^{\frac{2\pi\sqrt{-1}}{q}} \in \mathcal{T}_q$. So

$$\xi := (\eta')^i(\zeta')^j = e^{\frac{2\pi(ip+jq)\sqrt{-1}}{pq}} = e^{\frac{2\pi\sqrt{-1}}{l}} \in \mathcal{T}.$$

As ξ is an l-primitive root of 1, we obtain a contradiction to the maximality of p. So $p = h$ and $\mathcal{T}$ is the set of all h-roots of unity. □

Lemma 6.2.6 *Let $A \in \mathbb{R}_+^{n\times n}$ be irreducible, and assume that for a positive integer $h \ge 2$, A has $h-1$ distinct eigenvalues $\lambda_1, \ldots, \lambda_{h-1}$, which are distinct from $\rho(A)$, such that $|\lambda_1| = \ldots = |\lambda_{h-1}| = \rho(A)$. Then the conditions (5a-5c) of Theorem 6.2.1 hold. Moreover, $P^\top AP = B$, where B is of the form given in (5d) and P is a permutation matrix.*

Proof. Assume that $\zeta_i := \frac{\lambda_i}{\rho(A)} \in \mathrm{S}^1$ for $i = 1, \ldots, h-1$ and $\zeta_h = 1$. Apply Lemma 6.2.4 to $C = A$ and $\lambda = \zeta_i\rho(A)$ to deduce that $A = \zeta_i D_i A D_i^{-1}$ where D_i is a diagonal matrix such that $|D| = I$ for $i = 1, \ldots, h$. Hence, if λ is an eigenvalue of A then $\zeta_i\lambda$ is an eigenvalue of A, with an algebraic and geometrical multiplicity as λ. In particular, since $\rho(A)$ is an algebraically simple eigenvalue of A, $\lambda_i = \zeta_i\rho(A)$ is an algebraically simple of A for $i = 1, \ldots, h-1$. This establish (5a).

Let $\mathcal{T} = \{\zeta_1, \ldots, \zeta_h\}$. Note that

(6.2.8)
$$A = \zeta_i D_i A D_i^{-1} = \zeta_i D_i(\zeta_j D_j A D_j^{-1})D_i^{-1} = (\zeta_i\zeta_j)(D_iD_j)A(D_iD_j)^{-1}.$$

So $\zeta_i\zeta_j\rho(A)$ is an eigenvalue of A. Hence $\zeta_i\zeta_j \in \mathcal{T}$, i.e. $\mathcal{T}$ is a semigroup. Lemma 6.2.5 yields that $\{\zeta_1, \ldots, \zeta_n\}$ are all h roots of 1. Note that if $A\mathbf{z}_i = \lambda_i\mathbf{z}_i, \mathbf{z}_i \ne \mathbf{0}$, then $\mathbf{z}_i = tD_i\mathbf{u}$ for some $0 \ne t \in \mathbb{C}$, where $\mathbf{u} > \mathbf{0}$ is the Perron-Frobenius vector given in Theorem 6.2.1. This establish (5b) of Theorem 6.2.1.

Let $\zeta = e^{\frac{2\pi\sqrt{-1}}{h}} \in \mathcal{T}$. Then $A = \zeta DAD^{-1}$, where D is a diagonal matrix $D = (d_1, \ldots, d_n), |D| = I$. Since D can be replaced by $\bar{d}_1 D$, we can assume that $d_1 = 1$. (6.2.8) yields that $A = \zeta^h D^h AD^{-h} = IAI^{-1}$. Lemma 6.2.4 yields that $D^h = \mathrm{diag}(d_1^h, \ldots, d_n^h) = tI$. Since $d_1 = 1$ it follows that $D^h = I$. So all the diagonal entries of D are h-roots of unity. Let $P \in \mathcal{P}_n$ be a permutation matrix such that the diagonal matrix $E = P^\top DP$ is of the following block diagonal form

$$E = I_{n_1} \oplus \mu_1 I_{n_2} \oplus \ldots \oplus \mu_{l-1} I_{n_l}, \mu_i = e^{\frac{2\pi k_i \sqrt{-1}}{h}}, i = 1, \ldots, l-1, 1 \le k_1 < k_2 < \ldots < k_{l-1} \le h-1.$$

Note that $l \le h$ and equality holds if and only if $k_i = i$. Let $\mu_0 = 1$.

Let $B = P^\top AP$. Partition B to a block matrix $[B_{ij}]_{i=j=1}^l$ where $B_{ij} \in \mathbb{R}_+^{n_i \times n_j}$ for $i, j = 1, \ldots, l$. Then the equality $A = \zeta DAD^{-1}$ yields $B = \zeta EBE^{-1}$. The structure of B and E implies the equalities

$$B_{ij} = \zeta \frac{\mu_{i-1}}{\mu_{j-1}} B_{ij}, \quad i, j = 1, \ldots, l.$$

Since all the entries of B_{ij} are nonnegative we obtain that $B_{ij} = 0$ if $\zeta \frac{\mu_{i-1}}{\mu_{j-1}} \neq 1$. Hence $B_{ii} = 0$ for $i = 1, \ldots, l$ Since B is irreducible it follows that not all $B_{i1}, \ldots, B_{il}$ are zero matrices for each $i = 1, \ldots, l$. First start with $i = 1$. Since $\mu_0 = 1$ and $j_1 \ge 1$ it follows that $\mu_j \neq \zeta$ for $j > 1$. So $B_{1j} = 0$ for $j = 3, \ldots, l$. Hence $B_{12} \neq 0$, which implies that $\mu_1 = \zeta$, i.e. $k_1 = 1$. Now let $i = 2$ and consider $j = 1, \ldots, l$. As $k_i \in [k_1 + 1, h - 1]$ for $i > 1$, it follows that $B_{2j} = 0$ for $j \neq 3$. Hence $B_{23} \neq 0$ which yields that $k_2 = 2$. Applying these arguments for $i = 3, \ldots, l - 1$ we deduce that $B_{ij} = 0$ for $j \neq i + 1$, $B_{i(i+1)} \neq 0, k_i = i$ for $i = 1, \ldots, l - 1$. It is left to consider $i = l$. Note that

$$\frac{\zeta \mu_{l-1}}{\mu_{j-1}} = \frac{\zeta^l}{\zeta^{j-1}} = \zeta^{l-(j-1)}, \text{ which is different from 1 for } j \in [2, l].$$

Hence $B_{lj} = 0$ for $j > 1$. Since B is irreducible, $B_{11} \neq 0$. So $\zeta^l = 1$. Since $l \le h$ we deduce that $l = h$. Hence B has the block form given in (5d). □

Proposition 6.2.7 *Let $A \in \mathbb{R}_+^{n \times n}$ be irreducible. Suppose that $\mathbf{0} \lneq \mathbf{w} \in \mathbb{R}_+^n$ is an eigenvector of A, i.e. $A\mathbf{w} = \lambda \mathbf{w}$. Then $\lambda = \rho(A)$ and $\mathbf{w} > \mathbf{0}$.*

Proof. Let $\mathbf{v} > \mathbf{0}$ be the Perron-Frobenius vector of $A^\top$, i.e. $A^\top \mathbf{v} = \rho(A)\mathbf{v}$. Then

$$\mathbf{v}^\top A\mathbf{w} = \mathbf{v}^\top(\lambda \mathbf{w}) = \rho(A)\mathbf{v}^\top \mathbf{w} \Rightarrow (\rho(A) - \lambda)\mathbf{v}^\top \mathbf{w} = 0.$$

If $\rho(A) \neq \lambda$ we deduce that $\mathbf{v}^\top \mathbf{w} = 0$, which is impossible, since $\mathbf{v} > \mathbf{0}$ and $\mathbf{w} \gneq \mathbf{0}$. Hence $\lambda = \rho(A)$. Then $\mathbf{w}$ is the Perron-Frobenius eigenvector and $\mathbf{w} > \mathbf{0}$. □

Lemma 6.2.8 *Let $A \in \mathbb{R}_+^{n\times}$ be irreducible. Then A is primitive if and only if each eigenvalue λ of A different from $\rho(A)$ satisfies the inequality $|\lambda| < \rho(A)$. I.e. condition (4) of Theorem 6.2.2 holds.*

Proof. By considering $B = \frac{1}{\rho(A)}A$, it is enough to consider the case $\rho(A) = 1$. Assume first that if $\lambda \neq 1$ is an eigenvalue of A then $|\lambda| < 1$. Theorem 6.2.2 implies $A\mathbf{u} = \mathbf{u}, A^\top \mathbf{w} = \mathbf{w}$ for some $\mathbf{u}, \mathbf{w} > \mathbf{0}$. So $\mathbf{w}^\top \mathbf{u} > 0$. Let $\mathbf{v} := (\mathbf{w}^\top \mathbf{u})^{-1}\mathbf{w}$. Then $A^\top \mathbf{v} = \mathbf{v}$ and $\mathbf{v}^\top \mathbf{u} = 1$. Fact **??** yields that $\lim_{k\to\infty} A^k = \mathbf{u}\mathbf{v}^\top > 0$. So there exists integer $k_0 \geq 1$, such that $A^k > 0$ for $k \geq k_0$, i.e. A is primitive.

Assume now A is has exactly $h > 1$ distinct eigenvalues λ satisfying $|\lambda| = 1$. Lemma 6.2.6 implies that there exists a permutation matrix P such that $B = P^\top AP$ is of the form (5d) of Theorem 6.2.1. Note that B^h is a block diagonal matrix. Hence $B^{hj} = (B^h)^j$ is a block diagonal matrix for $j = 1, \ldots,$.... Hence, B^{hj} is never a positive matrix, so A^{hj} is never a positive matrix. In view of Problem 4, A is not primitive. □

Lemma 6.2.9 *Let $B \in \mathbb{R}_+^{n\times n}$ be an irreducible, imprimitive matrix, having $h > 1$ distinct eigenvalues λ satisfying $|\lambda| = \rho(B)$. Suppose furthermore that B has the form (5d) of Theorem 6.2.1. Then B^h is a block diagonal matrix, where each diagonal block is an irreducible primitive matrix whose spectral radius is $\rho(B)^h$. In particular, the last claim of (5d) of Theorem 6.2.1 holds.*

Proof. Let $D(B) = (\langle n\rangle, E)$ be the digraph associated with B. Let $p_0 = 0, p_1 = p_0 + n_1, \ldots, p_h = p_{h-1} + n_h = n$. Denote $V_i = \{p_{i-1}+1, \ldots, p_i\}$ for $i = 1, \ldots, h$, and let $V_{h+1} := V_1$. So $\langle n\rangle = \cup_{i=1}^h V_i$. The form of B implies that $E \subset \cup_{i=1}^h V_i \times V_{i+1}$. Thus, any walk that connects vertices $j, k \in V_i$ must be divisible by h. Observe next that $B^h = \mathrm{diag}(C_1, \ldots, C_h)$, where $C_1 = [c_{jk}^{(1)}]_{j=k=1}^{n_1}, \ldots, C_h = [c_{jk}^{(h)}]_{j=k=1}^{n_h}$ are defined in (6.2.1). Let $D(C_i) = (V_i, E_i)$ be the digraph associated with C_i for $i = 1, \ldots, h$. Then there exists a path of length h from j to k in V_i if and only if $c_{jk}^{(i)} > 0$. Since B is irreducible, $D(B)$ is strongly connected. Hence, each $D(C_i)$ is strongly connected. Thus, each C_i is irreducible.

Recall that $B\mathbf{u} = \rho(B)\mathbf{u}$ for the Perron-Frobenius vector $\mathbf{u}^\top = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_h^\top) > \mathbf{0}^\top, \mathbf{u}_i \in \mathbb{R}_+^{n_i}, i = 1, \ldots, h$. Thus, $B^h\mathbf{u} = \rho(B)^h\mathbf{u}$, which implies that $C_i\mathbf{u}_i = \rho(B)^h\mathbf{u}_i, i = 1, \ldots, h$. Since $\mathbf{u}_i > \mathbf{0}$ Proposition 6.2.7 yields that

$\rho(C_i) = \rho(B)^h, i = 1,\ldots,h$. Recall that the eigenvalues of B^h are the h power of the eigenvalues of B, i.e. $\boldsymbol{\lambda}(B) = (\lambda_1^h,\ldots,\lambda_n^h)$, where $\boldsymbol{\lambda}(B^h) = (\lambda_1,\ldots,\lambda_n)$. Furthermore, B has h simple eigenvalues $\rho(B)e^{\frac{2\pi i\sqrt{-1}}{h}}, i = 1,\ldots,h$ with $|\lambda| = \rho(B)$, and all other eigenvalues satisfy $|\lambda| < \rho(B)$. Hence B^h has one eigenvalues $\rho(B)^h$ of an algebraic multiplicity h and all other eigenvalues μ satisfy $|\mu| < \rho(B)^h$.

Since $B^h = \mathrm{diag}(C_1,\ldots,C_h)$, we deduce that $\boldsymbol{\lambda}(B^h) = (\boldsymbol{\lambda}(C_1),\ldots,\boldsymbol{\lambda}(C_h))$. As C_i is irreducible and $\rho(C_i) = \rho(B)^h$, we deduce that all other eigenvalues μ of C_i satisfy $|\mu| < \rho(C_i)$. Lemma 6.2.8 yields that C_i is primitive. □

Problems

1. Prove equality (6.2.3).

2. Show that if $A \in \mathbb{R}_+^{n\times n}$ is irreducible then can not have a zero row or column.

3. Assume that $A \in \mathbb{R}_+^{n\times n}$ is irreducible. Show

   (a) For each $\mathbf{x} \in \mathit{\Pi}_n$ the vector $(I + A)^{n-1}\mathbf{x}$ is positive.

   (b) The set $(I + A)^{n-1}\Pi_n := \{\mathbf{y} = (I + A)^{n-1}\mathbf{x}, \mathbf{x} \in \mathit{\Pi}_n\}$ is a compact set of positive vectors.

   (c) Show that $r(\mathbf{y})$ is a continuous function on $(I + A)^{n-1}\Pi_n$.

   (d) Show (6.2.7). **Hint**: use that $(A+I)^{n-1}(A\mathbf{x}-r(\mathbf{x})\mathbf{x})$ is a positive vector, unless $A\mathbf{x} = r(\mathbf{x})\mathbf{x}$.

4. Assume that $A \in \mathbb{R}_+^{n\times n}$ is irreducible. Show the following are equivalent

   (a) A is primitive.

   (b) There exists a positive integer k_0 such that for any integer $k \geq k_0$ $A^k > 0$.

5. Let $D = (\langle h\rangle, E)$ be the cycle $1 \to 2 \to \ldots \to h-1 \to h \to 1$.

   (a) Show that representation matrix $A(D)$ is a permutation matrix, which has the form of B given in (5d) of Theorem 6.2.1, where each nonzero block is 1×1 matrix $[1]$. $A(D)$ is called a *circulant* matrix.

   (b) Find all the eigenvalues and the corresponding eigenvectors of $A(D)$.

6. Let the assumptions of Lemma 6.2.9. Assume the notation of the proof of Lemma 6.2.9.
   (a) Show that the length of any closed walk in $D(B)$ is divisible by h.
   (b) Show that a length of any walk from a vertex in V_i to a vertex V_j, such that $1 \le i < j \le h$, minus $j - i$ is divisible by h.
   (c) What can you say on a length of any walk from a vertex in V_i to a vertex V_j, such that $1 \le j < i \le h$?
   (d) Show that each C_i is irreducible.

7. Let $D = (V, E)$ be a digraph and assume that V is a disjoint union of h nonempty sets $V_1, \ldots, V_h$. Denote $V_{h+1} := V_1$. Assume that $E \subset \cup_{i=1}^{h} V_i \times V_{i+1}$. Let $D_i = (V_i, E_i)$ be the following digraph. The diedge $(v, w) \in E_i$, if there is a path of length h in D from v to w in D.
   (a) Show that D is strongly connected if and only if D_i is strongly connected for $i = 1, \ldots, h$.
   (b) Assume that D is strongly connected. Let $1 \le i < j \le h$. Then D_i is primitive if and only if D_j is primitive.

8. Let $B \in \mathbb{R}_+^{n\times n}$ be a block matrix of the form given in (5d) of Theorem 6.2.1.
   (a) Show that B^h is a block diagonal matrix $\mathrm{diag}(C_1, \ldots, C_h)$, where C_i is given (6.2.1).
   (b) Show that B is irreducible if and only if C_i is irreducible for $i = 1, \ldots, h$.
   (c) Assume that B is irreducible.
      i. Let $1 \le i < j \le h$. Then C_i is primitive if and only if C_j is primitive.
      ii. B has h distinct eigenvalues on the circle $|z| = \rho(B)$ if and only if some C_i is primitive.

9. Assume the assumptions of Lemma 6.2.6. Let $A\mathbf{u} = \rho(A)\mathbf{u}, \mathbf{u} = (u_1, \ldots, u_n)^\top > \mathbf{0}$. Assume that η is an h-root of unity, and suppose that $A\mathbf{z} = \eta\mathbf{z}, \mathbf{z} = (z_1, \ldots, z_n)$, such that $|\mathbf{z}| = \mathbf{u}$. Assume that $z_i = u_i$ for a given $i \in \langle n\rangle$. (This is always possible by considering $\frac{\bar{z}_i}{|z_i|}\mathbf{z}$.) Show $z_j = \eta^{k(j)} u_j$, for a suitable integer $k(j)$, for $j = 1, \ldots, n$. Furthermore, given an integer k then there exists $j \in \langle n\rangle$ such that $z_j = \eta^k u_j$.

Hint: Use the proof of Lemma 6.2.6.

10. Let $B \in \mathbb{R}_+^{n\times n}$ be an irreducible block matrix of the form given in (5d) of Theorem 6.2.1. Let $C_1, \ldots, C_h$ be defined in (6.2.1). Suppose that B has more than h distinct eigenvalues on the circle $|z| = \rho(B)$. Then TFAE

    (a) B has qh eigenvalues the circle $|z| = \rho(B)$, for some $q > 1$.

    (b) Each C_i has $q > 1$ distinct eigenvalues on $|z| = \rho(C_i) = \rho(B)^h$.

    (c) Some C_i has $q > 1$ distinct eigenvalues on $|z| = \rho(C_i) = \rho(B)^h$.

    (d) Let $D(B) = (\langle n\rangle, E)$ and $V_i = \{p_{i-1}+1, \ldots, p_i\}$ for $i = 1, \ldots, h$ be defined as in the proof of Lemma 6.2.9. Then each V_i is a disjoint union of q nonempty sets $W_i, W_{i+h}, \ldots, W_{i+(q-1)h}$ for $i = 1, \ldots, h$, such that $E \subset \cup_{j=1}^{qh} W_j \times W_{j+1}$, where $W_{qh+1} := W_1$. Let $H_j = (W_j, F_j), F_j \subset W_j \times W_j$ be the following digraph. The diedge (v, w) is in F_j, if and only if there is a path of length qh in $D(B)$ from v to w in W_j. Then each digraph H_j is strongly connected and primitive.

    Hint: Use the structure of the eigenvalues λ of B on the circle $|\lambda| = \rho(B)$, and the corresponding eigenvector $\mathbf{z}$ to λ given in (5b) of Theorem 6.2.1.

11. For $A \in R_+^{n\times n}$ and $\mathbf{0} < \mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}_+^n$ let $R(\mathbf{x}) = \max_{i\in\langle n\rangle} \frac{(A\mathbf{x})_i}{x_i}$. Assume that A is irreducible. Show that $\inf_{\mathbf{x}>\mathbf{0}} R(\mathbf{x}) = \rho(A)$. I.e.

    $$\min_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \max_{i\in\langle n\rangle} \frac{(A\mathbf{x})_i}{x_i} = \rho(A). \tag{6.2.9}$$

    Furthermore, $R(\mathbf{x}) = \rho(A)$ if and only if $A\mathbf{x} = \rho(A)\mathbf{x}$.

    Hint: Mimic the proof of Theorem 6.2.2.

12. Let $n > 1$ and $D = (\langle n\rangle, E)$ be a strongly connected digraph. Show

    (a) If D has exactly one cycle, it must be a Hamiltonian cycle, i.e. the length of of this cycle is n. Then D is not primitive.

    (b) Suppose that D has exactly two directed cycles. Then the shortest cycle has length $n-1$ if and only if it is possible to rename the vertices so that the shortest cycle is of the form $1 \to 2 \to \ldots \to n-1 \to 1$ and the second cycle is a Hamiltonian cycle $1 \to 2 \to \ldots \to n-1 \to n \to 1$. In this case D is primitive. Moreover $A(D)^k > 0$ if and only if $k \geq n^2 - 2n + 2$.

    (c) Assume that D is primitive. Show that the shortest cycle of D has at most length $n-1$.

6.3 Index of primitivity

Theorem 6.3.1 *Let $A \in \mathbb{R}_+^{n\times n}$ be a primitive matrix. Let $s \geq 1$ be the length of the shortest cycle in the digraph $D(A) = (\langle n\rangle, E)$. Then $A^{s(n-2)+n} > 0$. In particular $A^{(n-1)^2+1} > 0$.*

Proof. For $n = 1$ we have that $s = 1$ and the theorem is trivial. Assume that $n > 1$. Note that since A is primitive $s \leq n-1$. (See Problem 12c.)

Suppose first that $s = 1$. So $D(A)$ contains a loop. Relabel the vertices of $D(A)$ to assume that $(1,1) \in E$. I.e. we can assume that $A = [a_{ij}]$ and $a_{11} > 0$. Recall that from 1 to $j > 1$ there exists a path of length $1 \leq l(j) \leq n-1$. By looping at 1 first $n-1-l(j)$ times we deduce the existence of a walk of length $n-1$ from 1 to $j > 1$. Clearly, there exists a walk of length $n-1$ from 1 to 1: $1 \to 1 \to \ldots \to 1$. Similarly, for each $j > 1$ there exists a walk of length $n-1$ from j to 1. Hence, the first row and the column of A^{n-1} is positive. Thus, $A^{2(n-1)} = A^{n-1}A^{n-1}$ is a positive matrix.

Assume now that $s \geq 2$. Relabel the vertices of $D(A)$ such that one has the cycle on vertices $c := \{1, 2, \ldots, s\}$: $1 \to \ldots \to s \to 1$. Then the first s diagonal entries of A^s are positive. Since A was primitive, Lemma 6.2.8 implies that A^s is primitive. Our previous arguments show that $(A^s)^{n-1}$ has the first s rows and columns positive. Let

$$A^{n-s} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, F_{11} \in \mathbb{R}_+^{s\times s}, F_{12}, F_{21}^\top \in \mathbb{R}_+^{s\times(n-s)}, F_{22} \in \mathbb{R}_+^{(n-s)\times(n-s)}.$$

Clearly, $F_{11} \geq ([a_{ij}]_{i=j=1}^s)^{n-s}$. Since $D(A)$ contains a cycle of length s on $\langle s\rangle$ it follows that each row and column of F_{11} is not zero. Clearly,

$$A^{s(n-2)+n} = A^{(n-s)}A^{s(n-1)}. \tag{6.3.1}$$

Hence the first s rows of $A^{s(n-2)+n}$ are positive. We claim that each row of F_{21} is nonzero. Indeed, take the shortest walk from $j \in U := \langle s+1, \ldots, n\rangle$ to the set of vertices $V := \{1, \ldots, s\}$. This shortest walk is a path which can contain at most $n-s$ vertices in U, before it ends in $i \in V$. Hence the length of this path is $m(j) \leq n-s$. After that continue take a walk on the cycle c of length $n-s-m(j)$, to deduce that there is a walk of length $n-s$ from j to V. Hence the $j-s$ row of F_{21} is nonzero. Use (6.3.1) and the fact that the first s rows of $(A^s)^{n-1}$ positive to deduce that $A^{s(n-2)+n} > 0$. □

Proof of Theorem 6.1.5. Problem 6.1.5.1 yields that the length L of any closed walk in D is a sum of lengthes of a number of cycles in D.

Hence ℓ divides L. Assume that D is primitive, i.e. $A^k > 0$ for any $k \ge k_0$. Hence for each $k \ge k_0$ there exists a closed walk in D of length k. Therefore $\ell = 1$.

Suppose now that $D = (V, E)$ is imprimitive, i.e. $A(D)$ is imprimitive. (5d) of Theorem 6.2.1 yields that there V decomposes to a nonempty disjoint sets $V_1, \ldots, V_h$, where $h > 1$. Moreover $E \subset \cup_{i=1}^{h} V_i \times V_{i+1}$, where $V_{h+1} = V_1$. So any closed walk must be divisible by $h > 1$. In particular, the length of each cycle is divisible by h. Thus $\ell \ge h > 1$. Hence D is primitive if and only if $\ell = 1$. Suppose that D is primite. Theorem 6.3.1 yields that $A(D)^{s(n-2)+n} > 0$, where s is the length of the shortest cycle. This proves part 1 of Theorem 6.1.5.

Assume now that D is imprimitive. So $A(D)$ has $h > 1$ distinct eigenvalues of modulus $\rho(A(D))$. Relabel the vertices of D so that $A(D)$ is of the form B given in (5d) of Theorem 6.2.1. As we pointed out, each cycle in D is divisible by h. It is left to show that the $\ell = h$. Let $D_i = (V_i, E_i)$ be defined as in the proof of Lemma 6.2.9. It is straightforward to see that each cycle in D_i corresponds of length L to a cycle in D of length hL. Since C_i is primitive, it follows from the first part of the proof, that the g.c.d of lengths of all cycles in C_i is 1. Hence, the g.c.d. of lengths of the corresponding cycles in D is h. □

6.4 Reducible matrices

Theorem 6.4.1 *Let $A \in \mathbb{R}_+^{n\times n}$. Then $\rho(A)$, the spectral radius of A, is an eigenvalue of A. There exists a probability vector $\mathbf{x} \in \Pi_n$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$.*

Proof. Let $J_n \in \{1\}^{n\times n}$ be a matrix whose entries are 1. For $\varepsilon > 0$ let $A(\varepsilon) = A + \varepsilon J_n$. Then $A(\varepsilon) > 0$. Hence,

$$(6.4.1) \qquad \rho(A(\varepsilon)) \in \text{spec}\ (\mathrm{A}(\varepsilon)) \text{ and } \mathrm{A}(\varepsilon)\mathbf{x}(\varepsilon),\ \mathbf{0} < \mathbf{x}(\varepsilon) \in \Pi_n \text{ for } \varepsilon > \mathrm{o}.$$

Since the coefficients of the characteristic polynomial of $A(\varepsilon)$ are polynomial in ε, it follows that the eigenvalues of $A(\varepsilon)$ are continuous function of ε. Hence

$$\lim_{\varepsilon\to 0} \text{spec}\ (\mathrm{A}(\varepsilon)) = \text{spec A}, \qquad \lim_{\mathrm{var}\epsilon\to 0} \rho(\mathrm{A}(\varepsilon)) = \rho(\mathrm{A}).$$

Combine that with (6.4.1) to deduce that $\rho(A) \in$ spec A. Choose $\varepsilon_k = \frac{1}{k}, k = 1, \ldots,$. Since Π_n is a compact set, there exists a subsequence $1 < k_1 < k_2 < \ldots$ such that $lim_{j\to\infty}\mathbf{x}(\epsilon_{k_j} = \mathbf{x} \in \Pi_n$. The second equality of (6.4.1) yields that $A\mathbf{x} = \rho(A)\mathbf{x}$. □

It is easy to have examples where $\rho(A) = 0$ for some $A \in \mathbb{R}_+^{n\times n}$, and $\rho(A)$ is not a geometrically simple eigenvalue. (I.e. the Jordan canonical form of A contains a Jordan block of order greater than one with the eigenvalue $\rho(A)$.)

Proposition 6.4.2 *Let $A \in \mathbb{R}_+^{n\times n}$.*

1. *Assume that $C \in \mathbb{R}_+^{n\times n}$ and $A \ge C$. Then $\rho(A) \ge \rho(C)$. If either A or C are irreducible then $\rho(A) = \rho(C)$ if and only if $A = C$.*

2. *Assume that $B \in \mathbb{R}_+^{m\times m}, 1 \le m < n$ is a principle submatrix of A, obtained by deleting $n - m$ rows and columns of A from a subset $J \subset \langle n\rangle$ of cardinality $n - m$. Then $\rho(B) \le \rho(A)$. If A is irreducible then $\rho(B) < \rho(A)$.*

Proof. 1. Suppose first that A is irreducible. Then Lemma 6.2.4 yields that $\rho(A) \ge \rho(C)$. Equality holds if and only if $A = C$. Suppose next that C is irreducible. Then A is irreducible. Hence $\rho(A) \ge \rho(C)$, and equality holds if and only if $C = A$.

Assume now that A is reducible. Let $A(\varepsilon), C(\varepsilon)$ be defined as in the proof of Theorem 6.4.1. For $\varepsilon > 0$ the above arguments show that $\rho(A(\varepsilon)) \ge \rho(C(\varepsilon))$. Letting $\epsilon \searrow 0$ we deduce that $\rho(A) \ge \rho(C)$.

2. By considering a matrix $A_1 = PAP^\top$ for a corresponding $P \in \mathcal{P}_n$ we may assume that $A_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, where $B = A_{11}$. Clearly, $\rho(A_1) = \rho(A)$, and A_1 irreducible if and only if A irreducible. Let $C = \begin{bmatrix} B & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{bmatrix}$. Then $A_1 \ge C$. 2. yields that $\rho(A) = \rho(A_1) \ge \rho(C)$. Suppose that A_1 is irreducible. Since C is reducible, $A_1 \ne C$. Hence $\rho(C) < \rho(A_1) = \rho(A)$. □

Lemma 6.4.3 *Let $A \in \mathbb{R}_+^{n\times n}$. Assume that $t > \rho(A)$. Then $(tI - A)^{-1} \ge 0$. Furthermore, $(tI - A)^{-1} > 0$ if and only if A is irreducible.*

Proof. Since $t > \rho(A)$ it follows that $\det\,(tI - A) \ne 0$. (Actually, $\det\,(tI - A) > 0$. See Problem 1.) So $(tI - A)^{-1}$ exists and $(tI - A)^{-1} = \frac{1}{t}(I - \frac{1}{t}A)^{-1}$. Since $\rho(\frac{1}{t}A) < 1$ we deduce the *Neumann expansion* [Neu77], which holds for bounded operators in Banach spaces,

$$(tI - A)^{-1} = \sum_{k=0}^{\infty} \frac{1}{t^{k+1}} A^k, \quad \text{for } |t| > \rho(A). \tag{6.4.2}$$

Since $A^k \ge 0$ we deduce that $(tI - A)^{-1} \ge 0$. Let $A^k = [a_{ij}^{(k)}]$. The the (i, j) entry of $(tI - A)^{-1}$ is positive, if and only if $a_{ij}^{(k)} > 0$ for some $k = k(i, j)$.

This shows that $(tI - A)^{-1} > 0$ if and only if A is primitive. $\Box$

Theorem 6.4.4 *Let $A \in \mathbb{R}_+^{n\times n}$ be a nonnegative matrix. Then there exists a permutation matrix $P \in \mathcal{P}_n$ such that $B = PAP^\top$ is of the following block upper triangular form.*

(6.4.3)
$$B = \begin{bmatrix} B_{11} & B_{12} & \ldots & B_{1t} & B_{1(t+1)} & B_{1(t+2)} & \ldots & B_{1(t+f)} \\ 0 & B_{22} & \ldots & B_{2t} & B_{2(t+1)} & B_{1(t+2)} & \ldots & B_{2(t+f)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & B_{tt} & B_{t(t+1)} & B_{t(t+2)} & \ldots & B_{t(t+f)} \\ 0 & 0 & \ldots & 0 & B_{(t+1)(t+1)} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \vdots & 0 & 0 & B_{(t+f)(t+f)} \end{bmatrix},$$
$$B_{ij} \in \mathbb{R}^{n_i \times n_j},\ i,j = 1,\ldots,t+f,\ n_1 + \ldots + n_{t+f} = n,\ t \geq 0,\ f \geq 1.$$

Each B_{ii} is irreducible, and the submatrix $B' := [B_{ij}]_{i=j=t+1}^{t+f}$ is block diagonal. If $t = 0$ then B is a block diagonal. If $t \geq 1$ then for each $i = 1,\ldots,t$ not all the matrices $B_{i(i+1)},\ldots,B_{i(i+f)}$ are zero matrices.

Proof. Let $D_r = (W,F)$ be the reduced graph of $D(A) = (\langle n\rangle, E)$. Then D_r is a diforest. Let $\ell \geq 1$ be the length of the longest path in the digraph D_r. For a given vertex $w \in \mathbb{W}$ let $\ell(w)$ be the length of the longest path in D_r from w. So $\ell(w) \in [0,\ell]$. For $j \in \{0,\ldots,\ell\}$ denote by W_j the set of of all vertices in W such that $\ell(w) = j$. Since D_r is diforest, it follows that $W_\ell,\ldots,W_0$ is a decomposition of W to nonempty set. Note if there is a diedge in D_r from W_i to W_j then $i > j$. Also we have always at least one diedge from W_i to W_{i-1} for $i = \ell,\ldots,1$, if $\ell > 0$.

Assume that $\#W_j = m_{1+\ell-j}$ for $j = 0,\ldots,\ell$. Let $M_0 = 0$ and $M_j = \sum_{i=1}^j m_i$ for $j = 1,\ldots,\ell$. Then we name the vertices of W_j as $\{M_{\ell-j} + 1,\ldots,M_{\ell-j} + m_{1+\ell-j}\}$ for $j = 0,\ldots,\ell$. Let $f := \#W_0 = m_{\ell+1}$ and $t := \#(\cup_{j=1}^{\ell} W_j) = \sum_{j=1}^{\ell} m_i$. Note that $f \geq 1$ and $t = 0$ if and only if $\ell = 0$. Hence the representation matrix $A(D_r)$ is strictly upper triangular. Furthermore the last f rows of $A(D_r)$ are zero rows.

Recall that each vertex in W corresponds to a maximal strongly connected component of $D(A)$. That is, to each $i \in W = \langle t+f\rangle$ one has a nonempty subset $V_i \subset \langle n\rangle$, which correspond to the maximal connected component of $D(A)$. Let $n_i := \#V_i$ for $i = 1,\ldots,t+f$. Let $N_0 = 0$ and $N_i = \sum_{j=1}^i n_i, i = 1,\ldots,t+s$. Rename vertices of $D(A)$ to satisfy

$V_i = \{N_{i-1}+1, \ldots, N_{i-1}+n_i\}$. Then $PAP^\top$ is of the form (6.4.3). Furthermore, the digraph induced by B_{ii} is a strongly connected component of $D(A)$. Hence B_{ii} is irreducible. Note $B_{ij} = 0$, for $j > i$ if and only if there is no biedge from the vertex i to the vertex j in D_r. Recall that for $i \le t$, the vertex i represents a vertex in W_k, for some $k \ge 1$. Hence, for $i \le t$ there exists $j > i$ such that $B_{ij} \ne 0$. □

Theorem 6.4.5 *Let $A \in \mathbb{R}_+^{n\times n}$. Then there exists a positive eigenvector $\mathbf{x} > 0$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$ if and only if the following conditions hold. Let B be the Frobenius normal form of A given in Theorem 6.4.4. Then*

1. *$\rho(B_{(t+1)(t+1)}) = \ldots = \rho(B_{(t+f)(t+f)})$;*

2. *$\rho(B_{ii}) < \rho(B_{(t+1)(t+1)})$ for $i = 1, \ldots, t$.*

Proof. Clearly, A has a positive eigenvector corresponding to $\rho(A)$, if and only if B has a positive eigenvector corresponding to $\rho(B) = \rho(A)$. Thus we may assume that $A = B$. Suppose first that $B\mathbf{x} = \rho(B)\mathbf{x}$ for $\mathbf{x} > \mathbf{0}$. Let $\mathbf{x}^\top = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_{t+f}^\top)$, where $\mathbf{0} < \mathbf{u}_i \in \mathbb{R}^{n_i}$ for $i = 1, \ldots, t+f$. Since $B' = [B_{ij}]_{i=j=t+1}^{t+f}$ is a block diagonal matrix we deduce that $B_{ii}\mathbf{u}_i = \rho(B)\mathbf{u}_i$ for $i = t+1, \ldots, t+f$. Proposition 6.2.7 yields the equality $\rho(B) = \rho(B_{(t+1)(t+1)}) = \ldots = \rho(B_{(t+f)(t+f)})$. Hence 1 holds. Furthermore,

$$B_{ii}\mathbf{u}_i + \sum_{j=i+1}^{t+f} B_{ij}\mathbf{u}_j = \rho(B)\mathbf{u}_i, \quad i = 1, \ldots, t. \tag{6.4.4}$$

Since for each $i \in [1,t]$ there exists an integer $j(i) \in [i+1, t+f]$ such that $B_{ij} \gneq 0$ we deduce that $B_{ii}\mathbf{u}_i \lneq \rho(B)\mathbf{u}_i$ for each $i \in [1,t]$. Use Problem 11 to deduce that $\rho(B_{ii}) \le \rho(B) = \rho(B_{(t+1)(t+1)})$ for $i \in [1,t]$.

Assume now that 1 and 2 holds. Let $r = \rho(B_{(t+1)(t+1)}) = \ldots = \rho(B_{(t+f)(t+f)})$. Then $B_{ii}\mathbf{u}_i = r\mathbf{u}_i, \mathbf{u}_i > \mathbf{0}$ for $i = t+1, \ldots, t+f$. Also, since $\rho(B) = \max_{i\in\langle t+f\rangle} \rho(B_{ii})$, we deduce that $\rho(B) = r$.

Consider the equality (6.4.4) for $i = t$. Rewrite it as

$$\mathbf{v}_t = \sum_{j=i+1} B_{tj}\mathbf{u}_j = (rI - B_{tt})\mathbf{u}_i.$$

Since some $B_{tj} \gneq \mathbf{0}$ it follows that $\mathbf{v}_j \gneq \mathbf{0}$. As $r > \rho(B_{tt})$ and B_{tt} is irreducible, Lemma 6.4.3 implies that $(rI - B_{tt})^{-1} > \mathbf{0}$. Hence $\mathbf{u}_t := (rI - B_{tt})^{-1}\mathbf{v}_t > \mathbf{0}$. Thus we showed that there exists $\mathbf{u}_t > 0$ so that equality (6.4.4) holds for $i = t$. Suppose we already showed that there exists $\mathbf{u}_t, \ldots, \mathbf{u}_k > \mathbf{0}$ such that (6.4.4) holds for $i = t, t-1, \ldots, k$. Consider the equality (6.4.4) for $i = k - 1$. Viewing this equality as a system of equations in $\mathbf{u}_{k-1}$, the above arguments show the existence of unique solution $\mathbf{u}_{k-1} > \mathbf{0}$. Let $\mathbf{x}^\top = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_{t+f}^\top)$. Then $B\mathbf{x} = r\mathbf{x}$. $\square$

Corollary 6.4.6 *Let $A \in \mathbb{R}_+^{n\times n}$ and assume that $A\mathbf{x} = \rho(A)\mathbf{x}, A^\top\mathbf{y} = \rho(A)\mathbf{y}$, where $\mathbf{x}, \mathbf{y} > \mathbf{0}$. Then the Frobenius normal form of A, given by (6.4.3) is block diagonal.*

Theorem 6.4.7 *Let $A \in \mathbb{R}_+^{n\times n}$. Assume that $A\mathbf{x} = \mathbf{x}$ for some $\mathbf{x} > 0$. $B = [B_{ij}]_{i=j=1}^{t+f}$ be the Frobenius normal form of A given by (6.4.3). Denote $B^k = [B_{ij}^{(k)}]_{i=j=1}^{t+f}$ for $k = 1, 2, \ldots$. Then the block matrix form of B^k is of the form (6.4.3). Furthermore, the following conditions hold.*

1. $\lim_{k\to\infty} B_{ij}^{(k)} = 0$ *for $i, j = 1, \ldots, t$.*
2. *$A^k, k = 1, 2, \ldots,$ converge to a limit if and only if the matrices B_{ii} are primitive for $i = t+1, \ldots, t+f$.*
3. *Assume that B_{ii} are primitive and $B_{ii}\mathbf{u}_i = \mathbf{u}_i, B_{ii}^\top\mathbf{v}_i = \mathbf{v}_i, \mathbf{u}_i, \mathbf{v}_i > \mathbf{0}, \mathbf{v}_i^\top\mathbf{u}_i = 1$ for $i = t+1, \ldots, t+f$. Then $\lim_{k\to\infty} B^k = E = [E_{ij}]_{i=j=1}^{f+t} \geq 0$, where E has the block matrix form (6.4.3). Furthermore*
   - (a) *E is a nonnegative projection, i.e $E^2 = E$.*
   - (b) *$E_{ii} = \mathbf{u}_i\mathbf{v}_i^\top$ for $i = t+1, \ldots, t+f$, and $E_{ij} = 0$ for $i, j = 1, \ldots, t$.*
   - (c) *For each $i \in \langle t\rangle$ and a given row r in matrices $E_{i(t+1)}, \ldots, E_{i(t+f)}$, there exists $j > t$, $j = j(r)$, such that E_{ij} has a positive element in row j.*

Proof. Since B is a block upper triangular, it follows that B^k is block upper triangular. Since $B_{ij} = 0$ for $j > i > t$ if follows that $B_{ij}^{(k)} = 0$ for $j > i > t$. (One can prove it by induction on k.) Let $\hat{B} := [B_{ij}]_{i=j=1}^{t}$. Since B is block upper triangular it follows that $\hat{B}^k = [B_{ij}^{(k)}]_{i=j=1}^{t}$. Furthermore, $\rho(\hat{B}) = \max_{i\in\langle t\rangle} \rho(B_{ii})$. As B has a positive eigenvector, we deduce from Theorem 6.4.5 $\rho(B_{ii}) < \rho(B_{(t+1)(t+1)}) = 1$ for $i \in \langle t\rangle$. Hence $\rho(\hat{B}) < 1$. Therefore $\lim_{k\to\infty} \hat{B}^k = 0$. This implies 1.

Clearly, $A^k, k = 1, 2, \ldots$ converges if and only if the sequence $B^k, k = 1, 2, \ldots$ converges. Assume that the second sequence converges. As $B_{ii}^{(k)} = B_{ii}^k$ for $k = t+1, \ldots, t+f$, we deduce that the sequences $B_{ii}^k, k = 1, 2, \ldots,$ converge for $i = [t+1, t+f]$. Since B_{ii} is irreducible and $\rho(B_{ii}) = 1$ for

$i \in [t+1, t+f]$, the convergence of $B_{ii}^k, k = 1, 2, \ldots$ implies that the only eigenvalue of B_{ii} on unit circle is 1. (Recall that the eigenvalues of B_{ii}^k are the k-powers of the eigenvalues of B_{ii}.) So each B_{ii} is primitive.

Assume now that each B_{ii} is primitive. Hence, the algebraic multiplicity of the eigenvalue of 1 of B is f. We claim that the geometric multiplicity of 1 is also f. Let $\mathbf{u}_{t+1}, \ldots, \mathbf{u}_{t+f}$ be defined as in 3. For any $a_{t+1}, \ldots, a_{t+f} > 0$ we have that $B_{ii}(a_i\mathbf{u}_i) = a_i\mathbf{u}_i$ for $i = t+1, \ldots, t+f$. From the proof of Theorem 6.4.5 it follows that B has a positive eigenvector $\mathbf{x}$, $B\mathbf{x} = \mathbf{x}$, such that $\mathbf{x}^\top = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_t^\top, a_{t+1}\mathbf{u}_{t+1}^\top, \ldots, a_{t+f}\mathbf{u}_{t+f}), \mathbf{x}_i \in \mathbb{R}_+^{n_i}, i = 1, \ldots, t$. Hence the subspace of eigenvectors of B corresponding to the eigenvalue 1 has dimension f at least f. Since the algebraic multiplicity of 1 is f it follows that the geometric multiplicity of 1 is f. As all other eigenvalues λ of B satisfy $|\lambda| < 1$ Fact **??** yields that $\lim_{k\to\infty} B^k = E$. This implies 2.

Since B^k has the same block upper triangular form as B it follows that $E = [E_{ij}]_{i=j=1}^{t+f}$ has the block triangular form. So $E_{ij} = 0$ for $j > i > t$. Furthermore, 1 implies that $E_{ij} = 0$ for $i, j \in \langle t \rangle$. Let $\mathbf{u}_i, \mathbf{v}_i, i = t+1, \ldots, t+f$ be defined as in 3. The proof of Lemma 6.2.8 yields that $\lim_{k\to\infty} B_{ii}^k = \mathbf{v}_i\mathbf{u}_i^\top = E_{ii}$ for $i > t$. Since $B^{2k} = B^kB^k$ we deduce that $E^2 = E$. 3c will be proved later. □

Theorem 6.4.8 *Let $F \in \mathbb{R}_+^{m\times m}$ be a projection, i.e. $F^2 = F$. Then P is permutationally similar to a nonnegative projection G, i.e. $G = PFP^\top$ for some $P \in \mathcal{P}_m$, of exactly one of the following forms.*

1. *$G = 0_{m\times m}$.*

2. *$G = E$, where $n = m$ and E has a block upper triangular form given in conditions 3a-3c of Theorem 6.4.7. That is, one of the following conditions hold.*

   (a) *$E = T$, where $T = \operatorname{diag}(\mathbf{u}_1\mathbf{v}_1^\top, \ldots, \mathbf{u}_t\mathbf{v}_t^\top)$, where $0 < \mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}_+^{m_i}, \mathbf{v}_i^\top\mathbf{u}_i = 1, i = 1, \ldots, t$.*

   (b) *$E = \begin{bmatrix} 0 & R \\ 0 & T \end{bmatrix}$, where T is of the form given in (2a), and each k-row of R is of the form $(r_{k1}\mathbf{v}_1^\top, \ldots, r_{kt}\mathbf{v}_k^\top)$, where $(r_{k1}, \ldots, r_{kt}) \gneq \mathbf{0}^\top$.*

3. *$G = \begin{bmatrix} E & H \\ 0 & 0 \end{bmatrix}$, where $E \in \mathbb{R}_+^{n\times n}$is of the form described in 2, where $1 \le n < m$. So each column of H either a zero column, or a nonzero nonnegative eigenvector of E corresponding to the eigenvalue 1.*

Proof. Recall that spec F $\subset \{0,1\}$ and F is similar to a diagonal matrix. Suppose that spec F $= \{0\}$. Then $F = 0$ and the condition 1 holds. Assume now that $1 \in$ spec F. So there exists $\mathbf{x} \gneq \mathbf{0}$ such that $F\mathbf{x} = \mathbf{x}$. Since $F(F\mathbf{1}) = (F\mathbf{1})$ it follows that $\mathbf{x} = F\mathbf{1}$ is an eigenvector of F corresponding to 1 with the maximal number of nonzero coordinates. Suppose first that $\mathbf{x}$ has no zero coordinates, i.e. F does not have zero rows. Let B be the Frobenius normal form of F as in Theorem 6.4.7. As $B^2 = B$ we deduce that $E = B$. So 2 holds.

Assume finally that F has exactly $m - n$ zero rows. So there exists $Q \in \mathcal{P}_n$ such that $QF\mathbf{1} = (\mathbf{y}^\top, \mathbf{0}_{n-m}^\top)^\top$. Thus $QFQ^\top = \left[\begin{array}{cc} F_1 & H_1 \\ 0 & 0 \end{array}\right]$, where $F_1^2 = F_1$ and $F_1\mathbf{y} = \mathbf{y}, \mathbf{y} > 0$. Use 2 for F_1 to deduce 3. □

Theorem 6.4.9 *Let $A = \mathbb{R}_+^{n\times n}$. Then*

$$\rho(A) = \limsup_{m\to\infty} (\operatorname{tr} A^m)^{\frac{1}{m}}. \tag{6.4.5}$$

(Here $\operatorname{tr} B$ *is the trace of a square matrix B, i.e. the sum of its diagonal entries.)*

Proof. Clearly, for any $B \in \mathbb{C}^{n\times n}$, $|\operatorname{tr} B| = |\sum_{i=1}^n \lambda_i(B)|$. Hence $|\operatorname{tr} B| \le n\rho(B)$. Therefore, $\operatorname{tr} A^m = |\operatorname{tr} A^m| \le n\rho(A^m) = n\rho(A)^m$. Thus, $(\operatorname{tr} A^m)^{\frac{1}{m}} \le n^{\frac{1}{m}}\rho(A)$. Therefore, $\limsup_{m\to\infty}(\operatorname{tr} A^m)^{\frac{1}{m}} \le \rho(A)$. It is left to show the opposite inequality.

Assume first that A is an irreducible and primitive. Let $A\mathbf{u} = \rho(A)\mathbf{u}, A^\top\mathbf{v} = \rho(A)\mathbf{v}, \mathbf{0} < \mathbf{u}, \mathbf{v}, \mathbf{v}^\top\mathbf{u} = 1$. Theorem 6.4.7 yields that $\lim_{m\to\infty} \frac{1}{\rho(A)^m} A^m = \mathbf{u}\mathbf{v}^\top$. Hence

$$\operatorname{tr} A^m \ge \rho(A)^m \frac{1}{2} \operatorname{tr} \mathbf{u}\mathbf{v}^\top = \frac{\rho(A)^m}{2} \Rightarrow \lim_{m\to\infty} (\operatorname{tr} A^m)^{\frac{1}{m}} = \rho(A).$$

Assume that A is an irreducible and imprimitive. If A 1×1 zero matrix, then (6.4.5) trivially holds. Assume that $n > 1$. Without loss of generality we can assume that A is of the form given in Theorem 6.2.1 part 5d. Then $A^h = \operatorname{diag}(B_1, \ldots, B_h)$, where each B_j is primitive and $\rho(B_j) = \rho(A)^h$, see Lemma 6.2.9. So $\operatorname{tr} A^{hk} = \sum_{j=1}^h \operatorname{tr} B_j^k$. Since each B_j is primitive and irreducible, we deduce from the previous arguments that $\lim_{k\to\infty}(\operatorname{tr} A^{hk})^{\frac{1}{hk}} = \rho(A)$. Hence (6.4.5) holds in this case too.

Assume now that A is not irreducible. Without loss of generality we can assume that A is in the Frobenius form (6.4.3). Then there exists $i \in \langle t+f \rangle$

such that $\rho(A) = \rho(B_{ii})$. Clearly

$$(\operatorname{tr} A^m)^{\frac{1}{m}} = (\sum_{j=1}^{t+f} \operatorname{tr} B_{jj}^m)^{\frac{1}{m}} \geq (\operatorname{tr} B_{ii}^m)^{\frac{1}{m}} \Rightarrow \limsup_{m\to\infty} (\operatorname{tr} A^m)^{\frac{1}{m}} \geq \limsup_{m\to\infty} (\operatorname{tr} B_{ii}^m)^{\frac{1}{m}} = \rho(A).$$

□

Problems

1. Let $A \in \mathbb{R}^{n\times n}$. Show that for $t > \rho(A)$ det $(tI - A) > 0$.

6.5 Stochastic matrices and Markov Chains

Definition 6.5.1 *A matrix S is a called a stochastic matrix if $S \in \mathbb{R}_+^{n\times n}$, for some integer $n \geq 1$, and $S\mathbf{1} = \mathbf{1}$. Denote by $\mathcal{S}_n \subset \mathbb{R}_+^{n\times n}$ the set of $n \times n$ stochastic matrices. A matrix A is called doubly stochastic if A and $A^\top$ is a stochastic matrix. Denote by $\Omega_n \subset \mathcal{S}_n$ the set of $n \times n$ doubly stochastic matrices.*

Note that $A \in \mathbb{R}^{n\times n}$ is a stochastic matrix, if and only if each row of A is a probability vector. Furthermore the $\mathcal{S}_n$ and Ω_n are compact semigroups with respect to the product of matrices, see Problem 1. The following lemma is straightforward, see Problem 2.

Lemma 6.5.2 *Let $A \in \mathbb{R}_+^{n\times n}$. Then $A = DSD^{-1}$ for some $S \in \mathcal{S}_n$ and a diagonal matrix $D \in \mathbb{R}_+^{n\times n}$ with positive diagonal entries if and only if $A\mathbf{x} = \mathbf{x}$ for some positive $\mathbf{x} \in \mathbb{R}^n$.*

Definition 6.5.3 *Let $S \in \mathcal{S}_n$ be irreducible. We will assume the normalization that the eigenvector of S and $S^\top$ corresponding to the eigenvalue 1 are of the form $\mathbf{1} = (\mathbf{1}_n) \in \mathbb{R}_+^n$ and $\pi \in \Pi_n$, respectively, unless stated otherwise. S is called aperiodic if it is primitive, and periodic if it is imprimitive.*

Theorem 6.5.4 *Let $A \in \mathcal{S}_n$. Denote by $B = [B_{ij}]_{i=j=1}^{t+f}$ the Frobenius normal form of A given by (6.4.3). Then $B, B_{(t+1)(t+1)}, \ldots, B_{(t+f)(t+f)}$ are stochastic. Furthermore the conditions 1-3b of Theorem 6.4.7 hold. The limit matrix E is stochastic. In the condition 3 we can assume that $\mathbf{u}_i = \mathbf{1}_{n_i}, \mathbf{v}_i \in \mathit{\Pi}_{n_i}$ for $i = t+1, \ldots, t+f$.*

Finally the condition 3c of Theorem 6.4.7 is replaced by the following stronger condition. For each $i \in \langle t\rangle$ the sum of the entries of a row r in matrices $E_{i(t+1)}, \ldots, E_{i(t+f)}$ is 1, for any given row r.

Proof. Since A is stochastic, in view of Problem 1b, B is stochastic. Since B is block upper triangular and $B' = [B_{ij}]_{i=j=t+1}^{t+f}$ is block diagonal, it follows that B_{ii} is stochastic for $i = t+1, \ldots, t+f$. Since B^k is stochastic, Problem 1c implies that $E = [E_{ij}]_{i=j=1}^{t+f}$ is stochastic. Since $\hat{E} = [E_{ij}]_{i=j=1}^{t} = 0$ we deduce that the last part of the theorem. □

Proof of condition 3c **of Theorem** 6.4.7. In view of Lemma 6.5.2 B is diagonally similar to a stochastic matrix,. Hence 3c follows from the last part of Theorem 6.5.4. □

We now recall the classical connection between the stochastic matrices and Markov chains. To each probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top \in \Pi_n$ we associated a random variable X, which takes values in the set $\langle n \rangle$, such that $\mathrm{P}(X = i) = \pi_i$ for $i = 1, \ldots, n$. Then $\pi = \pi(X)$ is called the *distribution* of X.

Assume that we are given a sequence of random variables $X_0, X_1, \ldots$ each taking values in the set $\langle n \rangle$. Let $s_{ij}^{(k)}$ be the conditional probability of $X_k = j$ given that $X_{k-1} = i$:

$$s_{ij}^{(k)} := \mathrm{P}(X_k = j | X_{k-1} = i), \quad i, j = 1, \ldots, n, \quad k = 1, \ldots \tag{6.5.1}$$

Clearly, $S_k = [s_{ij}^{(k)}]_{i=j=1}^{n}, k = 1, 2, \ldots$ is a stochastic matrix for $k = 1, \ldots$.

Definition 6.5.5 *Let $X_0, X_1, \ldots,$ be a sequence of random variables taking values in $\langle n \rangle$. Then*

1. *$X_0, X_1, \ldots$ is called a homogeneous Markov chain if*

   $\mathrm{P}(X_k = j_k | X_{k-1} = j_{k-1}, \ldots, X_0 = j_0) = \mathrm{P}(X_1 = j_k | X_0 = j_{k-1})$ *for* $k = 1, 2, \ldots.$

2. *$X_0, X_1, \ldots$ is called a nonhomogeneous Markov chain if*

   $\mathrm{P}(X_k = j_k | X_{k-1} = j_{k-1}, \ldots, X_0 = j_0) = \mathrm{P}(X_k = j_k | X_{k-1} = j_{k-1})$ *for* $k = 1, 2, \ldots.$

   (Note that a homogeneous Markov chain is a special case of nonhomogeneous Markov chain.)

3. *A nonhomogeneous Markov chain is said to have a limiting distribution if the limit $\boldsymbol{\pi}_\infty(\boldsymbol{\pi}_0) := lim_{k \to \infty} \boldsymbol{\pi}_k$ exists. If $\boldsymbol{\pi}_\infty$ does not depend on $\boldsymbol{\pi}_0$ then $\boldsymbol{\pi}_\infty$ is called the stationary distribution of the Markov chain.*

The following lemma is straightforward, see Problem **??**.

Lemma 6.5.6 *Let $X_0, X_1, \ldots$ be a sequence of random variables taking values in $\langle n \rangle$. Let $\boldsymbol{\pi}_k := \boldsymbol{\pi}(X_k)$ be the distribution of X_k for $k = 0, 1, \ldots$. If $X_0, X_1, \ldots,$ is a nonhomogeneous Markov chain then $\boldsymbol{\pi}_k^\top = \boldsymbol{\pi}_0^\top S_1 \ldots S_k$ for $k = 1, \ldots,$ where S_k are defined by (6.5.1). In particular, if $X_0, X_1, \ldots,$ is a homogeneous Markov process, i.e. $S_k = S, k = 1, 2, \ldots,$ then $\boldsymbol{\pi}_k^\top = \boldsymbol{\pi}_0^\top S^k$ for $k = 1, \ldots, .$*

Theorem 6.5.7 *Let $X_0, X_1, \ldots,$ be a homogeneous Markov chain on $\langle n \rangle$, given by a stochastic matrix $S = [s_{ij}] \in \mathcal{S}_n$. Let $D(S)$ and $D_r(S)$ be the digraph and the reduced digraph corresponding to S. Label the vertices of the reduced graph $D_r(S)$ by $\{1, \ldots, t+f\}$. Let $V_1, \ldots, V_{t+f}$ be the decomposition of $\langle n \rangle$ to the strongly connected components of the digraph $D(S)$. Assume that $B = PSP^\top, P \in \mathcal{P}_n$ is given by the form (6.4.3). The vertices, (states), in $\cup_{i=1}^t V_i$ are called the transient vertices, (states). (Note that if $t = 0$ then no transient vertices exist.) The vertices, (states), in $\cup_{i=t+1}^{t+f} V_i$ are called the final vertices, (states). $V_{t+1}, \ldots, V_{t+f}$ are called the final strongly connected components. Furthermore the following conditions hold.*

1. *For each $i \in \cup_{j=1}^t V_j$ $\lim_{k\to\infty} \mathrm{P}(X_k = i) = 0$.*

2. *$X_0, X_1, \ldots,$ have a limiting distribution if and only if each stochastic matrix corresponding to V_i is aperiodic for $i = t+1, \ldots, t+f$. I.e. the irreducible matrices B_{ii} are primitive for $i = t+1, \ldots, t+f$.*

3. *$X_0, X_1, \ldots$ have a stationary distribution if and only if $f = 1$. I.e. there exists only one final strongly connected component.*

Proof. . Without loss of generality we may assume that $S = B$. Let $\boldsymbol{\pi}_k^\top = (\boldsymbol{\pi}_{1,k}^\top, \ldots, \boldsymbol{\pi}_{t+f,k}^\top)$ for $k = 0, 1, \ldots$. From the proof of Theorem 6.4.7 we deduce that

$$\boldsymbol{\pi}_{i,k}^\top = \sum_{j=1}^{i} \boldsymbol{\pi}_{j,0} B_{ji}^{(k)}, \text{ for } i = 1, \ldots, t.$$

In view of part 1 of Theorem 6.4.7 we deduce that $\lim_{k\to\infty} \boldsymbol{\pi}_{i,k} = \mathbf{0}$ for $i = 1, \ldots, t$. This proves part 1.

Suppose that $\boldsymbol{\pi}_0 = (\pi_1, 0, \ldots,$ is supported only on V_{t+i} for some $i \in \langle f \rangle$. That is $\pi_{j,0} = 0$ if $j \notin V_{t+i}$. Then each $\boldsymbol{\pi}_k$ is supported on V_{t+i}. Furthermore, $\boldsymbol{\pi}_{t+i,k}^\top = \boldsymbol{\pi}_{t+i,0}^\top B_{(t+i)(t+i)}^k$. Assume that $B_{(t+i)(t+i)}$ is imprimitive. Choose $\boldsymbol{\pi}_{t+i,0}$ to have one nonzero coordinate to be 1 and all other coordinates to be zero. Assuming that $B_{(t+i)(t+i)}$ has the form given in 5d of Theorem 6.2.1, we see that there no is limit distribution. Hence, to have the limit distribution for each $\boldsymbol{\pi}_0$ we must assume that $B_{(t+i)(t+i)}$ is primitive for $i = 1, \ldots, f$.

Assume that $B_{(t+i)(t+i)}$ is primitive for $i = 1, \ldots, f$. Then Theorem 6.5.4 implies that $\boldsymbol{\pi}_\infty^\top = \boldsymbol{\pi}_0 E$, where $E = \lim_{k\to\infty} B^k$ is the stochastic projection. As we pointed out before if we assume that $\boldsymbol{\pi}_0$ is supported only on V_{t+i} then the limit probability is also supported on V_{t+i}. Hence, to have a stationary distribution we must have that $f = 1$.

Assume that $f = 1$. Then $\lim_{k\to\infty} B_{t+1}^k = \mathbf{1}_{t+1}\boldsymbol{\pi}_{t+1}^\top$. Observe that the limit probability is $\boldsymbol{\pi}_0^\top E = (\boldsymbol{\pi}_0^\top E)E$. Since $\boldsymbol{\pi}_0^\top E$ is supported only on V_{t+1} it follows that $\boldsymbol{\pi}_0^\top E^2 = (\underbrace{\mathbf{0}^\top, \ldots, \mathbf{0}^\top}_{t}, \boldsymbol{\pi}_{t+1}^\top)$, which is independent of $\boldsymbol{\pi}_0$. □

The proof of the above theorem yields the well known result.

Corollary 6.5.8 *Let $X_0, X_1, \ldots,$ be a homogeneous Markov chain on $\langle n\rangle$, given by an aperiodic stochastic matrix $S = [s_{ij}] \in \mathcal{S}_n$. Assume that $S^\top \boldsymbol{\pi} = \boldsymbol{\pi}$ for a unique $\mathbf{0} < \boldsymbol{\pi} \in \mathit{\Pi}_n$. Then this Markov process has a stationary distribution equal to $\boldsymbol{\pi}$.*

A stronger result is proven in [Fri06].

Theorem 6.5.9 *Let $X_0, X_1, \ldots,$ be a nonhomogeneous Markov chain on $\langle n\rangle$, given by the sequence of stochastic matrices $S_1, S_2, \ldots,$ defined in (6.5.1). Assume that $\lim_{k\to\infty} S_k = S$, where S is a stochastic matrix,. Suppose furthermore that the corresponding homogeneous Markov chain to S has a stationary distribution $\boldsymbol{\pi}$. Then the given nonhomogeneous Markov process has a stationary distribution equal to $\boldsymbol{\pi}$.*

We close this section with *Google's Page Ranking*. Let n be the current number of Web pages. (Currently around a few billions.) Then $S = [s_{ij}] \in \mathcal{S}_n$ is defined as follows. Let $\mathcal{A}(i) \subset \langle n\rangle$ be the set of all pages accessible from the Web page i. Assume first that i is a *dangling* Web page, i.e. $\mathcal{A}(i) = \emptyset$. Then $s_{ij} = \frac{1}{n}$ for $j = 1, \ldots, n$. Assume now that $n_i = \#\mathcal{A}(i) \ge 1$. Then $s_{ij} = \frac{1}{n_i}$ if $j \in \mathcal{A}(i)$ and otherwise $s_{ij} = 0$. Let $\mathbf{0} < \boldsymbol{\omega} \in \mathit{\Pi}_n, t \in (0, 1)$. Then the Google positive stochastic matrix is given by

$$G = tS + (1-t)\mathbf{1}\boldsymbol{\omega}^\top. \tag{6.5.2}$$

It is rumored that $t \sim 0.85$. Then the stationary distribution corresponding to G is given by $G^\top \boldsymbol{\pi} = \boldsymbol{\pi} \in \Pi_n$. The coordinates of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ constitute Google's popularity score of each Web page. I.e. if $\pi_i > \pi_j$ then Web page i is more popular than Web page j.

A reasonable choice of $\boldsymbol{\omega}$ would be the stationary distribution of yesterday Google stochastic matrix. To find the stationary distribution $\boldsymbol{\pi}$ one can iterate several times the equality

$$\boldsymbol{\pi}_k^\top = \boldsymbol{\pi}_{k-1}^\top G, \quad k = 1, \ldots, N. \tag{6.5.3}$$

Then $\boldsymbol{\pi}_N$ would be a good approximation of $\boldsymbol{\pi}$. One can choose $\boldsymbol{\pi}_0 = \boldsymbol{\omega}$.

Problems

1. Show
   (a) If $S_1, S_2 \in \mathcal{S}_n$ then $S_1S_2 \in \mathcal{S}_n$.
   (b) $P\mathcal{S}_n = \mathcal{S}_nP = \mathcal{S}_n$ for any $P \in \mathcal{P}_n$.
   (c) $\mathcal{S}_n$ is a compact set in $\mathbb{R}_+^{n\times n}$.
   (d) If $S_1, S_2 \in \Omega_n$ then $S_1S_2 \in \Omega_n$.
   (e) $P\Omega_n = \Omega_nP = \Omega_n$ for any $P \in \mathcal{P}_n$.
   (f) Ω_n is a compact set in $\mathbb{R}_+^{n\times n}$.
   (g) $\mathcal{P}_n$ is a group of doubly stochastic matrices of cardinality $n!$.
2. Prove Lemma 6.5.2.
3. Let $A, B \in \mathbb{C}^{n\times n}$, and assume that A and B are similar. I.e. $A = TBT^{-1}$ for some invertible T. Then the sequence $A^k, k = 1, 2, \ldots,$ converges if and only if $B^k, k = 1, 2, \ldots,$ converges.
4. Prove Lemma 6.5.6.

6.6 Friedland-Karlin results

Definition 6.6.1 *Let $B = [b_{ij}]_{i=j=1}^n \in \mathbb{R}^{n\times n}$. B is called a Z-matrix if $b_{ij} \le 0$ for each $i \ne j$. B is called an M-matrix if $B = rI - A$ where $A \in \mathbb{R}_+^{n\times n}$ and $r \ge \rho(A)$. For $r = \rho(A)$ B is called a singular M-matrix.*

The following result is straightforward, see Problem 1.

Lemma 6.6.2 *Let $B = [b_{ij}] \in \mathbb{R}^{n\times n}$ be a Z-matrix. Let $C = [c_{ij}] \in \mathbb{R}_+^{n\times n}$ be defined as follows. $c_{ij} = -b_{ij}$ for each $i \ne j$ and $c_{ii} = r_0 - b_{ii}$ for $i = 1, \ldots, n$, where $r_0 = \max_{i\in\langle n\rangle}$. Then $B = r_0I - C$. Furthermore, $B = rI - A$ for some $A \in \mathbb{R}_+^{n\times n}$ if and only if $r = r_0 + t, A = tI + C$ for some $t \ge 0$.*

Theorem 6.6.3 *Let $B \in \mathbb{R}^{n\times n}$ be a $Z - matrix$. Then TFAE.*

1. *B is an M-matrix.*

2. *All principal minors of B are nonnegative.*

3. *The sum of all $k \times k$ principal minors of B are nonnegative for $k = 1, \ldots, n$.*

4. *For each $t > 0$ there exists $\mathbf{0} < \mathbf{x} \in \mathbb{R}^n_+$, which may depend on t, such that $B\mathbf{x} \geq -t\mathbf{x}$.*

Proof. $1 \Rightarrow 2$. We first show that det $B \geq 0$. Let $\lambda_1(A), \ldots, \lambda_n(A)$ be the eigenvalues of A. Assume that $\lambda_i(A)$ is real. Then $r - \lambda_i(A) \geq \rho(A) - \lambda_i(A) \geq 0$. Assume that $\lambda_i(A)$ is complex. Since A is a real valued matrix, $\overline{\lambda_i(A)}$ is also an eigenvalue of A. Hence $(r - \lambda_i(A))(r - \overline{\lambda_i(A)}) = |r - \lambda_i(A)|^2 > 0$. Since det $B = \prod_{i=1}^n (r - \lambda_i(A))$, we deduce that det $B \geq 0$. Let B' be a principal submatrix of B. Then $B' = rI' - A'$, where A' is a corresponding principal submatrix of A and I' is the identity matrix of the corresponding order. Part 2 of Proposition 6.4.2 implies that $\rho(A) \geq \rho(A')$. So B' is an M-matrix, Hence det $B' \geq 0$.

$2 \Rightarrow 3$. Trivial.

$3 \Rightarrow 1$. Let det $(tI + B) = t^n + \sum_{i=k}^n \beta_k t^{n-k}$. Then β_k is the sum of all principal minors of B of order k. Hence $\beta_k \geq 0$ for $k = 1, \ldots, n$. Therefore $0 <$ det $(tI + B) =$ det $((t + r)I - A)$. Recall that det $(\rho(A)I - A) = 0$. Thus $t + r > \rho(A)$ for any $t > 0$. So $r \geq \rho(A)$, i.e. B is an M-matrix.

$1 \Rightarrow 4$. Let $t > 0$. Use the Neumann expansion (6.4.2) to deduce that $(tI + B)^{-1} \geq \frac{1}{t+r} I$. So for any $\mathbf{y} > \mathbf{0}$ $\mathbf{x} := (tI + B)^{-1}\mathbf{y} > \mathbf{0}$. So $\mathbf{y} = (tI + B)\mathbf{x} \geq \mathbf{0}$.

$4 \Rightarrow 1$. By considering $PBP^\top = rI - PAP^\top$ we may assume that $A = [A_{ij}]_{i=j=1}^{t+f}$ is in the Frobenius normal form (6.4.3). Let $B\mathbf{x} \geq -t\mathbf{x}$. Partition $\mathbf{x}^\top = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_{t+f}^\top)$. Hence $(t+r)\mathbf{x}_i \geq A_{ii}\mathbf{x}_i$. Problem 11 yields that $t + r \geq \rho(A_{ii})$ for $i = 1, \ldots, t + f$. Hence $t + r \geq \rho(A)$. Since $t > 0$ was arbitrary we deduce that $r \geq \rho(A)$. □

Corollary 6.6.4 *Let B be a Z-matrix. Assume that there exist $\mathbf{x} > \mathbf{0}$ such that $B\mathbf{x} \geq \mathbf{0}$. Then B is an M-matrix.*

Lemma 6.6.5 *Let $B = [f_{ij}]_{i=j=1}^n$ be real symmetric matrix with the eigenvalues $\lambda_1(B) \geq \ldots \geq \lambda_n(B)$. Assume that $\lambda_n(B)$ is a simple eigenvalue, i.e. $\lambda_{n-1}(B) > \lambda_n(B)$. Suppose that $B\mathbf{x} = \lambda_n(B)\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top\mathbf{x} = 1$. Let $\mathbf{U} \subset \mathbb{R}^n$ be a subspace which does not contain $\mathbf{x}$. Then*

$$\min_{\mathbf{y} \in \mathbf{U}, \mathbf{y}^\top\mathbf{y} = 1} \mathbf{y}^\top B \mathbf{y} > \lambda_n(B). \tag{6.6.1}$$

Proof. Recall the minimum characterization of $\lambda_n(B)$

$$\min_{\mathbf{z}\in\mathbb{R}^n,\mathbf{z}^\top\mathbf{z}=1} \mathbf{z}^\top B\mathbf{z} = \lambda_n(B). \tag{6.6.2}$$

See for example http://www.math.uic.edu/~friedlan/math310lec.pdf page 114. Equality holds if and only if $B\mathbf{z} = \lambda_n(B)\mathbf{z}$, where $\mathbf{z}^\top\mathbf{z} = 1$. Since $\lambda_n(B)$ is simple it follows that $\mathbf{z} = \pm\mathbf{x}$. Since $\mathbf{U}$ does not contain $\mathbf{x}$, hence it does not contain $-\mathbf{x}$, we deduce that the minimum in (6.6.1) is achieved for some $\mathbf{y}^* \neq \pm\mathbf{x}$. Hence this minimum is greater than $\lambda_n(B)$. □

Corollary 6.6.6 *Let $B \in \mathbb{R}^{n\times n}$ be an M-singular symmetric matrix of the form $B = \rho(C)I - C$, where C is a nonnegative irreducible symmetric matrix. Let $\mathbf{U} = \{\mathbf{y} \in \mathbb{R}^n, \mathbf{1}^\top\mathbf{y} = \mathbf{0}\}$. Then $\lambda_n(B) = 0$ is a simple eigenvalue, and (6.6.1) hold.*

As usual, we let $\|\mathbf{z}\| := \sqrt{\mathbf{z}^*\mathbf{z}}$ for any $\mathbf{z} \in \mathbb{C}^n$ be the Euclidean norm of $\mathbf{z}$.

Theorem 6.6.7 *Let $\mathcal{D} \subset \mathbb{R}^m$ be a bounded domain. ($\mathcal{D}$ is open and connected, $\partial\mathcal{D}$, the boundary of $\mathcal{D}$, is a compact set, so $\mathcal{D}\cup\partial\mathcal{D}$ is a compact set in $\mathbb{R}^m$.) Let $f \in \mathcal{D} \to \mathbb{R}$ be $\mathrm{C}^2(\mathcal{D})$. i.e. the function and is derivatives up the second order are continuous. Suppose that $f|\partial\mathcal{D} = \infty$, i.e. for each sequence $\mathbf{x}_i \in \mathcal{D}, i = 1,\ldots,$ such that $\lim_{i\to\infty}\mathbf{x}_i = \mathbf{x} \in \partial\mathcal{D}$, $\lim_{i\to\infty} f(\mathbf{x}_i) = \infty$. Assume furthermore, that for each critical point $\boldsymbol{\xi} \in$, i.e. $\nabla f(\boldsymbol{\xi}) = (\frac{\partial f}{\partial x_1}(\boldsymbol{\xi}),\ldots,\frac{\partial f}{\partial x_m}(\boldsymbol{\xi}))^\top = \mathbf{0}$, the eigenvalues of the Hessian $H(\boldsymbol{\xi}) = [\frac{\partial^2 f}{\partial x_i\partial x_j}(\boldsymbol{\xi})]_{i=j=1}^n$ are positive. Then f has a unique critical point $\boldsymbol{\xi} \in \mathcal{D}$, which is a global minimum, i.e $f(\mathbf{x}) > f(\boldsymbol{\xi})$ for any $\mathbf{x} \in \mathcal{D}\backslash\{\boldsymbol{\xi}\}$.*

Proof. Consider the negative gradient flow

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f(\mathbf{x}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathcal{D}. \tag{6.6.3}$$

Clearly, the fixed points of this flow are the critical points of f. Observe next that if $\mathbf{x}_0$ is not a critical point then $f(\mathbf{x}(t))$ decreases, as $\frac{df(\mathbf{x}(t)}{dt} = -\|\nabla f(\mathbf{x}(t))\|^2$. Since $f|\partial\mathcal{D} = \infty$, we deduce that all accumulations points of the flow $\mathbf{x}(t), t \in [t_0,\infty)$ are in $\mathcal{D}$, and are critical points of f. Consider the flow (6.6.3) in the neighborhood of a critical point $\boldsymbol{\xi} \in \mathcal{D}$. Let $\mathbf{x} = \mathbf{y} + \boldsymbol{\xi}, \mathbf{x}_0 = \mathbf{y}_0 + \boldsymbol{\xi}$. The for $\mathbf{x}_0$ close to $\boldsymbol{\xi}$ the flow (6.6.3) is of the form

$$\frac{d\mathbf{y}(t)}{dt} = -(H(\boldsymbol{\xi})\mathbf{y} + \mathrm{Er}(\mathbf{y})), \quad \mathbf{y}(t_0) = \mathbf{y}_0.$$

For a given $\delta >$ the exists $\varepsilon = \varepsilon(\delta) > 0$ such that for $\|\mathbf{y}\| < \varepsilon$, $\|\mathrm{Er}(\mathbf{y})\| < \delta\|\mathbf{y}\|$. Let $\alpha > 0$ be the smallest positive eigenvalue of $H(\boldsymbol{\xi})$. So $\mathbf{z}^\top H(\boldsymbol{\xi})\mathbf{z} \geq$

$\alpha\|\mathbf{z}\|^2$ for any $\mathbf{z} \in \mathbb{R}^m$. Choose $\varepsilon > 0$ so that $\|\mathrm{Er}(\mathbf{y}\| < \frac{\alpha}{2}\|\mathbf{y}\|$ for $\|\mathbf{y}\| < \varepsilon$. Thus

$$\frac{d\|\mathbf{y}(t)\|^2}{dt} = -2(\mathbf{y}(t)^\top H(\boldsymbol{\xi})\mathbf{y}(t) + \mathbf{y}(t)^\top \mathrm{Er}(\mathbf{y}))| \le -\alpha|\mathbf{y}\|^2 \text{ if } |\mathbf{y}(t)| < \varepsilon.$$

This shows that if $\|\mathbf{y}(t_o)\| < \varepsilon$ then for $t \ge t_0$ $\|\mathbf{y}(t)\|$ decreases. Moreover

$$\frac{d\log\|\mathbf{y}(t)\|^2}{dt} \le -\alpha \text{ for } t \ge t_0 \Rightarrow \|\mathbf{y}(t)\|^2 \le \|\mathbf{y}_o\|^2 e^{-\alpha(t-t_o)} \text{ for } t \ge t_o.$$

This shows that $\lim_{t\to\infty}\mathbf{y}(t) = \mathbf{0}$. Let $\beta \ge \alpha$ be the maximal eigenvalue of $H(\boldsymbol{\xi})$. Similar estimates show that if $\|\mathbf{y}_o\| < \varepsilon$ then $\|\mathbf{y}(t)\|^2 \ge \|\mathbf{y}_o\|^2 e^{-(2\beta+\alpha)(t-t_o)}$.

These results, combined with the continuous dependence of the flow (6.6.3) on the initial conditions $\mathbf{x}_o$, imply the following facts. Any flow (6.6.3) which starts at a noncritical point $\mathbf{x}_o$ must terminate at $t = \infty$ at some critical point $\boldsymbol{\xi}$, which may depend on $\mathbf{x}_o$. For a critical point $\boldsymbol{\xi}$, denote by the set $\mathcal{A}(\boldsymbol{\xi})$ all points $\mathbf{x}_o$ for which the flow (6.6.3) terminates at finite or infinite time at $\boldsymbol{\xi}$. (The termination at finite time can happen only if $\mathbf{x}_o = \boldsymbol{\xi}$.) Then $\mathcal{A}(\boldsymbol{\xi})$ is an open connected set of $\mathcal{D}$.

We claim that $\mathcal{A}(\boldsymbol{\xi}) = \mathcal{D}$. If not, there exists a point $\mathbf{x}_o \in \partial\mathcal{A}(\boldsymbol{\xi}) \cap \mathcal{D}$. Since $\mathcal{A}(\boldsymbol{\xi})$ is open, $\mathbf{x}_o \notin \mathcal{A}(\boldsymbol{\xi})$. As we showed above $\mathbf{x}_o \in \mathcal{A}(\boldsymbol{\xi}')$ for some another critical point $\boldsymbol{\xi}' \ne \boldsymbol{\xi}$. Clearly $\mathcal{A}(\boldsymbol{\xi}) \cap \mathcal{A}(\boldsymbol{\xi}') = \emptyset$. As $\mathcal{A}(\boldsymbol{\xi}')$ is open there exists an open neighborhood of $\mathbf{x}_o$ in $\mathcal{D}$ which belongs to $\mathcal{A}(\boldsymbol{\xi}')$. Hence $\mathbf{x}_o$ can not be a boundary point of $\mathcal{A}(\boldsymbol{\xi})$, which contradicts our assumption. Hence $\mathcal{A}(\boldsymbol{\xi}) = \mathcal{D}$, and $\boldsymbol{\xi}$ is a unique critical point of f in $\mathcal{D}$. Hence $\boldsymbol{\xi}$ is the unique minimal point of f. □

Theorem 6.6.8 *Let $A = [a_{ij}]_{i=j=1}^n \in \mathbb{R}_+^{n\times n}$ be an irreducible matrix. Suppose furthermore that $a_{ii} > 0$ for $i = 1, \ldots, n$. Let $\mathbf{w} = (w_1, \ldots, w_n)^\top > \mathbf{0}$. Define the following function*

$$f = f_{A,\mathbf{w}} = \sum_{i=1}^n w_i \log\frac{(A\mathbf{x})_i}{x_i}, \quad \mathbf{x} = (x_1, \ldots, x_n) > \mathbf{0}. \tag{6.6.4}$$

Let $\mathcal{D}$ be the interior of Π_n, the compact set of probability vectors in $\mathbb{R}^n$. ($\mathcal{D}$ can be viewed as an open connected bounded set in $\mathbb{R}^{n-1}$, see the proof.) Then f satisfies the assumptions of Theorem 6.6.7. Let $\mathbf{0} < \boldsymbol{\xi} \in \Pi_n$ be the unique critical point of f in $\mathcal{D}$. Then $f(\mathbf{x}) \ge f(\boldsymbol{\xi})$ for any $\mathbf{x} > \mathbf{0}$. Equality holds if and only if $\mathbf{x} = t\boldsymbol{\xi}$ for some $t > 0$.

Proof. Observe that any probability vector $\mathbf{p} = (p_1, \ldots, p_n)^\top$ can be written as $\mathbf{p} = \frac{1}{n}\mathbf{1} + \mathbf{y}$ where $\mathbf{y} \in \mathbb{R}^n, \mathbf{1}^\top\mathbf{y} = 0$ and $\mathbf{y} \ge -\frac{1}{n}\mathbf{1}$. Since any $\mathbf{y} \in \mathbb{R}^n, \mathbf{1}^\top \mathbf{y} = 0$ is of the form $\mathbf{y} = u_1(-1, 1, 0 \dots, 0)^\top + \dots + u_{n-1}(-1, 0, \dots, 0, 1)^\top$ we deduce that we can view Π_n as an compact connected set in $\mathbb{R}^{n-1}$, and its interior $\mathcal{D}$, i.e. $\mathbf{0} < \mathbf{p} \in \Pi_n$, is an open connected bounded set in $\mathbb{R}^{n-1}$.

We now claim that $f|\partial\Pi_n = \infty$. Let $\mathbf{0} < \mathbf{p}_k = (p_{1,k}, \dots, p_{n,k})^\top \in \Pi_n, k = 1, \dots,$ converge to $\mathbf{p} = (p_1, \dots, p_n)^\top \in \partial\Pi_n$. Let $\emptyset \neq Z(\mathbf{p}) \subset \langle n \rangle$ be the set of vanishing coordinates of $\mathbf{p}$. Observe first that $\frac{(A\mathbf{x})_i}{x_i} \geq a_{ii} > 0$ for $i = 1, \dots, n$. Since A is irreducible, it follows that there exists $l \in Z(\mathbf{p}), j \in \langle n \rangle \backslash Z(\mathbf{p})$ such that $a_{lj} > 0$. Hence

$$\lim_{k\to\infty} \frac{(A\mathbf{p}_k)_l}{p_{l,k}} \geq \lim_{k\to\infty} \frac{a_{lj} p_{j,k}}{p_{l,k}} = \infty.$$

Thus

$$\lim_{k\to\infty} f(\mathbf{p}_k) \geq \lim_{k\to\infty} \log \frac{(A\mathbf{p}_k)_l}{p_{l,k}} + \sum_{i\neq l} \log a_{ii} = \infty.$$

Observe next that $f(\mathbf{x})$ is a homogeneous function of degree 0 on $\mathbf{x} > \mathbf{0}$, i.e. $f(t\mathbf{x}) = f(\mathbf{x})$ for all $t > 0$. Hence $\frac{df(t\mathbf{x})}{dt} = 0$. Thus

$$\mathbf{x}^\top \nabla f(\mathbf{x}) = 0 \tag{6.6.5}$$

for all $\mathbf{x} > \mathbf{0}$. Let $\boldsymbol{\xi} \in \mathcal{D}$ be a critical point of $f|\mathcal{D}$. Then $\mathbf{y}^\top \nabla f(\boldsymbol{\xi}) = 0$ for each $\mathbf{y} \in \mathbb{R}^n, \mathbf{1}^\top \mathbf{y} = 0$. Combine this fact with (6.6.5) for $\mathbf{x} = \boldsymbol{\xi}$ to deduce that $\boldsymbol{\xi}$ is a a critical point of f in $\mathbb{R}^n_+$. So $\nabla f(\boldsymbol{\xi}) = \mathbf{0}$. Differentiate (6.6.5) with respect to $x_i, i = 1, \dots, n$ and evaluate these expressions at $\mathbf{x} = \boldsymbol{\xi}$. Since $\boldsymbol{\xi}$ is a critical point we deduce that $H(\boldsymbol{\xi})\boldsymbol{\xi} = \mathbf{0}$. We claim that $H(\boldsymbol{\xi})$ is a symmetric singular M-matrix. Indeed

$$\frac{\partial f}{\partial x_j}(\mathbf{x}) = -w_j \frac{1}{x_j} + \sum_{i=1}^n w_i \frac{a_{ij}}{(A\mathbf{x})_i}. \tag{6.6.6}$$

Hence for $l \neq j$

$$\frac{\partial^2 f}{\partial x_l \partial x_j}(\mathbf{x}) = -\sum_{i=1}^n w_i \frac{a_{ij} a_{il}}{(A\mathbf{x})_i^2} \leq 0.$$

So $H(\mathbf{x})$ is a Z-matrix for any $\mathbf{x} > \mathbf{0}$. Since $H(\boldsymbol{\xi})\boldsymbol{\xi} = \mathbf{0}$ Corollary 6.6.4 yields that $H(\boldsymbol{\xi})$ is a symmetric singular M-matrix. So $H(\boldsymbol{\xi}) = \rho(C)I - C, C = [c_{ij}]_{i=j=1}^n$. We claim that C is an irreducible matrix. Indeed assume that $a_{jl} > 0$. Then

$$c_{jl} = c_{lj} = -\frac{\partial^2 f}{\partial x_l \partial x_j} \geq w_j \frac{a_{jj} a_{jl}}{(A\boldsymbol{\xi})_j^2} > 0.$$

Since A is irreducible C is irreducible. Hence $0 = \lambda_n(H(\boldsymbol{\xi}))$is a simple eigenvalue of $H(\boldsymbol{\xi})$. The restriction of the quadratic form corresponding to the Hessian of $f|\Pi_n$ at ξ, corresponds to $\mathbf{y}^\top H(\boldsymbol{\xi})\mathbf{y}$ where $\mathbf{1}^\top \mathbf{y} = 0$. Corollary 6.6.5 implies that there exists $\alpha > 0$ such that $\mathbf{y}^\top H(\boldsymbol{\xi})\mathbf{y} \ge \alpha \|\mathbf{y}\|^2$ for all $\mathbf{1}^\top \mathbf{y} = 0$. Hence the Hessian of $f|\mathcal{D}$ at the critical point $\mathbf{0} < \boldsymbol{\xi} \in \mathit{\Pi}_n$ has positive eigenvalues. Theorem 6.6.7 yields that there exists a unique critical point $\boldsymbol{\xi} \in \mathcal{D}$ of $f|\mathcal{D}$ such that $f(\mathbf{p}) > f(\boldsymbol{\xi})$ for any $\mathbf{p} \in \mathcal{D}\backslash\{\boldsymbol{\xi}\}$. Since $f(\mathbf{x})$ is a homogeneous function of degree 0 we deduce that $f(\mathbf{x}) \ge f(\boldsymbol{\xi})$ for any $\mathbf{x} > \mathbf{0}$. Equality holds if and only if $\mathbf{x} = t\boldsymbol{\xi}$ for some $t > 0$. □

Theorem 6.6.9 *Let $A \in \mathbb{R}_+^{n\times n}$ and assume that*

$$A\mathbf{u} = \rho(A)\mathbf{u}, A^\top \mathbf{v} = \rho(A)\mathbf{v}, 0 < \rho(A), \mathbf{0} < \mathbf{u} = (u_1, \ldots, u_n)^\top, \mathbf{v} = (v_1, \ldots, v_n)^\top, \mathbf{v}^\top \mathbf{u} = 1.$$

Then

$$(6.6.7)\ \sum_{i=1}^n u_i v_i \log \frac{(A\mathbf{x})_i}{x_i} \ge \log \rho(A) \textit{ for any } \mathbf{x} = (x_1, \ldots, x_n)^\top > \mathbf{0},$$

$$(6.6.8)\ \rho(DA) \ge \rho(A) \prod_{i=1}^n d_i^{u_i v_i} \textit{ for any diagonal } D = \operatorname{diag}(d_1, \ldots, d_n) \ge 0.$$

Equality holds for $\mathbf{x} = t\mathbf{u}$ and $D = sI$, where $t > 0, s \ge 0$, respectively. Assume that A is irreducible and all the diagonal entries of A are positive. Then equality holds in (6.6.7) and (6.6.7) if and only if $\mathbf{x} = tu, D = sI$ for some $t > 0, s \ge 0$ respectively.

Proof. Assume that $A = [a_{ij}]_{i=j=1}^n \in \mathbb{R}_+^{n\times n}$ be irreducible and $a_{ii} > 0$ for $i = 1, \ldots, n$. Let $\mathbf{w} = (u_1 v_1, \ldots, u_n v_n)^\top$. Define $f(\mathbf{x})$ as in (6.6.4). We claim that $\mathbf{u}$ is a critical point of f. Indeed, (6.6.6) yields

$$\frac{\partial f}{\partial x_j}(\mathbf{u}) = -u_j v_j \frac{1}{u_j} + \sum_{i=1}^n u_i v_i \frac{a_{ij}}{(A\mathbf{u})_i} = -v_j + \frac{1}{\rho(A)}(A^\top \mathbf{v})_j = 0,\ j = 1, \ldots, n.$$

Similarly, $t\mathbf{u}$ is a critical point of f for any $t > 0$. In particular, $\boldsymbol{\xi} = t\mathbf{u} \in \mathit{\Pi}_n$ is a critical point of f in $\mathcal{D}$. Theorem 6.6.8 implies that $f(\mathbf{x}) \ge f(\mathbf{u}) = \log \rho(A)$ and equality holds if and only if $\mathbf{x} = t\mathbf{u}$ for some $t > 0$.

Let D be a diagonal matrix with positive diagonal entries. Then DA is irreducible, and $DA\mathbf{x} = \rho(DA)\mathbf{x}$ for some $\mathbf{x} = (x_1, \ldots, x_n)^\top > \mathbf{0}$. Note that since $f(\mathbf{u}) \le f(\mathbf{x})$ we deduce

$$\log \rho(A) \le \sum_{i=1}^n u_i v_i \log \frac{(A\mathbf{x})_i}{x_i} = \log \rho(DA) - \sum_{i=1}^n u_i v_i \log d_i.$$

The above inequality yields (6.6.8). Suppose that equality holds in (6.6.8). Then $\mathbf{x} = t\mathbf{u}$, which yields that $D = sI$ for some $s > 0$. Suppose that $D \geq 0$ and D has at least one zero diagonal element. Then the right-hand side of (6.6.8)is zero. Clearly $\rho(DA) \geq 0$. Since $d_i a_{ii}$ is a principle submatrix of DA, Lemma 6.4.2 yields that $\rho(DA) \geq \max_{i\in\langle n\rangle} d_i a_{ii}$. Hence $\rho(DA) = 0$ if and only if $D = 0$. These arguments prove the theorem when A is irreducible with positive diagonal entries.

Let us now consider the general case. For $\varepsilon > 0$ let $A(\varepsilon) := A + \varepsilon \mathbf{u}\mathbf{v}^\top$. Then $A(\varepsilon) > 0$ and $A(\varepsilon)\mathbf{u} = (\rho(A) + \epsilon)\mathbf{u}, A(\varepsilon)^\top \mathbf{v} = (\rho(A) + \epsilon)\mathbf{v}$. Hence inequalities (6.6.7) and (6.6.8) hold for $A(\varepsilon)$ and fixed $x > \mathbf{0}, D \geq$ o. Let $\varepsilon \searrow 0$ to deduce (6.6.7) and (6.6.8). For $\mathbf{x} = t\mathbf{u}, D = sI$, where $t > 0, s \geq 0$ one has equality. □

Corollary 6.6.10 *Let the assumptions of Theorem 6.6.9 hold. Then*

$$(6.6.9) \qquad \sum_{i=1}^{n} u_i v_i \frac{(A\mathbf{x})_i}{x_i} \geq \rho(A) \textit{ for any } \mathbf{x} = (x_1, \ldots, x_n)^\top > \mathbf{0}.$$

If A is irreducible and has positive diagonal entries then equality holds if and only if $\mathbf{x} = t\mathbf{u}$ for some $t > \mathbf{0}$.

Proof. Use the arithmetic-geometric inequality $\sum_{i=1} p_i c_i \geq \prod_{i=1} c_i^{p_i}$ for any $\mathbf{p} = (p_1, \ldots, p_n) \in \Pi_n$ and any $\mathbf{c} = (c_1, \ldots, c_n)^\top \geq \mathbf{0}$. □

Definition 6.6.11 *Let $\mathbf{x} = (x_1, \ldots, x_n)^\top, \mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{C}^n$. Denote by $D(\mathbf{x}) = \mathrm{diag}(\mathbf{x})$ the diagonal matrix $\mathrm{diag}(x_1, \ldots, x_n)$, by $e^{\mathbf{x}} := (e^{x_1}, \ldots, e^{x_n})^\top$, by $\mathbf{x}^{-1} := (\frac{1}{x_1}, \ldots, \frac{1}{x_n})^\top$ for $\mathbf{x} > \mathbf{0}$, and by $\mathbf{x} \circ \mathbf{y}$ the vector $(x_1 y_1, \ldots, x_n y_n)^\top$.*

For a square diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n) \in C^{n\times n}$ denote by $\mathbf{x}(D)$ the vector $(d_1, \ldots, d_n)^\top$.

Theorem 6.6.12 *Let $\mathbf{0} < \mathbf{u} = (u_1, \ldots, u_n)^\top, \mathbf{v} = (v_1, \ldots, v_n)^\top$. Let $\mathbf{0} < \mathbf{w} = \mathbf{u} \circ \mathbf{v}$. Assume that $A = [a_{ij}]_{i=j=1}^n \in \mathbb{R}_+^{n\times n}$ is irreducible. Then there exists two diagonal matrices $D_1, D_2 \in \mathbb{R}_+^{n\times n}$, with positive diagonal entries, such that $D_1 A D_2 \mathbf{u} = \mathbf{u}, (D_1 A D_2)^\top \mathbf{v} = \mathbf{v}$ if one of the following conditions hold. Under any of this conditions D_1, D_2 are unique up to the transformation $t^{-1} D_1, t D_2$ for some $t > 0$.*

1. *All diagonal entries of A are positive.*

2. *Let $\mathcal{N} \subset \langle n\rangle$ be a nonempty set of all $j \in \langle n\rangle$ such that $a_{jj} = 0$. Assume that all off-diagonal entries of A are positive and the following inequalities hold.*

$$\sum_{i\in\langle n\rangle\backslash\{j\}} w_i > w_j \textit{ for all } j \in \mathcal{N}. \tag{6.6.10}$$

(For $n = 2$ and $\mathcal{N} = \langle 2\rangle$ the above inequalities are not satisfied by any $\mathbf{w}$.)

Proof. We first observe that it is enough to consider the case where $\mathbf{u} = \mathbf{1}$, i.e. $B := D_1AD_2$ is a stochastic matrix. See Problem 2. In this case $\mathbf{w} = \mathbf{v}$.

1. Assume that all diagonal entries of A are positive. Let $f = f_{A,\mathbf{w}}$ be defined as in (6.6.4). The proof of Theorem 6.6.8 yields that f has a unique critical $\mathbf{0} < \boldsymbol{\xi}$ in Π_n. (6.6.6) implies that

$$\sum_{i=1}^{n} \frac{w_i a_{ij}}{(A\boldsymbol{\xi})_i} = \frac{w_j}{\xi_j}.$$

This is equivalent to the equality $(D(A\boldsymbol{\xi})^{-1}AD(\boldsymbol{\xi}))^\top\mathbf{w} = \mathbf{w}$. A straightforward calculation show that $D(A\boldsymbol{\xi})^{-1}AD(\boldsymbol{\xi})\mathbf{1} = \mathbf{1}$. Hence $D_1 = D(A\boldsymbol{\xi})^{-1}, D_2 = D(\boldsymbol{\xi})$.

Suppose that D_1, D_2 are diagonal matrices with positive diagonal entries so that $D_1AD_2\mathbf{1} = \mathbf{1}$ and $(D_1AD_2)^\top\mathbf{w} = \mathbf{w}$. Let $\mathbf{u} = D_2\mathbf{1}$. The equality $D_1AD_2\mathbf{1} = \mathbf{1}$ implies that $D_1 = D(A\mathbf{u})^{-1}$. The equality $(D_1AD_2)^\top\mathbf{w} = \mathbf{w}$ is equivalent to $(D(A(\mathbf{u}))^{-1}AD(\mathbf{u}))^\top\mathbf{w} = \mathbf{w}$. Hence $\mathbf{u}$ is a critical point of f. Therefore $\mathbf{u} = t\boldsymbol{\xi}$. So $D_2 = tD(\boldsymbol{\xi})$ and $D_1 = t^{-1}D(A\boldsymbol{\xi})^{-1}$.

2. As in the proof of Theorem 6.6.8, we show that $f = f_{\mathbf{w},A}$ is blows up to ∞ as $\mathbf{p}$ approaches $\partial\Pi_n$. Let $\mathbf{0} < \mathbf{p}_k = (p_{1,k}, \ldots, p_{n,k})^\top \in \mathit{\Pi}_n, k = 1, \ldots,$ converge to $\mathbf{p} = (p_1, \ldots, p_n)^\top \in \partial\mathit{\Pi}_n$. Let $\emptyset \neq Z(\mathbf{p}) \subset \langle n\rangle$ be the set of vanishing coordinates of $\mathbf{p}$. Since all off-diagonal entries of A are positive, it follows that $\lim_{k\to\infty} \frac{(A\mathbf{p}_k)_i}{p_{i,k}} = \infty$ for each $i \in Z(\mathbf{p})$. To show that $\lim_{k\to\infty} f(\mathbf{p}_k) = \infty$ it is enough to consider the case where $\lim_{k\to\infty} \frac{(A\mathbf{p}_k)_m}{p_{m,k}} = 0$ for some $m \notin Z(\mathbf{p})$. In view of the proof of Theorem 6.6.8 we deduce that $m \in \mathcal{N}$. Furthermore, $\#Z(\mathbf{p}) = n - 1$. Hence $\lim_{k\to\infty} p_{m,k} = 1$. Assume for simplicity of notation that $m = 1$, i.e. $Z(\mathbf{p}) = \{2, 3, \ldots, n\}$. Let $s_k = \max_{i\geq 2} p_{i,k}$. So $\lim_{k\to\infty} s_k = 0$. Let $a > 0$ be the value of the minimal off-diagonal entry of A. Then $\frac{(A\mathbf{p}_k)_i}{p_{i,k}} \geq \frac{ap_{1,k}}{s_k}$ for $i \geq 2$. Also $\frac{(A\mathbf{p}_k)_1}{p_{1,k}} \geq \frac{as_k}{p_{1,k}}$. Thus

$$f(\mathbf{p}_k) \geq w_1 \log \frac{as_k}{p_{1,k}} + \sum_{i\geq 2} w_i \log \frac{ap_{1,k}}{s_k} = (\sum_{i=1}^{n} w_i) \log a + (-w_1 + \sum_{i\geq 2} w_i) \log \frac{p_{1,k}}{s_k}.$$

(6.6.10) for $j = 1$ implies that $\lim_{k\to\infty} f(\mathbf{p}_k) = \infty$.

Let $\mathbf{0} < \boldsymbol{\xi} \in \mathit{\Pi}_n$ be a critical point of f. We claim that $H(\boldsymbol{\xi}) = \rho(C)I - C$, and $0 \le C$ is irreducible. Indeed, for $j \ne l$ $c_{jl} = \sum_{i=1}^n w_i \frac{a_{ij}a_{il}}{(A\boldsymbol{\xi})_i^2}$. Since $n \ge 3$ choose $i \ne j, l$ to deduce that $c_{jl} > 0$. So C has positive off-diagonal entries, hence irreducible. Hence $\mathbf{0} < \boldsymbol{\xi} \in \mathit{\Pi}_n$ is a unique critical point in Π_n. The arguments for 1 yield the existence of D_1, D_2, which are unique up to scaling. □

Problems

1. Prove Lemma 6.6.2.

2. Let $B \in \mathbb{R}_+^{n\times n}$ and assume that $B\mathbf{u} = \mathbf{u}, B^\top \mathbf{v}$, where $\mathbf{0} < \mathbf{u} = (u_1, \dots, u_n)^\top, \mathbf{v} = (v_1, \dots, v_n)^\top$. Then $C := D(\mathbf{u})^{-1} B D(\mathbf{u})$ satisfies the following $C\mathbf{1} = \mathbf{1}, C^\top \mathbf{w} = \mathbf{w}$, where $\mathbf{w} = \mathbf{u} \circ \mathbf{v}$.

3. Let $A \in \mathbb{R}_+^{n\times n}$ is called fully indecomposable if there exists $P \in \mathcal{P}_n$ such that PA is irreducible and have positive diagonal elements. Show that that if A is fully indecomposable, then there exists diagonal matrices D_1, D_2, with positive diagonal entries such that $D_1 A D_2$ is doubly stochastic. D_1, D_2 are unique up to a scalar factor $t^{-1}D_1, tD_2$.

6.7 Convexity and log-convexity

Definition 6.7.1

1. *For any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ denote by $(\mathbf{x}, \mathbf{y})$ and $[\mathbf{x}, \mathbf{y}]$, the open and the closed interval spanned by $\mathbf{x}, \mathbf{y}$ respectively. I.e. the set of points of the form $t\mathbf{x} + (1-t)\mathbf{y}$, where $t \in (0,1)$ and $[0,1]$, respectively.*

2. *Set $\mathrm{D} \subset \mathbb{R}^n$ is called convex if for any $\mathbf{x}, \mathbf{y} \in \mathrm{D}$ the open interval $(\mathbf{x}, \mathbf{y})$ is in D. (Note that a convex set is connected).*

3. *For $\mathbf{x} \le \mathbf{y} \in \mathbb{R}^n$ we denote $[\mathbf{x}, \mathbf{y}] := \{\mathbf{z} \in \mathbb{R}^n,\ \mathbf{x} \le \mathbf{z} \le \mathbf{y}\}$. Clearly, $[\mathbf{x}, \mathbf{y}]$ is a convex set.*

4. *Let $\mathrm{D} \subset \mathbb{R}^n$ be a convex set. Assume that $f : \mathrm{D} \to \mathbb{R}$. f is called convex if*
(6.7.1)
$$f(t\mathbf{x} + (1 - t\mathbf{y})) \le tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \text{ for all } t \in (0,1), \mathbf{x}, \mathbf{y} \in \mathrm{D}.$$

f is called log-convex if $f(\mathrm{D}) \geq 0$ *and*
(6.7.2)
$$f(t\mathbf{x} + (1 - t\mathbf{y})) \leq (f(\mathbf{x}))^t f(\mathbf{y})^{(1-t)} \textit{ for all } t \in (0,1), \mathbf{x}, \mathbf{y} \in \mathrm{D}.$$

f is called strictly convex or strictly log-convex if strict inequality holds in (6.7.1) and (6.7.2), respectively.

Note that if f is a positive function on D then f is log-convex if and only if $\log f$ is convex on D. See Problem 1. The following results are well known, e.g. [Roc70].

Fact 6.7.2

1. $\mathrm{H} = \mathbf{x} + \mathbb{U} := \{\mathbf{y} \in \mathbb{R}^n, \mathbf{y} = \mathbf{x} + \mathbf{u}, \mathbf{u} \in \mathbb{U}\}$, *is a convex set if* $\mathbb{U}$ *is a convex set. If* $\mathbb{U}$ *is a subspace, then* H *is called a hyperplane of dimension k, where k is the dimension of k.* 0*-dimensional hyperplane is a point.*

2. *For a given convex set* $\mathrm{D} \in \mathbb{R}^n$ *let* $\mathbb{U} = \mathrm{span}\ \{\mathbf{y} - \mathbf{x};\ \textit{ for all } \mathbf{y}, \mathbf{x} \in \mathrm{D}\}$. *Then for any* $\mathbf{x} \in \mathrm{D}$, *the hyperplane* $\mathbf{x} + \mathbb{U}$ *is the minimal hyperplane containing* D. *The dimension of* D *is defined as the dimension of of* $\mathbb{U}$, *denoted as* $\dim \mathrm{D} = \dim \mathbb{U}$. *The interior of* D, *denoted by* $\int \mathrm{D}$, *is the set of the interior points of* $\mathrm{D} - \mathbf{x}$, *for a fixed* $\mathbf{x} \in \mathrm{D}$, *viewed as a set in the* dim D *dimensional subspace* $\mathbb{U}$. *Then the closure of* D *is equal to the closure of* $\int \mathrm{D}$, *denoted as* $\mathrm{clo}\ \mathrm{D} = \mathrm{clo}\ \int \mathrm{D}$.

3. *Let* $\mathrm{D} \subset \mathbb{R}^n$ *be a convex set, and* $f : \mathrm{D} \to \mathbb{R}$ *a convex function. Then* $f : \int D \to \mathbb{R}$ *is a continuous function. Furthermore, at each* $\mathbf{x} \in \int D$, $\mathbf{f}$ *has a supporting hyperplane. That is there exists* $\mathbf{p} = \mathbf{p}(\mathbf{x}) \in \mathbb{R}^n$ *such that*

   $$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{p}^\top (\mathbf{y} - \mathbf{x}) \textit{ for any } \mathbf{y} \in \mathrm{D}. \tag{6.7.3}$$

   Assume furthermore that $\dim \mathrm{D} = n$. *Then the following conditions are equivalent.*

   (a) *f is differentiable at* $\mathbf{x} \in \int \mathrm{D}$. *I.e. the gradient of f, ∇f, at* $\mathbf{x}$ *exists and the following equality holds.*

   $$\|f(\mathbf{y}) - (f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}))\| = o(\|\mathbf{y} - \mathbf{x}\|). \tag{6.7.4}$$

   (b) *f has a unique supporting hyperplane at* $\mathbf{x}$.

   The set of points $\mathrm{Diff(f)} \subset \int(\mathrm{D})$, *where f is differentiable, is a dense set in* D *of the full Lebesgue measure, i.e.* $\mathrm{D}\backslash\mathrm{Diff(f)}$ *has zero Lebesgue measure. Furthermore,* ∇f *is continuous on* Diff(f).

4. *Let* $\mathrm{D} \subset \mathbb{R}^n$ *be a convex set, and* $f : \mathrm{D} \to \mathbb{R}$ *a function. Assume that* $f \in \mathrm{C}^2(\int \mathrm{D})$*, i.e.* f *and its first and second partial derivatives are continuous in* $\int \mathrm{D}$*. Then* $f : \int \mathrm{D} \to \mathbb{R}$ *is convex iff and only the Hessian matrix* $H(\mathbf{x}) = [\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})]_{i=j=1}^n$ *has nonnegative eigenvalues for each* $\mathbf{x} \in \int \mathrm{D}$*. If the Hessian* $H(\mathbf{x})$ *has positive eigenvalues for each* $\mathbf{x} \in \int D$*, then* f *is strictly convex on* $\int \mathrm{D}$*.*

5. *Let* $\mathrm{D} \subset \mathbb{R}^n$ *be a convex set.*

   (a) *If* f, g *are convex on* D *then* $\max(f, g)$*, where* $\max(f, g)(x) := \max(f(x), g(x))$*, is convex. Furthermore,* $af + bg$ *is convex for any* $a, b \geq 0$*.*

   (b) *Let* $f_i : \mathrm{D} \to \mathbb{R}$ *for* $i = 1, 2, \ldots$*. Denote by* $f := \limsup_i f_i$ *the function given by* $f(\mathbf{x}) := \limsup_i f(\mathbf{x})$ *for each* $\mathbf{x} \in \mathrm{D}$*. Assume that* $f : \mathrm{D} \to \mathbb{R}$*, i.e. the sequence* $f_i(x), i = 1, \ldots,$ *is bounded for each* x*. If each* f_i *is convex on* D *then* f *is convex on* D*.*

Theorem 6.7.3 *Let* $\mathrm{D} \in \mathbb{R}^m$ *be a convex set. Assume that* $a_{ij} : \mathrm{D} \to \mathbb{R}$ *are log-convex functions for* $i, j = 1, \ldots, n$*. Let* $A(\mathbf{x}) := [a_{ij}(\mathbf{x})]_{i=j=1}^n$ *be the induced nonnegative matrix function on* D*. Then* $\rho(A(\mathbf{x}))$ *is a log-convex function on* D*. Assume furthermore that each* $a_{ij}(\mathbf{x}) \in \mathrm{C}^{\mathrm{k}}(\int \mathrm{D})$*, for some* $k \geq 1$*. (All partial derivatives of* $a_{ij}(\mathbf{x})$ *of order less or equal to* k *are continuous in* $\int \mathrm{D}$*.) Suppose furthermore that* $A(\mathbf{x}_0)$ *is irreducible. Then* $0 < \rho(A(\mathbf{x})) \in \mathrm{C}^{\mathrm{k}}(\int \mathrm{D})$*.*

Proof. In view of Fact 6.7.2.5a each entry of the matrix $A(\mathbf{x})^m$ is log-convex. Hence $\operatorname{tr} A(\mathbf{x})^m$ is log-convex, which implies that $(\operatorname{tr} A(\mathbf{x})^m)^{\frac{1}{m}}$ is log-convex. Theorem 6.4.9 combined with Fact 6.7.2.5b yields that $\rho(A(\mathbf{x}))$ is log-convex.

Assume that $A(\mathbf{x}_0)$ is irreducible for some $\mathbf{x}_0 \in \int(\mathrm{D})$. Since each $a_{ij}(\mathbf{x})$ is continuous in $\int \mathrm{D}$, Problem 1 yields that the digraph $D(A(\mathbf{x}))$ is a constant digraph on $\int \mathrm{D}$. Since $D(A(\mathbf{x}_0))$ is strongly connected, it follows that $D(A(\mathbf{x})))$ is strongly connected for each $\mathbf{x} \in \int \mathrm{D}$. Hence $A(\mathbf{x})$ is irreducible for $\mathbf{x} \in \int \mathrm{D}$ and $\rho(A(\mathbf{x})) > 0$ is a simple root of its characteristic polynomial for $\mathbf{x} \in \int \mathrm{D}$. The implicit function theorem implies that $\rho(A(\mathbf{x})) \in \mathrm{C}^{\mathrm{k}}(\int \mathrm{D})$. $\square$

Theorem 6.7.4 *Let* $A \in \mathbb{R}_+^{n \times n}$*. Define* $A(\mathbf{x}) = D(e^{\mathbf{x}})A$ *for any* $\mathbf{x} \in \mathbb{R}^n$*. Then* $\rho(A(\mathbf{x}))$ *is a log-convex function. Suppose furthermore that* A *is irreducible. Then* $\log \rho(A(\mathbf{x}))$ *is a smooth convex function on* $\mathbb{R}^n$*, i.e.*

$\log \rho(A(\mathbf{x})) \in \mathrm{C}^\infty(\mathbb{R}^{\mathrm{n}})$. *Let*

$$A(\mathbf{x})\mathbf{u}(\mathbf{x}) = \rho(A(\mathbf{x}))\mathbf{u}(\mathbf{x}),\ A(\mathbf{x})^\top \mathbf{v}(\mathbf{x}) = \rho(A(\mathbf{x}))\mathbf{v}(x),$$

$$\mathbf{0} < \mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x}),\ \textit{with the normalization } \mathbf{w}(\mathbf{x}) =: \mathbf{u}(\mathbf{x}) \circ \mathbf{v}(\mathbf{x}) \in \Pi_n.$$

Then

$$\nabla \log \rho(A(\mathbf{x})) = \frac{1}{\rho(A(\mathbf{x}))} \nabla \rho(A(\mathbf{x})) = \mathbf{w}(\mathbf{x}). \tag{6.7.5}$$

That is, the inequality (6.6.8) corresponds to the standard inequality

$$\log \rho(A(\mathbf{y})) \geq \log \rho(A(\mathbf{0})) + \nabla \log \rho(A(\mathbf{0}))^\top \mathbf{y}, \tag{6.7.6}$$

for smooth convex functions.

Proof. Clearly, the function $f_i(\mathbf{x}) = e^{x_i}$ is a smooth log-convex function for $\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$. Since $A \geq 0$ it follows that each entry of $A(\mathbf{x})$ is a log-convex function. Theorem 6.7.3 yields that $\rho(A(\mathbf{x}))$ is log-convex.

Assume in addition that $A = A(\mathbf{0})$ is irreducible. Theorem 6.7.3 yields that $\log \rho(A(\mathbf{x}))$ is a smooth convex function on $\mathbb{R}^n$. Hence $\log \rho(A(\mathbf{x}))$ has a unique supporting hyperplane at each $\mathbf{x}$. For $\mathbf{x} = 0$ this supporting hyperplane is given by the right-hand side of (6.7.6). Consider the inequality (6.6.8). By letting $D = D(e^{\mathbf{y}})$ and taking the logarithm of this inequality we obtain that $\log \rho(A) + \mathbf{w}(\mathbf{0})^\top \mathbf{y}$ is also a supporting hyperplane for $\log \rho(A(\mathbf{x}))$ at $\mathbf{x} = \mathbf{0}$. Hence $\nabla \log \rho(A(\mathbf{0})) = \mathbf{w}(\mathbf{0})$. Similar arguments for any $\mathbf{x}$ proves the equality (6.7.5). □

Problems

1. Let $\mathrm{D} \subset \mathbb{R}^m$ be a convex set.
   (a) Show that if f is a continuous log-convex on D, then either f identically zero function or positive at each $\mathbf{x} \in \mathrm{D}$.
   (b) Assume that f is positive on D, i.e. $f(\mathbf{x}) > 0$ for each $\mathbf{x} \in \mathrm{D}$. Then f is log-convex on D if and only if $\log f$ is convex on D.
   (c) Assume that f is log-convex on D. Then f is continuous on $\int$ D.
2. Let $f : \mathrm{D} \to \mathbb{R}$ be a log-convex function. show that f is a convex function.
3. Let $\mathrm{D} \subset \mathbb{R}^n$ be a convex set.
   (a) If f, g are log-convex on D then $\max(f, g)$ is log-convex. Furthermore, $f^a g^b$ and $af + bg$ are log-convex for any $a, b \geq 0$.
   (b) Let $f_i : \mathrm{D} \to \mathbb{R}$ $i = 1, 2, \dots$ be log-convex. Assume that $f := \limsup_i f_i : \mathrm{D} \to \mathbb{R}$. Then f is log-convex on D.

6.8 Min-max characterizations of $\rho(A)$

Theorem 6.8.1 *Let $\Psi : \mathbb{R} \to \mathbb{R}$ be a differentiable convex nondecreasing function. Let $A \in \mathbb{R}_+^{n\times n}$, and assume that $\rho(A) > 0$. Then*

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n)^\top \in \Pi_n} \inf_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \sum_{i=1}^n p_i \Psi\left(\log \frac{(A\mathbf{x})_i}{x_i}\right) = \Psi(\log \rho(A)). \tag{6.8.1}$$

Suppose that $\Psi'(\log \rho(A)) > 0$ and A has a positive eigenvector $\mathbf{u}$ which corresponds to $\rho(A)$. If

$$\inf_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \sum_{i=1}^n p_i \Psi\left(\log \frac{(A\mathbf{x})_i}{x_i}\right) = \Psi(\log \rho(A)) \tag{6.8.2}$$

then the vector $\mathbf{v} = \mathbf{p} \circ \mathbf{u}^{-1}$ is a nonnegative eigenvector of $A^\top$ corresponding to $\rho(A)$. In particular, if A is irreducible, then $\mathbf{p}$ satisfying (6.8.2) is unique.

Proof. Let $\mu(A)$ be the left-hand side of (6.8.1). We first show that $\mu(A) \le \Psi(\log \rho(A))$. Suppose first that there exists $\mathbf{u} > \mathbf{0}$ such that $A\mathbf{u} = \rho(A)\mathbf{u}$. Then

$$\inf_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \sum_{i=1}^n p_i \Psi\left(\log \frac{(A\mathbf{x})_i}{x_i}\right) \le \sum_{i=1}^n p_i \Psi\left(\log \frac{(A\mathbf{u})_i}{u_i}\right) = \Psi(\log \rho(A)),$$

for any $\mathbf{p} \in \Pi_n$. Hence $\mu(A) \le \Psi(\log \rho(A))$.

Let $J_n \in \mathbb{R}^{n\times n}$ be the matrix whose all entries are equal to 1. For $\varepsilon > 0$ let $A(\varepsilon) := A + \varepsilon J_n$. As $A(\varepsilon)$ is positive, it has a positive Perron-Frobenius eigenvector. Hence $\mu(A(\varepsilon)) \le \Psi(\log \rho(A(\varepsilon)))$. Since Ψ is nondecreasing and $A(\varepsilon) > A$, it follows that $\mu(A) \le \mu(A(\varepsilon)) \le \Psi(\log \rho(A(\varepsilon)))$. Let $\varepsilon \searrow 0$, and use the continuity of $\Psi(t)$ to deduce $\mu(A) \le \Psi(\log \rho(A))$.

Assume now that $A \in \mathbb{R}_+^{n\times n}$ is irreducible. Let $\mathbf{u}, \mathbf{v} > \mathbf{0}$ be the the right and the left Perron-Frobenius eigenvectors of A, such that $\mathbf{p}^\star = (p_1^\star, \ldots, p_n^\star)^\top := \mathbf{u} \circ \mathbf{v} \in \Pi_n$. Suppose first that $\Psi(t) = t$. Theorem 6.6.9 yields the equality

$$\min_{\mathbf{x}=(x_1,\ldots,x_n)^\top>\mathbf{0}} \sum_{i=1}^n p_i^\star \log \frac{(A\mathbf{x})_i}{x_i} = \log \rho(A).$$

Hence we deduce that $\mu(A) \ge \log \rho(A)$. Combine that with the previous inequality $\mu(A) \le \log \rho(A)$ to deduce (6.8.1) in this case.

Suppose next that Ψ is a convex differentiable nondecreasing function on $\mathbb{R}$. Let $s := \Psi'(\log \rho(A))$. So $s \geq 0$, and

$$\Psi(t) \geq \Psi(\log \rho(A)) + (t - \log \rho(A))s, \text{ for any } t \in \mathbb{R}.$$

Thus

$$\sum_{i=1}^{n} p_i \Psi\left(\log \frac{(A\mathbf{x})_i}{x_i}\right) \geq \Psi(\log \rho(A)) - \log \rho(A))s + s \sum_{i=1}^{n} p_i \log \frac{(A\mathbf{x})_i}{x_i}.$$

Use the equality (6.8.1) for $\Phi(t) = t$ to deduce that $\mu(A) \geq \Psi(\log \rho(A))$. Combine that with the inequality $\mu(A) \leq \Psi(\log \rho(A))$ to deduce (6.8.1) for any irreducible A.

Suppose next that $A \in \mathbb{R}_+^{n\times n}$ is reducible and $\rho(A) > 0$. By applying a permutational similarity to A, if necessary, we may assume that $A = [a_{ij}]$ and $B = [a_{ij}]_{i=j=1}^{m} \in \mathbb{R}^{m\times m}, 1 \leq m < n$ is an irreducible submatrix of A with $\rho(B) = \rho(A)$. Clearly, for any $\mathbf{x} > \mathbf{0}$, $(A(x_1, \ldots, x_n)^\top)_i \geq (B(x_1, \ldots, x_m)^\top)_i$ for $i = 1, \ldots, m$. Since Ψ is nondecreasing we obtain the following set of inequalities

$$\mu(A) \geq \sup_{\mathbf{q}^\top \in \Pi_m} \inf_{\mathbf{0} < \mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} q_i \Psi\left(\log \frac{(A\mathbf{x})_i}{x_i}\right) \geq$$
$$\sup_{\mathbf{q}^\top \in \Pi_m} \inf_{\mathbf{0} < \mathbf{y} \in \mathbb{R}^m} \sum_{i=1}^{m} q_i \Psi\left(\log \frac{(B\mathbf{y})_i}{y_i}\right) = \Psi(\log \rho(B)).$$

Use the equality $\rho(A) = \rho(B)$ and the inequality $\mu(A) \leq \Psi(\log \rho(A))$ to deduce the theorem.

Assume now that

$$\rho(A) > 0,\ \Psi'(\log \rho(A)) > 0,\ A\mathbf{u} = \rho(A)\mathbf{u},\ \mathbf{u} > \mathbf{0},$$

and equality (6.8.2) holds. So the infimum is achieved at $\mathbf{x} = \mathbf{u}$. Since $\mathbf{x} = \mathbf{u}$ is a critical point we deduce that $A^\top \mathbf{p} \circ \mathbf{u}^{-1} = \rho(A)\mathbf{p} \circ \mathbf{u}^{-1}$. If A is irreducible then $\mathbf{p}$ is unique.

□

Corollary 6.8.2 *Let $A \in \mathbb{R}_+^{n\times n}$. Then*

$$\text{(6.8.3)} \qquad \sup_{\mathbf{p}=(p_1,\ldots,p_n)^\top \in \Pi_n} \inf_{\mathbf{x}=(x_1,\ldots,x_n) > \mathbf{0}} \sum_{i=1}^{n} p_i \frac{(A\mathbf{x})_i}{x_i} = \rho(A).$$

Suppose that $\rho(A) > 0$ and A has a positive eigenvector $\mathbf{u}$ which corresponds to $\rho(A)$. If

$$\text{(6.8.4)} \qquad \inf_{\mathbf{x}=(x_1,\dots,x_n)>\mathbf{0}} \sum_{i=1}^{n} p_i \frac{(A\mathbf{x})_i}{x_i} = \rho(A)$$

then the vector $\mathbf{v} = \mathbf{p} \circ \mathbf{u}^{-1}$ is a nonnegative eigenvector of $A^\top$ corresponding to $\rho(A)$. In particular, if A is irreducible, then $\mathbf{p}$ satisfying (6.8.3) is unique.

Proof. If $\rho(A) > 0$ the corollary follows from Theorem 6.8.1 by letting $\Psi(t) = e^t$. For $\rho(A) = 0$ apply the corollary to $A_1 = A + I$ to deduce the corollary in this case. □

Theorem 6.8.3 *Let $\mathcal{D}_{n,+}$ denote the convex set of all $n \times n$ nonnegative diagonal matrices. Assume that $A \in \mathbb{R}_+^{n\times n}$. Then*

$$\text{(6.8.5)} \qquad \rho(A + tD_1 + (1-t)D_2) \le t\rho(A + D_1) + (1-t)\rho(A + D_2)$$

for $t \in (0, 1)$ and $D_1, D_2 \in \mathcal{D}_{n,+}$. If A is irreducible then equality holds if and only if $D_1 - D_2 = aI$.

Proof. Let $\phi(\mathbf{p}) = \inf_{\mathbf{x}>\mathbf{0}} \sum_{i=1}^{n} p_i \frac{(A\mathbf{x})_i}{x_i}$ for $\mathbf{p} \in \Pi_n$. Since $\frac{((A+D)\mathbf{x})_i}{x_i} = d_i + \frac{(A\mathbf{x})_i}{x_i}$ for $D = \operatorname{diag}(d_1, \dots, d_n)$ we deduce that

$$\psi(D, \mathbf{p}) := \inf_{\mathbf{x}>\mathbf{0}} \sum_{i=1}^{n} p_i \frac{((A+D)\mathbf{x})_i}{x_i} = \sum_{i=1} p_i d_i + \phi(\mathbf{p}).$$

Thus $\psi(D, \mathbf{p})$ is an affine function, hence convex on $\mathcal{D}_{n,+}$. Therefore, $\rho(A + D) = \sup_{\mathbf{p}\in\Pi_n} \psi(D, \mathbf{p})$ is a convex function on $\mathcal{D}_{n,+}$. Hence (6.8.5) holds for any $t \in (0, 1)$ and $D_1, D_2 \in \mathcal{D}_{n,+}$.

Suppose that A is irreducible and equality holds in (6.8.5). Since $\rho(A + bI + D) = b + \rho(A + D)$ for any $b > 0$, we may assume without loss of generality that all the diagonal elements of A are positive. Let $A_0 = A + tD_1 + (1-t)D_2$. Since A_0 has a positive diagonal and is irreducible we deduce that $A_0\mathbf{u} = r\mathbf{u}, A_0^\top \mathbf{v} = r\mathbf{v}$ where $r > 0, \mathbf{u}, \mathbf{v} > \mathbf{0}, \mathbf{w} := \mathbf{v} \circ \mathbf{u} \in \Pi_n$. Corollary 6.8.2 yields that

$$\rho(A_0) = \psi(tD_1 + (1-t)D_2, \mathbf{w}) = t\psi(D_1, \mathbf{w}) + (1-t)\psi(D_2, \mathbf{w}).$$

Hence, equality in (6.8.5) implies that

$$\rho(A+D_1) = \psi(D_1,\mathbf{w}) = \rho(A_0) + (1-t)\sum_{i=1}^{n}(d_{1,i} - d_{2,i}),$$
$$\rho(A+D_2) = \psi(D_2,\mathbf{w}) = \rho(A_0) + t\sum_{i=1}^{n}(d_{2,i} - d_{1,i}),$$
$$D_1 = \mathrm{diag}(d_{1,1},\ldots,d_{n,1}), \quad D_2 = \mathrm{diag}(d_{1,2},\ldots,d_{n,2}).$$

Furthermore, the infima on $\mathbf{x} > \mathbf{0}$ in the $\psi(D_1,\mathbf{w})$ and $\psi(D_2,\mathbf{w})$ are achieved for $\mathbf{x} = \mathbf{u}$. Corollary 6.8.2 that $\mathbf{u}$ is the Perron-Frobenius eigenvector of $A+D_1$ and $A+D_2$. Hence $D_1 - D_2 = aI$. Clearly, if $D_1 - D_2 = aI$ then equality holds in (6.8.5). □

Theorem 6.8.4 *Let $A \in \mathbb{R}_+^{n\times n}$ be an inverse of an M-matrix. Then*

$$\rho((tD_1 + (1-t)D_2)A) \le t\rho(D_1A) + (1-t)\rho(D_2A), \tag{6.8.6}$$

for $t \in (0,1)$, $D_1, D_2 \in \mathcal{D}_{n,+}$. If $A > 0$ and D_1, D_2 have positive diagonal entries then equality holds if and only if $D_1 = aD_2$.

Proof. Let $A = B^{-1}$, where $B = rI - C, C \in \mathbb{R}_+^{n\times n}$ and $\rho(C) < r$. Use Neumann expansion to deduce that $A = \sum_{i=0}^{\infty} r^{-(i+1)}B^i$. Hence, $A > 0$ if and only if C is irreducible. Assume first that A is positive. Denote by $\mathcal{D}_{n,+}^o$ the set diagonal matrices with positive diagonal, i.e. the interior of $\mathcal{D}_{n,+}$. Clearly, $DA > 0$ for $D \in \mathcal{D}_{n,+}^o$. Thus, $\rho(DA) > 0$ is a simple eigenvalue of $\det(\lambda I - DA)$. Hence $\rho(DA)$ is an analytic function on $\mathcal{D}_{n,+}^o$. Denote by $\nabla\rho(DA) \in \mathbb{R}^n$ the gradient of $\rho(DA)$ as a function on $\mathcal{D}_{n,+}$. Since $\rho(DA) \in \mathrm{C}^2(\mathcal{D}_{\mathrm{n},+}^{\mathrm{o}}$ it follows that convexity of $\rho(DA)$ on $\mathcal{D}_{n,+}^o$ is equivalent to the following inequality.

$$\rho(D(\mathbf{d})A) \ge \rho(D(\mathbf{d}_0)A) + \nabla\rho(D(\mathbf{d}_0)A)^\top(\mathbf{d} - \mathbf{d}_0), \ \mathbf{d}, \mathbf{d}_0 > \mathbf{0}. \tag{6.8.7}$$

See Problem 1. We now show (6.8.7). Let

$$D_0A\mathbf{u} = \rho(D_0)\mathbf{u}, \mathbf{v}^\top D_0A = \rho(D_0A)\mathbf{v}^\top, \mathbf{u}, \mathbf{v} > 0, \mathbf{v}\circ\mathbf{u} \in \Pi_n, D_0 = D(\mathbf{d}_0).$$

Theorem 6.7.4 implies that $\nabla\rho(D_0A) = \rho(D_0A)\mathbf{v}\circ\mathbf{u}\circ\mathbf{d}_0^{-1}$. Hence, (6.8.7) is equivalent to

$$\rho(D(\mathbf{d})A) \ge \rho(D_0A)(\mathbf{v}\circ\mathbf{u})^\top(\mathbf{d}\circ\mathbf{d}_0^{-1}).$$

Let $D(\mathbf{d})A\mathbf{w} = \rho(D(\mathbf{d})A)\mathbf{w}, \mathbf{w} > \mathbf{0}$. Then the above inequality follows from the inequality

$$\rho(D_0A)^{-1} \ge (\mathbf{v}\circ\mathbf{u})^\top(\mathbf{w}\circ(D_0A\mathbf{w})^{-1}). \tag{6.8.8}$$

This inequality follows from the inf sup characterization of $\rho(D_0A)^{-1}$. See Problem 2. The equality case in (6.8.7) follows from the similar arguments for the equality case in (6.8.6). Since $\rho(DA)$ is a continuous function on $\mathcal{D}_{n,+}$, the convexity of $\rho(DA)$ on $\mathcal{D}_{n,+}^o$ yields the convexity of $\rho(DA)$ on $\mathcal{D}_{n,+}$.

Consider now the case where $A^{-1} = rI - B, B \in \mathbb{R}_+^{n\times n}, r > \rho(B)$, and B is reducible. Then there exists $b > 0$ such that for $\rho(B + b\mathbf{1}^\top\mathbf{1}) < r$. For $\varepsilon \in (0, b)$ let $A(\varepsilon) := (rI - (B+\varepsilon\mathbf{1}\mathbf{1}^\top))^{-1}$. Then the inequality (6.8.7) holds if A is replaced by $A(\varepsilon)$. Let $\varepsilon \searrow 0$ to deduce (6.8.7). □

Problems

1. (a) Let $f \in \mathrm{C}^2(\mathrm{a,b})$. Show that $f''(x) \ge 0$ for $x \in (a,b)$ if and only if $f(x) \ge f(x_0) + f'(x_0)(x - x_0)$ for each $x, x_0 \in (a,b)$.

   (b) Let $\mathrm{D} \subset \mathbb{R}^n$ be an open convex set. Assume that $f \in \mathrm{C}^2(\mathrm{D})$. Let $\mathrm{H}(f)(\mathbf{x}) = [\frac{\partial^2 f}{\partial x_i \partial x_j}](\mathbf{x}) \in \mathbf{S}(n,\mathbb{R})$ for $\mathbf{x}_\mathrm{o} \in \mathrm{D}$. Show that $\mathrm{H}(f)(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in \mathrm{D}$ if and only if $f(\mathbf{x}) \ge f(\mathbf{x}_\mathrm{o}) + \nabla f(\mathbf{x}_\mathrm{o})^\top(\mathbf{x} - x_\mathrm{o})$ for all $\mathbf{x}, \mathbf{x}_\mathrm{o} \in \mathrm{D}$. (**Hint**: Restrict f to an interval $(\mathbf{u}, \mathbf{v}) \subset \mathrm{D}$ and use part (a) of the problem.)

2.

   (a) Let $F \in \mathbb{R}_+^{n\times n}$ be an inverse of an M-matrix. Show

$$\frac{1}{\rho(F)} = \inf_{\mathbf{p}=(p_1,\ldots,p_n)^\top \in \Pi_n} \sup_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \sum_{i=1}^n p_i \frac{x_i}{(F\mathbf{x})_i}.$$

   Hint: Use Corollary 6.8.2.

   (b) Let $0 < F \in \mathbb{R}_+^{n\times n}$ be an inverse of an M-matrix. Assume that

$$F\mathbf{u} = \rho(F)\mathbf{u}, F^\top\mathbf{v} = \rho(F)\mathbf{v}, \mathbf{u}, \mathbf{v} > \mathbf{0}, \mathbf{v}\circ\mathbf{u} \in \Pi_n.$$

   Show

$$\frac{1}{\rho(F)} = \sup_{\mathbf{x}=(x_1,\ldots,x_n)>\mathbf{0}} \sum_{i=1}^n v_i u_i \frac{x_i}{(F\mathbf{x})_i}.$$

   Furthermore, $\frac{1}{\rho(F)} = \sum_{i=1}^n v_i u_i \frac{x_i}{(F\mathbf{x})_i}$ for $\mathbf{x} > \mathrm{o}$ if and only if $F\mathbf{x} = \rho(F)\mathbf{x}$.

   (c) Show (6.8.8). **Hint**: Use Corollary 6.6.4 to show that $A^{-1}D_0^{-1}$ is an M-matrix.

3. Let $P = [\delta_{i(j+1)}] \in \mathcal{P}_n$ be a cyclic permutation matrix. Show that $\rho(D(\mathbf{d})P) = (\prod_{i=1}^n d_i)^{\frac{1}{n}}$ for any $\mathbf{d} \in \mathbb{R}_+^n$.

4. Show that (6.8.7) does not hold for all $A \in \mathbb{R}_+^{n\times n}$.

5. Let $A \in \mathbb{R}_+^{n\times n}$ be an inverse of an M-matrix. Show that the convexity of $\rho(D(e^{\mathbf{x}})A)$ on $\mathbb{R}^n$ is implied by the convexity of $\rho(DA)$ on $\mathcal{D}_{n,+}$. **Hint**: Use the generalized arithmetic-geometric inequality.

6.9 Application to cellular communication

6.9.1 Introduction

Power control is used in cellular and ad-hoc networks to provide a high signal-to-noise ratio (SNR) for a reliable connection. A higher SNR also allows a wireless system that uses link adaptation to transmit at a higher data rate, thus leading to a greater spectral efficiency. Transmission rate adaptation by power control is an active research area in communication networks that can be used for both interference management and utility maximization [Sri03].

The motivation of the problems studied in this section comes from maximizing sum rate, (data throughput), in wireless communications. Due to the broadcast nature of radio transmission, data rates in a wireless network are affected by interference. This is particularly true in Code Division Multiple Access (CDMA) systems, where users transmit at the same time over the same frequency bands and their spreading codes are not perfectly orthogonal. Transmit power control is often used to control signal interference to maximize the total transmission rates of all users.

6.9.2 Statement of problems

Consider a wireless network, e.g., cellular network, with L logical transmitter/receiver pairs. Transmit powers are denoted as $p_1, \ldots, p_L$. Let $\mathbf{p} = (p_1, \ldots, p_L)^\top \geq \mathbf{0}$ be the power transmission vector. In many situation we will assume that $\mathbf{p} \leq \bar{\mathbf{p}} := (\bar{p}_1, \ldots, \bar{p}_l)^\top$, where $\bar{p}_l$ is the maximal transmit power of the user l. In the cellular uplink case, all logical receivers may reside in the same physical receiver, i.e., the base station. Let $G = [g_{ij}]_{i,j=1}^L > 0_{L\times L}$ representing the channel gain, where g_{ij} is the channel gain from the jth transmitter to the ith receiver, and n_l is the noise power for the lth receiver be given. The Signal-to-Interference Ratio (SIR) for the lth receiver is denoted by $\gamma_l = \gamma_l(\mathbf{p})$. The map $\mathbf{p} \mapsto \boldsymbol{\gamma}(\mathbf{p})$ is given

by
(6.9.1)

$$\gamma_l(\mathbf{p}) := \frac{g_{ll}p_l}{\sum_{j\neq l} g_{lj}p_j + n_l},\ l = 1,\ldots,L,\ \boldsymbol{\gamma}(\mathbf{p}) = (\gamma_1(\mathbf{p}),\ldots,\gamma_L(\mathbf{p}))^\top.$$

That is, the power p_l is amplified by the factor g_{ll}, and diminished by other users and the noise, inversely proportional to $\sum_{j\neq l} g_{lj}p_j + n_l$.

Define

$$F = [f_{ij}]_{i,j=1}^L, \text{ where } f_{ij} = \begin{cases} 0, & \text{if } i = j \\ \frac{g_{ij}}{g_{ii}}, & \text{if } i \neq j \end{cases} \tag{6.9.2}$$

and

$$\begin{gathered}\mathbf{g} = (g_{11},\ldots,g_{LL})^\top,\ \mathbf{n} = (n_1,\ldots,n_L)^\top, \\ \mathbf{s} = (\mathbf{s_1},\ldots,\mathbf{s_L})^\top := (\frac{\mathbf{n_1}}{\mathbf{g_{11}}},\frac{\mathbf{n_2}}{\mathbf{g_{22}}},\ldots,\frac{\mathbf{n_L}}{\mathbf{g_{LL}}})^\top.\end{gathered} \tag{6.9.3}$$

Then

$$\boldsymbol{\gamma}(\mathbf{p}) = \mathbf{p}\circ(F\mathbf{p}+\mathbf{s})^{-1}. \tag{6.9.4}$$

Let
(6.9.5)

$$\Phi_{\mathbf{w}}(\boldsymbol{\gamma}) := \sum_{i=1}^L w_i \log(1+\gamma_i), \text{ where } \mathbf{w} = (w_1,\ldots,w_n)^\top \in \Pi_n, \boldsymbol{\gamma} \in \mathbb{R}_+^L.$$

The function $\Phi_{\mathbf{w}}(\boldsymbol{\gamma}(\mathbf{p}))$ is the sum rate of the interference-limited channel.

We can study the following optimal problems in the power vector $\mathbf{p}$. The first problem is concerned with finding the optimal power that maximizes the minimal SIR for all users:

$$\max_{\mathbf{p}\in[\mathbf{0},\bar{\mathbf{p}}]} \min_{i\in\langle L\rangle} \gamma_l(\mathbf{p}) \tag{6.9.6}$$

Then second, more interesting problem, is the sum rate maximization problem in interference-limited channels

$$\max_{\mathbf{p}\in[\mathbf{0},\bar{\mathbf{p}}]} \Phi_{\mathbf{w}}(\boldsymbol{\gamma}(\mathbf{p})). \tag{6.9.7}$$

The exact solution to this problem is known to be NP-complete [Luo08]. Note that for a fixed $p_1,\ldots,p_{l-1},p_{l+1},\ldots,p_L$ each $\gamma_j(\mathbf{p}), j \neq l$ is a decreasing function of p_l, while $\gamma_l(\mathbf{p})$ is an increasing function of l. Thus, if $w_l = 0$ we can assume that in the maximal problem (6.9.7) we can choose $p_l = 0$. Hence, it is enough to study the maximal problem (6.9.7) in the case $\mathbf{w} > \mathbf{0}$.

6.9.3 Relaxations of optimal problems

In this subsection, we study several relaxed versions of (6.9.6) and (6.9.7). We will assume first that we do not have the restriction $\mathbf{p} \le \bar{\mathbf{p}}$. Let $\boldsymbol{\gamma}(\mathbf{p}, \mathbf{n})$ be given by (6.9.1). Note that since $\mathbf{n} > 0$ we obtain

$$\boldsymbol{\gamma}(t\mathbf{p}, \mathbf{n}) = \boldsymbol{\gamma}(\mathbf{p}, \frac{1}{t}\mathbf{n}) \Rightarrow \boldsymbol{\gamma}(t\mathbf{p}, \mathbf{n}) > \boldsymbol{\gamma}(\mathbf{p}, \mathbf{n}) \text{ for } t > 1.$$

Thus, to increase the values of the optimal problems in (6.9.6) and (6.9.7), we let $t \to \infty$, which is equivalent to the assumption in this subsection that $\mathbf{n} = \mathbf{0}$.

Theorem 6.9.1 *Let $F \in \mathbb{R}_+^{L\times L}, L \ge 2$ be a matrix with positive off-diagonal entries. Let $F\mathbf{u} = \rho(F)\mathbf{u}$ for a unique $\mathbf{0} < \mathbf{u} \in \Pi_L$. Then*

$$\max_{\mathbf{0}<\mathbf{p}\in\Pi_L} \min_{l\in\langle L\rangle} \frac{p_l}{\sum_{j=1}^L f_{lj}p_j} = \frac{1}{\rho(F)}, \tag{6.9.8}$$

which is achieved only for $\mathbf{p} = \mathbf{u}$. In particular, The value of the optimal problem given in (6.9.6) is less than $\frac{1}{\rho(F)}$.

Proof. Clearly, the left-hand side of 6.9.8 is equal to $(\min_{\mathbf{0}<\mathbf{p}} \max_{l\in\langle L\rangle} \frac{(F\mathbf{p})_l}{p_l})^{-1}$. Since F is irreducible, our theorem follows from Problem 11.

Clearly $\boldsymbol{\gamma}(\mathbf{p}, \mathbf{n}) < \boldsymbol{\gamma}(\mathbf{p}, \mathbf{0})$. Hence, for $\mathbf{p} > \mathbf{0}$ $\min_{l\in\langle L\rangle} \boldsymbol{\gamma}(\mathbf{p}, \mathbf{n}) < \min_{l\in\langle L\rangle} \boldsymbol{\gamma}(\mathbf{p}, \mathbf{0})$. Since $\mathbf{p} \in [\mathbf{0}, \bar{\mathbf{p}}] \subset \mathbb{R}_+^L$ we deduce that the value of the optimal problem given in (6.9.6) is less than $\frac{1}{\rho(F)}$. □

We now consider the relaxation problem of (6.9.7). We approximate $\log(1 + x)$ by $\log x$ for $x \ge 0$. Clearly, $\log(1 + x) \ge \log x$. Let

$$\Psi_{\mathbf{w}}(\boldsymbol{\gamma}) = \sum_{j=1}^L w_i \log \gamma_j, \quad \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^\top. \tag{6.9.9}$$

Theorem 6.9.2 *Let $F = [f_{ij}] \in \mathbb{R}_+^{L\times L}$ have positive off-diagonal elements and zero diagonal entries. Assume that $L \ge 3$, $\mathbf{w} = (w_1, \ldots, w_L)^\top > \mathbf{0}$, and suppose that $\mathbf{w}$ satisfies the inequalities (6.6.10) for each $j \in \langle L\rangle$, where $n = L$. Let $D_1 = \mathrm{diag}(d_{1,1}, \ldots, d_{L,1}), D_2 = \mathrm{diag}(d_{1,2}, \ldots, d_{L,2})$ be two diagonal matrices, with positive diagonal entries, such that $B = D_1FD_2, B\mathbf{1} = \mathbf{1}, B^\top\mathbf{w} = \mathbf{w}$.(As given by Theorem 6.6.12.) Then*

$$\max_{\mathbf{p}>\mathbf{0}} \Psi_{\mathbf{w}}(\mathbf{p}) = \sum_{j=1}^L w_j \log d_{j,1}d_{j,2}. \tag{6.9.10}$$

Equality holds if and only if $\mathbf{p} = tD_2^{-1}\mathbf{1}$ for some $t > 0$.

Proof. Let $\mathbf{p} = D_2\mathbf{x}$. Then

$$\Psi_{\mathbf{w}}(D_2\mathbf{x}) = \sum_{j=1}^{L} w_j \log d_{j,1}d_{j,2} - \sum_{j=1}^{L} w_j \frac{(B\mathbf{x})_j}{x_j}.$$

Use Theorem 6.6.9 to deduce that the above expression is not more than the right-hand side of (6.9.10). For $\mathbf{x} = \mathbf{1}$ equality holds. From the proof of the second part of Theorem 6.6.12 it follows that this minimum is achieved only for $\mathbf{x} = t\mathbf{1}$, which is equivalent to $\mathbf{p} = tD_2^{-1}\mathbf{1}$. □

6.9.4 Preliminary results

Claim 6.9.3 *Let $\mathbf{p} \geq \mathbf{0}$ be a nonnegative vector. Assume that $\boldsymbol{\gamma}(\mathbf{p})$ is defined by (6.9.1). Then $\rho(\mathrm{diag}(\boldsymbol{\gamma}(\mathbf{p}))F) < 1$, where F is defined by (6.9.2). Hence, for $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{p})$,*

$$\mathbf{p} = P(\boldsymbol{\gamma}) := (I - \mathrm{diag}(\boldsymbol{\gamma})F)^{-1}\,\mathrm{diag}(\boldsymbol{\gamma})\mathbf{v}. \tag{6.9.11}$$

Vice versa, if $\boldsymbol{\gamma}$ is in the set

$$\boldsymbol{\Gamma} := \{\boldsymbol{\gamma} \geq \mathbf{0},\ \rho(\mathrm{diag}(\boldsymbol{\gamma})F) < 1\}, \tag{6.9.12}$$

then the vector $\mathbf{p}$ defined by (6.9.11) is nonnegative. Furthermore, $\boldsymbol{\gamma}(P(\mathbf{p})) = \boldsymbol{\gamma}$. That is, $\boldsymbol{\gamma} : \mathbb{R}_+^L \to \boldsymbol{\Gamma}$, and $P : \boldsymbol{\Gamma} \to \mathbb{R}_+^L$ are inverse mappings.

Proof. Observe that (6.9.1) is equivalent to the equality

$$\mathbf{p} = \mathrm{diag}(\boldsymbol{\gamma})F\mathbf{p} + \mathrm{diag}(\boldsymbol{\gamma})\mathbf{v}. \tag{6.9.13}$$

Assume first that $\mathbf{p}$ is a positive vector, i.e., $\mathbf{p} > \mathbf{0}$. Hence, $\boldsymbol{\gamma}(\mathbf{p}) > \mathbf{0}$. Since all off-diagonal entries of $\mathbb{F}$ are positive it follows that the matrix $\mathrm{diag}(\boldsymbol{\gamma})F$ is irreducible. As $\mathbf{v} > \mathbf{0}$, we deduce that $\max_{i\in[1,n]} \frac{(\mathrm{diag}(\boldsymbol{\gamma})F\mathbf{p})_i}{p_i} < 1$. The min max characterization of Wielandt of $\rho(\mathrm{diag}(\boldsymbol{\gamma})F)$, [**?**] implies $\rho(\mathrm{diag}(\boldsymbol{\gamma})F) < 1$. Hence, $\boldsymbol{\gamma}(\mathbf{p}) \in \boldsymbol{\Gamma}$. Assume now that $\mathbf{p} \geq \mathbf{0}$. Note that $p_i > 0 \iff \gamma_i(\mathbf{p}) > 0$. So $\mathbf{p} = \mathbf{0} \iff \gamma(\mathbf{p}) = 0$. Clearly, $\rho(\boldsymbol{\gamma}(\mathbf{0})F) = \rho(0_{L\times L}) = 0 < 1$. Assume now that $\mathbf{p} \gneq \mathbf{0}$. Let $\mathcal{A} = \{i :\ p_i > 0\}$. Denote $\boldsymbol{\gamma}(\mathbf{p})(\mathcal{A})$ the vector composed of positive entries of $\boldsymbol{\gamma}(\mathbf{p})$. Let $F(\mathcal{A})$ be the principal submatrix of $\mathbb{F}$ with rows and columns in $\mathcal{A}$. It is straightforward to see that $\rho(\mathrm{diag}(\boldsymbol{\gamma}(\mathbf{p}))F) = \rho(\mathrm{diag}(\boldsymbol{\gamma}(p)(\mathcal{A})F(\mathcal{A}))$. The arguments above imply that

$$\rho(\mathrm{diag}(\boldsymbol{\gamma}(\mathbf{p}))F) = \rho(\mathrm{diag}(\boldsymbol{\gamma}(\mathbf{p})(\mathcal{A})\mathbb{F}(\mathcal{A})) < 1.$$

Assume now that $\boldsymbol{\gamma} \in \mathbf{\Gamma}$. Then

$$(I - \mathrm{diag}(\boldsymbol{\gamma})F)^{-1} = \sum_{k=0}^{\infty} (\mathrm{diag}(\boldsymbol{\gamma})F)^k \ge 0_{L \times L}. \tag{6.9.14}$$

Hence, $P(\boldsymbol{\gamma}) \ge \mathbf{0}$. The definition of $P(\boldsymbol{\gamma})$ implies that $\boldsymbol{\gamma}(P(\boldsymbol{\gamma})) = \boldsymbol{\gamma}$. □

Claim 6.9.4 *The set $\mathbf{\Gamma} \subset \mathbb{R}_+^L$ is monotonic with respect to the order $\ge$. That is if $\boldsymbol{\gamma} \in \mathbf{\Gamma}$ and $\boldsymbol{\gamma} \ge \boldsymbol{\beta} \ge \mathbf{0}$ then $\boldsymbol{\beta} \in \mathbf{\Gamma}$. Furthermore, the function $P(\boldsymbol{\gamma})$ is monotone on $\mathbf{\Gamma}$.*

$$P(\boldsymbol{\gamma}) \ge P(\boldsymbol{\beta}) \text{ if } \boldsymbol{\gamma} \in \mathbf{\Gamma} \text{ and } \boldsymbol{\gamma} \ge \boldsymbol{\beta} \ge \mathbf{0}. \tag{6.9.15}$$

Equality holds if and only if $\boldsymbol{\gamma} = \boldsymbol{\beta}$.

Proof. Clearly, if $\boldsymbol{\gamma} \ge \boldsymbol{\beta} \ge \mathbf{0}$ then $\mathrm{diag}(\boldsymbol{\gamma})F \ge \mathrm{diag}(\boldsymbol{\beta})F$ which implies $\rho(\mathrm{diag}(\boldsymbol{\gamma})F) \ge \rho(\mathrm{diag}(\boldsymbol{\beta})F)$. Hence, $\mathbf{\Gamma}$ is monotonic. Use the Neumann expansion (6.9.14) to deduce the monotonicity of P. The equality case is straightforward. □

Note that $\boldsymbol{\gamma}(\mathbf{p})$ is not monotonic in $\mathbf{p}$. Indeed, if one increases only the ith coordinate of $\mathbf{p}$, then one increases the ith coordinate of $\boldsymbol{\gamma}(\mathbf{p})$ and decreases all other coordinates of $\boldsymbol{\gamma}(\mathbf{p})$.

As usual, let $\mathbf{e}_i = (\delta_{i1}, \ldots, \delta_{iL})^\top$, $i = 1, \ldots, L$ be the standard basis in $\mathbb{R}^L$. In what follows, we need the following result.

Theorem 6.9.5 *Let $l \in [1, L]$ be an integer and $a > 0$. Denote $[0, a]_l \times \mathbb{R}_+^{L-1}$ the set of all $\mathbf{p} = (p_1, \ldots, p_L)^\top \in \mathbb{R}_+^L$ satisfying $p_l \le a$. Then the image of the set $[0, a]_l \times \mathbb{R}_+^{L-1}$ by the map $\boldsymbol{\gamma}$ (6.9.1), is given by*

$$\rho(\mathrm{diag}(\boldsymbol{\gamma})(F + \frac{1}{a}\mathbf{v}\mathbf{e}_l^\top)) \le 1, \ \mathbf{0} \le \boldsymbol{\gamma}. \tag{6.9.16}$$

Furthermore, $\mathbf{p} = (p_1, \ldots, p_L) \in \mathbb{R}_+^L$ satisfies the condition $p_l = a$ if and only if $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{p})$ satisfies

$$\rho(\mathrm{diag}(\boldsymbol{\gamma})(F + \frac{1}{a}\mathbf{v}\mathbf{e}_l^\top)) = 1. \tag{6.9.17}$$

Proof. Suppose that $\boldsymbol{\gamma}$ satisfies (6.9.16). We claim that $\boldsymbol{\gamma} \in \mathbf{\Gamma}$. Suppose first that $\boldsymbol{\gamma} > \mathbf{0}$. Then $\mathrm{diag}(\boldsymbol{\gamma})(F + t_1\mathbf{v}\mathbf{e}_l^\top) \lneq \mathrm{diag}(\boldsymbol{\gamma})(F + t_2\mathbf{v}\mathbf{e}_l^\top)$ for any $t_1 < t_2$. Lemma 6.2.4 yields

$$\begin{aligned}\rho(\mathrm{diag}(\boldsymbol{\gamma})F) < \rho(\mathrm{diag}(\boldsymbol{\gamma})(F + t_1\mathbf{v}\mathbf{e}_l^\top)) < \rho(\mathrm{diag}(\boldsymbol{\gamma})(F + t_2\mathbf{v}\mathbf{e}_l^\top)) < \\ \rho(\mathrm{diag}(\boldsymbol{\gamma})(F + \frac{1}{a}\mathbf{v}\mathbf{e}_l^\top)) \le 1 \text{ for } 0 < t_1 < t_2 < \frac{1}{a}.\end{aligned} \tag{6.9.18}$$

Thus $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$. Combine the above argument with the arguments of the proof of Claim 6.9.3 to deduce that $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ for $\boldsymbol{\gamma} \geq \mathbf{0}$.

We now show that $P(\boldsymbol{\gamma})_l \leq a$. The continuity of P implies that it is enough to consider the case $\boldsymbol{\gamma} > \mathbf{0}$. Combine the Perron-Frobenius theorem with (6.9.18) to deduce

$$0 < \det\,(I - \mathrm{diag}(\boldsymbol{\gamma})(F + t\mathbf{v}\mathbf{e}_l^\top)) \text{ for } t \in [0, a^{-1}). \tag{6.9.19}$$

We now expand the right-hand side of the above inequality. Let $B = \mathbf{x}\mathbf{y}^\top \in \mathbb{R}^{L\times L}$ be a rank one matrix. Then B has $L-1$ zero eigenvalues and one eigenvalue equal to $\mathbf{y}^\top\mathbf{x}$. Hence, $I - \mathbf{x}\mathbf{y}^\top$ has $L-1$ eigenvalues equal to 1 and one eigenvalue is $(1 - \mathbf{y}^\top\mathbf{x})$. Therefore, $\det\,(I - \mathbf{x}\mathbf{y}^\top) = 1 - \mathbf{y}^\top\mathbf{x}$. Since $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ we get that $(I - \mathrm{diag}(\boldsymbol{\gamma})F)$ is invertible. Thus, for any $t \in \mathbb{R}$

$$\begin{aligned} &\det\,(I - \mathrm{diag}(\boldsymbol{\gamma})(F + t\mathbf{v}\mathbf{e}_l^\top)) = \\ &\det\,(I - \mathrm{diag}(\boldsymbol{\gamma})F)\det\,(I - t((I - \mathrm{diag}(\boldsymbol{\gamma})F)^{-1}\,\mathrm{diag}(\boldsymbol{\gamma})\mathbf{v})\mathbf{e}_l^\top) \\ &\det\,(I - \mathrm{diag}(\boldsymbol{\gamma})F)(1 - t\mathbf{e}_l^\top(I - \mathrm{diag}(\boldsymbol{\gamma})F)^{-1}\,\mathrm{diag}(\boldsymbol{\gamma})\mathbf{v}). \end{aligned} \tag{6.9.20}$$

Combine (6.9.19) with the above identity to deduce

$$1 > t\mathbf{e}_l^\top(I - \mathrm{diag}(\boldsymbol{\gamma})F)^{-1}\,\mathrm{diag}(\boldsymbol{\gamma})\mathbf{v} = tP(\boldsymbol{\gamma})_l \text{ for } t \in [0, a^{-1}). \tag{6.9.21}$$

Letting $t \nearrow a^{-1}$, we deduce that $P(\boldsymbol{\gamma})_l \leq a$. Hence, the set of $\boldsymbol{\gamma}$ defined by (6.9.16) is a subset of $\boldsymbol{\gamma}([0,a]_l \times \mathbb{R}_+^{L-1})$.

Let $\mathbf{p} \in [0,a]_l \times \mathbb{R}_+^{L-1}$ and denote $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{p})$. We show that $\boldsymbol{\gamma}$ satisfies (6.9.16). Claim 6.9.3 implies that $\rho(\mathrm{diag}(\boldsymbol{\gamma})F) < 1$. Since $\mathbf{p} = P(\boldsymbol{\gamma})$ and $p_l \leq a$ we deduce (6.9.21). Use (6.9.20) to deduce (6.9.19). As $\rho(\mathrm{diag}(\boldsymbol{\gamma})F) < 1$, the inequality (6.9.19) implies that $\rho(\mathrm{diag}(\boldsymbol{\gamma})F + t\mathbf{v}^\top\mathbf{e}_l) < 1$ for $t \in (0, a^{-1})$. Hence, (6.9.16) holds.

It is left to show the condition (6.9.17) holds if and only if $P(\boldsymbol{\gamma})_l = a$. Assume that $\mathbf{p} = (p_1, \ldots, p_L)^\top \in \mathbb{R}_+^L$, $p_l = a$ and let $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{p})$. We claim that equality holds in (6.9.16). Assume to the contrary that $\rho(\mathrm{diag}(\boldsymbol{\gamma})(F + \frac{1}{a}\mathbf{v}\mathbf{e}_l^\top)) < 1$. Then, there exists $\boldsymbol{\beta} > \boldsymbol{\gamma}$ such that $\rho(\mathrm{diag}(\boldsymbol{\beta})(F + \frac{1}{a}\mathbf{v}\mathbf{e}_l^\top)) < 1$. Since P is monotonic $P(\boldsymbol{\beta})_l > p_l = a$. On the other hand, since $\boldsymbol{\beta}$ satisfies (6.9.16), we deduce that $P(\boldsymbol{\beta})_l \leq a$. This contradiction yields (6.9.17). Similarly, if $\boldsymbol{\gamma} \geq \mathbf{0}$ and (6.9.17) then $P(\boldsymbol{\gamma})_l = a$. □

Corollary 6.9.6 *Let $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_L)^\top > \mathbf{0}$ be a given positive vector. Then $\boldsymbol{\gamma}([\mathbf{0}, \bar{\mathbf{p}}])$, the image of the set $[\mathbf{0}, \bar{\mathbf{p}}]$ by the map $\boldsymbol{\gamma}$ (6.9.1), is given by*

$$\rho\left(\mathrm{diag}(\boldsymbol{\gamma})\left(F + \frac{1}{\bar{\mathbf{p}}_l}\mathbf{v}\mathbf{e}_l^\top\right)\right) \leq 1, \textit{ for } l = 1, \ldots, L, \textit{ and } \boldsymbol{\gamma} \in \mathbb{R}_+^L. \tag{6.9.22}$$

In particular, any $\boldsymbol{\gamma} \in \mathbb{R}_+^L$ *satisfying the conditions (6.9.22) satisfies the inequalities*

$$\boldsymbol{\gamma} \le \bar{\boldsymbol{\gamma}} = (\bar{\gamma}_1, \dots, \bar{\gamma}_L)^\top, \ \textit{where } \bar{\gamma}_l = \frac{\bar{p}_l}{v_l}, \ i = 1, \dots, L. \tag{6.9.23}$$

Proof. Theorem 6.9.5 yields that $\boldsymbol{\gamma}([\mathbf{0}, \bar{\mathbf{p}}])$ is given by (6.9.22). (6.9.4) yields

$$\gamma_l(\mathbf{p}) = \frac{p_l}{((F\mathbf{p})_l + v_l)} \le \frac{p_l}{v_l} \le \frac{\bar{p}_l}{v_l} \text{ for } \mathbf{p} \in [\mathbf{0}, \bar{\mathbf{p}}].$$

Note that equality holds for $\mathbf{p} = \bar{p}_l \mathbf{e}_l$. □

6.9.5 Reformulation of optimal problems

Theorem 6.9.7 *The maximum problem (6.9.7) is equivalent to the maximum problem.*

$$\begin{array}{ll} \textit{maximize} & \sum_l w_l \log(1 + \gamma_l) \\ \textit{subject to} & \rho(\operatorname{diag}(\boldsymbol{\gamma})(F + (1/\bar{p}_l)\mathbf{v}\mathbf{e}_\mathbf{l}^\top)) \le \mathbf{1} \quad \forall\, \mathbf{l} \in \langle \mathbf{L} \rangle, \\ \textit{variables:} & \gamma_l, \quad \forall\, l. \end{array} \tag{6.9.24}$$

$\boldsymbol{\gamma}^\star$ *is a maximal solution of the above problem if and only if* $P(\boldsymbol{\gamma}^\star)$ *is a maximal solution* $p^\star$ *of the problem (6.9.7). In particular, any maximal solution* $\boldsymbol{\gamma}^\star$ *satisfies the equality (6.9.22) for some integer* $l \in [1, L]$.

We now give the following simple necessary conditions for a maximal solution $\mathbf{p}^\star$ of (6.9.7). We first need the following result, which is obtained by straightforward differentiation.

Lemma 6.9.8 *Denote by*

$$\nabla \Phi_\mathbf{w}(\boldsymbol{\gamma}) = \left(\frac{w_1}{1 + \gamma_1}, \dots, \frac{w_L}{1 + \gamma_L} \right)^\top = \mathbf{w} \circ (\mathbf{1} + \boldsymbol{\gamma})^{-1}$$

the gradient of $\Phi_\mathbf{w}$. *Let* $\boldsymbol{\gamma}(\mathbf{p})$ *be defined as in (6.9.1). Then* $\mathrm{H}(\mathbf{p}) = [\frac{\partial \gamma_i}{\partial p_j}]_{i=j=1}^L$, *the Hessian matrix of* $\boldsymbol{\gamma}(\mathbf{p})$, *is given by*

$$H(\mathbf{p}) = \operatorname{diag}((F\mathbf{p} + \mathbf{v})^{-1})(-\operatorname{diag}(\boldsymbol{\gamma}(\mathbf{p}))F + I).$$

In particular,

$$\nabla_\mathbf{p} \Phi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p})) = \mathrm{H}(\mathbf{p})^\top \nabla \mathit{\Phi}_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p})).$$

Corollary 6.9.9 *Let* $\mathbf{p}^\star = (p_1^\star, \ldots, p_L^\star)^\top$ *be a maximal solution to the problem (6.9.7). Divide the set* $\langle L\rangle = \{1, \ldots, L\}$ *to the following three disjoint sets* $\mathrm{S}_{\max}, \mathrm{S}_{\mathrm{in}}, \mathrm{S}_0$*:*

$$\mathrm{S}_{\max} = \{\mathrm{i} \in \langle \mathrm{L}\rangle,\ \mathrm{p}_\mathrm{i}^\star = \bar{\mathrm{p}}_\mathrm{i}\},\ \mathrm{S}_{\mathrm{in}} = \{\mathrm{i} \in \langle \mathrm{L}\rangle,\ \mathrm{p}_\mathrm{i}^\star \in (0, \bar{\mathrm{p}}_\mathrm{i})\},\ \mathrm{S}_0 = \{\mathrm{i} \in \langle \mathrm{L}\rangle,\ \mathrm{p}_\mathrm{i}^\star = 0\}.$$

Then the following conditions hold.

$$\begin{aligned}&(\mathrm{H}(\mathbf{p}^\star)^\top \nabla \mathit{\Phi}_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}^\star)))_i \ge \mathrm{o}\ \textit{for}\ i \in \mathrm{S}_{\max},\\ &(\mathrm{H}(\mathbf{p}^\star)^\top \nabla \mathit{\Phi}_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}^\star)))_i = \mathrm{o}\ \textit{for}\ i \in \mathrm{S}_{\mathrm{in}},\\ &(\mathrm{H}(\mathbf{p}^\star)^\top \nabla \mathit{\Phi}_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}^\star)))_i \le \mathrm{o}\ \textit{for}\ i \in \mathrm{S}_0\end{aligned} \tag{6.9.25}$$

Proof. Assume that $p_i^\star = \bar{p}_i$. Then $\frac{\partial}{\partial p_i}\Phi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}))(\mathbf{p}^\star) \ge \mathrm{o}$. Assume that $0 < p_i^\star < \bar{p}_i$. Then $\frac{\partial}{\partial p_i}\Phi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}))(\mathbf{p}^\star) = \mathrm{o}$. Assume that $p_i^\star = 0$. Then $\frac{\partial}{\partial p_i}\Phi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}))(\mathbf{p}^\star) \le \mathrm{o}$. □

We now show that the maximum problem (6.9.24) can be restated as the maximum problem of convex function on a closed unbounded domain. For $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^\top > 0$ let $\tilde{\boldsymbol{\gamma}} = \log \boldsymbol{\gamma}$, i.e. $\boldsymbol{\gamma} = e^{\tilde{\boldsymbol{\gamma}}}$. Recall that for a nonnegative irreducible matrix $B \in \mathbb{R}_+^{L\times L}$ $\log \rho(e^{\mathbf{x}} B)$ is a convex function, Theorem 6.7.4. Furthermore, $\log(1+e^t)$ is a strict convex function in $t \in \mathbb{R}$. Hence, the maximum problem (6.9.24) is equivalent to the problem

$$\begin{array}{rl}\text{maximize} & \sum_l w_l \log(1 + e^{\tilde{\gamma}_l})\\ \text{subject to} & \log \rho(\mathrm{diag}(e^{\tilde{\boldsymbol{\gamma}}})(F + (1/\bar{p}_l)\mathbf{v}\mathbf{e}_\mathbf{l}^\top)) \le \mathbf{0}\quad \forall \mathbf{l} \in \langle \mathbf{L}\rangle,\\ \text{variables:} & \tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n)^\top \in \mathbb{R}^L.\end{array} \tag{6.9.26}$$

The unboundedness of the convex set in (6.9.26) is due to the identity $0 = e^{-\infty}$.

Theorem 6.9.10 *Let* $\mathbf{w} > \mathbf{0}$ *be a probability vector. Consider the maximum problem (6.9.7). Then any point* $\mathbf{0} \le \mathbf{p}^\star \le \bar{\mathbf{p}}$ *satisfying the conditions (6.9.25) is a local maximum.*

Proof. Since $\mathbf{w} > \mathbf{0}$, $\Phi_\mathbf{w}(e^{\tilde{\boldsymbol{\gamma}}})$ is a strict convex function in $\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^L$. Hence, the maximum of (6.9.26) is achieved exactly on the extreme points of the closed unbounded set specified in (6.9.26). (It may happen that some coordinate of the extreme point are $-\infty$.) Translating this observation to the maximal problem (6.9.7), we deduce the theorem. □

We now give simple lower and upper bounds on the value of (6.9.7).

Lemma 6.9.11 *Consider the maximal problem (6.9.7). Let* $B_l = (F + (1/\bar{p}_l)\mathbf{v}\mathbf{e}_\mathbf{l}^\top))$ *for* $l = 1, \ldots, L$. *Denote* $R = \max_{l\in\langle L\rangle} \rho(B_l)$. *Let* $\bar{\boldsymbol{\gamma}}$ *be defined by (6.9.23). Then*

$$\Phi_\mathbf{w}((1/R)\mathbf{1}) \le \max_{\mathbf{p}\in[\mathbf{0},\bar{\mathbf{p}}]} \varPhi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}) \le \varPhi_\mathbf{w}(\bar{\boldsymbol{\gamma}}).$$

Proof. By Corollary 6.9.6, $\boldsymbol{\gamma}(\mathbf{p}) \le \bar{\boldsymbol{\gamma}}$ for $\mathbf{p} \in [\mathbf{0}, \bar{\mathbf{p}}]$. Hence, the upper bounds holds. Clearly, for $\boldsymbol{\gamma} = (1/R)\mathbf{1}$, we have that $\rho(\mathrm{diag}(\boldsymbol{\gamma})B_l) \le 1$ for $l \in \langle L\rangle$. Then, from Theorem 6.9.7, $\Phi_\mathbf{w}((1/R)\mathbf{1})$ yields the lower bound. Equality is achieved in the lower bound when $\mathbf{p}^\star = \mathbf{tx}(\mathbf{B_i})$, where $i = \arg\max_{l\in\langle L\rangle} \rho(B_l)$, for some $t > 0$. □

We now show that the substitution $\mathbf{0} < \mathbf{p} = e^{\mathbf{q}}$, i.e. $p_l = e^{q_l}, l = 1, \ldots, L$, can be used to find an efficient algorithm to solve the optimal problem (6.9.6). As in §6.9.3 we can consider the inverse of the maxmin problem of (6.9.6). It is equivalent to the problem
(6.9.27)

$$\min_{\mathbf{q}\le\bar{\mathbf{q}}} g(\mathbf{q}), \quad g(\mathbf{q}) = \max_{l\in\langle L\rangle} s_l e^{-q_l} + \sum_{j=1}^{L} f_{lj} e^{q_j - q_l}, \quad \bar{\mathbf{q}} = (\log \bar{p}_1, \ldots, \log \bar{p}_L)^\top.$$

Note that $s_l e^{-q_l} + \sum_{j=1}^L f_{lj} e^{q_j - q_l}$ is a convex function. Fact 6.7.2.5a implies that $g(\mathbf{q})$ is a convex function. We have quite a good software and mathematical theory to find fast the minimum of a convex function in a convex set as $\mathbf{q} \le \bar{\mathbf{q}}$, i.e. [NoW99].

6.9.6 Algorithms for sum rate maximization

In this section, we outline three algorithms for finding and estimating the maximal sum rates. As above we assume that $\mathbf{w} > \mathbf{0}$. Theorem 6.9.10 gives rise to the following algorithm, which is the gradient algorithm in the variable $\mathbf{p}$ in the compact polyhedron $[\mathbf{0}, \bar{\mathbf{p}}]$.

Algorithm 6.9.12

1. *Choose* $\mathbf{p}_0 \in [\mathbf{0}, \bar{\mathbf{p}}]$*:*
   (a) *Either at random;*
   (b) *or* $\mathbf{p}_0 = \bar{\mathbf{p}}$.
2. *Given* $\mathbf{p}_k = (p_{1,k}, \ldots, p_{L,k})^\top \in [\mathbf{0}, \bar{\mathbf{p}}]$ *for* $k \ge 0$, *compute* $\mathbf{a} = (a_1, \ldots, a_L)^\top = \nabla_\mathbf{p} \varPhi_\mathbf{w}(\boldsymbol{\gamma}(\mathbf{p}_k))$. *If* $\mathbf{a}$ *satisfies the conditions (6.9.25) for* $\mathbf{p}^\star = \mathbf{p}_k$, *then* $\mathbf{p}_k$ *is the output. Otherwise let* $\mathbf{b} = (b_1, \ldots, b_L)^\top$ *be defined as follows.*

(a) $b_i = 0$ *if* $p_{i,k} = 0$ *and* $a_i < 0$*;*

(b) $b_i = 0$ *if* $p_{i,k} = \bar{p}_i$ *and* $a_i > 0$*;*

(c) $b_i = a_i$ *if* $0 < p_i < \bar{p}_i$*.*

Then $\mathbf{p}_{k+1} = \mathbf{p}_k + t_k\mathbf{b}$*, where* $t_k > 0$ *satisfies the conditions* $\mathbf{p}_{k+1} \in [\mathbf{0}, \bar{\mathbf{p}}]$ *and* $\Phi_{\mathbf{w}}(\boldsymbol{\gamma}(\mathbf{p}_k + t_k\mathbf{b}_k))$ *increases on the interval* $[0, t_k]$*.*

The problem with the gradient method, and its variations as a conjugate gradient method is that it is hard to choose the optimal value of t_k in each step, e.g. [Avr03]. We now use the reformulation of the maximal problem given by (6.9.26). Since $\mathbf{w} > \mathbf{0}$, the function $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$ is strictly convex. Thus, the maximum is achieved only on the boundary of the convex set

$$(6.9.28)\quad D(\{F\}) = \{\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^L, \quad \log \rho(\mathrm{diag}(e^{\tilde{\boldsymbol{\gamma}}})(F + (1/\bar{p}_l)\mathbf{v}\mathbf{e}_\mathbf{l}^\top)) \le \mathbf{0}, \ \ \forall\, \mathbf{l}\}.$$

If one wants to use numerical methods and software for finding the maximum value of convex functions on bounded closed convex sets , e.g., [NoW99], then one needs to consider the maximization problem (6.9.26) with additional constraints:

$$(6.9.29)\qquad D(\{F\}, K) = \{\tilde{\boldsymbol{\gamma}} \in D(\{F\}), \quad \tilde{\boldsymbol{\gamma}} \ge -K\mathbf{1}\}.$$

for a suitable $K \gg 1$. Note that the above closed set is compact and convex. The following lemma gives the description of the set $D(\{F\}, K)$.

Lemma 6.9.13 *Let* $\bar{\mathbf{p}} > \mathbf{0}$ *be given and let* R *be defined as in Lemma 6.9.11. Assume that* $K > \log R$*. Let* $\underline{\mathbf{p}} = P(e^{-K}\mathbf{1}) = (e^K I - F)^{-1}\mathbf{v}$*. Then* $D(\{F\}, K) \subseteq \log \boldsymbol{\gamma}([\underline{\mathbf{p}}, \bar{\mathbf{p}}])$*.*

Proof. From the definition of K, we have that $e^K > R$. Hence, $\rho(e^{-K}B_l) < 1$ for $l = 1, \ldots, L$. Thus $-K\mathbf{1} \in D(\{F\})$. Let $\underline{\boldsymbol{\gamma}} = e^{-K}\mathbf{1}$. Assume that $\tilde{\boldsymbol{\gamma}} \in D(\{F\}, K)$. Then $\tilde{\boldsymbol{\gamma}} \ge -K\mathbf{1}$. Hence, $\boldsymbol{\gamma} = e^{\tilde{\boldsymbol{\gamma}}} \ge \underline{\boldsymbol{\gamma}}$. Since $\rho(\mathrm{diag}(\boldsymbol{\gamma})F) < 1$, Claim 6.9.4 yields that $\mathbf{p} = P(\boldsymbol{\gamma}) \ge P(\underline{\boldsymbol{\gamma}}) = \underline{\mathbf{p}}$, where P is defined by (6.9.11). The inequality $P(\boldsymbol{\gamma}) \le \bar{\mathbf{p}}$ follows from Corollary 6.9.6. □

Thus, we can apply the numerical methods to find the maximum of the strictly convex function $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$ on the closed bounded set $D(\{F\}, K)$, e.g. [NoW99]. In particular, we can use the gradient method. It takes the given boundary point $\tilde{\boldsymbol{\gamma}}_k$ to another boundary point of $\tilde{\boldsymbol{\gamma}}_{k+1} \in D(\{F\}, K)$, in the direction induced by the gradient of $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$. However, the complicated boundary of $D(\{F\}, K)$ will make any algorithm expensive.

Furthermore, even though the constraint set in (6.9.24) can be transformed into a strict convex set, it is in general difficult to determine precisely the spectral radius of a given matrix [Var63]. To make the problem simpler and to enable fast algorithms, we approximate the convex set $D(\{F\},K)$ by a bigger polyhedral convex sets as follows. Choose a finite number of points $\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M$ on the boundary of $D(\{F\})$, which preferably lie in $D(\{F\},K)$. Let

$\mathrm{H}_1(\boldsymbol{\xi}),\dots,\mathrm{H}_N(\boldsymbol{\xi}),\boldsymbol{\xi}\in\mathbb{R}^L$ be the N supporting hyperplanes of $D(\{F\}$. (Note that we can have more than one supporting hyperplane at $\boldsymbol{\zeta}_i$, and at most L supporting hyperplanes.) So each $\boldsymbol{\xi}\in D(\{F\},K)$ satisfies the inequality $H_j(\boldsymbol{\xi})\le 0$ for $j=1,\dots,N$. Let $\bar{\boldsymbol{\gamma}}$ be defined by (6.9.23). Define
(6.9.30)

$$D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)=\{\boldsymbol{\xi}\in\mathbb{R}^L,\ -K\mathbf{1}\le\boldsymbol{\xi}\le\log\bar{\boldsymbol{\gamma}},\ \mathrm{H}_j(\boldsymbol{\xi})\le 0 \text{ for } j=1,\dots,N\}.$$

Hence, $D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$ is a polytope which contains $D(\{F\},K)$. Thus

$$\max_{\tilde{\gamma}\in D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)}\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})\ge \tag{6.9.31}$$

$$\max_{\tilde{\gamma}\in D(\{F\},K)}\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}}). \tag{6.9.32}$$

Since $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$ is strictly convex, the maximum in (6.9.31) is achieved only at the extreme points of $D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$. The maximal solution can be found using a variant of a simplex algorithm [**?**]. More precisely, one starts at some extreme point of $\boldsymbol{\xi}\in D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$. Replace the strictly convex function $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$ by its first order Taylor expansion $\Psi_{\boldsymbol{\xi}}$ at $\boldsymbol{\xi}$. Then we find another extreme point $\boldsymbol{\eta}$ of $D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$, such that $\Psi_{\boldsymbol{\xi}}(\boldsymbol{\eta})>\Psi_{\boldsymbol{\xi}}(\boldsymbol{\xi})=\Phi_{\mathbf{w}}(e^{\boldsymbol{\xi}})$. Then we replace $\Phi_{\mathbf{w}}(e^{\tilde{\boldsymbol{\gamma}}})$ by its first order Taylor expansion $\Psi_{\boldsymbol{\eta}}$ at $\boldsymbol{\eta}$ and continue the algorithm. Our second proposed algorithm for finding an optimal $\tilde{\boldsymbol{\gamma}}^\star$ that maximizes (6.9.31) is given as follows.

Algorithm 6.9.14

1. *Choose an arbitrarily extreme point* $\boldsymbol{\xi}_0\in D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$.
2. *Let* $\Psi_{\boldsymbol{\xi}_k}(\boldsymbol{\xi})=\Phi_{\mathbf{w}}(e^{\boldsymbol{\xi}_k})+(\mathbf{w}\circ(\mathbf{1}+\mathbf{e}^{\boldsymbol{\xi}_\mathbf{k}})^{-\mathbf{1}}\circ\mathbf{e}^{\boldsymbol{\xi}_\mathbf{k}})^\top(\boldsymbol{\xi}-\boldsymbol{\xi}_\mathbf{k})$. *Solve the linear program* $\max_{\boldsymbol{\xi}}\Psi_{\boldsymbol{\xi}_k}(\boldsymbol{\xi})$ *subject to* $\boldsymbol{\xi}\in D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$ *using the simplex algorithm in [?] by finding an extreme point* $\boldsymbol{\xi}_{k+1}$ *of* $D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)$, *such that* $\Psi_{\boldsymbol{\xi}_k}(\boldsymbol{\xi}_{k+1})>\Psi_{\boldsymbol{\xi}_k}(\boldsymbol{\xi}_k)=\Phi_{\mathbf{w}}(e^{\boldsymbol{\xi}_k})$.
3. *Compute* $\mathbf{p}_k=P(e^{\boldsymbol{\xi}_{k+1}})$. *If* $\mathbf{p}_k\in[0,\bar{\mathbf{p}}]$, *compute* $\mathbf{a}=(a_1,\dots,a_L)^\top=\nabla_{\mathbf{p}}\Phi_{\mathbf{w}}(\gamma(\mathbf{p}_k))$. *If* $\mathbf{a}$ *satisfies the conditions (6.9.25) for* $\mathbf{p}^\star=\mathbf{p}_k$, *then* $\mathbf{p}_k$ *is the output. Otherwise, go to Step 2 using* $\Psi_{\boldsymbol{\xi}_{k+1}}(\boldsymbol{\xi})$.

As in §6.9.3, it would be useful to consider the following related maximal problem:

$$\max_{\tilde{\gamma}\in D(\boldsymbol{\zeta}_1,\dots,\boldsymbol{\zeta}_M,K)}\mathbf{w}^\top\tilde{\boldsymbol{\gamma}}. \tag{6.9.33}$$

This problem given by (6.9.33) is a standard linear program, which can be solved in polynomial time by the classical ellipsoid algorithm [**?**]. Our third proposed algorithm for finding an optimal $\tilde{\boldsymbol{\gamma}}^\star$ that maximizes (6.9.33) is given as follows. Then $\mathbf{p}^\star = P(e^{\tilde{\boldsymbol{\gamma}}^\star})$.

Algorithm 6.9.15

1. *Solve the linear program* $\max_{\tilde{\boldsymbol{\gamma}}} \mathbf{w}^\top \tilde{\boldsymbol{\gamma}}$ *subject to* $\tilde{\boldsymbol{\gamma}} \in D(\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_M, K)$ *using the ellipsoid algorithm in [**?**].*

2. *Compute* $\mathbf{p} = P(e^{\tilde{\boldsymbol{\gamma}}})$. *If* $\mathbf{p} \in [\text{o}, \bar{\mathbf{p}}]$, *then* $\mathbf{p}$ *is the output. Otherwise, project* $\mathbf{p}$ *onto* $[0, \bar{\mathbf{p}}]$.

We note that $\tilde{\boldsymbol{\gamma}} \in D(\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_M, K)$ in Algorithm 6.9.15 can be replaced by the set of supporting hyperplane $D(\tilde{F}, K) = \{\tilde{\boldsymbol{\gamma}} \in \rho(\mathrm{diag}(e^{\tilde{\boldsymbol{\gamma}}})\tilde{F}) \le 1, \quad \tilde{\boldsymbol{\gamma}} \ge -K\mathbf{1}\}$ or, if $L \ge 3$ and $\mathbf{w}$ satisfies the conditions (6.6.10), $D(F, K) = \{\tilde{\boldsymbol{\gamma}} \in \rho(\mathrm{diag}(e^{\tilde{\boldsymbol{\gamma}}})F) \le 1, \quad \tilde{\boldsymbol{\gamma}} \ge -K\mathbf{1}\}$ based on the relaxed maximal problems in Section 4. Then Theorem 6.9.2 quantify the closed-form solution $\tilde{\boldsymbol{\gamma}}$ computed by Algorithm 6.9.15.

We conclude this section by showing how to compute the supporting hyperplanes $\mathrm{H}_j, j = 1, \dots, N$, which define $D(\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_M, K)$. To do that, we give a characterization of supporting hyperlanes of $D(\{F\})$ at a boundary point $\boldsymbol{\zeta} \in \partial D(\{F\})$.

Theorem 6.9.16 *Let* $\bar{\mathbf{p}} = (\bar{p}_1, \dots \bar{p}_L)^\top > \mathbf{0}$ *be given. Consider the convex set (6.9.28). Let* $\boldsymbol{\zeta}$ *be a boundary point of* $\partial D(\{F\})$. *Then* $\boldsymbol{\zeta} = \log \boldsymbol{\gamma}(\mathbf{p})$, *where* $\mathbf{0} \le \mathbf{p} = (p_1, \dots, p_L)^\top \le \bar{\mathbf{p}}$. *The set* $\mathcal{B} := \{l \in \langle L \rangle,\ p_l = \bar{p}_l\}$ *is nonempty. For each* $B_l = (F + (1/\bar{p}_l)\mathbf{v}\mathbf{e}_l^\top))$ *let* $\mathrm{H}_l(\boldsymbol{\zeta})$ *be the supporting hyperplane of* $\mathrm{diag}(e^{\mathbf{x}})B_l$ *at* $\boldsymbol{\zeta}$, *defined as in Theorem 6.7.4. Then* $\mathrm{H}_l \le 0$, *for* $l \in \mathcal{B}$, *are the supporting hyperplanes of* $D(\{F\})$ *at* $\boldsymbol{\zeta}$.

Proof. Let $\mathbf{p} = P(e^{\boldsymbol{\zeta}})$. Theorem 6.9.5 implies the set $\mathcal{B}$ is nonempty. Furthermore, $\rho(e^{\boldsymbol{\zeta}} B_l) = 1$ if and only if $p_l = \bar{p}_l$. Hence, ζ lies exactly at the intersection of the hypersurfaces $\log \rho(e^{\boldsymbol{\zeta}} B_l) = 0, l \in \mathcal{B}$. Theorem 6.7.4 implies that the supporting hyperplanes of $D(\{F\})$ at $\boldsymbol{\zeta}$ are $\mathrm{H}_l(\boldsymbol{\xi}) \le 0$ for $l \in \mathcal{B}$. □

We now show how to choose the boundary points $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_M \in \partial D(\{F\})$ and to compute the supporting hyperplanes of $D(\{F\})$ at each ζ_i. Let $\underline{\mathbf{p}} = P(e^{-K}\mathbf{1}) = (\underline{p_1}, \dots, \underline{p_L})^\top$ be defined as in Lemma 6.9.13. Choose $M_i \ge 2$ equidistant points in each interval $[\underline{p}_i, \bar{p}_i]$.

$$(6.9.34) \quad p_{j_i, i} = \frac{j_i \underline{p}_i + (M_i - j_i)\bar{p}_i}{M_i} \text{ for } j_i = 1, \dots, M_i, \text{ and } i = 1, \dots, L.$$

Let

$$\mathcal{P} = \{\mathbf{p}_{j_1,\ldots,j_L} = (p_{j_1,1},\ldots,p_{j_L,L})^\top,\ \min(\bar{p}_1 - p_{j_1,1},\ldots,\bar{p}_L - p_{j_L,L}) = 0\}.$$

That is, $\mathbf{p}_{j_1,\ldots,j_L} \in \mathcal{P}$ if and only $\mathbf{p}_{j_1,\ldots,j_L} \not< \bar{\mathbf{p}}$. Then

$$\{\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_M\} = \log \boldsymbol{\gamma}(\mathcal{P}).$$

The supporting hyperplanes of $D(\{F\})$ at each $\boldsymbol{\zeta}_i$ are given by Theorem 6.9.16.

Chapter 7

Convexity

7.1 Convex sets

In this chapter all vector spaces are finite dimensional.

Definition 7.1.1 *Let* $\mathbf{V}$ *be a finite dimensional vector space over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$.

1. *For* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ *denote*

$$[\mathbf{x}, \mathbf{y}] := \{\mathbf{z} : \quad \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \text{ for all } \alpha \in [0, 1]\},$$
$$(\mathbf{x}, \mathbf{y}) := \{\mathbf{z} : \quad \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \text{ for all } \alpha \in (0, 1)\}.$$

   $[\mathbf{x}, \mathbf{y}], (\mathbf{x}, \mathbf{y})$ *are called closed and open intervals respectively, with the end point* $\mathbf{x}, \mathbf{y}$.

2. *For a nonempty* $S \subset \mathbf{V}$ *denote* $\operatorname{conv} S = \cup_{\mathbf{x},\mathbf{y} \in S}[\mathbf{x}, \mathbf{y}]$, *called the convex hull of* S. *(*$\operatorname{conv} \emptyset = \emptyset$*.)*

3. *A set* $C \subset \mathbf{V}$ *is called convex if for each* $\mathbf{x}, \mathbf{y} \in \mathbf{V}$ $[\mathbf{x}, \mathbf{y}] \subset C$. *(*$\emptyset$ *is convex.)*

4. *Assume that* $C \subset \mathbf{V}$ *is a convex set and let* $\mathbf{x} \in C$. *Denote by* $C - \mathbf{x}$ *the set* $\{\mathbf{z} : \ \mathbf{z} = \mathbf{y} - \mathbf{x}, \ \mathbf{y} \in C\}$. *Let* $\mathbf{U} = \operatorname{span}_{\mathbb{R}}(\mathrm{C} - \mathbf{x})$, *i.e. the set of all linear combinations of elements of* $C - \mathbf{x}$ *with real coefficients. Then* $\mathbf{U}$ *is a finite dimensional real space. The dimension of* C, *denoted by* $\dim C$, *is the dimension of the vector space* $\mathbf{U}$. *(*$\dim \emptyset = -1$.*)* $C - \mathbf{x}$ *has interior as a subset of* $\mathbf{U}$, *which is called the* relative interior *and denoted by* $\operatorname{ri}(C - \mathbf{x})$. *Then the relative interior of* C *is defined as* $\operatorname{ri} C$ *equal to* $\operatorname{ri}(C - \mathbf{x}) + \mathbf{x}$.

5. *A point* $\mathbf{x}$ *in a convex set* C *is called an extreme point if there are no two points* $\mathbf{y}, \mathbf{z} \in C\backslash\{\mathbf{x}\}$ *such that* $\mathbf{x} \in [\mathbf{y}, \mathbf{z}]$*. Denote by* $\mathcal{E}(C)$ *the set of the extreme points of the convex set* C*.*

6. *Let* C *be a convex set and* $\mathcal{E}(C)$ *the set of its extreme points. For* k *distinct extreme points* $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{E}(C)$ *the convex set* $\mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ *is called the* k*-face of* C *if the following property holds. Let* $\mathbf{x}, \mathbf{y} \in C$ *and assume that* $(\mathbf{x}, \mathbf{y}) \cap \mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \neq \emptyset$*. Then* $[\mathbf{x}, \mathbf{y}] \subset \mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$*.*

7. *Let* C *be a convex set in a finite dimensional vector space* $\mathbf{V}$*. For* $\mathbf{f} \in \mathbf{V}^*$ *and* $\mathbf{x} \in \mathbf{V}$ *denote*

$$\begin{aligned}
\mathrm{H}_0(\mathbf{f}, \mathbf{x}) &:= \{\mathbf{y} \in \mathbf{V},\ \Re\mathbf{f}(\mathbf{y}) = \Re\mathbf{f}(\mathbf{x})\},\\
\mathrm{H}_+(\mathbf{f}, \mathbf{x}) &:= \{\mathbf{y} \in \mathbf{V},\ \Re\mathbf{f}(\mathbf{y}) \geq \Re\mathbf{f}(\mathbf{x})\},\\
\mathrm{H}_-(\mathbf{f}, \mathbf{x}) &:= \{\mathbf{y} \in \mathbf{V},\ \Re\mathbf{f}(\mathbf{y}) \leq \Re\mathbf{f}(\mathbf{x})\}.
\end{aligned}$$

   $\mathrm{H}_0(\mathbf{f}, \mathbf{x})$ *is called the (real) hyperplane,* $\mathrm{H}_+(\mathbf{f}, \mathbf{x}), \mathrm{H}_-(\mathbf{f}, \mathbf{x})$ *are called the upper and the lower half spaces respectively, or simply the half spaces.*

It is straightforward to show that $\dim C$, $\mathrm{ri}\, C$ do not depend on the choice of $\mathbf{x} \in C$. Furthermore $\mathrm{ri}\, C$ is convex. See Problem 3 or [Roc70].

Assume that $\mathbf{V}$ is a complex finite dimensional subspace, of dimension n. Then $\mathbf{V}$ can be viewed as a real vector space $\mathbf{V}_{\mathbb{R}}$ of dimension $2n$. A convex set $C \subset \mathbf{V}$ is a convex set $C_{\mathbb{R}} \subset \mathbf{V}_{\mathbb{R}}$. However, as we see later, sometimes it is natural to consider convex sets as subsets of complex vector space $\mathbf{V}$, rather then subsets of $\mathbf{V}_{\mathbb{R}}$.

Clearly, $\mathrm{H}_0(\mathbf{f}, \mathbf{x}), \mathrm{H}_+(\mathbf{f}, \mathbf{x}), \mathrm{H}_-(\mathbf{f}, \mathbf{x})$ are convex sets. Note also that $\mathrm{H}_-(\mathbf{f}, \mathbf{x}) = \mathrm{H}_+(-\mathbf{f}, \mathbf{x})$.

Definition 7.1.2 *An intersection of a finite number of half spaces* $\cap_{i=1}^m \mathrm{H}_+(\mathbf{f}_i, \mathbf{x}_i)$ *is called a polyhedron. A nonempty compact polyhedron is called polytope.*

Clearly, a polyhedron is a closed convex set. Given a polyhedron C, it is a natural problem to find if this polyhedron is empty or not empty. The complexity of finding out if this polyhedron is empty or not depends polynomially on: the dimension of $\mathbf{V}$, and the complexity of all the half spaces in the characterizing C. This is not a trivial fact, which is obtained using an ellipsoid method. See [Kha79, Kar84, Lov86].

It is well known that any polytope has a finite number of extreme points, and is equal to the convex hull of its extreme points [Roc70, p'12]. The following result is a generalization of this fact [Roc70, Part IV,§17-18].

Theorem 7.1.3 *Let C be a compact convex set in a finite dimensional vector space $\mathbf{V}$ of dimension d. Then $\mathcal{E} := \mathcal{E}(C)$ is nonempty set and $C = \operatorname{conv}\mathcal{E}$. Furthermore for each $\mathbf{x} \in C$ there exists at most $d+1$ extreme points $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{E}, k \le d+1$, such that $\mathbf{x} \in \operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$.*

In general it is a difficult problem to find explicitly all the extreme points of the given compact convex set, or even of a given polyhedron. The following example is a classic in matrix theory.

Theorem 7.1.4 *Let $\mathrm{H}_{n,+,1} \subset \mathbb{C}^{n\times n}$ be the convex set of nonnegative definite hermitian matrices with trace 1. Then*

$$\mathcal{E}(\mathrm{H}_{n,+,1}) = \{\mathbf{x}\mathbf{x}^*,\ \mathbf{x} \in \mathbb{C}^n, \mathbf{x}^*\mathbf{x} = 1\}, \tag{7.1.1}$$

$$\mathcal{E}(\mathrm{H}_{n,+,1} \cap \mathrm{S}(\mathrm{n},\mathbb{R})) = \{\mathbf{x}\mathbf{x}^\top,\ \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top\mathbf{x} = 1\}. \tag{7.1.2}$$

Each matrix in $\mathrm{H}_{n,+,1}$ or $\mathrm{H}_{n,+,1} \cap \mathrm{S}(\mathrm{n},\mathbb{R})$ is a convex combination of at most n extreme points.

Proof. Let $A = \mathbf{x}\mathbf{x}^*, \mathbf{x} \in \mathbb{C}^n, \mathbf{x}^*\mathbf{x} = 1$. Clearly $A \in \mathrm{H}_{n,+,1}$. Suppose that $A = aB + (1-a)C$ for some $B, C \in \mathrm{H}_{n,+,1}$ and $a \in (0,1)$. Hence $A \succcurlyeq aB \succcurlyeq 0$. Since $\mathbf{y}^*A\mathbf{y} \ge a\mathbf{y}^*B\mathbf{y} \ge 0$ it follows that $\mathbf{y}^*B\mathbf{y} = 0$ for $\mathbf{y}^*\mathbf{x} = 0$. Hence $B\mathbf{y} = \mathbf{0}$ for $\mathbf{y}^*\mathbf{x} = 0$. Thus B is a rank one nonnegative definite matrix of the form $t\mathbf{x}\mathbf{x}^*$ where $t > 0$. Since $\operatorname{tr} B = 1$ we deduce that $t = 1$ and $B = A$. Similarly $C = A$. Hence A is an extremal point.

Let $F \in \mathrm{H}_{n,+,1}$. Then the spectral decomposition of F yields that $F = \sum_{i=1}^n \lambda_i \mathbf{x}_i\mathbf{x}_i^*$, where $\mathbf{x}_i^*\mathbf{x}_j = \delta_{ij}, i, j = 1, \ldots, n$. Furthermore, since F is nonnegative definite of trace 1, $\lambda_1, \ldots, \lambda_n$, the eigenvalues of F, are nonnegative and sum to 1. So $F \in \operatorname{conv}\{\mathbf{x}_1\mathbf{x}_1^*, \ldots, \mathbf{x}_n\mathbf{x}_n^*\}$. Similar arguments apply to nonnegative real symmetric matrices of rank 1. □

Definition 7.1.5 *Let $C_1, C_2 \subset \mathbf{V}$, where $\mathbf{V}$ is a finite dimensional vector space over $\mathbb{F}$. C_1, C_2 are called hyperplane separated if there exists $\mathbf{f} \in \mathbf{V}^*$ and $\mathbf{x} \in \mathbf{V}$ such that $C_1 \subset \mathrm{H}_+(\mathbf{f},\mathbf{x}), C_2 \subset \mathrm{H}_-(\mathbf{f},\mathbf{x})$. $\mathrm{H}_0(\mathbf{f},\mathbf{x})$ is called the separating (real) hyperplane. $\mathrm{H}_0(\mathbf{f},\mathbf{x})$ is said to separate C_1 and C_2 properly if $\mathrm{H}_0(\mathbf{f},\mathbf{x})$ separates C_1 and C_2 and $\mathrm{H}_0(\mathbf{f},\mathbf{x})$ does contain C_1 and C_2.*

The following result is well known [Roc70, Theorems 11.3].

Theorem 7.1.6 *Let C_1, C_2 be nonempty convex sets in a finite dimensional vector space $\mathbf{V}$. Then there exists a hyperplane separating C_1 and C_2 properly if and only $\operatorname{ri} C_1 \cap \operatorname{ri} C_2 = \emptyset$.*

Corollary 7.1.7 *Let C_1 be a compact convex set in a finite dimensional vector space $\mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that C_1 contains more than one point. Let $\mathbf{x}$ be a an extreme point of C. Then there exists a hyperplane which supports properly C_1 at $\mathbf{x}$. I.e., there exists $\mathbf{0} \neq f \in \mathbf{V}^*$, such that $\Re f(\mathbf{x}) \le \Re f(\mathbf{y})$ for each $\mathbf{y} \in C$. Furthermore, there exists $\mathbf{y} \in C$ such that $\Re f(\mathbf{x}) < \Re f(\mathbf{y})$.*

Proof. Let $C_2 = \{\mathbf{x}\}$. So C_2 is a convex set. Problem 4 yields that $\mathrm{ri}\, C_1 \cap \mathrm{ri}\, C_2 = \emptyset$. Use Theorem 7.1.4 to deduce the Corollary. □

Definition 7.1.8 *A point $\mathbf{x}$ of a convex set C in a finite dimensional vector space $\mathbf{V}$ is called exposed, if there there exist a linear functional $\mathbf{f} \in \mathbf{V}^*$ such that $\Re\mathbf{f}(\mathbf{x}) > \Re\mathbf{f}(\mathbf{y})$ for any $\mathbf{y} \in \mathrm{C}\backslash\{\mathbf{x}\}$.*

Clearly, an exposed point of C is an extreme point, (Problem 5). There exist compact convex sets with extreme points which are not exposed. See Problem 6. In what follows we need Straszewiz [Str35].

Theorem 7.1.9 *. Let C be a closed convex set. Then the set of exposed points of C is a dense subset of extreme points of C. Thus every extreme point is the limit of some sequence of exposed points.*

Corollary 7.1.10 *Let C be a closed convex set. Let $\mathbf{x} \in \mathrm{C}$ be an isolated extreme point. (I.e. there is a neighborhood of $\mathbf{x}$, where $\mathbf{x}$ is the only extreme point of C.) Then $\mathbf{x}$ is an exposed point.*

Problems

1. Show

   (a) For any nonempty subset S of a finite dimensional vector space $\mathbf{V}$ over $\mathbb{F}$, $\mathrm{conv}\, S$ is a convex set.

   (b) Furthermore, if S is compact, then $\mathrm{conv}\, S$ is compact and $\mathcal{E}(C) \subset C$.

2. Let C be a convex set in a finite dimensional subspace, with the set of extreme points $\mathcal{E}(C)$. Let $E_1 \subset \mathcal{E}(C)$ and $C_1 = \mathrm{conv}\, E_1$. Show that $\mathcal{E}(C_1) = E_1$.

3. Let $\mathbf{V}$ be a finite dimensional space and $C \subset \mathbf{V}$ a nonempty convex set. Let $\mathbf{x} \in C$. Show

   a. The subspace $\mathbf{U} := \mathrm{span}\ (\mathrm{C} - \mathbf{x})$ does not depend on $\mathbf{x} \in C$.

   b. $C - \mathbf{x}$ has a nonempty convex interior in $\mathbf{U}$

4. Let C be a convex set in a finite dimensional vector space $\mathbf{V}$. Assume that C contains at least two distinct points. Show

   (a) Show that $\dim C \geq 1$.

   (b) Show that $\mathrm{ri}\ C \cap \mathcal{E}(C) = \emptyset$.

5. Let $\mathbf{x} \in \mathrm{C}$ be an exposed point. Show that $\mathbf{x}$ is an extreme point of C.

6. Consider the convex set $\mathrm{C} \in \mathbb{R}^2$, which is a union of the three convex sets:

$$\mathrm{C}_1 := \{(\mathrm{x},\mathrm{y})^\top, |\mathrm{x}| \leq 1, |\mathrm{y}| \leq 1\},\ \mathrm{C}_2 = \{(\mathrm{x},\mathrm{y})^\top, (\mathrm{x}-1)^2 + \mathrm{y}^2 \leq 1\},$$
$$\mathrm{C}_3 = \{(\mathrm{x},\mathrm{y})^\top, (\mathrm{x}+1)^2 + \mathrm{y}^2 \leq 1\}.$$

   Show that C has exactly 4 extreme points $(\pm 1, \pm 1)^\top$ which are not exposed points.

7.2 Doubly stochastic matrices

Definition 7.2.1 *$A \in \mathbb{R}_+^{n\times n}$ is called doubly stochastic matrix if the sum of each row and column of A is equal to* 1. *Denote by $\Omega_n \subset \mathbb{R}_+^{n\times n}$ the set of doubly stochastic matrices. Denote by $\frac{1}{n}J_n$ the $n\times n$ doubly stochastic matrix whose all entries are equal to $\frac{1}{n}$, i.e. $J_n \in \mathbb{R}_+^{n\times n}$ is the matrix whose each entry is* 1.

Definition 7.2.2 *$P \in \mathbb{R}_+^{n\times n}$ is called a permutation matrix if each row and column of P a contains exactly one nonzero element which is equal to* 1. *Denote by $\mathcal{P}_n$ the set of $n \times n$ permutation matrices.*

Lemma 7.2.3 *The following properties hold.*

1. *$A \in \mathbb{R}_+^{n\times n}$ is double stochastic if and only if $A\mathbf{1} = A^\top\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = (1,\ldots,1)^\top \in \mathbb{R}^n$.*

2. *$\Omega_1 = \{1\}$.*

3. *$A, B \in \Omega_n \Rightarrow tA + (1-t)B \in \Omega_n$ for each $t \in [0,1]$.*

4. $A, B \in \Omega_n \Rightarrow AB \in \Omega_n$.

5. $\mathcal{P}_n \subset \Omega_n$.

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

7. $A \in \Omega_l,\ B \in \Omega_m \rightarrow A \oplus B \in \Omega_{l+m}$.

See Problem 1.

Theorem 7.2.4 *The set* Ω_n *is a polytope of dimension* $(n-1)^2$*, whose extreme points is the set of permutation matrices* $\mathcal{P}_n$.

Proof. Clearly, Ω_n is a nonempty compact convex set in $\mathbb{R}^{n\times n}$. Ω_n is a polytope since it is intersection of $4n + n^2$ half spaces

$$(7.2.1) \sum_{k=1}^n x_{kj} \ge 1, \sum_{k=1}^n -x_{kj} \ge -1, \sum_{k=1}^n x_{jk} \ge 1, \sum_{k=1}^n -x_{jk} \ge -1,$$
$$x_{ij} \ge 0, \quad i = 1, \dots, n, \quad j = 1, \dots, n,$$

where $X = [x_{ij}]_{i=j=1}^n$.

Let $\Omega_{n,0} = \Omega - \{\frac{1}{n}J_n\}$, i.e. $\Omega_{n,0}$ is the set of all matrices of the form $A - \frac{1}{n}J_n$, where $A \in \Omega_n$. Denote

$$(7.2.2) \qquad \mathcal{X}_n = \{X \in \mathbb{R}^{n\times n},\ X\mathbf{1} = X^\top\mathbf{1} = \mathbf{0}\}.$$

Clearly $\mathcal{X}_n$ is a subspace of $\mathbb{R}^{n\times n}$, which contains $\Omega_{n,0}$. Let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal matrix whose first column is the vector$\frac{1}{\sqrt{n}}\mathbf{1}$. We claim that $X \in \mathcal{X}_n$ if and only if $Q^\top XQ = [0] \oplus Y$ for some $Y \in \mathbb{R}^{(n-1)\times(n-1)}$. Indeed, observe that $Z = [0] \oplus Y$ if and only if $Z\mathbf{e}_1 = Z^\top\mathbf{e}_1 = \mathbf{0}$, where $\mathbf{e}_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^n$. Clearly, $Q\mathbf{e}_1 = \frac{1}{\sqrt{n}}\mathbf{1}$, hence $Q^\top XQ = [0] \oplus Y$ if and only if $X\mathbf{1} = X^\top\mathbf{1} = \mathbf{0}$. So $\mathcal{X}_n = Q([0] \oplus \mathbb{R}^{(n-1)\times(n-1)})Q^\top$, hence $\dim \mathcal{X}_n = \dim \mathbb{R}^{(n-1)\times(n-1)} = (n-1)^2$.

Let

$$(7.2.3) \qquad \mathrm{B}(0, r) = \{X \in \mathbb{R}^{n\times n},\ \operatorname{tr} X^\top X \le r^2\}$$

be the closed ball of radius r in Frobenius, i.e. Euclidean norm, in $\mathbb{R}^{n\times n}$ centered in the origin. We claim that $\mathrm{B}(0, \frac{1}{n}) \cap \mathcal{X}_n \subset \Omega_{n,0}$. Indeed, let $X = [x_{ij}]_{i=j=1}^n \in \mathrm{B}(0, \frac{1}{n}) \cap \mathcal{X}_n$. Then $|x_{ij}| \le \frac{1}{n}$ and $X\mathbf{1} = X^\top\mathbf{1} = \mathbf{0}$. Let $A = X + \frac{1}{n}J_n$. So $X \ge 0$ and $A\mathbf{1} = A^\top\mathbf{1} = \mathbf{1}$ and $A \in \Omega_n$. Hence $\dim \Omega_n = (n-1)^2$.

We next observer that any permutation matrix $P = [p_{ij}] \in \mathcal{P}_n$ is an extreme point of Ω_n. Indeed, assume that $P = tA + (1-t)B$ for some $A = [a_{ij}], B = [b_{ij}] \in \Omega_n$ and $t \in (0,1)$. Hence, $a_{ij} = b_{ij} = 0$ if $p_{ij} = 0$. So A and B have at most one nonzero element in each row i. As $A, B \in \Omega_n$ it follows that $a_{ij} = b_{ij} = p_{ij}$ if $p_{ij} = 1$. Thus $A = B = P$ and P is an extremal point.

It is left to show that $\mathcal{E}(\Omega_n) = \mathcal{P}_n$. This is equivalent to the statement that $A \in \mathbb{R}_+^{n\times n}$ is doubly stochastic if and only

$$A = \sum_{P\in\mathcal{P}_n} a_P P \text{ for some } a_P \geq 0,\ P \in \mathcal{P}_n,\ \sum_{P\in\mathcal{P}_n} a_P = 1. \tag{7.2.4}$$

We now show by induction on n that any $A \in \Omega_n$ is of the form (7.2.4). For $n = 1$ the result trivially holds. Assume that the result holds for $n = m-1$ and assume that $n = m$. Let $A = (a_{ij}) \in \Omega_n$. Denote by $l(A)$ be the number of nonzero entries of A. Since each row sum of A is 1 it follows that $l(A) \geq n$. Suppose first $l(A) \leq 2n-1$. Then there exists a row i of A which has exactly one nonzero element, which must be 1. Hence there exists $i, j \in \langle n\rangle$ such that $a_{ij} = 1$. Then all other elements of A on the row i and column j are zero. Denote by $A_{ij} \in \mathbb{R}_+^{(n-1)\times(n-1)}$ the matrix obtained from A by deleting the row and column j. Clearly $A_{ij} \in \Omega_{n-1}$. Use the induction hypothesis on A_{ij} to deduce (7.2.4), where $a_P = 0$ if the entry (i,j) of P is not 1.

We now show by induction on $l(A) \geq 2n-1$ that A is of the form (7.2.4). Suppose that any $A \in \Omega_n$ such that $l(A) \leq l-1, l \geq 2n$ is of the form (7.2.4). Assume that $l(A) = l$. Let $\mathcal{S} \subset \langle n\rangle \times \langle n\rangle$ be the set of all indices $(i,j) \in \langle n\rangle \times \langle n\rangle$ where $a_{ij} > 0$. Note $\#\mathcal{S} = l(A) \geq 2n$. Consider the following system of equations in n^2 variables, which are the entries $X = (x_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$:

$$\sum_{j=1}^n x_{ij} = \sum_{j=1}^n x_{ji} = 0, \quad i = 1, \ldots, n.$$

Since the sum of all rows of X is equal to the sum of all columns of X we deduce that the above system has at most $2n-1$ linear independent equations. Assume furthermore the conditions $x_{ij} = 0$ for $(i,j) \notin \mathcal{S}$. Since we have at least $2n$ variables it follows that there exist $X \neq 0_{n\times n}$ satisfying the above conditions. Note that X has zero entry in the places where A has zero entry. Furthermore, X has at least one positive and one negative entry. Therefore the exists $b, c > 0$ such that $A - bX, A + cX \in \Omega_n$ and $l(A - bX), l(A + cX) < l$. So $A - bX, A + cX$ are of the form (7.2.4). As $A = \frac{c}{b+c}(A - bX) + \frac{b}{b+c}(A + cX)$ we deduce that A is of the form (7.2.4). □

Definition 7.2.5 *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 $\mathbf{x} = (x_1, ..., x_n)^T, \mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$. *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*

$$\sum_{i=1}^{k} \bar{x}_i \leq \sum_{i=1}^{k} \bar{y}_i, \quad k = 1, ..., n. \tag{7.2.5}$$

$\mathbf{x}$ *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* $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

Theorem 7.2.6 . *For* $\mathbf{y} \in \mathbb{R}^n$ *let*

$$\mathcal{M}(\mathbf{y}) := \{\mathbf{x} \in \mathbb{R}^n, \ \mathbf{x} \prec \mathbf{y}\}. \tag{7.2.6}$$

Then $\mathcal{M}(\mathbf{y})$ *is a polyhedron whose extreme points are* $P\mathbf{y}$ *for all* $P \in \mathcal{P}_n$. *In particular,* $\mathbf{x} \prec \mathbf{y}$ *if and only if there exists* $A \in \Omega_n$ *such that* $\mathbf{x} = A\mathbf{y}$.

Proof. Observe first that $\mathbf{x} = (x_1, \ldots, x_n)^\top \prec \mathbf{y} = (y_1, \ldots, y_n)^\top$ is equivalent to the following conditions

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1, \tag{7.2.7}$$

$$\sum_{i=1}^{k} (P\mathbf{x})_i \leq \sum_{i=1}^{k} \bar{y}_i, \ k = 1, \ldots, n-1, \ \text{for each } P \in \mathcal{P}_n. \tag{7.2.8}$$

Clearly, $\mathcal{M}(\mathbf{y})$ is a closed convex set. Also $x_i \leq \bar{\mathbf{y}}_1$ for $i = 1, \ldots, n$. Hence

$$x_i = y_i + \sum_{j=1, \neq i}^{n} (y_j - x_j) \geq -(n-1)\bar{y}_1 + \sum_{j=1}^{n} y_j.$$

Thus $\mathcal{M}(\mathbf{y})$ is a compact convex set containing $Q\mathbf{y}$ for all $Q \in \mathcal{P}_n$. Hence $\mathcal{M}(\mathbf{y})$ is a polytope.

Clearly

$$P\mathcal{M}(Q\mathbf{y}) = \mathcal{M}(\mathbf{y}) \text{ for each } P, Q \in \mathcal{P}_n. \tag{7.2.9}$$

Assume that $\mathbf{x}$ is an extremal point of $\mathcal{M}(\mathbf{x})$. Then the above equality implies that $P\mathbf{x}$ is also an extreme point of $\mathcal{M}(\mathbf{x})$ for each $P \in \mathcal{P}_n$. Let $P\mathbf{x} = \bar{\mathbf{x}}, Q\mathbf{y} = \bar{\mathbf{y}}$ for some $P, Q \in \mathcal{P}_n$. We claim that $\bar{\mathbf{x}} = \bar{\mathbf{y}}$. Without loss of generality we consider the case that $\mathbf{x} = \bar{\mathbf{x}}, \mathbf{y} = \bar{\mathbf{y}}$. We prove this claim by induction on n. For $n = 1$ this claim is trivial. Assume that this claim holds for any $m \le n - 1$.

Let $m = n$. Assume to the contrary that $\mathbf{x} \ne \mathbf{y}$. Suppose first that for some $1 \le k \le n - 1$ we have the equality $\sum_{i=1}^k x_i = \sum_{i=1}^k y_i$. Let

$$\mathbf{x}_1 = (x_1, \ldots, x_k)^\top, \mathbf{y}_1 = (y_1, \ldots, y_k)^\top \in \mathbb{R}^k,$$
$$\mathbf{x}_2 = (x_{k+1}, \ldots, x_n)^\top, \mathbf{y}_2 = (y_{k+1}, \ldots, y_n)^\top \in \mathbb{R}^{n-k}.$$

Then $\mathbf{x}_1 \prec \mathbf{y}_1, \mathbf{x}_2 \prec \mathbf{y}_2$. Use the induction hypothesis that $\mathbf{x}_i = \mathbf{y}_i, i = 1, 2$. Hence $\mathbf{x} = \mathbf{y}$ contrary to our assumption.

It is left to consider the case where strict inequalities hold in (7.2.5) for $k = 1, \ldots, n - 1$. In particular $y_1 > y_n$ and $y_1 > x_1, x_n > y_n$. Assume first that $x_1 = \ldots = x_n$. Then $\mathbf{x} = \sum_{i=1} \frac{1}{n!} P\mathbf{y}$ and $\mathbf{x}$ can not be an extremal point in $\mathcal{M}(\mathbf{y})$ contrary to our assumption. Hence, there exists and integer $k \in [1, n-1]$ such that $x_1 = \ldots = x_k > x_{k+1}$. For $t \in \mathbb{R}$ define $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^\top$, where

$$x_i(t) = x_i + t \text{ for } i = 1, \ldots, k, \quad x_i(t) = x_i - \frac{k}{n-k} t \text{ for } i = k+1, \ldots, n.$$

It is straightforward to see that there exists $\varepsilon > 0$ such that for each $t \in [\varepsilon, -\varepsilon]$ $\mathbf{x}(t) \in \mathbb{R}^n_{\searrow} \cap \mathcal{M}(\mathbf{y})$. As $\mathbf{x} = \frac{1}{2}\mathbf{x}(\varepsilon) + \frac{1}{2}\mathbf{x}(-\varepsilon)$ we deduce that $\mathbf{x}$ is not an extremal point, contrary to our assumption. Hence $\mathcal{E}(\mathcal{M}(\mathbf{y})) = \cup_{P \in \mathcal{P}_n} P\mathbf{y}$.

Let $\mathbf{x} \in \mathcal{M}(\mathbf{y})$. Then

$$\mathbf{x} = \sum_{P \in \mathcal{P}_n} a_P (P\mathbf{y}) = (\sum_{P \in \mathcal{P}_n} a_P P)\mathbf{y}, \tag{7.2.10}$$
$$\text{where } a_P \ge 0, P \in \mathcal{P}_n, \sum_{P \in \mathcal{P}_n} a_P = 1.$$

Hence $\mathbf{x} = A\mathbf{y}$ for a corresponding $A \in \Omega_n$. Vice versa, any $A \in \Omega_n$ is a convex combination of the permutation matrices. Therefore $A\mathbf{x} \in \mathcal{M}(\mathbf{y})$ for any doubly stochastic matrix. □

Problems

1. Prove Lemma 7.2.3.

2. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Show that $\mathbf{x} \prec \mathbf{y} \iff -\mathbf{y} \prec -\mathbf{x}$.

3. Let $m, n \in \mathbb{N}$. Denote by $\Omega_{m,n} \subset \mathbb{R}_+^{m\times n}$ the set of stochastic matrices, i.e. each row is a probability vector, such each column has sum $\frac{m}{n}$. Assume that $m \neq n$. Show

   (a) $\Omega_{m,n}$ is a polytope.

   (b) $\dim \Omega_{m,n} = (m-1)(n-1)$.

   (c) Each extreme point of $\Omega_{m,n}$ at most $m+n-1$ nonzero elements.

4. Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Recall that one needs $O(n \log n)$ swaps to obtain the coordinates of $\bar{\mathbf{x}}$. Deduce that for a given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ one needs $O(n \log n)$ swaps and $2n^2$ additions of entries of $\mathbf{x}$ and $\mathbf{y}$ to determine if $\mathbf{x}$ is or is not in the set $\mathcal{M}(\mathbf{y})$.

7.3 Convex functions

Definition 7.3.1 *Let C be a convex set in a finite dimensional subspace $\mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. A function $\phi : C \to \mathbb{R}$ is called convex if for any $\mathbf{x}, \mathbf{y} \in C$ and $t \in [0,1]$*

$$\phi(t\mathbf{x} + (1-t)\mathbf{y}) \le t\phi(\mathbf{x}) + (1-t)\phi(\mathbf{y}). \tag{7.3.1}$$

ϕ is called strictly convex on C if for any $\mathbf{x}, \mathbf{y} \in C, \mathbf{x} \neq \mathbf{y}$ and $t \in (0,1)$ strict inequality holds in (7.3.1). A function $\psi : C \to \mathbb{R}$ is called concave or strictly concave if the function $-\psi$ is convex or strictly convex respectively.

We remark that in [Roc70] a convex function on ϕ on a convex set is allowed to have the values $\pm\infty$. To avoid the complications, we restrict our attention to convex function with finite values. The following result is well known [Roc70, Theorem 10.1]

Theorem 7.3.2 *Let C be a convex set in a finite dimensional subspace $\mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\phi : C \to \mathbb{R}$ is convex. The $\phi : \mathrm{ri}\, C \to \mathbb{R}$ is continuous.*

For any set $T \in \mathbf{V}$ we let $\mathrm{Cl}\, T$ be the closure of T in the standard topology in $\mathbf{V}$ (which is identified with the standard topology of $\mathbb{R}^{\dim_{\mathbb{R}} \mathbf{V}}$).

Proposition 7.3.3 *Let $C \subset \mathbf{V}$ be convex. Then $\mathrm{Cl}\, C$ is convex. Assume that* $\mathrm{ri}\, C$ *is an open set in $\mathbf{V}$ and $f \in \mathrm{C}^0(\mathrm{Cl\, C})$. Then f is convex in $\mathrm{Cl}\, C$ if and only if f is convex in C.*

See Problem 3.

Theorem 7.3.4 *Let C be a compact convex set in a finite dimensional subspace $\mathbf{V}$ over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\phi : C \to \mathbb{R}$ is continuous. Then*

$$\max_{\mathbf{x}\in C} \phi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{E}(C)} \phi(\mathbf{y}). \tag{7.3.2}$$

Assume in addition that ϕ is strictly convex. If ϕ achieves its maximum at $\mathbf{x}^\star$, then $\mathbf{x}^\star$ is an extreme point of ϕ.

Proof. Assume that $\max_{\mathbf{x}\in C} \phi(\mathbf{x}) = \phi(\mathbf{x}^\star)$. Suppose first that $\mathbf{x}^\star$ is an extreme point of C. Then (7.3.2) trivially holds. Assume now that $\mathbf{x}^\star$ is not an extreme point of C. Theorem 7.1.3 yields that $\mathbf{x}^\star = \sum_{i=1}^m a_i\mathbf{x}_i$, where $a_i \in (0,1), \mathbf{x}_i \in \mathcal{E}(C), i = 1, \ldots, m$ and $m \geq 2$. The convexity of ϕ and 7.3.8 yield that

$$\phi(\mathbf{x}^\star) \leq \sum_{i=1}^m a_i\phi(\mathbf{x}_i) \leq \max_{i=1,\ldots,m} \phi(\mathbf{x}_i) = \phi(\mathbf{x}_j) \text{ for some } j \in \langle m\rangle.$$

Since ϕ achieves its maximum at $\mathbf{x}^\star$ we deduce that $\phi(\mathbf{x}^\star) = \phi(\mathbf{x}_j)$. Hence (7.3.2) holds.

Suppose now that ϕ is strictly convex. Then Problem 1b implies that strict inequality holds in the above inequality. Hence $\phi(\mathbf{x}^\star) < \phi(\mathbf{x}_j)$, which contradicts the maximality $\mathbf{x}^\star$. □

Theorem 7.3.5 *Let $\mathbf{x} = (x_1, ..., x_n)^\top, \mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$ and assume that $\mathbf{x} \prec \mathbf{y}$. Let $\phi : [\bar{y}_n, \bar{y}_1] \to \mathbb{R}$ be a convex function. Then*

$$\sum_{i=1}^n \phi(x_i) \leq \sum_{i=1}^n \phi(y_i). \tag{7.3.3}$$

If ϕ is strictly convex on $[\bar{y}_n, \bar{y}_1]$ and $P\mathbf{x} \neq \mathbf{y}$ for all $P \in \mathcal{P}_n$ then strict inequality holds in the above inequality.

Proof. Define $\psi : \mathcal{M}(\mathbf{y}) \to \mathbb{R}$ by the equality $\psi((x_1, \ldots, x_n)) = \sum_{i=1}^n \phi(x_i)$. Since for any $(x_1, \ldots, x_n)^\top \in \mathcal{M}(\mathbf{y})$ $x_i \in [\bar{y}_n, \bar{y}_1]$ it follows that ψ is well defined on $\mathcal{M}(\mathbf{y})$. Clearly for each $P \in \mathcal{P}_n$ we have the equality $\psi(P\mathbf{y}) = \psi(\mathbf{y})$. (7.2.10) and the convexity of ϕ yield

$$\phi(x_i) = \phi(\sum_{P\in\mathcal{P}_n} a_P(P\mathbf{y})_i) \leq \sum_{P\in\mathcal{P}_n} a_P\phi((P\mathbf{y})_i).$$

Sum on $i = 1, \ldots, n$ to deduce that

$$\psi(\mathbf{x}) \leq \sum_{P\in\mathcal{P}_n} a_P\psi(P\mathbf{y}) = \sum_{P\in\mathcal{P}_n} a_P\psi(\mathbf{y}) = \psi(\mathbf{y}).$$

If ϕ is strictly convex and $\mathbf{x} \notin \mathcal{E}(\mathcal{M}(\mathbf{y}))$ then the above inequalities are strict. □

The following result is well known. (See Problems 2-3)

Theorem 7.3.6 *Let $C \subset \mathbf{V}$ be a convex set. Assume that* $\mathrm{ri}\, C$ *is an open set in $\mathbf{V}$ and $\phi \in \mathrm{C}^2(\mathrm{C})$. Then ϕ is convex if and only if the symmetric matrix $H(\phi) := (\frac{\partial^2 \phi}{\partial x_i x_j})_{i,j=1}^d$ is nonnegative definite for each $\mathbf{y} \in C$. Furthermore, if $H(\phi)$ is positive definite for each $\mathbf{y} \in C$ then ϕ is strictly convex.*

Let $\phi : C \to \mathbb{R}$ be a convex function. Then ϕ has the following continuity and differentiability properties: In the one dimensional case where $C = (a, b) \subset \mathbb{R}$ ϕ is continuous on C and ϕ has a derivative $\phi'(x)$ at all but a countable set of points. $\phi'(x)$ is an nondecreasing function (where defined). In particular ϕ has left and right derivatives at each x, which is given as the left and the right limits of $\phi'(x)$ (where defined).

In the general case $C \subset \mathbf{V}$, ϕ is continuous function in $\mathrm{ri}\, C$, has a differential $D\phi$ in a dense set C_1 of $\mathrm{ri}\, C$, the complement of C_1 in $\mathrm{ri}\, C$ has a zero measure, and $D\phi$ is continuous in C_1. Furthermore at each $\mathbf{x} \in \mathrm{ri}\, C$ ϕ has a subdifferential $\mathbf{f} \in \mathrm{Hom}(\mathbf{V}, \mathbb{R})$ such that

$$\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + f(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in C. \tag{7.3.4}$$

See for example [Roc70].

Definition 7.3.7 *Let $C \subset \mathbf{V}$ be a convex set. Then $f : C \to \mathbb{R}$ is called an affine function if for each $\mathbf{x}, \mathbf{y} \in C$ $f(t\mathbf{x} + (1-t)\mathbf{y}) = tf(\mathbf{x}) + (1-t)f(\mathbf{y})$ for each $t \in [0, 1]$*

Clearly an affine function f on a convex set the functions f and $-f$ are convex. Theorem 7.3.4 yield.

Corollary 7.3.8 *Let $C \subset \mathbf{V}$ be a compact convex set. Assume that $f : C \to \mathbb{R}$ be an affine function. Then*

$$\max_{\mathbf{x} \in \mathbb{C}} f(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{E}(C)} f(\mathbf{y}), \quad \min_{\mathbf{x} \in \mathbb{C}} f(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{E}(C)} f(\mathbf{y}).$$

Let $C \subset \mathbf{V}$ be a polytope. Then finding the maximum or the minimum of of an affine function on C is called the *linear programming*. It is known that the complexity of the linear programming is polynomial in the dimension of $\mathbf{V}$, and the complexity of all the half spaces in the characterizing C and f [Kha79, Kar84].

We now give a simple example. Let $A = [a_{ij}] \in \mathbb{R}^{n\times n}$. Denote by $\mathrm{S_n}$ the group of permutations $\pi : \langle n\rangle \to \langle n\rangle$. Consider the maximal problem to maximize the sum of a generalized diagonal of A:

$$\mu(A) := \max_{\pi\in \mathrm{S_n}} \sum_{i=1}^{n} a_{i\pi(i)}. \tag{7.3.5}$$

Since $\#\mathrm{S_n} = \mathrm{n!}$, a brute force algorithm to try all the permutation will need $n! \approx (\frac{n}{e})^n$ computations times. (We ignore the complexity of computing the sum $\sum_{i=1}^{n} a_{i\pi(i)}$.) However, $\mu(A)$ can be computed polynomially in n. Define an affine $f : \Omega_n \to \mathbb{R}$ by $f(X) = \operatorname{tr} AX$, for any doubly stochastic X. It is straightforward to show that

$$\mu(A) = \max_{X\in\Omega_n} \operatorname{tr} AX. \tag{7.3.6}$$

Since Ω_n is a polytope given by at most $4n + n^2$ inequalities, and the complexity of f is n^2 times the complexity of entries, we see that the complexity of computing the maximum of $f(X)$ is polynomial in n.

Definition 7.3.9 *Let $\mathbf{V}_1, \mathbf{V}_2$ be finite dimensional vector spaces. Assume that $C_1 \subset \mathbf{V}_1, C_2 \subset \mathbf{V}_2$ are convex sets. A functions $\phi : C_1 \times C_2 \to \mathbb{R}$ is called concave-convex if the functions $\phi(\cdot, \mathbf{y}) : C_1 \to \mathbb{R}, \phi(\mathbf{x}, \cdot) : C_2 \to \mathbb{R}$ are concave for each $\mathbf{y} \in C_2$ and convex for each $\mathbf{x} \in C_1$ respectively.*

The following result is known as minimax theorem [Roc70, Cor. 37.6.2].

Theorem 7.3.10 *Let C_i be a compact convex set in a finite dimensional vector space $\mathbf{V}_i$ for $= 1, 2$. Assume that $\phi : C_1 \times C_2 \to \mathbb{R}$ be a continuous concave-convex function. Then*

$$\min_{\mathbf{y}\in C_2} \max_{\mathbf{x}\in C_1} \phi(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{x}\in C_1} \min_{\mathbf{y}\in C_2} \phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}^\star, \mathbf{y}^\star) \tag{7.3.7}$$

$$\textit{for some } \mathbf{x}^\star \in C_1, \mathbf{y}^\star \in C_2.$$

The point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is called a *saddle point*. More general types of the minimax theorems are can be found in [Roc70].

Problems

1. Let $C \subseteq \mathbf{V}$ be a convex set and assume that $\phi : C \to \mathbb{R}$ is convex. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in C, m \geq 3$. Show

   (a) Let $a_1, \ldots, a_m \in [0, 1]$ and assume that $\sum_{i=1}^{m} a_i = 1$. Then

   $$\phi(\sum_{i=1}^{m} a_i x_i) \leq \sum_{i=1}^{m} a_i \phi(x_i). \tag{7.3.8}$$

(b) Assume in addition that ϕ is strictly convex, $\mathbf{x}_i \neq \mathbf{x}_j$ for $i \neq j$ and $a_1, \ldots, a_m > 0$. Then strict inequality holds in (7.3.8).

2. (a) Let $f \in \mathrm{C}^1(\mathrm{a}, \mathrm{b})$. Show that f is convex on (a, b) if and only if $f'(x)$ is nondecreasing on (a, b). Show that if $f'(x)$ is increasing on (a, b) then f is strictly convex on (a, b).

   (b) Let $f \in \mathrm{C}[\mathrm{a}, \mathrm{b}] \cap \mathrm{C}^1(\mathrm{a}, \mathrm{b})$. Show that f is convex in $[a, b]$ if and only if f is convex in (a, b). Show that if $f'(x)$ is increasing on (a, b) then f is strictly convex on $[a, b]$.

   (c) Let $f \in \mathrm{C}^2(\mathrm{a}, \mathrm{b})$. Show that f is convex on (a, b) if and only if f'' is a nonnegative function on (a, b). Show that if $f''(x) > 0$ for each $x \in (a, b)$ then f is strictly convex on (a, b).

   (d) Prove Theorem 7.3.6.

3. Prove Proposition 7.3.3.

4. Let $C_i \subset \mathbf{V}_i$ be a compact set in a finite dimensional vector space for $i = 1, 2$. Let $\phi : C_1 \times C_2 \to \mathbb{R}$ be a continuous function. Show the inequality

$$\min_{\mathbf{y} \in C_2} \max_{\mathbf{x} \in C_1} \phi(\mathbf{x}, \mathbf{y}) \geq \max_{\mathbf{x} \in C_1} \min_{\mathbf{y} \in C_2} \phi(\mathbf{x}, \mathbf{y}). \tag{7.3.9}$$

7.4 Norms over vector spaces

In this Chapter we assume that $\mathbb{F} = \mathbb{R}, \mathbb{C}$ unless stated otherwise.

Definition 7.4.1 *Let* $\mathbf{V}$ *be a vector space over* $\mathbb{F}$. *A continuous function* $\| \cdot \| : \mathbf{V} \to [0, \infty)$ *is called a norm if the following conditions are satisfied:*

1. *Positivity:* $\|\mathbf{v}\| = 0$ *if and only if* $\mathbf{v} = \mathbf{0}$.

2. *Homogeneity:* $\|a\mathbf{v}\| = |a| \, \|\mathbf{v}\|$ *for each* $a \in \mathbb{F}$ *and* $\mathbf{v} \in \mathbf{V}$.

3. *Subadditivity:* $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ *for all* $\mathbf{u}, \mathbf{v} \in \mathbf{V}$.

A continuous function $\| \cdot \| : \mathbf{V} \to [0, \infty)$ *which satisfies the conditions 2 and 3 is called a seminorm. The sets*

$$\mathrm{B}_{\|\cdot\|} := \{\mathbf{v} \in \mathbf{V}, \ \|\mathbf{v}\| \leq 1\}, \ \mathrm{B}^o_{\|\cdot\|} := \{\mathbf{v} \in \mathbf{V}, \ \|\mathbf{v}\| < 1\},$$
$$\mathrm{S}_{\|\cdot\|} := \{\mathbf{v} \in \mathbf{V}, \ \|\mathbf{v}\| = 1\},$$

are called the (closed) unit ball, the open unit ball and the unit sphere of the norm respectively. For $\mathbf{a} \in \mathbf{V}$ *and* $r > 0$ *we let*

$$\mathrm{B}_{\|\cdot\|}(\mathbf{a}, r) = \{\mathbf{x} \in \mathbf{V} : \ \|\mathbf{x} - \mathbf{a}\| \leq r\}, \quad \mathrm{B}^o_{\|\cdot\|}(\mathbf{a}, r) = \{\mathbf{x} \in \mathbf{V} : \ \|\mathbf{x} - \mathbf{a}\| < r\}$$

be the closed and the open ball of radius r *centered at* $\mathbf{a}$ *respectively. If the norm* $\|\cdot\|$ *is fixed, we use the notation*

$$\mathrm{B}(\mathbf{a}, r) = \mathrm{B}_{\|\cdot\|}(\mathbf{a}, r), \quad \mathrm{B}^o(\mathbf{a}, r) = \mathrm{B}^o_{\|\cdot\|}(\mathbf{a}, r).$$

See Problem 2 for the properties of unit balls. The standard norms on $\mathbb{F}^n$ are the l_p norms:

$$\begin{aligned} \|(x_1, \ldots, x_n)^\top\|_p &= \big(\sum_{i=1}^n |x_i|^p\big)^{\frac{1}{p}}, \quad p \in [1, \infty), \\ \|(x_1, \ldots, x_n)^\top\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned} \tag{7.4.1}$$

See Problem 8.

Definition 7.4.2 *Let* $\mathbf{V}$ *be a finite dimensional vector space over* $\mathbb{F}$*. Denote by* $\mathbf{V}^*$ *the set of all linear functionals* $f : \mathbf{V} \to \mathbb{F}$*. Assume that* $\|\cdot\|$ *is a norm on* $\mathbf{V}$*. The conjugate norm* $\|\cdot\| : \mathbf{V}^* \to \mathbb{F}$ *is defined as*

$$\|\mathbf{f}\|^* = \max_{\mathbf{x} \in \mathrm{B}_{\|\cdot\|}} |\mathbf{f}(\mathbf{x})|, \ \textit{for } \mathbf{f} \in \mathbf{V}^*.$$

For a norm $\|\cdot\|$ *on* $\mathbb{F}^n$ *the conjugate norm* $\|\cdot\|$ *on* $\mathbb{F}^n$ *is given by*

$$\|\mathbf{x}\|^* = \max_{\mathbf{y} \in \mathrm{B}_{\|\cdot\|}} |\mathbf{y}^\top \mathbf{x}| \ \textit{for } \mathbf{x} \in \mathbb{F}^n. \tag{7.4.2}$$

A norm $\|\cdot\|$ *on* $\mathbf{V}$ *is called strictly convex if for any two distinct points* $\mathbf{x}, \mathbf{y} \in S_{\|\cdot\|}$ *and* $t \in (0, 1)$ *the inequality* $\|t\mathbf{x} + (1 - t)\mathbf{y}\| < 1$ *holds. A norm* $\|\cdot\|$ *on* $\mathbb{F}^n$ *is called* C^k*, for* $k \in \mathbb{N}$*, if the sphere* $S_{\|\cdot\|}$ *is a* C^k *manifold.* $\|\cdot\|$ *is called smooth if it is* C^k *for each* $k \in \mathbb{N}$*.*

For $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{C}^n$ *let* $\mathrm{abs}\,\mathbf{x} = (|x_1|, \ldots, |x_n|)^\top$*. A norm* $\|\cdot\|$ *on* $\mathbb{F}^n$ *is called absolute if* $\|\mathbf{x}\| = \|\mathrm{abs}\,\mathbf{x}\|$ *for each* $\mathbf{x} \in \mathbb{F}^n$*. A norm* $|||\cdot|||$ *on* $\mathbb{F}^n$ *is called a transform absolute if there exists an absolute norm* $\|\cdot\|$ *on* $\mathbb{F}^n$ *and* $P \in \mathbf{GL}(n, \mathbb{F})$ *such that* $|||\mathbf{x}||| = \|P\mathbf{x}\|$ *for each* $\mathbf{x} \in \mathbb{F}^n$*.*

A norm $\|\cdot\|$ *on* $\mathbb{F}^n$ *is called symmetric if the function* $\|(x_1, \ldots, x_n)^\top\|$ *is a symmetric function in* $x_1, \ldots, x_n$*. (I.e. for each permutation* $\pi : \langle n \rangle \to \langle n \rangle$ *and each* $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{F}^n$ *equality* $\|(x_{\pi(1)}, \ldots, x_{\pi(n)})^\top\| = \|(x_1, \ldots, x_n)^\top\|$ *holds.*

Theorem 7.4.3 *Let $\|\cdot\|$ be a norm on $\mathbb{F}^n$. Then the following are equivalent.*

1. *$\|\cdot\|$ is an absolute norm.*

2. *$\|\cdot\|^*$ is an absolute norm.*

3. *There exists a compact set $L \subset \mathbb{F}^n$ not contained in the hyperplane $H_i = \{(y_1, \ldots, y_n)^\top \in \mathbb{F}^n,\ y_i = 0\}$ for $i = 1, \ldots, n$, such that $||\mathbf{x}|| = \max_{\mathbf{y}\in L}(\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x}$ for each $\mathbf{x} \in \mathbb{F}^n$.*

4. *$\|\mathbf{x}\| \le \|\mathbf{z}\|$ if* $\mathrm{abs}\,\mathbf{x} \le \mathrm{abs}\,\mathbf{z}$.

Proof. *1⇒2.* Assume that $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$. Then there exists $\mathbf{z} \in \mathbb{F}^n, \mathrm{abs}\,\mathbf{z} = \mathrm{abs}\,\mathbf{y}$ such that $|\mathbf{z}^\top \mathbf{x}| = (\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x}$. Since $\|\cdot\|$ is absolute $\|\mathbf{z}\| = \|\mathbf{y}\|$. Clearly $|\mathbf{y}^\top \mathbf{x}| \le (\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x}$. The characterization (7.4.2) yields that

$$||\mathbf{x}||^* = \max_{\mathbf{y}\in \mathrm{B}_{\|\cdot\|}} (\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x}. \tag{7.4.3}$$

Clearly $\|\mathbf{x}\| = \|\mathrm{abs}\,\mathbf{x}\|$.

2⇒3. The equality $(\|\cdot\|^*)^* = \|\cdot\|$, see Problem 3, and the equality (7.4.3) implies *3* with $L = \mathrm{B}_{\|\cdot\|^*}$. Clearly $\mathrm{B}_{\|\cdot\|^*}$ contains a vector whose all coordinates are different from zero.

3⇒4. Assume that $\mathrm{abs}\,\mathbf{x} \le \mathrm{abs}\,\mathbf{z}$. Then $(\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x} \le (\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{z}$ for any $\mathbf{y} \in \mathbb{F}^n$. In view of the characterization of the absolute norm given in *3* we deduce *4*.

4⇒1. Assume that $\mathrm{abs}\,\mathbf{x} = \mathrm{abs}\,\mathbf{y}$. Since $\mathrm{abs}\,\mathbf{x} \le \mathrm{abs}\,\mathbf{y}$ we deduce that $\|\mathbf{x}\| \le \|\mathbf{y}\|$. Similarly $\|\mathbf{x}\| \ge \|\mathbf{y}\|$. Hence *1* holds. □

Definition 7.4.4 *A set $L \subset \mathbb{F}^n$ is called symmetric if for each $\mathbf{y} = (y_1, \ldots, y_n)^\top$ in L the vector $(y_{\pi(1)}, \ldots, y_{\pi(n)})^\top$ is in L, for each permutation $\pi : \langle n\rangle \to \langle n\rangle$.*

Corollary 7.4.5 *Let $\|\cdot\|$ be a norm on $\mathbb{F}^n$. Then the following are equivalent.*

1. *$\|\cdot\|$ is an absolute symmetric norm.*

2. *$\|\cdot\|^*$ is an absolute symmetric norm.*

3. *There exists a compact symmetric set $L \subset \mathbb{F}^n$, not contained in the hyperplane $H_i = \{(y_1, \ldots, y_n)^\top \in \mathbb{F}^n,\ y_i = 0\}$ for $i = 1, \ldots, n$, such that $||\mathbf{x}|| = \max_{\mathbf{y}\in L}(\mathrm{abs}\,\mathbf{y})^\top \mathrm{abs}\,\mathbf{x}$ for each $\mathbf{x} \in \mathbb{F}^n$.*

See Problem 6.

Proposition 7.4.6 *Assume that $\|\cdot\|$ is a symmetric absolute norm on $\mathbb{R}^2$. Then*

$$\|\mathbf{x}\|_2^2 \le \|\mathbf{x}\|^*\|\mathbf{x}\| \le \|\mathbf{x}\|_\infty\|\mathbf{x}\|_1 \text{ for any } \mathbf{x} \in \mathbb{R}^2. \tag{7.4.4}$$

Proof. The lower bound follows from the Problem 9. We claim that for any two points on $\mathbf{x}, \mathbf{y}$ satisfying the condition $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ the following inequality holds

$$(\|\mathbf{x}\|_1 - \|\mathbf{y}\|_1)(\|\mathbf{x}\|_\infty - \|\mathbf{y}\|_\infty) \le 0. \tag{7.4.5}$$

Since $\|\cdot\|$ is symmetric and absolute it is enough to prove the above inequality in the case that $\mathbf{x} = (x_1, x_2)^\top, x_1 \ge x_2 \ge 0, \mathbf{y} = (y_1, y_2)^\top, y_1 \ge y_2 \ge 0$. View $\mathrm{B}_{\|\cdot\|}$ as a convex balanced set in $\mathbb{R}^2$, which is symmetric with respect to the line $z_1 = z_2$. The symmetricity of $\|\cdot\|$ implies that all the points $(z_1, z_2)^\top \in B_{\|\cdot\|}$ satisfy the inequality $z_1 + z_2 \le 2c$, where $\|(c, c)^\top\| = 1, c > 0$. Let C, D be the intersection of $\mathrm{B}_{\|\cdot\|}, \mathrm{C}_{\|\cdot\|}$ the with octant $K = \{(z_1, z_2)^\top \in \mathbb{R}^2,\ z_1 \ge z_2 \ge 0\}$ respectively. Observe that the line $z_1 + z_2 = 2c$ may intersect D at an interval. However the line $z_1 + z_2 = 2t$ will intersect D at a unique point $(z_1(t), z_2(t))^\top$ for $t \in [b, c)$, where $\|(2b, 0)\| = 1, b > 0$. Furthermore $z_1(t), -z_2(t)$ are decreasing in (b, c). Hence, if $x_1 + x_2 > y_1 + y_2$ it follows that $y_1 > x_1$. Similarly, $x_1 + x_2 < y_1 + y_2$ it follows that $y_1 < x_1$. This proves (7.4.5).

To show the right-hand side of (7.4.4) we may assume that $\|\mathbf{x}\| = 1$. So $\|\mathbf{x}\|^* = |\mathbf{y}^\top\mathbf{x}|$ for some $\mathbf{y} \in \mathrm{S}_{\|\cdot\|}$. Hence $\|\mathbf{x}\|\,\|\mathbf{x}\|^* = |\mathbf{y}^\top\mathbf{x}|$. Clearly

$$|\mathbf{y}^\top\mathbf{x}| \le \min(\|\mathbf{x}\|_1\|\mathbf{y}\|_\infty, \|\mathbf{x}\|_\infty\|\mathbf{y}\|_1).$$

Suppose that $\|\mathbf{y}\|_1 \le \|\mathbf{x}\|_1$. Then the right-hand side of (7.4.4) follows. Assume that $|\mathbf{y}|_1 > \|\mathbf{x}\|_1$. Then Then (7.4.5) yields that $\|\mathbf{y}\|_\infty \le \|\mathbf{x}\|_\infty$ and the right-hand side of (7.4.4) follows. □

A norm $\|\cdot\| : \mathbb{F}^{m\times n} \to \mathbb{R}_+$ is called a *matrix* norm. A standard example of matrix norm is the *Frobenius* norm of $A = [a_{ij}] \in \mathbb{F}^{m\times n}$:

$$\|A\|_F := \sqrt{\sum_{i=j=1}^{m,n} |a_{ij}|^2}\quad . \tag{7.4.6}$$

Recall that $\|A\|_F = (\sum_{i=1}^m \sigma_i(A)^2)^{\frac{1}{2}}$, where $\sigma_i(A), i = 1, \ldots,$ are the singular values of A. More generally, for each $q \in [1, \infty]$

$$\|A\|_{q,S} := (\sum_{i=1}^{m} \sigma_i(A)^q)^{\frac{1}{q}} \tag{7.4.7}$$

is a norm on $\mathbb{F}^{m\times n}$, which is called the q-Schatten norm of A. Furthermore, for any integer $p \in [1, m]$ and $w_1 \geq \ldots \geq w_p > 0$, the function $f(A)$ given in Corollary 4.10.5 is a norm on $\mathbb{F}^{m\times n}$. See Problem 4.10.4. We denote $\|\cdot\|_{\infty,S} = \sigma_1(\cdot)$ as the $\|\cdot\|_2$ operator norm:

$$\|A\|_{2,2} := \sigma_1(A) \text{ for } A \in \mathbb{C}^{m\times n}. \tag{7.4.8}$$

See §7.7.

Definition 7.4.7 *A norm $\|\cdot\|$ on $\mathbb{C}^{m\times n}$ is called a unitary invariant if $\|UAV\| = \|A\|$ for any $A \in \mathbb{C}^{m\times n}$ and unitary $U \in \mathbf{U}(m), V \in \mathbf{V}(n)$.*

Clearly, any p-Schatten norm on $\mathbb{C}^{m\times n}$ is unitary invariant.

Theorem 7.4.8 *For positive integers m, n let $l = \min(m, n)$. For $A \in \mathbb{C}^{m\times n}$ let $\boldsymbol{\sigma}(A) := (\sigma_1(A), \ldots, \sigma_l(A))^\top$. Then $\|\cdot\|$ is a unitary invariant norm on $\mathbb{C}^{m\times n}$ if and only if there exists an absolute symmetric norm $|||\cdot|||$ on $\mathbb{C}^l$ such that $\|A\| = |||\boldsymbol{\sigma}(A)|||$ for any $A \in \mathbb{C}^{m\times n}$.*

Proof. Let $\mathbf{D}(m, n) \subset \mathbb{C}^{m\times n}$ be the subspace of diagonal matrices. Clearly, $\mathbf{D}(m, n)$ is isomorphic to $\mathbb{C}^l$. Each $D \in \mathbf{D}(m, n)$ is of the form $\mathrm{diag}(\mathbf{x}), \mathbf{x} = (x_1, \ldots, x_l)^\top$, where $x_1, \ldots, x_l$ are the diagonal entries of D. Assume that $\|\cdot\|$ is a norm on $\mathbb{C}^{m\times n}$. Then the restriction of $\|\cdot\|$ to $\mathbf{D}(m, n)$ induces a norm $|||\cdot|||$ on $\mathbb{C}^l$ given by $|||\mathbf{x}||| := \|\mathrm{diag}(\mathbf{x})\|$. Assume now that $\|\cdot\|$ is a unitary invariant norm. For a given $\mathbf{x} \in \mathbb{C}^l$, there exists a diagonal unitary matrix such that $U \mathrm{diag}(\mathbf{x}) = \mathrm{diag}(\mathrm{abs}\, \mathbf{x})$. Hence

$$|||\mathbf{x}||| = \|\mathrm{diag}(\mathbf{x})\| = \|U \mathrm{diag}(\mathbf{x})\| = \|\mathrm{diag}(\mathrm{abs}\, \mathbf{x})\| = |||\mathrm{abs}\, \mathbf{x}|||.$$

Let $\pi : \langle l\rangle \to \langle l\rangle$ be a permutation. Denote $\mathbf{x}_\pi := (x_{\pi(1)}, \ldots, x_{\pi(l)})^\top$. Clearly there exists two permutation matrices $P \in \mathbf{U}(m), Q \in \mathbf{U}(n)$ such that $\mathrm{diag}(\mathbf{x}_\pi) = U \mathrm{diag}(\mathbf{x}) V$. Hence $|||\mathbf{x}_\pi||| = |||\mathbf{x}|||$, and $|||\cdot|||$ is absolute symmetric. Clearly, there exists unitary U, V such that $A = U \mathrm{diag}(\boldsymbol{\sigma}(A)) V$. Hence $\|A\| = |||\boldsymbol{\sigma}(A)|||$.

Assume now that $|||\cdot|||$ is an absolute symmetric norm on $\mathbb{C}^l$. Set $\|A\| = |||\boldsymbol{\sigma}(A)|||$ for any A. Clearly $\|\cdot\| : \mathbb{C}^{m\times n} \to \mathbb{R}_+$ is a continuous function, which satisfies the properties 1–2 of Definition 7.4.1. it is left to show that $\|\cdot\|$ satisfies the triangle inequality. Since $|||\cdot|||$ is an absolute symmetric and $\sigma_1(A)\ge\ldots\ge\sigma_l(A)\ge 0$, Corollary 7.4.5 yields that

$$\|A\| = \max_{\mathbf{y}=(y_1,\ldots,y_l)^\top\in L, |y_1|\ge\ldots\ge|y_l|} (\mathrm{abs}\,\mathbf{y})^\top\boldsymbol{\sigma}(A) \tag{7.4.9}$$

for a corresponding a compact symmetric set $L\subset\mathbb{C}^n$, not contained in the hyperplane $H_i=\{(y_1,\ldots,y_n)^\top\in\mathbb{F}^n,\ y_i=0\}$ for $i=1,\ldots,n$. Use Problem 4.10.6b to deduce from (7.4.9) that $\|A+B\|\le\|A\|+\|B\|$. □

Definition 7.4.9 *A norm on $\|\cdot\|$ on $\mathbb{F}^{n\times n}$ is called a spectral dominant norm if $\|A\|$ is not less than $\rho(A)$, the spectral radius of A, for every $A\in\mathbb{F}^{n\times n}$.*

Since $\sigma_1(A)\ge\rho(A)$, see (4.10.14) for $k=1$, we deduce that any q-Schatten norm is spectral dominant.

Problems

1. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Show that a seminorm $\|\cdot\|:\mathbf{V}\to\mathbb{R}_+$ is a convex function.

2. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. $X\subseteq\mathbf{V}$ is called balanced if $tX=X$ for every $t\in\mathbb{F}$ such that $|t|=1$. Identify $\mathbf{V}$ with $\mathbb{F}^{\dim\mathbf{V}}$. Then the topology on $\mathbf{V}$ is the topology induced by open sets in $\mathbb{F}^n$. Assume that $\|\cdot\|$ is a norm on $\mathbf{V}$. Show

   (a) $\mathrm{B}_{\|\cdot\|}$ is convex and compact.

   (b) $\mathrm{B}_{\|\cdot\|}$ is balanced.

   (c) $\mathbf{0}$ is an interior point of $\mathrm{B}_{\|\cdot\|}$.

3. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Let $X\subset\mathbf{V}$ be a compact convex set balanced set such $\mathbf{0}$ is its interior point. For each $\mathbf{x}\in\mathbf{V}\backslash\{\mathbf{0}\}$ let $f(\mathbf{x})=\min\{r>o:\ \frac{1}{r}\mathbf{x}\in X\}$. Set $f(\mathbf{0})=o$. Show that f is a norm on $\mathbf{V}$ whose unit ball is X.

4. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$ with a norm $\|\cdot\|$. Show

   (a) $\|\mathbf{f}\|^*=\max_{\mathbf{y}\in\mathrm{S}_{\|\cdot\|}}|\mathbf{f}(\mathbf{x})|$ for any $\mathbf{f}\in\mathbf{V}^*$.

   (b) Show that for any $\mathbf{x}\in\mathbf{V}$ and $\mathbf{f}\in\mathbf{V}^*$ the inequality $|\mathbf{f}(\mathbf{x})|\le\|\mathbf{f}\|^*\|\mathbf{x}\|$.

(c) Identify $(\mathbf{V}^*)^*$ with $\mathbf{V}$, i.e. any linear functional on $\lambda : \mathbf{V}^* \to \mathbb{F}$ is of the form $\lambda(\mathbf{f}) = f(\mathbf{x})$ for some $\mathbf{x} \in \mathbf{V}$. Then $(\|\mathbf{x}\|^*)^* = \|\mathbf{x}\|$.

(d) Let $L \subset \mathbf{V}^*$ be a compact set which contains a basis of $\mathbf{V}^*$. Define $\|\mathbf{x}\|_L = \max_{\mathbf{f}\in L} |\mathbf{f}(\mathbf{x})|$. Then $\|\mathbf{x}\|_L$ is a norm on $\mathbf{V}$.

(e) Show that $\|\cdot\| = \|\cdot\|_L$ for a corresponding compact set $L \subset \mathbf{V}^*$. Give a simple choice of L.

5. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$, $\dim \mathbf{V} \geq 1$, with a norm $\|\cdot\|$. Show

(a) $\mathcal{E}(\mathrm{B}_{\|\cdot\|}) \subset \mathrm{S}_{\|\cdot\|}$.

(b) $\mathcal{E}(\mathrm{B}_{\|\cdot\|}) = \mathrm{S}_{\|\cdot\|}$ if and only if for any $\mathbf{x} \neq \mathbf{y} \in \mathrm{S}_{\|\cdot\|}$ $(\mathbf{x}, \mathbf{y}) \subset \mathrm{B}^o_{\|\cdot\|}$.

(c) For each $\mathbf{x} \in \mathrm{S}_{\|\cdot\|}$ there exists $\mathbf{f} \in \mathrm{S}_{\|\cdot\|^*}$ such that $1 = \mathbf{f}(\mathbf{x}) \geq |\mathbf{f}(\mathbf{y})|$ for any $\mathbf{y} \in \mathrm{B}_{\|\cdot\|}$.

(d) Each $\mathbf{f} \in \mathrm{S}_{\|\cdot\|^*}$ is a proper supporting hyperplane of $\mathrm{B}_{\|\cdot\|}$ at some point $\mathbf{x} \in \mathrm{S}_{\|\cdot\|}$.

6. Prove Corollary 7.4.5.

7. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F}$. Assume that $\|\cdot\|_1, \|\cdot\|_2$ are two norms on $\mathbf{V}$. Show

(a) $\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbf{V}$ if and only if $\mathrm{B}_{\|\cdot\|_1} \supseteq \mathrm{B}_{\|\cdot\|_2}$.

(b) Show that there exists $C \geq c > 0$ such that $c\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq C\|\mathbf{x}\|_1$ for all $\mathbf{x} \in \mathbf{V}$.

8. For any $p \in [1, \infty]$ define the conjugate $p^* = q \in [1, \infty]$ to satisfy the equality $\frac{1}{p} + \frac{1}{q} = 1$. Show

(a) Hölder's inequality: $|\mathbf{y}^*\mathbf{x}| \leq (\mathrm{abs}\ \mathbf{y})^\top \mathrm{abs}\ \mathbf{x} \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_{p^*}$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n \backslash \{\mathbf{0}\}$ and $p \in [1, \infty]$. (For $p = 2$ this inequality is called the *Cauchy-Schwarz* inequality.) Furthermore, equalities hold in all inequalities if and only if $\mathbf{y} = a\mathbf{x}$ for some $a \in \mathbb{C}\backslash\{0\}$. (Prove Hölder's inequality for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$.)

(b) $\|\mathbf{x}\|_p$ is a norm on $\mathbb{C}^n$ for $p \in [1, \infty]$.

(c) $\|\mathbf{x}\|_p$ is strictly convex if and only if $p \in (1, \infty)$.

(d) For $p \in (1, \infty)$ $\mathcal{E}(\mathrm{B}_{\|\cdot\|_p}) = \mathrm{S}_{\|\cdot\|_\mathrm{p}}$.

(e) Characterize $\mathcal{E}(\mathrm{B}_{\|\cdot\|_p})$ for $p = 1, \infty$ for $\mathbb{F} = \mathbb{R}, \mathbb{C}$.

(f) For each $\mathbf{x} \in \mathbb{C}^n$ the function $\|\mathbf{x}\|_p$ is a nonincreasing function for $p \in [1, \infty]$.

9. Show that for any norm $\|\cdot\|$ on $\mathbb{F}^n$ the inequality

$$\|\mathbf{x}\|_2^2 \le \min(\|\bar{\mathbf{x}}\|^*\|\mathbf{x}\|, \|\mathbf{x}\|^*\|\bar{\mathbf{x}}\|) \text{ for any } \mathbf{x} \in \mathbb{F}^n.$$

   In particular, if $\|\cdot\|$ is absolute then $\|\mathbf{x}\|_2^2 \le \|\mathbf{x}\|^*\|\mathbf{x}\|$. (**Hint**: use the equality $\|\mathbf{x}\|_2^2 = \bar{\mathbf{x}}^\top\mathbf{x}$.)

10. Let $L \subset \mathbb{F}^n$ satisfy the assumptions of condition *3* of Theorem 7.4.3. Let $\nu(\mathbf{x}) = \max_{\mathbf{y}\in L}(\text{abs}\,\mathbf{y})^\top\text{abs}\,\mathbf{x}$ for each $\mathbf{x} \in \mathbb{F}^n$. Show that $\nu(\mathbf{x})$ is an absolute norm on $\mathbb{F}^n$.

11. Let $\|\cdot\|$ be an absolute norm on $\mathbb{R}^n$. Show that it extends in a unique way to an absolute norm on $\mathbb{C}^n$.

12. Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$.

    (a) Assume that $\|\cdot\|$ is a seminorm on $\mathbf{V}$. Let $\mathbf{W} := \{\mathbf{x} \in \mathbb{V}, \|\mathbf{x}\| = 0\}$. Show that $\mathbf{W}$ is a subspace of $\mathbf{V}$, and for each $\mathbf{x} \in \mathbb{V}$ the function $\|\cdot\|$ is a constant function on $\mathbf{x} + \mathbf{W}$.

    (b) Let $\mathbf{W}$ be defined as above. Let $\hat{\mathbf{U}}$ be the quotient space $\mathbf{V}/\mathbf{W}$. So $\hat{\mathbf{v}} \in \hat{\mathbb{V}}$ is viewed as any $\mathbf{y} \in \mathbf{v} + \mathbf{V}$ for a corresponding $\mathbf{v} \in \mathbf{V}$ Define the function $|||\cdot||| : \hat{\mathbf{V}} \to \mathbb{R}_+$ by $|||\hat{\mathbf{v}}||| = \|\mathbf{y}\|$. Show that $|||\cdot|||$ is a norm on $\hat{\mathbf{V}}$.

    (c) Let $\mathbf{U}_1, \mathbf{U}_2$ are finite dimensional vector spaces over $\mathbb{F}$. Assume that $|||\cdot||| : \mathbf{U}_1 \to \mathbb{R}_+$ is a norm. Let $\mathbf{V} = \mathbf{U}_1 \oplus \mathbf{U}_2$. Define $\|\mathbf{u}_1 \oplus \mathbf{u}_2\| = |||\mathbf{u}_1|||$ for each $\mathbf{u}_i \in \mathbf{U}_i, i = 1, 2$. Show that $\|\cdot\|$ is a seminorm on $\mathbf{V}$. Furthermore, the subspace $\mathbf{0} \oplus \mathbf{U}_2$ is the set where $\|\cdot\|$ vanishes.

13. Show that for any $A \in \mathbb{C}^{m\times n}$ $\|A\|_{2,2} = \sigma(A) = \max_{\|\mathbf{x}\|_2=1}\|A\mathbf{x}\|_2$. (**Hint**: Observe that $\|A\mathbf{x}\|_2^2 = \mathbf{x}^*(A^*A)\mathbf{x}$.)

14. For $\mathbb{F} = \mathbb{R}, \mathbb{C}$, identify $(\mathbb{F}^{m\times n})^*$ with $\mathbb{F}^{m\times n}$ by letting $\phi_A : \mathbb{C}^{m\times n} \to \mathbb{F}$ be $\text{tr}(A^\top X)$ for any $A \in \mathbb{F}^{m\times n}$. Show that for any $p \in [1, \infty]$ the conjugate of the p-Schatten norm $\|\cdot\|_{p,S}$ is the q-Schatten norm on $\mathbb{F}^{m\times n}$, where $\frac{1}{p} + \frac{1}{q} = 1$.

7.5 Numerical ranges and radii

Let $\mathrm{S}^{2\mathrm{n}-1} := \{\mathbf{x} \in \mathbb{C}^n,\ \mathbf{x}^*\mathbf{x} = 1\}$ be the unit sphere of the ℓ_2 norm on $\mathbb{C}^n$.

Definition 7.5.1 *A map ϕ from $\mathrm{S}^{2\mathrm{n}-1}$ to $2^{\mathbb{C}^n}$, the set of all subsets of $\mathbb{C}^n$, is called a ν-*map*, if the following conditions hold.*

1. *For each* $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ *the set* $\phi(\mathbf{x})$ *is a nonempty compact set.*
2. *The set* $\cup_{\mathbf{x}\in\mathrm{S}^{2\mathrm{n}-1}}\phi(\mathbf{x})$ *is compact.*
3. *Let* $\mathbf{x}_k \in \mathrm{S}^{2\mathrm{n}-1}, \mathbf{y}_k \in \phi(\mathbf{x}_k)$ *for* $k \in \mathbb{N}$. *Assume that* $\lim_{k\to\infty}\mathbf{x}_k = \mathbf{x}$ *and* $\lim_{k\to\infty}\mathbf{y}_k = \mathbf{y}$. *(Note that* $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$*.) Then* $\mathbf{y} \in \phi(\mathbf{x})$.
4. $\mathbf{y}^\top\mathbf{x} = 1$ *for each* $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ *and* $\mathbf{y} \in \phi(\mathbf{x})$.

Assume that ϕ *from* $\mathrm{S}^{\mathrm{n}-1}$ *to* $2^{\mathbb{C}^n}$ *is* ν*-map. Then for* $A \in \mathbb{C}^{n\times n}$

$$\omega_\phi(A) := \cup_{\mathbf{x}\in\mathrm{S}^{2\mathrm{n}-1}}\cup_{\mathbf{y}\in\phi(\mathbf{x})}\{\mathbf{y}^\top A\mathbf{x}\}, \tag{7.5.1}$$

$$r_\phi(A) = \max_{\mathbf{x}\in\phi(\mathbf{x}),\mathbf{y}\in\phi(\mathbf{x})}|\mathbf{y}^\top A\mathbf{x}| \tag{7.5.2}$$

are called the ϕ*-numerical range and the* ϕ*-numerical radius respectively.*

It is straightforward to show that r_ϕ is a seminorm on $\mathbb{C}^{n\times n}$, see Problem 1.

Lemma 7.5.2 *Let* $\phi : \mathrm{S}^{\mathrm{n}-1} \to 2^{\mathbb{C}^{\mathrm{n}}}$ *be a* ν*-map. Then* spec (A)*, the spectrum of A, is contained in the* ϕ*-numerical range of A. In particular* $r_\phi(A) \ge \rho(A)$.

Proof. Let $A \in \mathbb{C}^{n\times n}$ and and assume that λ is an eigenvalue of A. Then there exists an eigenvector $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. Choose $\mathbf{y} \in \phi(\mathbf{x})$. Then $\mathbf{y}^\top A\mathbf{x} = \lambda\mathbf{y}^\top\mathbf{x} = \lambda$. Hence $\lambda \in \omega_\phi(A)$. Thus $r_\phi(A) \ge |\lambda|$. □

Lemma 7.5.3 *Let* $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}_+$ *be a seminorm, which is spectral dominant, i.e.* $\|A\| \ge \rho(A)$ *for any* $A \in \mathbb{C}^{n\times n}$. *Then* $\|\cdot\|$ *is a norm on* $\mathbb{C}^{n\times n}$.

Proof. Assume to the contrary that $\|\cdot\|$ is not a norm. Hence there exists $0 \ne A \in \mathbb{C}^{n\times n}$ such that $\|A\| = 0$. Since $0 = \|A\| \ge \rho(A)$ we deduce that A is a nonzero nilpotent matrix. Hence, $T^{-1}AT = \oplus_{i=1}^k J_i$, where each J_i a nilpotent Jordan block and $T \in \mathbf{GL}(n,\mathbb{C})$. Since $A \ne 0$ we may assume that $J_1 \in \mathbb{C}^{l\times l}$, has an upper diagonal equal to 1, all other entries equal to 0 and $l \ge 2$. Let $B = \oplus_{i=1}^k B_i$ where each B_i has the same dimensions as J_i. Assume that B_i are zero matrices for $i > 1$, if $k > 1$. Let $B_1 = [b_{ij,1}] \in \mathbb{C}^{l\times l}$, where $b_{21,1} = 1$ and all other entries of B_1 equal to 0. It is straightforward to show that the matrix $\oplus_{i=1}^k(J_i + tB_i)$ has two nonzero eigenvalues $\pm\sqrt{t}$ for $t > 0$. Let $C := T(\oplus_{i=1}^k B_i)T^{-1}$. Then $\rho(A+tB) = \sqrt{t}$ for $t > 0$. Hence for $t > 0$ we obtain the inequalities

$$\rho(A+tB) = \sqrt{t} \le \|A+tB\| \le \|A\| + \|tB\| = t\|B\| \Rightarrow \|B\| \ge \frac{1}{\sqrt{t}}.$$

The above inequality cannot hold for an arbitrary small positive t. This contradiction implies the lemma. □

Use the above Lemmas and Problem 1 to deduce.

Theorem 7.5.4 *Let $\phi : \mathrm{S}^{\mathrm{n}-1} \to 2^{\mathbb{C}^{\mathrm{n}}}$ be a ν-map. Then $r_\phi(\cdot)$ is a spectral dominant norm on $\mathbb{C}^{n\times n}$.*

We now consider a few examples of ν-maps.

Example 7.5.5 *The function $\phi_2 : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^{\mathrm{n}}}$ given by $\phi_2(\mathbf{x}) := \{\bar{\mathbf{x}}\}$ is a ν-map. The corresponding numerical range and numerical radius of $A \in \mathbb{C}^{n\times n}$ are given by*

$$\omega_2(A) = \{z = \mathbf{x}^* A\mathbf{x}, \text{ for all } \mathbf{x} \in \mathbb{C}^n \text{ satisfying } \mathbf{x}^*\mathbf{x} = 1\} \subset \mathbb{C},$$
$$r_2(A) := \max_{\mathbf{x}\in\mathbb{C}^n, \mathbf{x}^*\mathbf{x}=1} |\mathbf{x}^* A\mathbf{x}|.$$

It is called the classical numerical range and numerical radius of A, or simply the numerical range and numerical radius of A.

More general

Example 7.5.6 *For $p \in (1,\infty)$ the function $\phi_p : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^{\mathrm{n}}}$ given by $\phi_p((x_1,\ldots,x_n)^\top) := \{\|\mathbf{x}\|_p^{-p}(|x_1|^{p-2}\bar{x}_1,\ldots,|x_n|^{p-2}\bar{x}_n)^\top\}$ is a ν-map. The corresponding numerical range and numerical radius of $A \in \mathbb{C}^{n\times n}$ are denoted by $\omega_p(A)$ and $r_p(A)$ respectively.*

The most general example related to a norm on $\mathbb{C}^n$ is as follows.

Example 7.5.7 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^n$. For each $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ let $\phi_{\|\cdot\|}(\mathbf{x})$ be the set of all $\mathbf{y} \in \mathbb{C}^n$ with the dual norm $\|\mathbf{y}\|^* = \frac{1}{\|\mathbf{x}\|}$ satisfying $\mathbf{y}^\top\mathbf{x} = 1$. Then $\phi_{\|\cdot\|}$ is a ν-map. (See Problem 6.) The corresponding numerical range $\omega_{\|\cdot\|}(A)$ and the numerical radius $r_{\|\cdot\|}(A)$ is called the Bauer numerical range and the Bauer numerical radius respectively of $A \in \mathbb{C}^{n\times n}$.*

Definition 7.5.8 *A norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ is called stable if there exists $K > 0$ such that $\|A^m\| \le K\|A\|^m$ for all $A \in \mathbb{C}^{n\times n}$.*

Clearly, $\|\cdot\|$ is stable if and only if the unit ball $\mathrm{B}_{\|\cdot\|} \subset \mathbb{C}^{n\times n}$ is power bounded, see Definition 3.4.5.

Theorem 7.5.9 *Let $\phi : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^{\mathrm{n}}}$ be a ν-map. Set $c := \max_{U\in\mathbf{U}(n)} r_\phi(U)$. Then*

$$\|(zI - A)^{-1}\|_2 \le \frac{c}{|z|-1} \text{ for all } |z| > 1,\ r_\phi(A) \le 1. \tag{7.5.3}$$

In particular, a ϕ-numerical radius is a stable norm.

Proof. Fix $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ We first note that $\|\mathbf{y}\|_2 \le c$ for each $\mathbf{y} \in \phi(\mathbf{x})$. Let $\mathbf{z} = \frac{1}{\|\mathbf{y}\|_2}\bar{\mathbf{y}} \in \mathrm{S}^{2\mathrm{n}-1}$. Then there exists $U \in \mathbf{U}(n)$ such that $U\mathbf{x} = \mathbf{z}$. Hence $\|\mathbf{y}\|_2 = \mathbf{y}^\top U\mathbf{x} \le r_\phi(U) \le c$. Assume next that $r_\phi(A) \le 1$. Hence $\rho(A) \le r_\phi(A) \le 1$. So $(zI - A)^{-1}$ is defined for $|z| > 1$. Let $\mathbf{v} := (zI - A)^{-1}\mathbf{x}, \mathbf{v}_1 := \frac{1}{\|\mathbf{v}\|_2}\mathbf{v}$. Then for $\mathbf{y} \in \phi(\mathbf{v}_1)$ we have

$$\frac{|\mathbf{y}^\top \mathbf{x}|}{\|\mathbf{v}\|_2} = |\mathbf{y}^\top(zI - A)\mathbf{v}_1| = |z - \mathbf{y}^\top A\mathbf{v}_1| \ge |z| - 1.$$

On the other hand

$$|\mathbf{y}^\top \mathbf{x}| \le \|\mathbf{y}\|_2\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 \le c.$$

Combine the above inequalities to deduce $\|(zI - A)\mathbf{x}\|_2 \le \frac{c}{|z|-1}$ for all $\|\mathbf{x}\|_2 = 1$. Use Problem 7.4.13 to deduce (7.5.3). Theorem 3.4.9 yields that the unit ball corresponding to the norm $r_\phi(\cdot)$ is a power bounded set, i.e. the norm $r_\phi(\cdot)$ is stable. □

Theorem 7.5.10 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^{n\times n}$. Then $\|\cdot\|$ is stable if and only if it is spectral dominant.*

Proof. Assume first that $\|\cdot\|$ is stable. So $B_{\|\cdot\|}$ is a power bounded set. Theorem 3.3.2 yields that each $A \in B_{\|\cdot\|}$ satisfies $\rho(A) \le 1$. So if $A \ne 0$ we get that $\rho(\frac{1}{\|A\|}A) \le 1$, i.e. $\rho(A) \le \|A\|$ for any $A \ne 0$. Clearly $\rho(0) = \|0\| = 0$. Hence a stable norm is spectral dominant.

Assume now that $\|\cdot\|$ is a spectral dominant norm on $\mathbb{C}^{n\times n}$. Recall that $\mathrm{B}_{\|\cdot\|}$ is a convex compact balanced set, and $\mathbf{0}$ is an interior point. Define a new set

$$\mathcal{A} := \{B \in \mathbb{C}^{n\times n},\ B = (1-a)A + zI,\ a \in [0,1],\ z \in \mathbb{C}, |z| \le a,\ A \in \mathrm{B}_{\|\cdot\|}\}.$$

It is straightforward to show that $\mathcal{A}$ is a convex compact balanced set. Note that by choosing $a = 1$ we deduce that $I \in \mathcal{A}$. Furthermore, by choosing $a = 0$ we deduce that $\mathrm{B}_{\|\cdot\|} \subseteq \mathcal{A}$. So $\mathbf{0}$ is an interior point of $\mathcal{A}$. Problem 7.4.3 yields that there exists a norm $|||\cdot|||$ on $\mathbb{C}^{n\times n}$ such that $\mathrm{B}_{|||\cdot|||} = \mathcal{A}$. Since $\mathrm{S}_{\|\cdot\|} \subset \mathrm{B}_{|||\cdot|||}$ it follows $|||A||| \le \|A\|$ for each $A \in \mathbb{C}^{n\times n}$. We claim that $|||\cdot|||$ is spectral dominant. Assume that $|||B||| = 1$. So $B = (1-a)A + zI$ for some $a \in [0,1]$, $z \in \mathbb{C}, |z| \le a$ and $A \in \mathrm{B}_{\|\cdot\|}$. Since $\|\cdot\|$ is spectral dominant it follows that $\rho(A) \le \|A\| \le 1$. Note that spec (B) $= (1-\mathrm{a})$spec (A) $+ \mathrm{z}$. Hence $\rho(B) \le (1-a)\rho(A) + |z| \le (1-a) + a = 1$. So $|||\cdot|||$ is spectral dominant. Since $|||I|||| \le 1$ and $|||I||| \ge \rho(I) = 1$ we deduce that $|||I||| = 1$. Hence, for any $z \in \mathbb{C}, |z| < 1$ we have $|||zI||| < 1$.

For $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ let

$$\mathrm{C}(\mathbf{x}) = \{\mathbf{u} \in \mathbb{C}^n,\ \mathbf{u} = B\mathbf{x},\ |||B||| < 1\}.$$

Clearly $\mathrm{C}(\mathbf{x})$ is a convex set in $\mathbb{C}^n$. Since for each $|||B||| < 1$ we have that $\rho(B) < 1$ it follows that $\mathbf{x} \not\in \mathrm{C}(\mathbf{x})$. The hyperplane separation theorem Theorem 7.1.6 implies the existence of $\mathbf{y} \in \mathbb{C}^n$ such that

$$\Re(\mathbf{y}^\top \mathbf{x}) > \Re(\mathbf{y}^\top B\mathbf{x}) \text{ for all } |||B||| < 1. \tag{7.5.4}$$

Substitute in the above inequality $B = zI, |z| < 1$ we deduce that $\Re(\mathbf{y}^\top \mathbf{x}) > \Re(z\mathbf{y}^\top \mathbf{x})$. By choosing an appropriate argument of z we deduce $\Re(\mathbf{y}^\top \mathbf{x}) > |z||\mathbf{y}^\top \mathbf{x}|$. Hence $\Re(\mathbf{y}^\top \mathbf{x}) \ge |\mathbf{y}^\top \mathbf{x}|$. In view of the strict inequality in (7.5.4) we deduce that $\mathbf{y}^\top \mathbf{x}$ is real and positive. Thus we can renormalize $\mathbf{y}$ so that $\mathbf{y}^\top \mathbf{x}$. Let $\phi(\mathbf{x})$ be the set of all $\mathbf{w} \in \mathbb{C}^n$ such that

$$\mathbf{w}^\top \mathbf{x} = 1, \quad \max_{|||B||| \le 1} |\mathbf{w}^\top B\mathbf{x}| = 1.$$

Clearly, $\mathbf{y} \in \phi(\mathbf{x})$. It is straightforward to show that $\phi : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^\mathrm{n}}$ is a ν-map.

As $\mathbf{w}^\top B\mathbf{x} = \operatorname{tr} B(\mathbf{x}\mathbf{w}^\top)$ we deduce that $|||\mathbf{x}\mathbf{w}^\top|||^* = 1$, where $||| \cdot |||^*$ is the dual norm of $||| \cdot |||$ on $\mathbb{C}^{n\times n}$:

$$|||C|||^* = \max_{B \in \mathrm{B}_{|||\cdot|||}} |\operatorname{tr} BC| = \max_{B \in \mathrm{S}_{|||\cdot|||}} |\operatorname{tr} BC|. \tag{7.5.5}$$

Let $\mathcal{R}(1, n, n) \subset \mathbb{C}^{n\times n}$ be the variety of all matrices of rank one at most. Clearly, $\mathcal{R}(1, n, n)$ is a closed set consisting of all matrices of rank one and $0_{n\times n}$. Hence $\mathcal{R}(1, n, n) \cap \mathrm{S}_{|||\cdot|||^*}$ is a compact set consisting of all $\mathbf{x}\mathbf{w}^\top$, where $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ and $\mathbf{w} \in \phi(\mathbf{x})$. Since $(||| \cdot |||^*)^* = ||| \cdot |||$ it follows that

$$r_\phi(B) = \max_{\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}, \mathbf{w} \in \phi(\mathbf{x})} |\mathbf{w}^\top B\mathbf{x}| = \max_{\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}, \mathbf{w} \in \phi(\mathbf{x})} |\operatorname{tr} B(\mathbf{x}\mathbf{w}^\top)| \le$$
$$\max_{C \in \mathrm{S}_{|||\cdot|||^*}} |\operatorname{tr} BC| = |||B||| \le \|B\|.$$

Hence $\mathrm{B}_{\|\cdot\|} \subseteq \mathrm{B}_{|||\cdot|||} \subseteq \mathrm{B}_{r_\phi(\cdot)}$. Theorem 7.5.9 yields that $r_\phi(\cdot)$ is a stable norm. Hence $\| \cdot \|$ and $||| \cdot |||$ are stable norms. □

Use Theorem 7.5.10 and Problem 3 to deduce.

Corollary 7.5.11 *Let $\mathcal{A} \subset \mathbb{C}^{n\times n}$ be a compact, convex, balanced set, whose interior contains $\mathbf{0}$. Then $\mathcal{A}$ is stable if and only $\rho(A) \le 1$ for each $A \in \mathcal{A}$.*

Definition 7.5.12 *Let $\mathbb{F}$ be field. A subspace $\mathbf{0} \neq \mathbf{U} \subset \mathbb{F}^{n\times n}$ is called stable if there exists an integer $k \in [1, n]$ such that the dimension of the subspace $\mathbf{U}\mathbf{x} \subset \mathbb{F}^n$ is k for any $\mathbf{0} \neq \mathbf{x} \in \mathbb{F}^n$. $\mathbf{U}$ is called maximally stable if $k = n$.*

The following result is a generalization of Theorem 7.5.10.

Theorem 7.5.13 *Let $\mathcal{A} \subset \mathbb{C}^{n\times n}$ be a compact convex balanced set, (see Problem 7.4.2 for the definition of a convex balanced set), which contains the identity matrix. Assume that $\mathcal{L} :=$ span $\mathcal{A}$ is a stable subspace. Then $\mathcal{A}$ is a stable set if and only if $\rho(A) \le 1$ for each $A \in \mathcal{A}$.*

Proof. Clearly, if $\mathcal{A}$ is stable, then each $A \in \mathcal{A}$ is power bounded, hence $\rho(A) \le 1$. Assume now that $\mathcal{A}$ is a compact convex balanced set containing identity such that $\mathcal{L}$ is a stable subspace. Let $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ and consider the subspace $\mathcal{L}\mathbf{x}$ of dimension k. Since $\mathcal{A}$ is a compact convex balanced set it follows that $\mathcal{A}\mathbf{x}$ is a compact convex balanced set in $\mathcal{L}$ since span $\mathcal{A}\mathbf{x} = \mathcal{L}\mathbf{x}$ it follows that ri $\mathcal{A}\mathbf{x}$. Hence $\mathcal{A}\mathbf{x}$ is a unit ball of the norm $\|\cdot\|_{\mathbf{x}}$ on the subspace $\mathcal{L}\mathbf{x}$. Since $I \in \mathcal{A}$ it follows that $\xi \in \mathcal{A}\mathbf{x}$. We claim that $\|\mathbf{x}\|_{\mathbf{x}} = 1$. Assume to the contrary $\|\mathbf{x}\|_{\mathbf{x}} < 1$. Then $(1+\varepsilon)\mathbf{x} \in \mathcal{A}\mathbf{x}$ for some $\varepsilon > 0$. So there exists $A \in \mathcal{A}$ such that $A\mathbf{x} = (1+\varepsilon)\mathbf{x}$. Hence $\rho(A) \ge (1+\varepsilon)$ contrary to our assumptions.

Identify $(\mathcal{L}\mathbf{x})^*$ with $\mathcal{L}\mathbf{x}$. So a linear functional $f : \mathcal{L}\mathbf{x} \to \mathbb{C}$ is given by $f(\mathbf{y}) = \mathbf{z}^\top\mathbf{y}$ for some $\mathbf{z} \in \mathcal{L}\mathbf{x}$. Let $\|\cdot\|_{\mathbf{x}}^*$ be the conjugate norm on $\mathcal{L}\mathbf{x}$. Denote by $\mathcal{B}(\mathbf{x}) \subset \mathcal{L}\mathbf{x}$ the unit ball of the norm $\|\cdot\|_{\mathbf{x}}^*$. Since $\|\mathbf{x}\|_{\mathbf{x}}^* = 1$ it follows that there exists $\mathbf{z}(\mathbf{x}) \in \mathcal{L}\mathbf{x}$ such that $\mathbf{z}(\mathbf{x})^\top\mathbf{x} = 1$ and $\|\mathbf{z}(\mathbf{x})\|_{\mathbf{x}}^* = 1$. We claim that $\cup_{\mathbf{x}\in\mathrm{S}^{2\mathrm{n}-1}}\mathcal{B}(\mathbf{x})$ is a compact set in $\mathbb{C}^n$.

Indeed, since $\mathcal{L}\mathbf{x}$ has a fixed dimension k for each for each $\mathbf{x} \in \mathbb{C}^n$ we can view $\mathcal{L}\mathbf{x}$ of the form $U(\mathbf{x})\mathbf{W}$ for some fixed subspace $\mathbf{W} \subset \mathbb{C}^n$ of dimension k and a unitary matrix $U(\mathbf{x})$. ($U(\mathbf{x})$ maps an orthonormal basis of $\mathbf{W}$ to an orthonormal basis of $\mathcal{L}\mathbf{x}$.) Hence the set $\mathcal{C}(\mathbf{x}) := U^*(\mathbf{x})\mathcal{A}\mathbf{x}$ is a compact convex balanced set in $\mathbf{W}$, with $\mathbf{0}$ an interior point. Since $\mathcal{A}$ is compact, it follows that $\mathcal{C}(\mathbf{x})$ varies continuously on $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$. Therefore the set $\mathcal{D}(\mathbf{x}) := U^*\mathcal{B}(\mathbf{x}) \subset \mathbf{W}$ varies continuously with $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$. Hence $\cup_{\mathbf{x}\in\mathrm{S}^{2\mathrm{n}-1}}\mathcal{D}(\mathbf{x})$ is a compact set in $\mathbf{W}$, which yields that $\cup_{\mathbf{x}\in\mathrm{S}^{2\mathrm{n}-1}}\mathcal{B}(\mathbf{x})$ is a compact set in $\mathbb{C}^n$. In particular, there exists a constant K such that $\|\mathbf{z}(\mathbf{x})\|_2 \le K$. We now claim that $\mathcal{A}$ satisfies the condition 3.4.13 of Theorem 3.4.9. Indeed, for $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}1-1}, \mathrm{A} \in \mathcal{A}$ we have that $A\mathbf{x} \in \mathcal{A}(\mathbf{x})$, hence $\|A\mathbf{x}\|_{\mathbf{x}} \le 1$. Hence for $|\lambda| > 1$ we have

$$\begin{aligned}\|(\lambda I - A)\mathbf{x}\|_2 \ge \frac{\|\mathbf{z}(\mathbf{x})^\top(\lambda I - A)\mathbf{x}\|_2}{K} = \frac{|\lambda - \mathbf{z}(\mathbf{x})^\top A\mathbf{x}|}{K} \ge \\ \frac{|\lambda| - |\mathbf{z}(\mathbf{x})^\top A\mathbf{x}|}{K} \ge \frac{|\lambda| - |\mathbf{z}(\mathbf{x})^\top A\mathbf{x}|}{K} \ge \\ \frac{|\lambda| - \|A\mathbf{x}\|_{\mathbf{x}}\|\mathbf{z}(\mathbf{x})\|_{\mathbf{x}}^*}{K} \ge \frac{|\lambda| - 1}{K} = \frac{|\lambda| - 1}{K}\|\mathbf{x}\|_2.\end{aligned}$$

Thus for each $|\lambda| > 1$ and $\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n$ we have the inequality $\|(\lambda I - A)\mathbf{x}\|_2 \geq \frac{|\lambda|-1}{K}\|\mathbf{x}\|_2$. Choose $\mathbf{x} = (\lambda I - A)^{-1}\mathbf{y}$ to deduce the inequality

$$\frac{\|(\lambda I - A)\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \leq \frac{K}{|\lambda| - 1}.$$

So $\sigma_1(\lambda I - A)^{-1}) \leq \frac{K}{|\lambda|-1}$. Hence $\mathcal{A}$ satisfies the condition 3.4.13 of Theorem 3.4.9 with the norm $\sigma_1(\cdot)$. Theorem 3.4.9 yields that $\mathcal{A}$ is stable. □

Problem 7 shows that in Theorem 7.5.13 the assumption that $\mathcal{L}$ is a stable subspace can not be dropped. In Chapter ? [Fri84] we show the following result.

Theorem 7.5.14 *Let $n \geq 2, d \in [2n-1, n^2]$ be integers. Then a generic subspace $\mathcal{L}$ of $\mathbb{C}^{n\times n}$ of dimension d is maximally stable.*

Problems

1. Let $\phi : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^\mathrm{n}}$ be a ν-map. Show that $r_\phi : \mathbb{C}^{n\times n} \to \mathbb{R}_+$ is a seminorm.

2. Let $\phi : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^\mathrm{n}}$ be a ν-map. Show that $r_\phi(I_n) = 1$.

3. Show that for any $p \in (1, \infty)$ the map ϕ_p given in Example 7.5.6 is a ν-map.

4. Show

   (a) For any unitary $U \in \mathbb{C}^{n\times n}$ and $A \in \mathbb{C}^{n\times n}$ $\omega_2(U^*AU) = \omega_2(A)$ and $r_2(U^*AU) = r_2(A)$.

   (b) For a normal $A \in \mathbb{C}^{n\times n}$ the numerical range $\omega_2(A)$ is a convex hull of the eigenvalues of A. In particular $r_2(A) = \rho(A)$ for a normal A.

   (c) $\omega_2(A)$ is a convex set for any $A \in \mathbb{C}^{n\times n}$. (Observe that it is enough to prove this claim only for $n = 2$.)

5. Let $A = J_4(0) \in \mathbb{C}^{4\times 4}$ be a nilpotent Jordan block of order 4. Show that $r_2(A) < 1, r_2(A^2) = \frac{1}{2}, r_2(A^3) = \frac{1}{2}$. Hence the inequality $r_2(A^3) \leq r_2(A)r_2(A^2)$ does not hold in general.

6. Show that the map $\phi_{\|\cdot\|} : \mathrm{S}^{2\mathrm{n}-1} \to 2^{\mathbb{C}^\mathrm{n}}$ given in Example 7.5.7 is a ν-map.

7. Let $\|\cdot\|$ be the norm on $\mathbb{C}^{n\times n}$ given by $\|[a_{ij}]_{i=j=1}^n\| := \max_{i,j\in[1,n]} |a_{ij}|$. Denote by $\mathcal{U}_n \subset \mathbb{C}^{n\times n}$ the subspace of upper triangular matrices. For $n \ge 2$ show.

   (a) $\mathcal{U}_n$ is not a stable subspace of $\mathbb{C}^{n\times n}$.

   (b) $\mathcal{U}_n \cap \mathrm{B}_{\|\cdot\|}$ is a compact, convex, balanced set.

   (c) $\rho(A) \le 1$ for each $A \in \mathcal{U}_n \cap \mathrm{B}_{\|\cdot\|}$.

   (d) $\mathcal{U}_n \cap \mathrm{B}_{\|\cdot\|}$ is not a stable set.

7.6 Superstable norms

Definition 7.6.1 *A norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ is called superstable if $\|A^k\| \le \|A\|^k$ for $k = 2, \ldots,$ and each $A \in \mathbb{C}^{n\times n}$.*

Clearly, any operator norm on $\mathbb{C}^{n\times n}$ is a superstable norm, see §7.7.

Theorem 7.6.2 *The standard numerical radius $r_2(A) = \max_{\mathbf{x}\in\mathbb{C}^n,\|\mathbf{x}\|_2=1} |\mathbf{x}^* A\mathbf{x}|$ is a superstable norm.*

To prove the theorem we need the following lemma.

Lemma 7.6.3 *Assume that $A \in \mathbb{C}^{n\times n}, \rho(A) < 1$ and $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$. Let $z_j = e^{\frac{2\pi j\sqrt{-1}}{m}}$ and assume that $I_n - z_jA \in \mathbf{GL}(n,\mathbb{C})$ for $j = 1,\ldots,m$. Then*

$$1 - \mathbf{x}^* A^m \mathbf{x} = \frac{1}{m}\sum_{j=1}^m \|\mathbf{x}_j\|_2^2(1 - z_j\mathbf{y}_j^* A\mathbf{y}_j), \tag{7.6.1}$$

where $\mathbf{x}_j = (\prod_{k\in\langle m\rangle\backslash\{j\}} (1 - z_kA))\mathbf{x},\ \mathbf{y}_j = \frac{1}{\|\mathbf{x}_j\|\mathbf{x}_j},\ j = 1,\ldots,m.$

Proof. Observe the following two polynomial identities in z variable

$$1 - z^m = \prod_{k=1}^m (1 - z_kz), \quad 1 = \frac{1}{m}\sum_{j=1}^m \prod_{k\in\langle m\rangle\backslash\{j\}} (1 - z_kz).$$

Replace the variable z by A obtain the identities

$$I_n - A^m = \prod_{k=1}^m (I_n - z_kA), \quad I_n = \frac{1}{m}\sum_{j=1}^m \prod_{k\in\langle m\rangle\backslash\{j\}} (I_n - z_kA). \tag{7.6.2}$$

Multiply the second identity from the right by $\mathbf{x}$ to get the identity $\mathbf{x} = \frac{1}{m}\sum_{j=1}\mathbf{x}_j$. Multiply the first identity by $\mathbf{x}$ from the right respectively to obtain $\mathbf{x} - A^m\mathbf{x} = (\prod_{k=1}^m (I_n - z_kA))\mathbf{x}$. Since $I_n - z_kA$ for $k = 1, \dots, m$ commute, we deduce that for each k, $\mathbf{x} - A^m\mathbf{x} = (I_n - z_kA)\mathbf{x}_k$. Hence

$$1 - \mathbf{x}^*A^m\mathbf{x} = \mathbf{x}^*(\mathbf{x} - A^m\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \mathbf{x}_j^*(\mathbf{x} - A^m\mathbf{x}) =$$

$$\frac{1}{m}\sum_{j=1}^m \mathbf{x}_j^*(I_n - z_jA)\mathbf{x}_j = \frac{1}{m}\sum_{j=1}^m \|\mathbf{x}_j\|_2^2(1 - z_j\mathbf{y}_j^*A\mathbf{y}_j).$$

□

Proof of Theorem 7.6.2. From the the proof of Lemma 7.6.3 it follows that (7.6.1) holds for any $A \in \mathbb{C}^{n\times n}$ and some $\mathbf{y}_1, \dots, \mathbf{y}_m \in \mathrm{S}^{2\mathrm{n}-1}$, since for $\mathbf{x}_j = 0$ in (7.6.1) we can choose any $\mathbf{y}_j \in \mathrm{S}^{2\mathrm{n}-1}$. Suppose that that $r_2(A) = 1$. Let $\zeta \in \mathbb{C}, |\zeta| = 1$. Apply the equality (7.6.1) to ζA and $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ to deduce

$$1 - \zeta^m\mathbf{x}^*A^m\mathbf{x} = \frac{1}{m}\sum_{j=1}^m \|\mathbf{u}_j\|_2^2(1 - z_j\zeta\mathbf{w}_j^*A\mathbf{w}_j)$$

for corresponding $\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathrm{S}^{2\mathrm{n}-1}$. Choose ζ such that $\zeta^m\mathbf{x}^*A^m\mathbf{x} = |\mathbf{x}^*A^m\mathbf{x}|$. Since $r_2(z_k\zeta A) = r_2(A) = 1$, it follows that $\Re(1 - z_j\zeta\mathbf{w}_j^*A\mathbf{w}_j) \geq 0$. Hence, the above displayed equality yields $1 - |\mathbf{x}^*A^m\mathbf{x}| \geq 0$, i.e $1 \leq |\mathbf{x}^*A^m\mathbf{x}|$. Since $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ is arbitrary, it follows that $r_2(A^m) \leq 1$ if $r_2(A) = 1$. Hence $r_2(\cdot)$ is a superstable norm. □

Definition 7.6.4 *For an integer $n \geq 2$ and $p \in [1, \infty]$ let $K_{p,n} \geq 1$ be the smallest constant satisfying $r_p(A^m) \leq K_{p,n}r_p(A)^m$ for all $A \in \mathbb{C}^{n\times n}$.*

Theorem 7.6.2 is equivalent to the equality $K_{2,n} = 1$. Problem 1b shows that $K_{1,n} = K_{\infty,n} = 1$. It is an open problem if $\sup_{n\in\mathbb{N}} \max_{p\in\infty[1,\infty]} K_{p,n} < \infty$.

Theorem 7.6.5 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^{n\times n}$ which is invariant under the similarity by unitary matrices, i.e. $\|UAU^{-1}\| = \|A\|$ for each $A \in \mathbb{C}^{n\times n}$ and $U \in \mathbf{U}(n)$. Then $\|\cdot\|$ is spectral dominant if and only if $\|A\| \geq r_2(A)$ for any $A \in \mathbb{C}^{n\times n}$.*

To prove the theorem we bring the following two lemmas which are of independent interest.

Lemma 7.6.6 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^{n\times n}$. Assume that $\|\cdot\|$ is invariant under the similarity by $U \in \mathbf{GL}(n,\mathbb{C})$, i.e. $\|A\| = \|UAU^{-1}\|$ for each $A \in \mathbb{C}^{n\times n}$. Then U is similar to diagonal matrix Λ:*

$$(7.6.3) \qquad \Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n), \quad |\lambda_1| = \ldots = |\lambda_n| > 0.$$

Proof. Let $\lambda, \mu \in \mathbb{C}$ be two distinct eigenvalues of U. So there are two corresponding nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ such that $U\mathbf{x} = \lambda\mathbf{x}, U^\top\mathbf{y} = \mu\mathbf{y}$. For $A = \mathbf{x}\mathbf{y}^\top$ we deduce that $UAU^{-1} = \frac{\lambda}{\nu}A$. Since $\|A\| = \|UAU^{-1}\| > 0$ it follows that $|\lambda| = |\mu|$. Hence all the eigenvalues of U have the same modulus.

it is left to show that U is diagonable. Assume to the contrary that U is not diagonable. Then there exists an invertible matrix T and upper triangular matrix $V = [v_{ij}]_{i=j=1}^n$ such that $v_{11} = v_{22} = \lambda \neq 0, v_{12} = 1$, and $V = TVT^{-1}$. Choose $A = TBT^{-1}$, where $B = [b_{ij}]_{i=j=1}^n$, where $b_{22} = \frac{1}{\lambda^2}$. Since $\|U^kAU^{-k}\| = \|A\|$ for $k =\in \mathbb{N}$ it follows that the sequence of matrices $U^kAU^{-k}, k \in \mathbb{N}$ is bounded. A straightforward calculation shows that the $(1,2)$ entry of $T^{-1}(U^kAU^{-k})T$ is k^2. Hence the sequence $U^kAU^{-k}, k \in \mathbb{N}$ is not bounded, contrary to our previous claim. The above contradiction establishes lemma. □

Lemma 7.6.7 *Let $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n) \in \mathbb{C}^{n\times n}$ and assume that $|\lambda_1| = \ldots = |\lambda_n| > 0$ and $\lambda_i \neq \lambda_j$ for $i \neq j$. Suppose that $\|\cdot\|$ is a norm on $\mathbb{C}^{n\times n}$ which is invariant under the similarity by Λ. Then*

$$(7.6.4) \qquad \|\mathrm{diag}(A)\| \le \|A\|.$$

Proof. Λ-similarity invariance implies

$$\|\frac{1}{m+1}\sum_{k=0}^m \Lambda^kA\Lambda^{-k}\| \le \frac{1}{m+1}\sum_{k=0}^m \|\Lambda^kA\Lambda^{-k}\| = \|A\|.$$

For $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ let

$$A_m = [a_{ij,m}] = \frac{1}{m+1}\sum_{k=0}^m \Lambda^kA\Lambda^{-k},$$
$$\text{where } a_{ii,m} = a_{ii},\ a_{ij,m} = a_{ij}\frac{1-\frac{\lambda_i^{m+1}}{\lambda_j^{m+1}}}{(m+1)(1-\frac{\lambda_i}{\lambda_j})} \text{ for } i \neq j.$$

Hence $\lim_{m\to\infty} A_m = \mathrm{diag}(A)$. Since $\|A_m\| \le \|A\|$ we deduce the inequality (7.6.4). □

Proof of Theorem 7.6.5. Assume first that $\|A\| \ge r_2(A)$ for each $A \in \mathbb{C}^{n\times n}$. Clearly, $\|\cdot\|$ is spectral dominant. Assume now that $\|\cdot\|$ is invariant under similarity by any unitary matrix U, and $\|\cdot\|$ is spectral dominant. Since $\|\cdot\|$ is invariant under the similarity by a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $|\lambda_1| = \ldots = |\lambda_n|$ and $\lambda_i \ne \lambda_j$ for $i \ne j$, Lemma 7.6.7 yields (7.6.4). Let $A = [a_{ij}]$. Since $\mathrm{diag}(A) = \mathrm{diag}(a_{11}, \ldots, a_{nn})$ and $\|\cdot\|$ is spectral dominant we obtain that

$$\|A\| \ge \|\,\mathrm{diag}(A)\| \ge \rho(\mathrm{diag}(A)) = \max_{i\in\langle n\rangle} |a_{ii}|.$$

Let $V \in \mathbf{U}(n)$. Then the first column of V is $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$. Furthermore, the $(1,1)$ entry of V^*AV is $\mathbf{x}^*A\mathbf{x}$. Since $\|\cdot\|$ is invariant under unitary similarity we obtain $\|A\| = \|V^*AV\| \ge |\mathbf{x}^*A\mathbf{x}|$. As for any $\mathbf{x} \in \mathrm{S}^{2\mathrm{n}-1}$ there exists a unitary V with the first column $\mathbf{x}$ we deduce that $\|A\| \ge r_2(A)$. $\square$

Problems

1. (a) Describe the ν-maps $\phi_{\|\cdot\|_1}, \phi_{\|\cdot\|_\infty}$.

   (b) Show that for for $p = 1, \infty$ $r_p(A)$ is equal to the operator norm of $\|A\|_p$, for $A \in \mathbb{C}^{n\times n}$ viewed as a linear operator $A : \mathbb{C}^n \to \mathbb{C}^n$. (See §7.7). Hence $K_{1,n} = K_{\infty,n} = 1$, where $K_{p,n}$ is given Definition 7.6.4.

   (c) Show that for each $p \in (1, \infty)$ and integer $n \ge 2$ there exists $A \in \mathbb{C}^{n\times n}$ such that $r_p(A) < \|A\|_p$, where $\|A\|_p$ the operator norm of A.

7.7 Operator norms

Let $\mathbf{V}_a, \mathbf{V}_b$ be two finite dimensional vector spaces over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\|\cdot\|_a, \|\cdot\|_b$ are norms on $\mathbf{V}_a, \mathbf{V}_b$ respectively. Let $T : \mathbf{V}_a \to \mathbf{V}_b$ be a linear transformation. Then

$$\|T\|_{a,b} := \max_{\mathbf{0}\ne\mathbf{x}\in\mathbf{V}_a} \frac{\|T\mathbf{x}\|_b}{\|\mathbf{x}\|_a}, \tag{7.7.1}$$

is called the *operator norm* of T. Clearly

$$\|T\|_{a,b} = \max_{\|\mathbf{x}\|_a\le 1} \|T\mathbf{x}\|_b = \max_{\|\mathbf{x}\|_a=1} \|T\mathbf{x}\|_b \tag{7.7.2}$$

See Problem 1. Let $\mathbf{V}_c$ be a third finite dimensional vector space over $\mathbb{F}$ with a norm $\|\cdot\|_c$. Assume that $Q:\mathbf{V}_b\to\mathbf{V}_c$ is a linear transformation. The we have the well known inequality

$$\|QT\|_{a,c}\le\|Q\|_{b,c}\|T\|_{a,b}. \tag{7.7.3}$$

See Problem 2.

Assume now that $\mathbf{V}_a=\mathbf{V}_b=\mathbf{V}$ and $\|\cdot\|_a=\|\cdot\|_b=\|\cdot\|_c=\|\cdot\|$. We then denote $\|T\|:=\|T\|_{a,b}$ and $\|Q\|:=\|Q\|_{b,c}$. Let $\mathrm{Id}:\mathbf{V}\to\mathbf{V}$ be the identity operator. Hence

$$\|\mathrm{Id}\|=1,\quad \|QT\|\le\|Q\|\,\|T\|,\quad \|T^m\|\le\|T\|^m \text{ for } m=2,\ldots. \tag{7.7.4}$$

Assume that $\mathbf{V}_a=\mathbb{F}^n,\mathbf{V}_b=\mathbb{F}^m$. Then $T:\mathbb{F}^n\to\mathbb{F}^m$ is represented by a matrix $A\in\mathbb{F}^{m\times n}$. Thus $\|A\|_{a,b}$ is the operator norm of A. For $m=n$ and $\|\cdot\|_a=\|\cdot\|_b=\|\cdot\|$ we denote by $\|A\|$ the *operator* norm. Assume that $s,t\in[1,\infty]$. Then for $A\in\mathbb{F}^{m\times n}$ we denote by $\|A\|_{s,t}$ the operator norm of A, where $\mathbb{F}^n,\mathbb{F}^m$ are equipped with the norms ℓ_s,ℓ_t respectively. Note that $\|A\|_{2,2}=\sigma_1(A)$, see Problem 7.4.13. For $m=n$ and $s=t=p$ we denote by $\|A\|_p$ the ℓ_p operator norm of A.

Lemma 7.7.1 *Let $A=[a_{ij}]\in\mathbb{F}^{m\times n}$ and $\|\cdot\|_a,\|\cdot\|_b$ be norms on $\mathbb{C}^n,\mathbb{C}^m$ respectively. If $\|\cdot\|_b$ is an absolute norm then*

$$\|A\|_{a,b}\le\|(\|(a_{11},\ldots,a_{1n})^\top\|_a^*,\ldots,\|(a_{m1},\ldots,a_{mn})^\top\|_a^*)^\top\|_b. \tag{7.7.5}$$

If $\|\cdot\|_a$ is an absolute norm then

$$\|A\|_{a,b}\le\|(\|(a_{11},\ldots,a_{m1})^\top\|_b,\ldots,\|(a_{1n},\ldots,a_{mn})^\top\|_b)^\top\|_a^*. \tag{7.7.6}$$

In both inequalities, equality holds for matrices of rank one.

Proof. Let $\mathbf{x}\in\mathbb{F}^n$. Then

$$(A\mathbf{x})_i=\sum_{j=1}^n a_{ij}x_j\Rightarrow|(A\mathbf{x})_i|\le\|(a_{i1},\ldots,a_{in})^\top\|_a^*\|\mathbf{x}\|_a,\quad i=1,\ldots,m\Rightarrow$$
$$|A\mathbf{x}|\le\|\mathbf{x}\|_a(\|(a_{11},\ldots,a_{1n})^\top\|_a^*,\ldots,\|(a_{m1},\ldots,a_{mn})^\top\|_a^*)^\top.$$

Assume that $\|\cdot\|_b$ is an absolute norm. Then

$$\|A\mathbf{x}\|_b\le\|\mathbf{x}\|_a\|(\|(a_{11},\ldots,a_{1n})^\top\|_a^*,\ldots,\|(a_{m1},\ldots,a_{mn})^\top\|_a^*)^\top\|_b,$$

which yields (7.7.5). Suppose that A is rank one matrix. So $A=\mathbf{u}\mathbf{v}^\top$, where $\mathbf{0}\ne\mathbf{u}\in\mathbb{F}^m,\mathbf{0}\ne\mathbf{v}\in\mathbb{F}^n$. There exists $\mathbf{0}\ne\mathbf{x}\in\mathbb{F}^n$ such that

$\mathbf{v}^\top\mathbf{x} = \|\mathbf{v}\|_a^*\|\mathbf{x}\|_a$. For this $\mathbf{x}$ we have that $|A\mathbf{x}| = \|\mathbf{v}\|_a^*\|\mathbf{x}\|_a|\mathbf{u}|$. Hence $\frac{\|A\mathbf{x}\|_b}{\|\mathbf{x}\|_a} = \|\mathbf{v}\|_a^*\|\mathbf{u}\|_b$. Thus $\|A\|_{a,b} \ge \|\mathbf{v}\|_a^*\|\mathbf{u}\|_b$. On the other hand the right-hand of (7.7.5) is $\|\mathbf{v}\|_a^*\|\mathbf{u}\|_b$. This shows that (7.7.5) is sharp for rank one matrices.

Assume now that $\|\cdot\|_a$ is an absolute norm. Theorem 7.4.3 claims that $\|\cdot\|_a^*$ is an absolute norm. Apply (7.7.5) to $\|A^\top\|_{b^*,a^*}$ and use Problem 1e to deduce the inequality (7.7.6). Assume that A is rank one matrix. Then $A^\top$ is a rank one matrix and equality holds in (7.7.6). □

Theorem 7.7.2 *Let $m, n \ge 2$ be integers. Assume that $s, t \in [1, \infty]$ and suppose $\mathbb{F}^n, \mathbb{F}^m$ are endowed with Hölder with norms $\|\cdot\|_s, \|\cdot\|_t$ respectively. Let s^* be defined by the equality $\frac{1}{s} + \frac{1}{s^*} = 1$. Then for $A = [a_{ij}]_{i=j=1}^{m,n}\mathbb{F}^{m\times n}$ the following hold.*

$$(7.7.7)\quad \|A\|_{s,t} \le \min\Big(\big(\sum_{i=1}^{m}\big(\sum_{j=1}^{n}|a_{ij}|^{s^*}\big)^{\frac{t}{s^*}}\big)^{\frac{1}{t}}, \big(\sum_{j=1}^{n}\big(\sum_{i=1}^{m}|a_{ij}|^{t}\big)^{\frac{s^*}{t}}\big)^{\frac{1}{s^*}}\Big),$$

$$(7.7.8)\quad \|A\|_{\infty,1} \le \sum_{i=j=1}^{m,n}|a_{ij}|,$$

$$(7.7.9)\quad \|A\|_{1,1} = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|,$$

$$(7.7.10)\quad \|A\|_{\infty,\infty} = \max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|,$$

$$(7.7.11)\quad \|A\|_{1,\infty} = \max_{1\le i\le m, 1\le j\le n}|a_{ij}|.$$

Proof. Since $\|\cdot\|_s, \|\cdot\|_t$ are absolute norm the inequalities (7.7.5) and (7.7.6) hold. As $\|\cdot\|_s^* = \|\cdot\|_{s^*}$ we deduce (7.7.7). For $s = \infty$ we have $s^* = 1$, and for $t = 1$ (7.7.7) yields (7.7.8).

Assume that $s = t = 1$. So $s^* = \infty$. The second part of the inequality (7.7.7) yields the inequality $\|A\|_{1,1} \le \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$. Let $\mathbf{e}_j = (\delta_{j1}, \ldots, \delta_{jn})^\top$. Clearly, $\|\mathbf{e}_j\|_1 = 1$ and $\|A\mathbf{e}_j\|_1 = \sum_{i=1}^m |a_{ij}|$. Hence $\|A\|_{1,1} \ge \sum_{i=1}^m |a_{ij}|$. So $\|A\|_{1,1} \ge \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$, which yields (7.7.9). Since $\|A\|_{\infty,\infty} = \|A^\top\|_{1,1}$, see Problem 1e, we deduce (7.7.10) from (7.7.9).

Let $s = 1, t = \infty$. Then (7.7.7) yields the inequality
$\|A\|_{1,\infty} \le \max_{1\le i\le m, 1\le j\le n}|a_{ij}| = |a_{i_1j_1}|$. Clearly, $\|A\mathbf{e}_{j_1}\|_\infty = |a_{i_1j_1}|$. Hence $\|A\|_{1,\infty} \ge |a_{i_1j_1}|$, which proves (7.7.11). □

Theorem 7.7.3 *Let* $\mathbf{V}$ *be a finite dimensional vector space over* $\mathbb{C}$ *with an norm* $\|\cdot\|$*. Let* $\|\cdot\|$ *be the induced operator norm* Hom $(\mathbf{V},\mathbf{V})$*. Then for* $A \in$ Hom $(\mathbf{V},\mathbf{V})$ *the inequality* $\rho(A) \le \|A\|$ *holds.*

Proof. Assume that $\lambda \in$ spec A. Then there exists $\mathbf{0} \ne \mathbf{x} \in \mathbf{V}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. So $\|A\| \ge \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = |\lambda|$. □

Problems

1. Show

   (a) The equality (7.7.2).

   (b) For each $t > 0$ there exists a vector $\mathbf{y} \in \mathbf{V}_a$, $\|\mathbf{y}\|_b = t$ such that $\|T\|_{a,b} = \frac{\|T\mathbf{y}\|_b}{\|\mathbf{y}\|_a}$.

   (c) $\|T\|_{a,b} = \max_{\mathbf{f}\in \mathrm{S}_{\|\cdot\|_{\mathrm{b}}^*}, \mathbf{x}\in\mathrm{S}_{\|\cdot\|}} |\mathbf{f}(T\mathbf{x})|$.

   (d) Denote by $\|T^*\|_{b^*,a^*}$ the operator norm of $T^* : \mathbf{V}_b^* \to \mathbf{V}_a^*$ with respect to the norms $\|\cdot\|_b^*, \|\cdot\|_a^*$ respectively. Show that $\|T^*\|_{b^*,a^*} = \|T\|_{a,b}$.

   (e) $\|A^\top\|_{b^*,a^*} = \|A\|_{a,b}$ for $A \in \mathbb{F}^{m\times n}$.

2. Show the inequality (7.7.3).

3. Let $T : \mathbf{V} \to \mathbf{V}$ be a linear transformation on a finite dimensional vector space $\mathbb{V}$ over $\mathbb{C}$ with a norm $\|\cdot\|$. Show that $\rho(T) \le \|T\|$.

4. Let $A \in \mathbb{C}^{n\times n}$. Then $A = Q^{-1}(\Lambda + N)Q$, where Λ is a diagonal matrix, N strictly upper triangular and $\Lambda + N$ is the Jordan canonical form of A. Show

   (a) A is similar to $\Lambda + tN$ for any $0 \ne t \in \mathbb{C}$, i.e. $A = Q_t^{-1}(\Lambda + tN)Q_t$. Show that $Q_t = QD_t$ for an appropriate diagonal matrix D_t.

   (b) Let $\varepsilon > 0$ be given. Show that one can choose a norm $\|\cdot\|_t$ on $\mathbb{C}^n$ of the form $\|\mathbf{x}\|_t := \|Q_t\mathbf{x}\|_2$ for $|t|$-small enough such that $\|A\|_t \le \rho(A) + \varepsilon$. **Hint**: Note that $\|\Lambda + tN\|_2 \le \|\Lambda\|_2 + |t|\|N\|_2 = \rho(A) + t\|N\|_2$.)

   (c) If $N = 0$ then $\|A\| = \rho(A)$ where $\|\mathbf{x}\| = |Q\mathbf{x}|_2$.

   (d) Suppose that each eigenvalue λ of modulus $\rho(A)$ is geometrically simple. Then there exists $|t|$ small enough such that $\|A\|_t = \rho(A)$.

5. Let $A \in \mathbb{C}^{n\times n}$. Then there exists a norm on $\|\cdot\|$ on $\mathbb{C}^n$ such that $\rho(A) = \|A\|$ if and only if A each eigenvalue λ of modulus $\rho(A)$ is geometrically simple. **Hint**: Note that if $\rho(A) = 1$ and there is an eigenvalue $\lambda, |\lambda| = 1$ which is not geometrically simple, then $A^m, m \in \mathbb{N}$ is not a bounded sequence.

6. Assume that $\|\cdots\|_a, \|\cdot\|_b$ are two absolute norm on $\mathbb{C}^n$ and $\mathbb{C}^m$. Assume that $Q_1\mathbb{C}^{m\times m}, Q_2 \in \mathbb{C}^{n\times n}$ are two diagonal matrices such that the absolute value of each diagonal entry is 1. Show that for any $A \in \mathbb{C}^{m\times n}$ $\|Q_1AQ_2\|_{a,b} = \|A\|_{a,b}$.

7. Show

   (a) Show that if $A \in \mathbb{R}_+^{m\times n}$ or $-A \in \mathbb{R}_+^{m\times n}$ then $\|A\|_{\infty,1} = \sum_{i=j=1}^{m,n} |a_{ij}|$.

   (b) Let $\mathbf{x} = (x_1, \dots, x_n), \mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$. We say that $\mathbf{y}$ has a weak sign pattern as $\mathbf{x}$ if $y_i = 0$ for $x_i = 0$ and $y_ix_i \ge 0$ for $x_i \ne 0$. Let $A \in \mathbb{R}^{m\times n}$. Assume that there exists $\mathbf{x} \in \mathbb{R}^n$ such that each row $\mathbf{r}_i$ of A either $\mathbf{r}_i$ or $-\mathbf{r}_i$ has a weak sign pattern as $\mathbf{x}$. Then $\|A\|_{\infty,1} = \sum_{i=j=1}^{m,n} |a_{ij}|$.

   (c) Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. Assume that $a_{11}, a_{12}, a_{21}, -a_{22} > 0$. Show that $\|A\|_{\infty,1} < a_{11} + a_{12} + a_{21} - a_{22}$.

   (d) Generalize the results of Problem 7c to $A \in \mathbb{C}^{m\times n}$.

7.8 Tensor products of convex sets

Definition 7.8.1 *Denote $\mathbf{V}_i$be a finite dimensional vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ for $i = 1, \dots, m$. Let $\mathbf{V} = \otimes_{i=1}\mathbf{V}_i$. Assume that $X_i \subset \mathbf{V}_i$ for $i = 1, \dots, m$. Denote*

$$\odot_{i=1}^m X_i := \{\otimes_{i=1}^m \mathbf{x}_i, \textit{ for all } \mathbf{x}_i \in X_i,\ i = 1, \dots, m\}.$$

We call $\odot_{i=1}^m X_i$ set tensor product of $X_1, \dots, X_m$, or simply tensor product of $X_1, \dots, X_m$.

Lemma 7.8.2 *Let C_i be a compact convex set in a finite dimensional vector space $\mathbf{V}_i$ for $i = 1, \dots, m$. Then* $\operatorname{conv} \odot_{i=1}^m C_i$ *is a compact convex set in $\otimes_{i=1}^m \mathbf{V}_i$, whose extreme points are contained in $\odot_{i=1}^m \mathcal{E}(C_i)$. In particular, if C_1 and C_2 are polytopes then* $\operatorname{conv} C_1 \odot C_2$ *is a polytope.*

Proof. Since C_i is compact for $i = 1, \dots, m$, it is straightforward to show that $C := \odot_{i=1}^m C_i$ is a compact set in $\mathbf{V} = \otimes_{i=1}^m \mathbf{V}_i$. Hence $\operatorname{conv} C$ is compact. Let $\dim \mathbf{V}_i = d_i$. Since C_i is compact, Theorem 7.1.3 implies that each $\mathbf{x}_i \in C_i$ is of the form $\mathbf{x}_i = \sum_{j_i=1}^{d_i+1} a_{ij_i}\mathbf{y}_{ij_i}$ where $a_{ij_i} \ge 0, \mathbf{y}_{ij_i} \in \mathcal{E}(C_i)$ for $j_i = 1, \ldots, d_i + 1$, and $\sum_{j_i=1}^{d_i+1} a_{ij_i} = 1$. Hence

$$\otimes_{i=1}^m \mathbf{x}_i = \sum_{j_1=\ldots=j_m=1}^{d_1+1,\ldots,d_m+1} (\prod_{i=1}^m a_{ij_i})(\otimes_{i=1}^m \mathbf{y}_{ij_i}).$$

Note that $\prod_{i=1}^m a_{ij_i} \ge 0$ and $\sum_{j_1=\ldots=j_m=1}^{d_1+1,\ldots,d_m+1} \prod_{i=1}^m a_{ij_i} = 1$. Hence $C \subset \operatorname{conv} \odot_{i=1}^m \mathcal{E}(C_i)$, which implies $\operatorname{conv} C \subset \operatorname{conv} \odot_{i=1}^m \mathcal{E}(C_i) \subset \operatorname{conv} C$.

Assume that C_1, C_2 are polytopes. Then $\operatorname{conv} C_1 \odot C_2$ is nonempty, compact and convex whose set of extreme points is finite. Hence $\operatorname{conv} C_1 \odot C_2$ is a polytope. □

See Problem 1 for en example where $\mathcal{E}(C_1 \odot C_2)$ is strictly contained in $\mathcal{E}(C_1) \odot \mathcal{E}(C_2)$. The next two examples give two important cases where $\mathcal{E}(C_1 \odot C_2) = \mathcal{E}(C_1) \odot \mathcal{E}(C_2)$. In these two examples we view $\mathbb{C}^{p\times q} \otimes \mathbb{C}^{m\times n}$ is viewed as a subspace of $\mathbb{C}^{pm\times qn}$, where the tensor product of two matrices $A \otimes B$ is the Kronecker tensor product.

Proposition 7.8.3 *Let $m, n \ge 2$ be integers. Then $\Omega_m \odot \Omega_n \subset \Omega_{mn}$, and $\mathcal{E}(\operatorname{conv} \Omega_m \odot \Omega_n) = \mathcal{P}_m \odot \mathcal{P}_n$.*

Proof. Let $A = [a_{ij}] \in \Omega_n, B = [b_{pq}] \in \Omega_n$. The the entries of $A \otimes B = [c_{(i,p)(j,q)}] \in \mathbb{R}_+^{mn\times mn}$, where $c_{(i,p)(j,q)} = a_{ij}b_{pq}$. Clearly

$$\sum_{j=1}^m c_{(i,p)(j,q)} = \sum_{j=1}^m a_{ij}b_{pq} = b_{pq}, \quad \sum_{i=1}^m c_{(i,p)(j,q)} = \sum_{i=1}^m a_{ij}b_{pq} = b_{pq},$$
$$\sum_{q=1}^n c_{(i,p)(j,q)} = \sum_{q=1}^n a_{ij}b_{pq} = a_{ij}, \quad \sum_{p=1}^n c_{(i,p)(j,q)} = \sum_{p=1}^n a_{ij}b_{pq} = a_{ij}$$

Hence $A \otimes B \in \Omega_{mn}$, where we identify the set $\langle m\rangle \times \langle n\rangle$ with $\langle mn\rangle$. Since Ω_{mn} is convex it follows that $\operatorname{conv} \Omega_m \odot \Omega_n \subset \Omega_{mn}$. Recall that $\mathcal{E}(\Omega_{mn}) = \mathcal{P}_{mn}$. Clearly $\mathcal{P}_m \odot \mathcal{P}_n \subset \mathcal{P}_{mn}$. Problem 7.1.2 yields that $\mathcal{E}(\operatorname{conv} \Omega_m \odot \Omega_n) = \mathcal{P}_m \odot \mathcal{P}_n$. □

Proposition 7.8.4 *Let $m, n \ge 2$ be integers. Then $\mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1} \subset \mathrm{H}_{mn,+,1}$, and $\mathcal{E}(\operatorname{conv} \mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1}) = \mathcal{E}(\mathrm{H}_{m,+,1}) \odot \mathcal{E}(\mathrm{H}_{n,+,1})$.*

Proof. Let $A \in \mathrm{H}_m, B \in H_n$ be nonnegative definite hermitian matrices. Then $A \otimes B$ is nonnegative definite. Since $\operatorname{tr} A \otimes B = (\operatorname{tr} A)(\operatorname{tr} B)$ it follows that $\mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1} \subset \mathrm{H}_{mn,+,1}$. Hence $\mathcal{E}(\mathrm{conv}\,\mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1}) \subset \mathcal{E}(\mathrm{H}_{m,+,1}) \odot \mathcal{E}(\mathrm{H}_{n,+,1})$.

Recall that $\mathcal{E}(\mathrm{H}_{m,+,1}), \mathcal{E}(\mathrm{H}_{n,+,1}), \mathcal{E}(\mathrm{H}_{mn,+,1})$ are hermitian rank one matrix of trace 1 of corresponding orders. Since $A \otimes B$ is a rank one matrix if A and B is a rank one matrix it follows that $\mathcal{E}(\mathrm{H}_{m,+,1}) \odot \mathcal{E}(\mathrm{H}_{n,+,1}) \subset \mathcal{E}(\mathrm{H}_{mn,+,1}$. Hence $\mathcal{E}(\mathrm{conv}\,\mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1}) = \mathcal{E}(\mathrm{H}_{m,+,1}) \odot \mathcal{E}(\mathrm{H}_{n,+,1})$. □

Problem 7.8.5 *Let C_i be a compact convex set in a finite dimensional space $\mathbf{V}_i$ for $i = 1, 2$. Suppose that $C_i = \cap_{\alpha \in \mathcal{F}_i} \mathrm{H}(f_\alpha, \mathbf{x}_\alpha)$, were $\mathcal{F}_i$ is the set of all supporting hyperplanes of C_i which characterize C_i, for $i = 1, 2$. ($\mathcal{F}_i$ may not be finite or countable.) The problem is to characterize the set of all supporting hyperplanes of* $\mathrm{conv}\, C_1 \odot C_2$.

Equivalently, suppose we know how to decide if $\mathbf{x}_i$ belongs or does not belong to C_i for $i = 1, 2$. How do we determine if $\mathbf{x}$ belongs or does not belong $\mathrm{conv}\, C_1 \odot C_2$*?*

It seems that the complexity of characterization $\mathrm{conv}\, C_1 \odot C_2$ can be much more complex then the complexity of C_1 and C_2. We will explain this remark in the two examples discussed in Propositions 7.8.3 and 7.8.4.

Consider first $H_{m,+1,} \odot H_{n,+,1}$. So any element in $H_{m,+1,} \odot H_{n,+,1}$ is of the form $A \otimes B$, where A and B are nonnegative definite hermitian matrices of trace one. The matrix $A \otimes B$ is called a *pure state* in quantum mechanics. A matrix $C \in \mathrm{conv}\, H_{m,+1,} \odot H_{n,+,1}$ is called a separable state. So $\mathrm{conv}\, H_{m,+1,} \odot H_{n,+,1}$ is the convex set of separable states. The set $\mathrm{H}_{mn,+1,1} \setminus \mathrm{conv}\, H_{m,+1,} \odot H_{n,+,1}$ the set of entangled states. See for example [BeZ06]. The following result is due to L. Gurvits [Gur03]

Theorem 7.8.6 *For general positive integers m, n and $A \in \mathrm{H}_{mn,+,1}$ the problem of decision if A is separable, i.e. $A \in \mathrm{conv}\, H_{m,+1,} \odot H_{n,+,1}$, is NP-Hard.*

On the other hand, given a hermitian matrix $A \in \mathrm{H}_n$, it well known that one can determine in polynomial time if A belongs or does not belong to $\mathrm{H}_{n,+,1}$. See Problem 3. We will discuss the similar situation for $\mathrm{conv}\, \Omega_m \odot \Omega_n$ in the §7.11.

Definition 7.8.7 *Let $\mathbf{V}_i$ be a finite dimensional vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ with a norm $\|\cdot\|_i$ for $i = 1, \ldots, k$. Let $\mathbf{V} := \otimes_{i=1}^k \mathbf{V}_i$ with the norm $\|\cdot\|$. Then $\|\cdot\|$ is called a cross norm if*

$$\|\otimes_{i=1}^k \mathbf{x}_i\| = \prod_{i=1}^k \|\mathbf{x}_i\|_i \tag{7.8.1}$$

for all rank one tensors.

Identify $\otimes_{i=1}^{k}\mathbf{V}_i^*$ *with* $\mathbf{V}^*$, *where* $(\otimes_{i=1}^{k}\mathbf{f}_i)(\otimes_{i=1}^{k}\mathbf{x}_i)=\prod_{i=1}^{k}\mathbf{f}_i(\mathbf{x}_i)$. *Then* $\|\cdot\|$ *is called a normal cross norm if the norm* $\|\cdot\|^*$ *on* $\mathbf{V}^*$ *is a cross norm with respect to the norms* $\|\cdot\|_i^*$ *on* $\mathbf{V}_i^*$ *for* $i=1,\ldots,k$.

See [Sch50] for properties of the cross norms. We discuss the following known results needed in the sequel.

Theorem 7.8.8 *Let* $\mathbf{V}_i$ *be a finite dimensional vector space over* $\mathbb{F}=\mathbb{R},\mathbb{C}$ *with a norm* $\|\cdot\|_i$ *for* $i=1,\ldots,k$. *Let* $\mathbf{V}:=\otimes_{i=1}^{k}\mathbf{V}_i$. *Then there exists a norm* $\|\cdot\|$ *on* $\mathbf{V}$ *satisfying (7.8.7) Furthermore, there exist unique norms* $\|\cdot\|_{\max},\|\cdot\|_{min}$ *satisfying the following properties. First,* $\|\cdot\|_{\max}$ *and* $\|\cdot\|_{min}$ *are normal cross norms. Moreover* $\|\mathbf{z}\|_{\min}\le\|\mathbf{z}\|_{\max}$ *for all* $\mathbf{z}\in\mathbf{V}$. *Any cross norm* $\|\cdot\|$ *on* $\mathbf{V}$ *satisfies the inequality* $\|\mathbf{z}\|\le\|\mathbf{z}\|_{\max}$ *for all* $\mathbf{z}\in\mathbf{V}$. *Third, assume that* $\|\cdot\|_a$ *on* $\mathbf{V}$ *satisfies the equality*

$$\|\otimes_{i=1}^{k}\mathbf{f}_i\|_a^*=\prod_{i=1}^{k}\|\mathbf{f}_i\|_i^* \textit{ for all } \mathbf{f}_i\in\mathbf{V}_i^*, i=1,\ldots,k. \tag{7.8.2}$$

I.e. $\|\cdot\|_a^*$ *is a cross norm with respect to the norms* $\|\cdot\|_i^*$ *on* $\mathbf{V}_i^*$ *for* $i=1,\ldots,k$. *Then* $\|\mathbf{z}\|_{\min}\le\|\mathbf{z}\|_a$ *for all* $\mathbf{z}\in\mathbf{V}$. *More precisely,*

$$\mathrm{B}_{\|\cdot\|_{\max}}=\mathrm{conv}\,\mathrm{B}_{\|\cdot\|_1}\odot\ldots\odot\mathrm{B}_{\|\cdot\|_k},\quad \mathrm{B}_{\|\cdot\|_{\min}^*}=\mathrm{conv}\,\mathrm{B}_{\|\cdot\|_1^*}\odot\ldots\odot\mathrm{B}_{\|\cdot\|_k^*}. \tag{7.8.3}$$

Proof. For simplicity of the exposition we let $k=2$. Define the set $\mathcal{B}:=\mathrm{conv}\,\mathrm{B}_{\|\cdot\|_1}\odot\mathrm{B}_{\|\cdot\|_2}$. Clearly, $\mathcal{B}$ is a compact convex balanced that $\mathbf{0}$ is in its interior. Hence there exists a norm $\|\cdot\|_{\max}$ such that $\mathcal{B}=\mathrm{B}_{\|\cdot\|_{\max}}$. We claim that $\|\mathbf{x}\otimes\mathbf{y}\|_{\max}=\|\mathbf{x}\|_1\|\mathbf{y}\|_2$. Clearly, to show that it is enough to assume that $\|\mathbf{x}\|_1=\|\mathbf{y}\|_2=1$. Since $\mathbf{x}\otimes\mathbf{y}\in\mathrm{B}_{\|\cdot\|_{\max}}$ we deduce that $\|\mathbf{x}\otimes\mathbf{y}\|_{\max}\le 1$. Problem 7.4.5c yields that there exists $\mathbf{f}\in\mathrm{S}_{\|\cdot\|_1^*},\mathbf{g}\in\mathrm{S}_{\|\cdot\|_2^*}$ such that

$$1=\mathbf{f}(\mathbf{x})\ge|\mathbf{f}(\mathbf{x}_1)|,\ \forall\mathbf{x}_1\in\mathrm{B}_{\|\cdot\|_1},\quad 1=\mathbf{g}(\mathbf{y})\ge|\mathbf{f}(\mathbf{y}_1)|,\ \forall\mathbf{y}_1\in\mathrm{B}_{\|\cdot\|_2}.$$

Hence

$$1=(\mathbf{f}\otimes\mathbf{g})(\mathbf{x}\otimes\mathbf{y})\ge|(\mathbf{f}\otimes\mathbf{g})(\mathbf{z})| \text{ for all } \mathbf{z}\in\mathrm{B}_{\|\cdot\|_{\max}}. \tag{7.8.4}$$

Hence $\mathbf{x}\otimes\mathbf{y}\in\partial\mathrm{B}_{\|\cdot\|_{\max}}=\mathrm{S}_{\|\cdot\|_{\max}}$, i.e. $\|\mathbf{x}\otimes\mathbf{y}\|_{\max}=1$. Therefore the norm $\|\cdot\|_{\max}$ satisfies (7.8.1). Let $\|\cdot\|$ be another norm on $\mathbf{V}$ satisfying (7.8.1). Hence $\mathrm{B}_{\|\cdot\|_1}\odot\mathrm{B}_{\|\cdot\|_2}\subset\mathrm{B}_{\|\cdot\|}$, which yields $\mathrm{B}_{\|\cdot\|_{\max}}=\mathrm{conv}\,\mathrm{B}_{\|\cdot\|_1}\odot\mathrm{B}_{\|\cdot\|_2}\subset\mathrm{B}_{\|\cdot\|}$. Therefore, $\|\mathbf{z}\|\le\|\mathbf{z}\|_{\max}$.

We next observe that that $\|\cdot\|_c:=\|\cdot\|_{\max}^*$ on $\mathbf{V}^*$ satisfies the equality

$$\|\mathbf{f}\otimes\mathbf{g}\|_c=\|\mathbf{f}\|_1^*\|\mathbf{g}\|_2^* \text{ for all } \mathbf{f}\in\mathbf{V}_1^*,\mathbf{g}\in\mathbf{V}_2^*. \tag{7.8.5}$$

Recall that $\|\mathbf{f}\otimes\mathbf{g}\|_c = \max_{\mathbf{z}\in \mathrm{B}_{\|\cdot\|}} |(\mathbf{f}\otimes\mathbf{g})(\mathbf{z})|$. Use the (7.8.4) to deduce (7.8.5). Hence $\|\mathbf{z}\|_{\max}$ is a normal cross norm.

Let $\|\cdot\|_b$ be the norm given by the unit ball $\operatorname{conv} \mathrm{B}_{\|\cdot\|_1^*} \odot \mathrm{B}_{\|\cdot\|_2^*}$ on $\mathbf{V}^*$. Hence the above results show that $\|\cdot\|_b$ is a normal cross norm. Recall that for any norm $\|\cdot\|_a^*$ on $\mathbf{V}^*$ satisfying (7.8.5) we showed the inequality $\|\mathbf{h}\|_a^* \le \|\mathbf{h}\|_b$ for any $\mathbf{h}\in\mathbf{V}^*$. Define $\|\cdot\|_{\min} := \|\cdot\|_b^*$. Hence $\|\mathbf{z}\|_{\min} \le \|\mathbf{z}\|_a$. The previous arguments show that $\|\cdot\|_{\min}$ satisfies the equality (7.8.1). Hence $\|\mathbf{z}\|_{\min} \le \|\mathbf{z}\|_{\max}$ for all $\mathbf{z}\in\mathbf{V}$. Also $\|\mathbf{z}\|_{\min}$ is a normal cross norm. □

Proposition 7.8.9 *Let $k > 1$ be an integer. Assume that $\mathbf{V}_1,\ldots,\mathbf{V}_k$ are inner product spaces over $\mathbb{F} = \mathbb{R},\mathbb{C}$. Let $\mathbf{V} = \otimes_{i=1}^k \mathbf{V}_i$ with the inner product induced by the inner products on $\mathbf{V}_1,\ldots,\mathbf{V}_k$. Assume that $\|\cdot\|_1,\ldots,\|\cdot\|_k, \|\cdot\|$ are the induced norms by the corresponding inner products on $\mathbf{V}_1,\ldots,\mathbf{V}_k,\mathbf{V}$. Then $\|\cdot\|$ is a normal cross norm. If $\dim \mathbf{V}_i > 1$ for $i = 1,\ldots,k$ then $\|\cdot\|$ is different from $\|\cdot\|_{max}$ and $\|\cdot\|_{min}$.*

See Problem 7.

Theorem 7.8.10 *Let $\mathbf{U},\mathbf{V}$ be finite dimensional vector spaces over $\mathbb{F} = \mathbb{R},\mathbb{C}$ with norms $\|\cdot\|, |||\cdot|||$ respectively. Identify $\mathbf{W} = \mathbf{V}\otimes\mathbf{U}^*$ with* Hom $(\mathbf{U},\mathbf{V})$*, via isomorphism $\theta : \mathbf{W}\to$* Hom $(\mathbf{U},\mathbf{V})$*, where $\theta(\mathbf{v}\otimes\mathbf{f})(\mathbf{u}) = \mathbf{f}(\mathbf{u})\mathbf{v}$ for any $\mathbf{f}\in\mathbf{U}^*$. Then the minimal cross norm on $\|\cdot\|_{\min}$ on $\mathbf{W}$ is the operator norm on* Hom $(\mathbf{U},\mathbf{V})$*, where the norms on $\mathbf{U}^*$ and $\mathbf{V}$ are $\|\cdot\|^*$ and $|||\cdot|||$ respectively. Identify $\mathbf{W}^*$ with $\mathbf{V}^*\otimes\mathbf{U} \sim$* Hom $(\mathbf{U}^*,\mathbf{V}^*)$*. Then the maximal cross norm $\|\cdot\|_{\max}$ on $\mathbf{W}$ is the conjugate to the operator norm on* Hom $(\mathbf{U}^*,\mathbf{V}^*)$*, which is identified with $\mathbf{W}^*$.*

Proof. Let $T\in$ Hom $(\mathbf{U},\mathbf{V})$. Then

$$\|T\| = \max_{\mathbf{u}\in \mathrm{S}_{\|\cdot\|}} |||T(\mathbf{u})||| = \max_{\mathbf{g}\in \mathrm{S}_{|||\cdot|||^*}, \mathbf{u}\in \mathrm{S}_{\|\cdot\|}} |\mathbf{g}(T(\mathbf{u}))|.$$

Let θ^{-1} : Hom $(\mathbf{U},\mathbf{V})\to\mathbf{V}\otimes\mathbf{U}^*$ be the isomorphism given in the theorem. Then $\mathbf{g}(T(\mathbf{u})) = (\mathbf{g}\otimes\mathbf{u})(\theta^{-1}(T))$. Let $\|\cdot\|_b$ be the norm given by the unit ball $\operatorname{conv} \mathrm{B}_{|||\cdot|||^*}\odot \mathrm{B}_{\|\cdot\|}$ on $\mathbf{V}^*\otimes\mathbf{U}\sim\mathbf{W}^*$, as in the proof of Theorem 7.8.8. Then

$$\|\theta^{-1}(T)\|_b^* = \max_{\mathbf{g}\in \mathrm{S}_{|||\cdot|||^*}, \mathbf{u}\in \mathrm{S}_{\|\cdot\|}} = |(\mathbf{g}\otimes\mathbf{u})(\theta^{-1}(T))|.$$

Use the proof of Theorem 7.8.8 to deduce that $\|T\| = \|\phi(T)\|_{\min}$.

Similar arguments show that the conjugate norm to the operator norm of Hom $(\mathbf{U}^*,\mathbf{V}^*)$, identified with $\mathbf{V}^*\otimes\mathbf{U}\sim\mathbf{W}^*$ gives the norm $\|\cdot\|_{\max}$ on $\mathbf{W}$. □

Use the above theorem and Problem 7.4.14 to deduce.

Corollary 7.8.11 *Let* $\mathbf{U},\mathbf{V}$ *be finite dimensional inner product spaces over* $\mathbb{F}=\mathbb{R},\mathbb{C}$*, with the corresponding induced norms. Identify* $\mathbf{W}=\mathbf{V}\otimes\mathbf{U}^*$ *with* $\mathrm{Hom}\,(\mathbf{U},\mathbf{V})$ *as in Theorem 7.8.10. Then* $\|T\|_{\min}=\sigma_1(T)$ *and* $\|T\|_{\max}=\sum_{i=1}^{\dim\mathbf{V}_1}\sigma_i(T)$.

More generally, given finite dimensional vectors spaces $\mathbf{U}_i,\mathbf{V}_i$ over $\mathbb{F}=\mathbb{R},\mathbb{C}$ for $i=1,\ldots,k$ we identify the tensor spaces $\otimes_{i=1}^k\mathrm{Hom}\,(\mathbf{U}_i,\mathbf{U}_i)$ with $\mathrm{Hom}\,(\otimes_{i=1}^k\mathbf{U}_i,\otimes_{i=1}^k\mathbf{V}_i)$ using isomorphism

$$\iota:\otimes_{i=1}^k\mathrm{Hom}\,(\mathbf{U}_i,\mathbf{V}_i)\to\mathrm{Hom}\,(\otimes_{i=1}^k\mathbf{U}_i,\otimes_{i=1}^k\mathbf{V}_i)\text{ satisfying}$$
$$\iota(\otimes_{i=1}^kT_i)(\otimes_{i=1}^k\mathbf{u}_i)=\otimes_{i=1}^k(T_i\mathbf{u}_i) \tag{7.8.6}$$
$$\text{where }T_i\in\mathrm{Hom}\,(\mathbf{U}_i,\mathbf{V}_i),\mathbf{u}_i\in\mathbf{U}_i,i=1,\ldots,k.$$

Theorem 7.8.12 *Let* $\mathbf{U}_i,\mathbf{V}_i$ *are finite dimensional vector spaces over* $\mathbb{F}=\mathbb{R},\mathbb{C}$ *with the norms* $\|\cdot\|_i,|||\cdot|||_i$ *respectively for* $i=1,\ldots,k$*. Let* $\mathrm{N}_i(\cdot)$ *be the operator on* $\mathrm{Hom}\,(\mathbf{U}_i,\mathbf{V}_i)$ *for* $i=1,\ldots,k$*. Let* $\|\cdot\|_{\max}$ *be the maximal cross norms on* $\mathbf{U}:=\otimes_{i=1}^k\mathbf{U}_i$ *and* $|||\cdot|||$ *be any cross norm on* $\mathbf{V}:=\otimes_{i=1}^k\mathbf{V}_i$*. Then the operator norm* $\mathrm{N}(\cdot)$ *on* $\mathrm{Hom}\,(\mathbf{U},\mathbf{V})$*, identified with* $\otimes_{i=1}^k\mathrm{Hom}\,(\mathbf{U}_i,\mathbf{V}_i)$*, is a cross norm with respect to the norms* $\mathrm{N}_i(\cdot),i=1,\ldots,k$.

Proof. Since $\mathrm{B}_{\|\cdot\|_{\max}}=\odot_{i=1}^k\mathrm{B}_{\|\cdot\|_i}$ we deduce that for any $T\in\mathrm{Hom}(\mathbf{U},\mathbf{V})$ one has

$$\mathrm{N}(T)=\max_{\mathbf{u}_i\in\mathrm{B}_{\|\cdot\|_i},i\in\langle k\rangle}|||T(\otimes_{i=1}^k\mathbf{u}_i)|||.$$

Let $T=\otimes_{i=1}^kT_i$. Since $|||\cdot|||$ is a cross norm on $\mathbf{V}$ we deduce

$$\mathrm{N}(T)=\max_{\mathbf{u}_i\in\mathrm{B}_{\|\cdot\|_i},i\in\langle k\rangle}|||\otimes_{i=1}^kT_i(\mathbf{u}_i)|||=\max_{\mathbf{u}_i\in\mathrm{B}_{\|\cdot\|_i},i\in\langle k\rangle}\prod_{i=1}^k|||T_i(\mathbf{u}_i)|||_i=\prod_{i=1}^k\mathrm{N}_i(T_i).$$

□

Problems

1. Let $\mathbf{V}_1,\mathbf{V}_2$ be one dimensional subspaces with bases $\mathbf{v}_1,\mathbf{v}_2$ respectively. Let $C_1=[-\mathbf{e}_1,2\mathbf{e}_1],C_2=[-e_2,3e_2]$. Show that $C_1\odot C_2=[-3(\mathbf{e}_1\otimes\mathbf{e}_2),6\mathbf{e}_1\otimes\mathbf{e}_2]$. Hence $\mathcal{E}(C_1\odot C_2)$ is contained strictly in $\mathcal{E}(C_1)\odot\mathcal{E}(C_2)$. Lemma 7.8.2 yields that $\mathrm{conv}\,C_1\odot C_2=\mathrm{conv}\,\mathcal{P}_m\odot\mathcal{P}_n$.

2. Index the entries of $C \in \Omega_{mn}$ as $c_{(i,p),(j,q)}$. Show

   (a) Assume that $C \in \operatorname{conv}\Omega_m \odot \Omega_n$. Then the entries of C satisfy

   $$\sum_{i=1}^{m} c_{(i,p)(j,q)} = \sum_{j=1}^{m} c_{(i,p)(j,q)} \text{ for each } i,j \in \langle m\rangle,\ p,q \in \langle n\rangle,$$
   $$\sum_{p=1}^{n} c_{(i,p)(j,q)} = \sum_{q=1}^{n} c_{(i,p)(j,q)} \text{ for each } i,j \in \langle m\rangle\ p,q \in \langle n\rangle.$$

   (b) For $m = n = 2$ the standard conditions for 4×4 doubly stochastic matrices, and the conditions in part 2a characterize the set $\operatorname{conv}\Omega_2 \odot \Omega_2$.

   (c) For $m = n \geq 4$ the standard conditions for $n^2 \times n^2$ doubly stochastic matrices, and the conditions in part 2a gives a set which contains strictly $\operatorname{conv}\Omega_n \odot \Omega_n$. **Hint**: Consult with [Fri08].

3. Show

   (a) $A \in \mathrm{H}_{n,+,1}$ if and only if $\operatorname{tr} A = 1$ and $A \geq 0$.

   (b) $A \in \mathrm{H}_{n,+} \Rightarrow \det A \geq 0$.

   (c) Assume that $A = [a_{ij}]_{i=j=1}^{n} \in \mathrm{H}_n$ and $\det A > 0$. Then $A \geq 0$ if and only $\det [a_{ij}]_{i=j=1}^{p} > 0$ for $p = 1,\ldots,n-1$.

   (d) Assume that $A = [a_{ij}]_{i=j=1}^{n} \in \mathrm{H}_n$ and $\det A = 0$. Find $\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n$ such that $A\mathbf{x} = \mathbf{0}$. Let $\mathbf{x}_n = \frac{1}{\|\mathbf{x}\|}\mathbf{x}$ and complete $\mathbf{x}_n$ to an orthonormal basis $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Let $A_{n-1} = [\mathbf{x}_i^* A\mathbf{x}_j]_{i=j=1}^{n-1}$. Then $A_{n-1} \in \mathrm{H}_{n-1}$. Furthermore, $A \geq 0$ if and only if $A_{n-1} \geq 0$.

4. Let $\tau : \mathbb{C}^{n\times n}$ be the transpose map: $\tau(A) = A^\top$. Show

   (a) $\tau(A)$ is similar to A for any $A \in \mathbb{C}^{n\times n}$.

   (b) τ leaves invariant the following subsets of $\mathbb{C}^{n\times n}$:

   $$\mathbb{R}^{n\times n}, \mathrm{S_n(\mathbb{R})}, \mathrm{S_n(\mathbb{C})}, \mathrm{O(n,\mathbb{R})}, \mathrm{O(n,\mathbb{C})},$$
   $$\mathrm{U}(n,\mathbb{R}), \mathrm{N}(n,\mathbb{R}), \mathrm{N}(n,\mathbb{C}), \mathrm{H}_n, \mathrm{H}_{n,+}, \mathrm{H}_{n,+,1}.$$

5. On $\mathbb{C}^{mn\times mn}$ viewed as $\mathbb{C}^{m\times m}\otimes\mathbb{C}^{n\times n}$ we define the partial transpose τ_{par} as follows. Let $C = [c_{(i,p),(j,q)}]_{i=j=p=q=1}^{m,m,n,n}] \in \mathbb{C}^{m\times m}\otimes\mathbb{C}^{n\times n}$. Them $\tau_{\mathrm{par}}(C) = [\tilde{c}_{(i,p),(j,q)}]_{i=j=p=q=1}^{m,m,n,n}]$, where $\tilde{c}_{(i,p),(j,q)} = c_{(i,q),(j,p)}$ for $i,j \in \langle m\rangle, p,q \in \langle n\rangle$. Equivalently, τ_{par} is uniquely determined by the condition $\tau_{\mathrm{par}}(A\otimes B) = A\otimes B^\top$ for any $A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{n\times n}$. Show

(a) τ_{par} leaves the following subsets of $\mathbb{C}^{mn\times mn}$ invariant: $\mathrm{S_{mn}}(\mathbb{R}), \mathrm{H_{mn}}$, and all the set of the form $X \odot Y$, where $X \subset \mathbb{C}^{m\times m}$ and $Y \subset \mathbb{C}^{n\times n}$ are given in Problem 4b. In particular the convex set of separable states $\operatorname{conv} \mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1}$ is invariant under the partial transpose.

(b) Show that for $m = 2$ and $n = 2, 3$ $C \in \mathrm{H}_{mn}$ is a separable state, i.e. $C \in \operatorname{conv} \mathrm{H}_{m,+,1} \odot \mathrm{H}_{n,+,1}$, if and only $C, \tau_{\mathrm{par}}(C) \in \mathrm{H}_{mn,+,1}$. (This is the Horodecki-Peres condition [Hor96, Per96].)

6. Let the assumptions of Theorem 7.8.8 hold. Show

(a) Each $\mathbf{z} \in \mathbf{V}$ can be decomposed, usually in many ways, as a sum of rank one tensors

$$\mathbf{z} = \sum_{j=1}^{N} \otimes_{i=1}^{k} \mathbf{x}_{j,i},\ \mathbf{x}_{j,i} \in \mathbf{V}_i,\ i = 1, \ldots, k,\ j = 1, \ldots, N,$$

where $N = \prod_{i=1}^{k} \dim \mathbf{V}_i$. Then $\|\mathbf{z}\|_{\max}$ is the minimum of $\sum_{j=1}^{N} \prod_{i=1}^{k} \|\mathbf{x}_{j,i}\|_i$ over all the above decompositions of $\mathbf{z}$.

(b)
$$\|\mathbf{z}\|_{\min} = \max_{\mathbf{f}_i \in \mathrm{B}_{\|\cdot\|_i^*}, i=1,\ldots,k} |(\otimes_{i=1}^{k} \mathbf{f}_i)(\mathbf{z})|.$$

7. Prove Proposiiton 7.8.9. **Hint:** To prove the first part of the problem choose orthonormal bases in $\mathbf{V}_1, \ldots, \mathbf{V}_k$. To prove the second part observe that $\|\cdot\|$ is smooth, while $\|\cdot\|_{\min}, \|\cdot\|_{\max}$ are not smooth if $\dim \mathbf{V}_i > 1$ for $i = 1, \ldots, k > 1$.

7.9 Variation of tensor powers and spectra

Definition 7.9.1 *Let $\mathbf{V}_1, \mathbf{V}_2$ be finite dimensional vector spaces over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ with norms $\|\cdot\|_1, \|\cdot\|_2$ respectively. Let $\mu : \mathbf{V}_1 \to \mathbf{V}_2$ be a nonlinear map. The map μ has a Fréchet derivative at $\mathbf{x} \in \mathbf{U}$, or simply differentiable at $\mathbf{x}$, if there exists a linear transformation $T_{\mathbf{x}} \in \operatorname{Hom}(\mathbf{U}, \mathbf{V})$ such that*

$$\mu(\mathbf{x} + \mathbf{u}) = \mu(\mathbf{x}) + T_{\mathbf{x}}\mathbf{u} + o(\mathbf{u})\|\mathbf{u}\|_1,$$

where $\|o(\mathbf{u})\|_2 \to 0$ uniformly as $\|\mathbf{u}\|_1 \to 0$. Denote $\mathrm{D}\mu(\mathbf{x}) := T_{\mathbf{x}}$. μ is differentiable, if it has the Fréchet derivative at each $\mathbf{x} \in \mathbf{V}_1$, and $\mathrm{D}\mu(\mathbf{x})$ is continuous on $\mathbf{V}_1$. (Note that by choosing fixed bases in $\mathbf{V}_1, \mathbf{V}_2$ each $\mathrm{D}\mu(\mathbf{x})$ is represented by a matrix $A(\mathbf{x}) = [a_{ij}(\mathbf{x})] \in \mathbb{F}^{m\times n}$, where $n = \dim \mathbf{V}_1, m = \dim \mathbf{V}_2$. Then $a_{ij}(\mathbf{x})$ is continuous on $\mathbf{V}_1$ for each $i \in \langle m \rangle, j \in \langle n \rangle$.)

Since all norms on a finite dimensional vector space are equivalent, it is straightforward to show that the notion of Fréchet derivative depend only on the standard topologies in $\mathbf{V}_1, \mathbf{V}_2$. See Problem 1. For properties of the Fréchet derivative consult with [Die69].

Proposition 7.9.2 *Let $\mathbf{V}_1, \mathbf{V}_2$ be finite dimensional vector spaces over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ with norms $\|\cdot\|_1, \|\cdot\|_2$ respectively. Assume that $\mu : \mathbf{V}_1 \to \mathbf{V}_2$ is differentiable. Then for any $\mathbf{x}, \mathbf{y} \in \mathbf{V}_1$ the following equality and inequalities holds.*

$$\mu(\mathbf{y}) - \mu(\mathbf{x}) = \int_0^1 \mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})(\mathbf{y} - \mathbf{x})dt, \tag{7.9.1}$$

$$\begin{aligned} \|\mu(\mathbf{y}) - \mu(\mathbf{x})\|_2 &\le \|\mathbf{y} - \mathbf{x}\|_1 \int_0^1 \|\mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})\|_{1,2} dt \\ &\le \|\mathbf{y} - \mathbf{x}\|_1 \max_{t\in[0,1]} \|\mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})\|_{1,2}. \end{aligned} \tag{7.9.2}$$

(7.9.1) and (7.9.2) are called here the *mean value theorem* and the *mean value inequalities* respectively.

Proof. Let $\mathbf{x}, \mathbf{u} \in \mathbf{V}_1$ be fixed. Clearly, the function $\mu(\mathbf{x} + t\mathbf{u})$ is a differentiable function from $\mathbb{R}$ to $\mathbf{V}_2$, where

$$\frac{d\mu(\mathbf{x} + t\mathbf{u})}{dt} = \mathrm{D}\mu(\mathbf{x} + t\mathbf{u})\mathbf{u}. \tag{7.9.3}$$

Letting $\mathbf{u} = \mathbf{y} - \mathbf{x}$ and integrating the above inequality for $t \in [0, 1]$ we get (7.9.1). Replacing the integration in (7.9.1) by the limiting summation and using the triangle inequality we obtain

$$\begin{aligned} \|\mu(\mathbf{y}) - \mu(\mathbf{x})\|_2 \le \int_0^1 \|\mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})(\mathbf{y} - \mathbf{x})\|_1 dt \le \\ \int_0^1 \|\mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})\|_{1,2} \|(\mathbf{y} - \mathbf{x})\|_1 dt \le \\ \|\mathbf{y} - \mathbf{x}\|_1 \max_{t\in[0,1]} \|\mathrm{D}\mu((1-t)\mathbf{x} + t\mathbf{y})\|_{1,2}. \end{aligned}$$

□

Theorem 7.9.3 *Let $\mathbf{V}$ be a finite dimensional vector space. Let $k \in \mathbb{N}$. Denote $\mathbf{V}^{\otimes k} := \underbrace{\mathbf{V} \otimes \ldots \otimes \mathbf{V}}_{k}$. Consider the map $\delta_k : \mathbf{V} \to \mathbf{V}^{\otimes k}$, where $\delta_k(\mathbf{x}) = \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k}$. Then*

$$\mathrm{D}\delta_k(\mathbf{x})(\mathbf{u}) = \mathbf{u} \otimes \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-1} + \mathbf{x} \otimes \mathbf{u} \otimes \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-2} + \ldots + \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-1} \otimes \mathbf{u}. \tag{7.9.4}$$

Let $\| \cdot \|$ *be a norm on* $\mathbf{V}$ *and assume that* $\| \cdot \|_k$ *is a cross norm on* $\mathbf{V}^{\otimes k}$*:*

$$\|\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \ldots \otimes \mathbf{x}_k\|_k = \prod_{i=1}^{k} \|\mathbf{x}_i\| \textit{ for } \mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbf{V}. \tag{7.9.5}$$

Denote by $\mathrm{N}_k(T) := \|T\|_{\|\cdot\|,\|\cdot\|_k}$ *the operator norm of* $T \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V}^{\otimes k})$*. Then*

$$\mathrm{N}_k(\mathrm{D}\delta_k(\mathbf{x})) = k\|\mathbf{x}\|^{k-1}. \tag{7.9.6}$$

Proof. Fix $\mathbf{x}, \mathbf{u} \in \mathbf{V}$. For $t \in \mathbb{R}$ expand the vector $\delta_k(\mathbf{x}+t\mathbf{u})$ in powers of t. Then

$$\delta_k(\mathbf{x}+t\mathbf{u}) = \delta_k(\mathbf{x}) + t(\mathbf{u} \otimes \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-1} + \mathbf{x} \otimes \mathbf{u} \otimes \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-2} +$$
$$\ldots + \underbrace{\mathbf{x} \otimes \ldots \otimes \mathbf{x}}_{k-1} \otimes \mathbf{u}) + \text{ higher order terms in } t.$$

Hence (7.9.4) holds. Apply the triangle inequality to (7.9.4) and use the assumption that $\|\cdot\|_k$ is a cross norms to deduce the inequality $\|\mathrm{D}\delta_k(\mathbf{x})(\mathbf{u})\|_k \le k\|\mathbf{x}\|^{k-1}\|\mathbf{u}\|$. Hence $\mathrm{N}_k(\mathrm{D}\delta_k(\mathbf{x})) \le k\|\mathbf{x}\|^{k-1}$. Clearly, equality holds if $\mathbf{x} = \mathbf{0}$. Suppose that $\mathbf{x} \ne \mathbf{0}$. Then $\|\mathrm{D}\delta_k(\mathbf{x})(\mathbf{x})\|_k = k\|\mathbf{x}\|^k$. Hence $\mathrm{N}_k(\mathrm{D}\delta_k(\mathbf{x})) \ge k\|\mathbf{x}\|^{k-1}$, which establishes (7.9.6). □

Theorem 7.9.4 *Let* $\mathbf{U}$ *be a finite dimensional vector space over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$*. For an integer* $k > 1$ *consider the map* $\delta_k : \mathrm{Hom}\,(\mathbf{U}, \mathbf{U}) \to \mathrm{Hom}\,(\mathbf{U}, \mathbf{U})^{\otimes k} \sim \mathrm{Hom}\,(\mathbf{U}^{\otimes k}, \mathbf{U}^{\otimes k})$ *given by* $\delta_k(T) = \underbrace{T \otimes \ldots \otimes T}_{k}$*. Let* $\mathbf{W}_k \subset \mathbf{U}^{\otimes k}$ *be a subspace which is invariant for each* $\delta_k(T), T \in \mathrm{Hom}\,(\mathbf{U}, \mathbf{U})$*. Denote by* $\hat{\delta}_k : \mathrm{Hom}\,(\mathbf{U}, \mathbf{U}) \to \mathrm{Hom}\,(\mathbf{W}_k, \mathbf{W}_k)$ *the restriction map* $\delta_k(T)|\mathbf{W}_k$*. Assume that* $\| \cdot \|$ *is a norm on* $\mathbf{U}$*. Let* $\| \cdot \|_k$ *be the maximal cross norm* $\mathbf{U}^{\otimes k}$ *Let* $\| \cdot \|, \| \cdot \|_k, ||| \cdot |||_k$ *be the induced operator norms on* $\mathrm{Hom}\,(\mathbf{U}, \mathbf{U})$*,* $\mathrm{Hom}\,(\mathbf{U}, \mathbf{U})^{\otimes k}, \mathrm{Hom}\,(\mathbf{W}_k, \mathbf{W}_k)$ *respectively. Let* $N_k(\cdot), \hat{\mathrm{N}}_k(\cdot)$ *be the operator norm on* $\mathrm{Hom}\,(\mathrm{Hom}\,(\mathbf{U}, \mathbf{U}), \mathrm{Hom}\,(\mathbf{U}, \mathbf{U})^{\otimes k})$*,*
$\mathrm{Hom}\,(\mathrm{Hom}\,(\mathbf{U}, \mathbf{U}), \mathrm{Hom}\,(\mathbf{W}_k, \mathbf{W}_k))$ *respectively. Then*

$$\mathrm{N}_k(\mathrm{D}\delta_k(T)) = k\|T\|^{k-1}, \quad \hat{\mathrm{N}}_k(\mathrm{D}\hat{\delta}_k)(T)) \le k\|T\|^{k-1} \tag{7.9.7}$$

for any $T \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})$*.*

Proof. Theorem 7.8.12 yields that the operator norm $\|\cdot\|_k$ on $\mathrm{Hom}\,(\mathbf{U}^{\otimes k}, \mathbf{U}^{\otimes k})$, identified with $\mathrm{Hom}\,(\mathbf{U}, \mathbf{U})^{\otimes k}$, is a cross norm with respect to the operator norm on Hom $(\mathbf{U},\mathbf{U})$. Theorem 7.9.3 yields the equality in (7.9.7). Observe next that $\mathrm{D}\hat{\delta}_k(T)$ is $\mathrm{D}\delta_k(T)|\mathbf{W}_k$. Hence $\hat{\mathrm{N}}_k(\mathrm{D}\hat{\delta}_k(T)) \le \mathrm{N}(\mathrm{D}\delta_k(T))$, which implies the inequality in (7.9.7). □

A simple example of $\mathbf{W}_k$ is the subspace $\bigwedge^k \mathbf{U}$. See Problem 3.

Theorem 7.9.5 *Let* $\mathbf{U}$ *be an* n*-dimensional vector space over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$, *with a norm* $\|\cdot\|$. *Denote by* $\|\cdot\|$ *the induced operator norm on* Hom $(\mathbf{V},\mathbf{V})$. *Then for* $A, B \in$ Hom $(\mathbf{U},\mathbf{U})$
(7.9.8)

$$|\det A - \det B| \le \|A-B\| \frac{\|A\|^n - \|B\|^n}{\|A\| - \|B\|} \le n\|A-B\|[\max(\|A\|,\|B\|)]^{n-1}.$$

Here $\frac{a^n - a^n}{a-a} := na^{n-1}$ *for any* $a \in \mathbb{C}$. *The first inequality is sharp for* $A = aI_n, B = bI_n$ *for* $a, b \ge 0$. *The constant* n *in the second inequality is sharp.*

Proof. In $\mathbf{U}^{\otimes n}$ consider the one dimensional invariant subspace $\mathbf{W}_n := \bigwedge^n \mathbf{U}$ for each $\delta_n(T), T \in$ Hom $(\mathbf{U},\mathbf{U})$. See Problem 3. Let $\mathbf{e}_1,\ldots,\mathbf{e}_n$ be a basis in $\mathbf{U}$. Then $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n$ is a basis vector in $\bigwedge^n \mathbf{U}$. Furthermore

$$\delta_n(T)(\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n) = (\det T)\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n.$$

See Proposition 5.2.7. Note that $\hat{\delta}_n(T) := \delta_n(T)|\bigwedge^n \mathbf{U}$ is the above operator. Observe next that any $Q \in$ Hom $(\bigwedge^n \mathbf{U}, \bigwedge^n \mathbf{U})$, is of the from

$$Q(\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n) = t\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \ldots \wedge \mathbf{e}_n.$$

Hence the operator norm of Q is $|t|$. We now apply Theorem 7.9.4 to this case. The inequality in (7.9.7) yields

$$\hat{\mathrm{N}}_n(\mathrm{D}\hat{\delta}_n(T)) \le n\|T\|^{n-1}.$$

Next we apply Proposition 7.9.2, where $\mathbf{V}_1 :=$ Hom $(\mathbf{U},\mathbf{U})$ and $\mathbf{V}_2 =$ Hom $(\bigwedge^n \mathbf{U}, \bigwedge^n \mathbf{U})$ equipped with the operator norms, and $\mu(T) = \hat{\delta}_n(T)$. So $\|\mu(A) - \mu(B)\|_2 = |\det A - \det B|$. The inequality (7.9.2) combined with the inequality in (7.9.7) yield

$$|\det A - \det B| \le n\|A-B\| \int_0^1 \|(1-t)A + tB\|^{n-1} dt \le$$

$$n\|A-B\| \int_0^1 ((1-t)\|A\| + t\|B\|)^{n-1} dt =$$

$$\|A-B\| \frac{\|A\|^n - \|B\|^n}{\|A\| - \|B\|} = \|A-B\| \sum_{i=0}^{n-1} \|A\|^{n-1-i}\|B\|^i =$$

$$n\|A-B\|[\max(\|A\|,\|B\|)]^{n-1}.$$

This shows (7.9.8). Recall that $\|xI\| = |x|$ for any $x \in \mathbb{F}$. Hence, for $A = aI, B = bI$ and $a, b \geq 0$ equality holds in the first inequality of (7.9.8). To show that the constant n can not be improved let $A = (1+x)I, B = I$, where $x > 0$. Then (7.9.8) is equivalent to the inequality $(1+x)^n - 1 \leq nx(1+x)^{n-1}$. Since $\lim_{x \searrow 0} \frac{(1+x)^n - 1}{x(1+x)^{n-1}} = n$ the constant n can not be improved. □

Definition 7.9.6 *Let Σ_n be the group of permutations $\sigma : \langle n\rangle \to \langle n\rangle$. Let $\mathrm{S} = \{\lambda_1, \ldots, \lambda_\mathrm{n}\}, \mathrm{T} = \{\mu_1, \ldots, \mu_\mathrm{n}\}$ be two multisets in $\mathbb{C}$ containing n elements each. Let*

$$\begin{aligned}\mathrm{dist}(\mathrm{S},\mathrm{T}) &= \max_{\mathrm{j}\in\langle\mathrm{n}\rangle}\min_{\mathrm{i}\in\langle\mathrm{n}\rangle}|\lambda_\mathrm{j} - \mu_\mathrm{i}|,\\ \mathrm{hdist}(\mathrm{S},\mathrm{T}) &= \max(\mathrm{dist}(\mathrm{S},\mathrm{T}), \mathrm{dist}(\mathrm{T},\mathrm{S})),\\ \mathrm{pdist}(\mathrm{S},\mathrm{T}) &= \min_{\sigma\in\Sigma_\mathrm{n}}\max_{\mathrm{i}\in\langle\mathrm{n}\rangle}|\lambda_\mathrm{i} - \mu_{\sigma(\mathrm{i})}|.\end{aligned}$$

Note: $\mathrm{dist}(\mathrm{S},\mathrm{T})$ is the distance from S to T, viewed as sets; $\mathrm{hdist}(\mathrm{S},\mathrm{T})$ is the Hausdorff distance between S and T, viewed as sets; $\mathrm{pdist}(\mathrm{S},\mathrm{T})$ is called *permutational* distance between two multisets of cardinality n. Clearly

$$\begin{aligned}&\mathrm{hdist}(\mathrm{S},\mathrm{T}) = \mathrm{hdist}(\mathrm{T},\mathrm{S}), \quad \mathrm{pdist}(\mathrm{S},\mathrm{T}) = \mathrm{pdist}(\mathrm{T},\mathrm{S}),\\ &\mathrm{dist}(\mathrm{S},\mathrm{T}) \leq \mathrm{hdist}(\mathrm{S},\mathrm{T}) \leq \mathrm{pdist}(\mathrm{S},\mathrm{T}).\end{aligned} \tag{7.9.9}$$

See Problem 4.

Theorem 7.9.7 *Let $\mathbf{U}$ be an n-dimensional vector space of $\mathbb{C}$ with the norm $\|\cdot\|$. Let $\|\cdot\|$ be the induced operator norm on* Hom $(\mathbf{U}, \mathbf{U})$*. For $A, B \in$* Hom $(\mathbf{U}, \mathbf{U})$ *let $\mathrm{S(A)}, \mathrm{S(B)}$ be the eigenvalue multisets of A, B of cardinality n respectively. Then*

$$\mathrm{pdist}(\mathrm{S(A)},\mathrm{S(B)}) \leq 4\mathrm{e}^{\frac{1}{2\mathrm{e}}}\mathrm{n}\|\mathrm{A}-\mathrm{B}\|^{\frac{1}{\mathrm{n}}}[\max(\|\mathrm{A}\|,\|\mathrm{B}\|)]^{\frac{\mathrm{n}-1}{\mathrm{n}}}. \tag{7.9.10}$$

To prove the theorem we need the following lemma.

Lemma 7.9.8 *Let the assumptions of Theorem 7.9.7 holds. Define*

$$h(A,B) := \max_{t\in[0,1]}\mathrm{dist}(\mathrm{S}((1-\mathrm{t})\mathrm{A}+\mathrm{tB}),\mathrm{S}(\mathrm{B})). \tag{7.9.11}$$

Then

$$\mathrm{pdist}(\mathrm{S(A)},\mathrm{S(B)}) \leq (2\mathrm{n}-1)\mathrm{h}(\mathrm{A},\mathrm{B}). \tag{7.9.12}$$

Proof. Let $\mathrm{S(A)} = \{\lambda_1(\mathrm{A}), \ldots, \lambda_\mathrm{n}(\mathrm{A})\}, \mathrm{S(B)} = \{\lambda_1(\mathrm{B}), \ldots, \lambda_\mathrm{n}(\mathrm{B})\}$. Let $\mathrm{D}(z, r) := \{w \in \mathbb{C}, |w - z| \le r\}$. Denote $\mathrm{K_B} = \cup_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{D}(\lambda_\mathrm{i}(\mathrm{B}), \mathrm{h(A,B)})$. Then $\mathrm{K_B}$ is a closed compact set, which decomposes as union of a $k \in \langle n\rangle$ connected components. Let $A(t) = (1-t)A+tB$. Since $\mathrm{dist(S(A(t)), S(B))} \le \mathrm{h(A,B)}$ we deduce that $\mathrm{K_B}$ contains $\mathrm{S(A(t))}$ for each $t \in [0,1]$. As $\mathrm{S(A(t))}$ various continuously for $t \in [0,1]$, each connected component of $\mathrm{K_B}$ contains a fixed number of the eigenvalues of $\mathrm{S(A(t))}$ counting with their multiplicities. Since $A(1) = B$, each connected component of $\mathrm{K_B}$ contains a fixed number of the eigenvalues of A and B counting with their multiplicities. Rename the eigenvalues of B such that indices of the eigenvalues of A and B are the same in each component of $\mathrm{K_B}$.

Let $\mathrm{C} = \cup_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{D(z_i, h(A,B))}$ be such a connected component, where $z_1, \ldots, z_p$ are p distinct eigenvalues of B. C contains exactly $q \ge p$ eigenvalues of A and B respectively. We claim that if $\lambda \in \mathrm{S(A)} \cap \mathrm{C}$ then $\max_{j \in \langle p\rangle} |\lambda - z_j| \le (2p-1)h(A,B)$. Consider a simple graph $G = (V, E)$, where $V = \langle p\rangle$ and $(i,j) \in E$ if and only if $|z_i - z_j| \le 2h(A,B)$. Since C is connected it follows that G is connected hence the maximal distance between two distinct point in G is $p-1$. So $|z_i - z_j| \le 2(p-1)h(A,B)$. Since $|\lambda - z_i| \le h(A,B)$ for some $i \in \langle p\rangle$, it follows that $|\lambda - z_j| \le (2p-1)h(A,B) \le (2n-1)h(A,B)$. Therefore for this particular renaming of the eigenvalues of B we have the inequality $|\lambda_i(A) - \lambda_i(B)| \le (2n-1)h(A,B), i = 1, \ldots, n$. □

Problem 5 shows that the inequality (7.9.12) is sharp.

Proof of Theorem 7.9.7. First observe that

$$\mathrm{dist(S(A), S(B))^n} \le \max_{\mathrm{i}\in\langle \mathrm{n}\rangle} |\prod_{\mathrm{j}=1}^{\mathrm{n}} (\lambda_\mathrm{i}(\mathrm{A}) - \lambda_\mathrm{j}(\mathrm{B}))| =$$
$$\max_{i\in\langle n\rangle} |\det\,(\lambda_i(A)I - B) - \det\,(\lambda_i(A)I - A)| \le$$
$$\max_{z\in\mathbb{C}, |z|\le\rho(A)} |\det\,(zI - B) - \det\,(zI - A)|.$$

We now apply (7.9.8) to deduce that for $|z| \le \rho(A) \le \|A\|$ we have

$$|\det\,(zI - B) - \det\,(zI - A)| \le n\|A - B\|[\max(\|zI - A\|, \|zI - B\|)]^{n-1}$$
$$\le n\|A - B\|[\max(|z| + \|A\|, (|z| + \|B\|)]^{n-1} \le$$
$$n\|A - B\|[\max(2\|A\|, \|A\| + \|B\|)]^{n-1}.$$

Thus

$$\mathrm{dist(S(A), S(B))} \le \mathrm{n}^{\frac{1}{\mathrm{n}}} \|\mathrm{A} - \mathrm{B}\|^{\frac{1}{\mathrm{n}}} [\max(2\|\mathrm{A}\|, \|\mathrm{A}\| + \|\mathrm{B}\|)]^{\frac{\mathrm{n}-1}{\mathrm{n}}}. \tag{7.9.13}$$

We apply the above inequality to $A(t)$ for each $t \in [0,1]$. Clearly, for $t \in [0,1]$

$$\begin{aligned}\|A(t)\| \le (1-t)\|A\| + t\|B\| \le \max(\|A\|, \|B\|) \Rightarrow \\ \max(2\|A(t)\|, \|A(t)\| + \|B\|) \le 2\max(\|A\|, \|B\|).\end{aligned}$$

Also $\|A(t) - B\| = (1-t)\|A - B\| \le \|A - B\|$. Hence we deduce

$$h(A,B) \le n^{\frac{1}{n}} \|A - B\|^{\frac{1}{n}} [2\max(|A\|, \|B\|)]^{\frac{n-1}{n}}. \tag{7.9.14}$$

Use (7.9.12) to obtain

$$\mathrm{pdist}(\mathrm{S(A)}, \mathrm{S(B)}) \le (2\mathrm{n} - 1)\mathrm{n}^{\frac{1}{\mathrm{n}}} \|\mathrm{A} - \mathrm{B}\|^{\frac{1}{\mathrm{n}}} [2\max(|\mathrm{A}\|, \|\mathrm{B}\|)]^{\frac{\mathrm{n}-1}{\mathrm{n}}}. \tag{7.9.15}$$

Use the inequality

$$(2n-1)2\big(\frac{n}{2}\big)^{\frac{1}{n}} \le 4n\big(\frac{n}{2}\big)^{\frac{1}{n}} \le 4ne^{\frac{1}{2e}} \text{ for } n \in \mathbb{N}, \tag{7.9.16}$$

to deduce (7.9.10). (See Problem 6.) □

The inequality (7.9.10) can be improved by fact 2 using the following theorem [EJRS83]. (See Problem 7.)

Theorem 7.9.9 *Let* $\mathbf{U}$ *be an* n*-dimensional vector space of* $\mathbb{C}$. *For* $A, B \in \mathrm{Hom}\,(\mathbf{U}, \mathbf{U})$ *let* $\mathrm{S(A)}, \mathrm{S(B)}$ *be the eigenvalue multisets of* A, B *of cardinality* n *respectively. Then*

$$\mathrm{pdist}(\mathrm{S(A)}, \mathrm{S(B)}) \le (2\lfloor\frac{\mathrm{n}+1}{2}\rfloor - 1)\max(\mathrm{h(A,B)}, \mathrm{h(B,A)}). \tag{7.9.17}$$

The above constant are sharp.

Problems

1. Let $\mu : \mathbf{V}_1 \to \mathbf{V}_2$ be a nonlinear map. Show

   (a) Assume that μ has a Fréchet derivative at $\mathbf{x}$ with respect to given two norms $\|\cdot\|_1, , \|\cdot\|_2$. Then μ has a Fréchet derivative at $\mathbf{x}$ with respect to any two norms $\|\cdot\|_a, , \|\cdot\|_b$.

   (b) Suppose that μ has a Fréchet derivative at $\mathbf{x}$. Then μ is continuous at $\mathbf{x}$ with respect to the standard topologies on $\mathbf{V}_1, \mathbf{V}_2$.

(c) Assume that μ has a Fréchet derivative at each point of a compact set $O \subset \mathbf{V}_1$. Then $\mu : O \to \mathbf{V}_2$ is uniformly continuous.

2. Let $\mathbf{U}_1, \ldots, \mathbf{U}_k, \mathbf{V}_1, \ldots, \mathbf{V}_k$ be finite dimensional inner product vector space overs $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\mathbf{U} := \otimes_{i=1}^k \mathbf{U}_i, \mathbf{V} := \otimes_{i=1}^k \mathbf{V}_i$ have the induced inner product. Identify Hom $(\mathbf{U}, \mathbf{V})$ with

   $\prod_{i=1}^k$ Hom $(\mathbf{U}_i, \mathbf{V}_i)$. Show

   (a) The operator norm on Hom $(\mathbf{U}, \mathbf{V})$, with respect to Hilbert norms, is a normal cross norm with respect to the operator norms on Hom $(\mathbf{U}_i, \mathbf{V}_i)$, the Hilbert norms, for $i = 1, \ldots, k$. **Hint:** Express the operator norm on Hom $(\mathbf{U}_i, \mathbf{V}_i)$ and its conjugate norm in terms of singular values of $T_i \in$ Hom $(\mathbf{U}_i, \mathbf{V}_i)$ for $i = 1, \ldots, k$.

   (b) Assume that $\mathbf{U}_1 = \mathbf{V}_1 = \ldots = \mathbf{U}_k = \mathbf{V}_k$. Let $\delta_k :$ Hom $(\mathbf{U}_1, \mathbf{U}_1) \to$ Hom $(\mathbf{U}, \mathbf{U})$. Then $\mathrm{N}_k(\delta_k(T)) = k\|T\|^{k-1}$, where $\|\cdot\|$ is the operator norm on Hom $(\mathbf{U}_1, \mathbf{U}_1)$.

3. Let $\mathbf{U}$ be a vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ of dimension $n > 1$. Let $k \in [2, n]$ be an integer. Show

   (a) $\mathbf{W}_k := \bigwedge^k \mathbf{U}$ is an invariant subspace for each $\delta_k(T)$ given in Theorem 7.9.4.

   (b) Assume that $\mathbf{U}$ is an inner product space. Let $T \in$ Hom $(\mathbf{U}, \mathbf{U})$, and denote by $\|T\| = \sigma_1(T) \geq \ldots \geq \sigma_k(T) \geq \ldots$ the singular values of T. Then

   $$\hat{\mathrm{N}}_k(\mathrm{D}\hat{\delta}_k(T)) = \sum_{i=1}^k \sigma_1(T) \ldots \sigma_{i-1}(T)\sigma_{i+1}(T) \ldots \sigma_k(T).$$

   In particular, $\hat{\mathrm{N}}_k(\mathrm{D}\hat{\delta}_k(T)) \leq k\sigma_1(T)^{k-1} = k\|T\|^{k-1} = \mathrm{N}_k(\mathrm{D}\delta_k(T))$. Equality holds if and only if $\sigma_1(T) = \ldots = \sigma_k(T)$. **Hint:** Consult with [BhF81].

4. Prove (7.9.9).

5. Let $A = \mathrm{diag}(0, 2, 4, \ldots, 2n-2), B = (2n-1)I_n \in \mathbb{R}^{n\times n}$. Show that in this case equality holds in (7.9.12)

6. Using the fact that $\min_{t\in[0,1]} -t\log t = \frac{1}{e}$ deduce the last part of (7.9.16).

7. Show

(a) Let $n = 2k+1$ and

$$A = \mathrm{diag}(\underbrace{0,\ldots,0}_{k+1}, 2, 4, \ldots, 2k),$$
$$B = \mathrm{diag}(1, 3, \ldots, 2k-1, \underbrace{2k+1, \ldots, 2k+1}_{k+1}).$$

Then equality holds in (7.9.17).

(b) Let $n = 2k$ and

$$A = \mathrm{diag}(\underbrace{0,\ldots,0}_{k}, 2, 4, \ldots, 2k),$$
$$B = \mathrm{diag}(1, 3, \ldots, 2k-1, \underbrace{2k+1, \ldots, 2k+1}_{k}).$$

Then equality holds in (7.9.17).

(c) $\max(h(A,B), h(B,A))$ is bounded above by the right-hand side of (7.9.14).

(d) Deduce from (7.9.17) and the previous part of the problem the improved version of (7.9.10).
(7.9.18)
$$\mathrm{pdist(S(A), S(B)) \le 2e^{\frac{1}{2e}} n \|A - B\|^{\frac{1}{n}} [\max(\|A\|, \|B\|)]^{\frac{n-1}{n}}}.$$

7.10 Variation of permanents

Definition 7.10.1 *For $A = [a_{ij}] \in \mathbb{D}^{n\times n}$ the permanent of A, denoted as* perm A

$$\mathrm{perm\ A} = \sum_{\sigma\in\Sigma_\mathrm{n}} \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}_{\mathrm{i}\sigma(\mathrm{i})},$$

where Σ_n is the group of permutations $\sigma : \langle n\rangle \to \langle n\rangle$.

The determinant and the permanent share some common properties as mulitlinear functions on $\mathbb{D}^{n\times n}$, as Laplace expansions. However, from the computational point of view the determinants are easy to compute while permanents are hard to compute over all fields, except the fields of characteristic 2. (Over the field of characteristic 2 $\mathrm{perm\ A = det\ A}$.) For $A \in \mathbb{Z}_+^{n\times n}$ the permanent of A has a fundamental importance in combinatorics, and usually is hard to evaluate [Val79]. The main aim of this section is to generalize the inequality (7.9.8) to the permanents of matrices. The analog of (7.9.8) holds for the norms $\ell_p, p \in [1,\infty]$ [BhE90]. However, it was also shown in [BhE90] that the analog of fails for some operator norm on $\mathbb{C}^{n\times n}$.

Theorem 7.10.2 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^n$, and denote by $\|\cdot\|$ the induced operator norm on $\mathbb{C}^n$. Let $A, B \in \mathbb{C}^{n\times n}$. Then*

$$\text{(7.10.1)} \qquad |\text{perm A} - \text{perm B}| \le \frac{\|A-B\|(\|A\|^n - \|B\|^n)}{2(\|A\|-\|B\|)} + \frac{\|A^*-B^*\|(\|A^*\|^n - \|B^*\|^n)}{2(\|A^*\|-\|B^*\|)}.$$

To prove this theorem we need two lemmas. The first lemma gives the following formula for the standard numerical radius of a square matrix with complex entries.

Lemma 7.10.3 *Let $A \in \mathbb{C}^{n\times n}$. Then*

$$\text{(7.10.2)} \qquad r_2(A) = \max_{|z|=1} \rho(\frac{zA + \bar{z}A}{2}).$$

In particular $r_2(A) \le \frac{1}{2}(\|A\| + \|A^*\|)$ for any operator norm on $\mathbb{C}^n$.

Proof. Let $z \in \mathrm{S}^1, \mathrm{B} = \mathrm{zA}$. Assume that $\mathbf{x}$ is an eigenvector of $\frac{1}{2}(B + B^*)$ of length one corresponding to the eigenvalue λ. Then

$$|\lambda| = |\Re(\mathbf{x}^*(zA)\mathbf{x})| \le |\mathbf{x}^*(zA)\mathbf{x}| = |\mathbf{x}^*A\mathbf{x}| \le r_2(A).$$

Hence the right-hand side of (7.10.2) is not bigger its left-hand side. On the other hand there exists $\mathbf{x} \in \mathbb{C}^n, \mathbf{x}^*\mathbf{x} = 1$ and $z \in \mathbb{C}, |z| = 1$ such that $r_2(A) = |\mathbf{x}^*A\mathbf{x}| = \mathbf{x}^*(zA)\mathbf{x}$. For this value of z we have that

$$r_2(A) \le \lambda_1(\frac{zA + \bar{z}A}{2}) \le \rho(\frac{zA + \bar{z}A}{2}).$$

Clearly,

$$\rho(\frac{zA + \bar{z}A}{2}) \le \|\frac{zA + \bar{z}A}{2}\| \le \frac{\|A\| + \|A\|}{2}.$$

Hence $r_2(A) \le \frac{1}{2}(\|A\| + \|A^*\|)$. □

For $A \in \mathbb{C}^{n\times n}$, view the matrix $\otimes^n A$ as a linear operator on $\otimes^n \mathbb{C}^n$, which is identified with $\mathbb{C}^{n^n}$. $\omega_2(\otimes^n A), r_2(\otimes^n A)$ are the numerical range and the numerical radius of $\otimes^n A$ corresponding to the inner product $\langle\cdot,\cdot\rangle$ on $\otimes^n \mathbb{C}^n$ induced by the standard inner product $\mathbf{y}^*\mathbf{x}$ on $\mathbb{C}^n$.

Lemma 7.10.4 *Let* $A \in \mathbb{C}^{n\times n}$*. Then* $\mathrm{perm\ A} \in \omega_2(\otimes^{\mathrm{n}}\mathrm{A})$.

Proof. Assume that $\mathbf{e}_i = (\boldsymbol{\delta} ta_{i1}, \ldots, \delta_{in})^\top, i \in \langle an\rangle$ is a standard basis in $\mathbb{C}^n$. Then $\otimes_{j=1}^n e_{i_j}$, where $i_j \in \langle n\rangle, j = 1, \ldots, n$, is the standard basis in $\otimes^n\mathbb{C}^n$. A straightforward calculation shows that

$$(7.10.3) \qquad \langle \otimes^n A\mathbf{x}, \mathbf{x}\rangle = \mathrm{perm\ A},\ \mathbf{x} = \frac{1}{\sqrt{n!}} \sum_{\sigma\in\Sigma_n} \otimes_{i=1}^n \mathbf{e}_{\sigma(i)}, \langle \mathbf{x}, \mathbf{x}\rangle = 1,$$

where Σ_n is the set of permutations $\sigma : \langle n\rangle \to \langle n\rangle$. (See Problem 1.) Hence $\mathrm{perm\ A} \in \omega_2(\otimes^{\mathrm{n}}\mathrm{A})$. □

Proof of Theorem 7.10.2. Since $\mathbf{x}$ given in (7.10.3) does not depend on A we deduce $\mathrm{perm\ A} - \mathrm{perm\ B} \in \omega_2(\otimes^{\mathrm{n}}\mathrm{A} - \otimes^{\mathrm{n}}\mathrm{B})$. Let $|||\cdot|||$ the maximal cross norm on $\otimes^n\mathbb{C}^n$ induced by the norm $\|\cdot\|$ on $\mathbb{C}^n$. Denote by $|||\cdot|||$ the operator norm on $|||\cdot|||$ on $\otimes^n\mathbb{C}^{n\times n}$, induced by the norm $|||\cdot|||$ on $\otimes^n\mathbb{C}^n$. Use the definition of $r_2(\otimes^n A - \otimes^n B)$ and Lemma 7.10.3 to deduce

$$(7.10.4) \qquad |\mathrm{perm\ A} - \mathrm{perm\ B}| \le \mathrm{r}_2(\otimes^{\mathrm{n}}\mathrm{A} - \otimes^{\mathrm{n}}\mathrm{B} \le$$
$$\frac{1}{2}(|||\otimes^n A - \otimes^n B||| + |||\otimes^n A^* - \otimes^n B^*|||).$$

Theorem 7.8.12 implies that the operator norm $|||\cdot|||$ on $\otimes^n\mathbb{C}^{n\times n}$, induced by the norm $|||\cdot|||$ on $\otimes^n\mathbb{C}^n$, is a cross norm with respect to the operator norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$. Observe next

$$\otimes^n A - \otimes^n B = \sum_{i=0}^{n-1} (\otimes^i B) \otimes (A - B) \otimes^{n-1-i} A.$$

(Here $\otimes^0 A$ means that this term does not appear at all.) Use the triangular inequality and the fact that the operator norm $|||\cdot|||$ is a cross norm we deduce

$$|||\otimes^n A - \otimes^n B||| \le \sum_{i=0}^{n-1} |||(\otimes^i B) \otimes (A - B) \otimes^{n-1-i} A||| \le$$
$$\sum_{i=0}^{n-1-i} \|B\|^i \|A - B\|\ \|A\|^{n-1-i} = \frac{\|A - B\|(\|A\|^n - \|B\|^n)}{\|A\| - \|B\|}.$$

Apply the above inequality to (7.10.4) to deduce the theorem. □

Problems

1. Prove (7.10.3).

2. Let the assumptions of Theorem 7.10.2 hold. Show that

$$\text{(7.10.5)} \qquad |\mathrm{perm\ A - perm\ B}| \le \|\mathrm{A - B}\| \frac{\|\mathrm{A}\|^{\mathrm{n}} - \|\mathrm{B}\|^{\mathrm{n}}}{\|\mathrm{A}\| - \|\mathrm{B}\|}$$

   in the following cases.

   (a) $A = A^*, B = B^*$.

   (b) The norm $\|\cdot\|$ on $\mathbb{C}^n$ is $\|\cdot\|_2$.

3. Show that (7.10.5) holds for the norms $\|\cdot\|_1, \|\cdot\|_\infty$ using the following steps.

   (a) For $A = [a_{ij}] \in \mathbb{C}^n$ denote $|A| := [|a_{ij}|] \in \mathbb{R}_+^{n\times n}$. Show

$$|\mathrm{perm\ A}| \le \mathrm{perm\ |A|} \le \prod_{\mathrm{i=1}}^{\mathrm{n}} \sum_{\mathrm{j=1}}^{\mathrm{n}} |\mathrm{a_{ji}}|$$

   (b) Let $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n], B = [\mathbf{b}_1, \ldots, \mathbf{b}_n] \in \mathbb{C}^{n\times n}$, where $\mathbf{a}_i, \mathbf{b}_i$ are the $i - th$ columns of A, B respectively, for $i = 1, \ldots, n$. Let

$$C_0 = [\mathbf{a}_1 - \mathbf{b}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n],\ C_{n-1} = [\mathbf{b}_1, \ldots, \mathbf{b}_{n-1}, \mathbf{a}_n - \mathbf{b}_n],$$
$$C_i = [\mathbf{b}_1, \ldots, \mathbf{b}_i, \mathbf{a}_{i+1} - \mathbf{b}_{i+1}, \mathbf{a}_{i+2}, \ldots, \mathbf{a}_n],\ \text{for } i = 1, \ldots, n-2.$$

   Then

$$\mathrm{perm\ A - perm\ B} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{perm\ C_{i-1}} \Rightarrow$$
$$|\mathrm{perm\ A - perm\ B}| \le \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{perm\ |C_{i-1}|}.$$

   (c) Recall that $\|A\|_1 = \|\ |A|\ \| = \max_{i\in\langle n\rangle} \|\mathbf{a}_i\|_1$. Then

$$\|\ C_{i-1}\ \| \le \|\mathbf{a}_i - \mathbf{b}_i\|_1 \|B\|_1^{i-1} \|A\|_1^{n-i} \le \|A - B\|_1 \|B\|_1^{i-1} \|A\|_1^{n-i},$$

   for $i = 1, \ldots, n$. Hence, (7.10.5) holds for $\|\cdot\|_1$ norm.

   (d) Use the equalities $\mathrm{perm\ A^\top = perm\ A}, \|\mathrm{A}\|_\infty = |\mathrm{A}^\top|_1$ deduce that (7.10.5) holds for $\|\cdot\|_\infty$ norm.

7.11 The complexity of $\mathrm{conv}\,\Omega_n \odot \Omega_m$

In this section we show that there a linear programming problem on $\Omega_{n,m} := \mathrm{conv}\,\Omega_n \odot \Omega_m$, whose solution gives an answer to the *subgraph isomorphism problem*, that will be stated precisely below. The subgraph isomorphism problem belongs to the class of NP-complete problems [GaJ79]. This shows, in our opinion, that the number of half spaces characterizing $\Omega_{n,m}$ is probably not polynomial in $\max(m,n)$, which is analogous to Theorem 7.8.6.

By graph G in this section we mean an undirected simple graph on the set of vertices V and the set of edges E. Here E is a subset of unordered pairs $\mathrm{P}(V) : \{(u,v), u \ne v \in V\}$, where (u,v) and (v,u) are identified. We will denote by $G = (V,E)$ the graph to emphasize the set of vertices and edges of G. The degree of $v \in V$, denoted by deg v is the number of edges that re connected to v, i.e. $\#\{u,\ (v,u) \in E\|$. Clearly, $\sum_{v\in V} \deg v = 2\#E$. A vertex $v \in V$ is called *isolated* if $\deg v = 0$. Denote by V_{iso} the set of isolated vertices in V. A subgraph of $G_1 = (V_1, E_1)$ of G is given by the condition $V_1 \subset V, E_1 \subset E \cap \mathrm{P}(V)$.

Definition 7.11.1 *Let $G = (V,E), G' = (V',E')$ be two undirected simple graphs. Then G and G' are called isomorphic if the following condition hold. There is a bijection $\phi : V\backslash\{V_{\mathrm{iso}}\} \to V'\backslash\{V'_{\mathrm{iso}}\}$ such that $(u,v) \in E$ if and only if $(\phi(u),\phi(v)) \in E'$. G' is called isomorphic to a subgraph of G if there exists a subgraph G_1 of G such that G' is isomorphic to G_1.*

We note that our definition of isomorphisms of two graphs are slightly different from the standard definition of graph isomorphism. Since the set of isolated vertices in graph are easily identified, i.e. $(\#V)^2$ steps, from the complexity point of view our definition is equivalent to the standard definition of graph and subgraph isomorphisms. We recall that the subgraphs isomorphism problem, which asks asking if G' is isomorphic to a subgraph of G, is an NP-complete problem [GaJ79].

We now relate the SGIP to certain linear programming problems on $\Omega_{m,n}$. We first recall the notion adjacency matrix of G. Assume that $\#V = m$ and label the vertices in V as $1,\ldots,m$, i.e. we let $\#V = \langle m\rangle$. Then the incidence matrix $A(G) = [a_{ij}]_{i=j=1}^{m} \in \{0,1\}^{m\times m}$ is a symmetric matrix with zero diagonal such that $a_{ij} = 1$ if and only the edge (i,j) is in E. Note that a different labeling of the elements of V gives rise to the adjacency matrix $A' = PA(G)P^\top$ for some permutation matrix $P \in \mathcal{P}_m$. Thus the graph G gives rise to the conjugacy class of matrices $\mathcal{A}(G) = \{PA(G)P^\top,\ P \in \mathcal{P}_m\}$. The following result is straightforward, see Problem 1.

Lemma 7.11.2 *Let $G=(V,E), G'=(V',E')$ are two undirected graphs. Assume that $\#V=\#V'$. Then G and G' are isomorphic if and only if $\mathcal{A}(G)=\mathcal{A}(G')$.*

We next introduce the notion of the vertex-edge incidence matrix $B(G)\in\{0,1\}^{\#V\times\#E}$. Assume that $G=(V,E)$ and let $m=\#V, n\#E$. Label the vertices of V and E by $\langle m\rangle$ and $\langle n\rangle$ respectively. Then $B(G)=[b_{ij}]_{i=j=1}^{m,n}\in\{0,1\}^{m\times n}$, such that $b_{ij}=1$ if and only the edge j contain the vertex i. A different labeling of V and E gives rise to the vertex-edge incidence matrix $B'=PB(A)Q$ for some $P\in\mathcal{P}_m, Q\in\mathcal{P}_n$. Thus the graph G gives rise to the equivalence class of matrices $\mathcal{B}(G)=\{PB(G)Q,\ P\in\mathcal{P}_m, Q\in\mathcal{P}_n\}$.

Lemma 7.11.3 *Let $G=(V,E), G'=(V',E')$ are two undirected graphs. Assume that $\#V=\#V', \#E=\#E'$. Then G and G' are isomorphic if and only if $\mathcal{B}(G)=\mathcal{B}(G')$.*

We now restate the SGIP in terms of bilinear programming on $\Omega_m\times\Omega_n$. It is enough to consider the following case.

Lemma 7.11.4 *Let $G'=(V',E'), G=(V,E)$ and assume $m':=\#V'\le m:=\#V, n':=\#E'\le n:=\#E$. Let $B(G')\in\{0,1\}^{m'\times n'}, B(G)\in\{0,1\}^{m\times n}$ be the vertex-edges incidence matrices of G' and G. Denote by $C(G')\in\{0,1\}^{m\times n}$ the matrix obtained from $B(G')$ by adding additional $m-m'$ and $n-n'$ zero rows and columns respectively. Then*

$$\max_{P\in\mathcal{P}_m, Q\in\mathcal{P}_n}\operatorname{tr}(C(G')QB(A)^\top P)\le 2n'. \tag{7.11.1}$$

Equality holds if and only if G' is isomorphic to a subgraph of G.

Proof. Let $B_1=[b_{ij,1}]_{i=j=1}^{m\times n}:=P^\top B(A)Q^\top\in\mathcal{B}(G)$. Note that B_1 has exactly the same number of ones as $B(G)$, namely 2n, since each edge is connected is connected to two vertices. Similarly $C(G')=[c_{ij}]_{i=j=1}^{m,n}$ has exactly the same number of ones as $B(G')$, namely $2n'$. Hence the $\langle C(G'),B_1\rangle=\operatorname{tr}(C(G_1)B_1^\top)\le 2n'$. Assume that $\operatorname{tr}(C(G_1)B_1^\top)=2n'$. So we can delete $2(n-n')$ ones in B_1 to obtain $C(G')$. Note that deleting $2(n-n')$ from B_1, means to delete $n-n'$ edges from the graph G. Indeed, assume $c_{ij}=b_{ij,1}=1$. So the vertex i is connected to the edge j. Hence there exists another vertex $i'\ne i$ such that $c_{i'j}=1$. As $\operatorname{tr}(C(G')B_1^\top)=2n'$ we deduce that $b_{i'j,1}=1$. Hence, if we rename the vertices and the edges of G corresponding to B_1 we deduce that G', represented by the matrix $B(G')$, is a subgraph of G. □

We now show how to translate the maximum in (7.11.1) to linear programming problem on $\Omega_{m,n}$. As in §2.8 for $F\in\mathbb{R}^{n\times m}$ let $\hat{F}\in\mathbb{R}^{nm}$

be a vector composed of the columns of F, i.e. first we have the coordinates of the first column, then the coordinates of the second column, and the last n coordinates are the coordinates of the last column. Hence $\widehat{XFY} = (Y^\top \otimes X)\hat{F}$, where $Y^\top \otimes X$ is the Kronecker tensor product.

Lemma 7.11.5 *Let $C, B \in \mathbb{R}^{m\times n}$. Then*

$$\max_{P\in\mathcal{P}_m, Q\in\mathcal{P}_n} \operatorname{tr}(CQB^\top P) = \max_{Z\in\Omega_{n,m}} (\hat{C})^\top Z\hat{B}. \tag{7.11.2}$$

Proof. Since $\Omega^\top = \Omega$ and $\mathcal{E}(\Omega_m) = \mathcal{P}_m$ we deduce

$$\max_{P\in\mathcal{P}_m, Q\in\mathcal{P}_n} \operatorname{tr}(CQB^\top P) = \max_{X\in\Omega_m, Y\in\Omega_n} \operatorname{tr}(CYB^\top X).$$

Observe next that

$$\operatorname{tr}(CYB^\top X) = \operatorname{tr}(C^\top(X^\top BY^\top)) = (\hat{C})^\top (Y\otimes X^\top)\hat{B}.$$

As $\Omega_{n,m} = \operatorname{conv}\Omega_n \odot \Omega_m$ we deduce (7.11.2). □

In summary we showed that if we can solve exactly the linear programming problem (7.11.2), using Lemma 7.11.4 we can determine if G' is isomorphic to a subgraph of G. Since the SGIP is NP-complete, we believe that this implies that for general m, n the number of half spaces characterizing $\Omega_{n,m}$ can not be polynomial.

Problems

1. Prove Lemma 7.11.2.

2. Prove Lemma 7.11.3.

7.12 Vivanti-Pringsheim theorem and applications

We start with the following basic result on the power series in one complex variable, which is usually called the Cauchy-Hadamard formula on power series [Rem98, §4.1].

Theorem 7.12.1 *Let*

$$f(z) = \sum_{i=0}^{\infty} a_i z^i, \quad a_i \in \mathbb{C}, i = 0, 1, \ldots, \textit{ and } z \in \mathbb{C}, \tag{7.12.1}$$

be power series. Define

$$R = R(f) := \frac{1}{\limsup_i |a_i|^{\frac{1}{i}}} \in [0, \infty]. \tag{7.12.2}$$

(R-is called the radius of convergence of the series.) Then

1. *For $R = 0$ the series converge only for $z = 0$.*
2. *For $R = \infty$ the series converge absolutely and uniformly for each $z \in \mathbb{C}$, and $f(z)$ is an entire function, i.e. analytic on $\mathbb{C}$.*
3. *For $R \in (0, \infty)$ the series converge absolutely and uniformly to an analytic function for each $z, |z| < R$, and diverge for each $|z| > R$. Furthermore, there exist $\zeta, |\zeta| = R$, such that $f(z)$ can not be extended to an analytic function in any neighborhood of ζ. (ζ is called a singular point of f.)*

Consider the Taylor series for the function complex valued $\frac{1}{1-z}$

$$\frac{1}{1-z} = \sum_{i=0}^{\infty} z^i.$$

Then $R = 1$, the function $\frac{1}{1-z}$ is analytic in $\mathbb{C}\backslash\{1\}$, and has a singular point at $z = 1$. Vivanti-Pringsheim theorem is an extension of this example [Viv93, Pri94].

Theorem 7.12.2 *Let the power series $f(z) = \sum_{i=0}^{\infty} a_i z^i$ have positive finite radius of convergence R, and suppose that the sequence $a_i, i = 0, 1, \ldots,$ is eventually nonnegative. (I.e. all but finitely many of its coefficients are real and nonnegative.) Then $\zeta := R$ is a singular point of f.*

See [Rem98, §8.1] for a proof. In what follows we need a stronger version of this theorem for rational functions, e.g. [Fri78b, Thm 2]. Assume that $f(z)$ is a rational function with 0 as a point of analyticity. So f has power series (7.12.1). Assume that f is not polynomial, i.e. $R(f) \in (0, \infty)$. Then f has the following form.

$$f(z) = P(z) + \sum_{i=1}^{N} \sum_{j=1}^{p_i} \frac{b_{j,i}}{(1 - \lambda_i z)^j}, \tag{7.12.3}$$
$$P \in \mathbb{C}[z], \lambda_i, b_{p_i,i} \in \mathbb{C}\backslash\{0\}, \lambda_i \neq \lambda_{i'} \text{ for } i \neq i'.$$

Note that

$$R(f) = \frac{1}{\max_i |\lambda_i|}. \tag{7.12.4}$$

Definition 7.12.3 *Let $f(z)$ be a rational function of the form (7.12.3). Let $p := \max_{|\lambda_i|=R(f)^{-1}} p_i$. Denote*

$$f_{\mathrm{prin}} = \sum_{i:|\lambda_i|=R(f)^{-1} \text{ and } p_i=p} \frac{b_{p,i}}{(1-\lambda_i z)^p}.$$

f_{prin} is called the principle part of f, i.e. $f - f_{\mathrm{prin}}$ does not have poles of order p on $|z| = R(f)$.

Theorem 7.12.4 *Let $f(z)$ be a rational function of the form (7.12.3). Assume that the sequence of coefficients in the power expansion (7.12.1) is eventually nonnegative. Then*

1. *The set $\{\lambda_1, \ldots, \lambda_N\}$ is symmetric with respect to $\mathbb{R}$. That is, for each $i \in \langle N\rangle$ there exists $i' \in \langle N\rangle$ such that $\bar\lambda_i = \lambda_{i'}$. Furthermore $p_i = p_{i'}$, and $\bar b_{j,i} = b_{j,i'}$ for $j = 1, \ldots p_i$ and $i = 1, \ldots, N$.*

2. *After renaming the indices in $\langle N\rangle$ we have: $\lambda_1 = \frac{1}{R(f)}$, $|\lambda_i| = \lambda_1$ for $i = 2, \ldots, M$, and $|\lambda_i| > \lambda_1$ for $i > M$. (Here $M \in [1, N]$.).*

3. *Let $p := p_1$. There exists an integer $L \in [1, M]$ such that $p_i = p$ for $i \in [2, L]$, and $p_i < p$ for $i > L$.*

4. *$b_{p,1} > 0$ and there exists $m \in [1, L]$ such that $|b_{p,i}| = b_{p,1}$ for $i = 1, \ldots, m$ and $|b_{p,i}| < b_{p,1}$ for $i \in [m+1, L]$.*

5. *Let $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$. After renaming the indices $2, \ldots, m$, $\lambda_i = \zeta^{i-1}\lambda_1$ for $i = 2, \ldots, m$. Furthermore there exists an integer $l \in [1, m]$ such that $b_{p,i} = \zeta^{l(i-1)} b_{p,1}$ for $i = 2, \ldots, m$.*

6. *$f_{\mathrm{prin}}(\zeta z) = \zeta^{-l} f_{\mathrm{prin}}(z)$.*

Proof. We outline the major steps in the proof of this theorem. For all details see the proof of [Fri78b, Thm 2]. By considering $g(z) = f(z) + P_1$ for some polynomial P_1, we may assume that the MacLaurin coefficients of g are real and nonnegative. As $g_{\mathrm{prin}} = f_{\mathrm{prin}}$, without loss of generality we may assume that the MacLaurin coefficients of f real and nonnegative. Hence $\overline{f(\bar z)} = f(z)$ for each z where f is defined. This shows part 1.

Part 2 follows from Theorem 7.12.2. For simplicity of the exposition assume that $R(f) = 1$. Recall that for each singular point $\lambda_i^{-1} = \bar{\lambda}_i$ of f on the circle $|z| = 1$ we have the equality

$$b_{p_i,i} = \lim_{r \nearrow 1} (1 - r)^{p_i} f(\bar{\lambda}_i r). \tag{7.12.5}$$

In particular $b_{p_1,1} = \lim_{r \nearrow 1}(1 - r)^{p_1} f(r) \geq 0$. Since $b_{p_1,1} \neq 0$ we obtain that $b_{p_1,1} > 0$. Let $p := p_1$. Since all the MacLaurin coefficients of f are nonnegative we have the inequality $|f(z)| \leq f(|z|)$ for all $|z| < 1$. Hence $\limsup_{r \nearrow 1}(1-r)^p |f(\bar{\lambda}_i r)| \leq b_{p,1}$. This inequality and (7.12.5) implies parts 3– 4.

For $m = 1$ parts 5– 6 are trivial. Assume that $m > 1$. Let $b_{p,i} = \eta_i b_{p,1}, |\eta_i| = 1$ for $i = 2, \ldots, m$. In view of the part 1 for each integer $i \in [2, m]$ there exists integer $i' \in [2, m]$ such that $\bar{\lambda}_i = \lambda_{i'}, \bar{\eta}_i = \eta_{i'}$. Consider the function

$$g(z) = 2f(z) - \eta_i f(\lambda_i z) - \bar{\eta}_i f(\bar{\lambda}_i z) = \sum_{j=0}^{\infty} 2(1 - \Re(\eta_i \lambda_i^j)) a_j z^j.$$

So the MacLaurin coefficients of g are nonnegative. Clearly $R(g) \geq 1$, and if g has a pole at 1, its order is at most $p - 1$. This implies the equality

$$2f_{\mathrm{prin}}(z) - \eta_i f_{\mathrm{prin}}(\lambda_i z) - \bar{\eta}_i f_{\mathrm{prin}}(\bar{\lambda}_i z) = 0.$$

Therefore the set $\{\lambda_1, \ldots, \lambda_m\}$ form a multiplicative group of order m. Hence, it is a group of of all m-roots of unity. So we can rename the indices $2, \ldots, m$ such that $\lambda_i = \zeta^{(i-1)}$ for $i = 1, \ldots, m$. Similarly, $\eta_1 = 1, \eta_2, \ldots, \eta_m$ form a multiplicative group, which must be a subgroup of m roots of 1. Furthermore $\eta_i \mapsto \lambda_i$ is a group homomorphism. This shows part 5. Part 5 straightforward implies part 6. □

Definition 7.12.5 *Let* $\mathrm{S} = \{\lambda_1, \ldots, \lambda_\mathrm{n}\} \subset \mathbb{C}$ *be a finite multiset. I.e. a point* $z \in \mathrm{S}$ *appears exactly* $m(z) \geq 1$ *times in* S. *Denote*

1. $r(\mathrm{S}) := \max_{\mathrm{z} \in \mathrm{S}} |\mathrm{z}|$.

2. *For any* $t \geq 0$ *denote by* $\mathrm{S(t)}$ *the multiset* $\mathrm{S} \cap \{\mathrm{z} \in \mathbb{C}, |\mathrm{z}| = \mathrm{t}\}$.

3. *For an integer* $k \in \mathbb{N}$ *denote by* $s_k(\mathrm{S}) := \sum_{\mathrm{i}=1}^{\mathrm{n}} \lambda_\mathrm{i}^\mathrm{k}$ *the* $k-th$ *moment of* S. *Let* $s_0(\mathrm{S}) = \mathrm{n}$.

4. *For* $\mathbf{z} = (z_1, \ldots, z_N)^\top \in \mathbb{C}^N$ *denote by* $\sigma_k(\mathbf{z}) = \sum_{1 \leq i_1 < \ldots i_k \leq N} z_{i_1} \ldots z_{i_N}$, *for* $k = 1, \ldots, N$ *the elementary symmetric polynomials in* $z_1, \ldots, z_N$.

5. *Denote* $\mathbf{z}(\mathrm{S}) = (\lambda_1, \ldots, \lambda_\mathrm{n})^\top \in \mathbb{C}^\mathrm{n}$. *Then* $\sigma_k(\mathrm{S}) := \sigma_\mathrm{k}(\mathbf{z}(\mathrm{S}))$ *for* $k = 1, \ldots, n$ *are called the elementary symmetric polynomials of* S.

S *is called a Frobenius multiset if the following conditions hold.*

1. $\bar{\mathrm{S}} = \mathrm{S}$.
2. $r(\mathrm{S}) \in \mathrm{S}$.
3. $m(z) = 1$ *for each* $z \in \mathrm{S}(\mathrm{r}(\mathrm{S}))$.
4. *Assume that* $\#\mathrm{S}(\mathrm{r}(\mathrm{S})) = \mathrm{m}$. *Then* $\zeta\mathrm{S} = \mathrm{S}$ *for* $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$.

A simple example of Frobenius multiset is the set of eigenvalues, counted with their mulitplicities, of a square nonnegative irreducible matrix.

Theorem 7.12.6 *Let* $\mathrm{S} \subset \mathbb{C}$ *be a multiset. Assume that the moments* $s_k(\mathrm{S}), \mathrm{k} \in \mathbb{N}$ *are eventually nonnegative. Then the following conditions hold.*

1. $r(\mathrm{S}) \in \mathrm{S}$.
2. *Denote* $\mu := m(r(\mathrm{S}))$. *Assume that* $\lambda \in \mathrm{S}(\mathrm{r}(\mathrm{S}))$. *Then* $m(\lambda) \le \mu$.
3. *Assume that* $r(\mathrm{S}) > 0$ *and suppose that* $\lambda_1 = r(\mathrm{S}), \lambda_2, \ldots, \lambda_\mathrm{m}$ *are all the distinct elements of* S *satisfying the conditions* $|\lambda_i| = r(\mathrm{S}), \mathrm{m}(\lambda_\mathrm{i}) = \mu$ *for* $i = 1, \ldots, m$. *Then* $\frac{\lambda_i}{r(\mathrm{S})}, i = 1, \ldots, m$ *are the* m *distinct roots of* 1.
4. *Let* $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$. *Then* $\zeta\mathrm{S}(\mathrm{r}(\mathrm{S})) = \mathrm{S}(\mathrm{r}(\mathrm{S}))$.
5. *If* $r(\mathrm{S}) > 0, \mu = 1$ *and none of the other elements of* S *are positive, then* S *is a Frobenius multiset.*

Proof. For a finite multiset $\mathrm{S} \subset \mathbb{C}$ define

$$f_\mathrm{S}(z) := \sum_{\lambda \in \mathrm{S}} \frac{1}{1 - \lambda z} = \sum_{k=0}^{\infty} s_k(\mathrm{S})\mathrm{z}^\mathrm{k}. \tag{7.12.6}$$

Apply Theorem 7.12.4 to deduce the parts 1–4.

Assume that $r(\mathrm{S})$ is the only positive element of S and $\mu = 1$. If $m = 1$ then S is a Frobenius set. Suppose the $m > 1$. Consider the function $g(z) = 2f_\mathrm{S}(z) - f_\mathrm{S}(\zeta z) - f_\mathrm{S}(\bar{\zeta} z)$. We claim that g is the zero function. Suppose to the contrary that $g \ne 0$. In view of 4 we deduce $R(f) < R(g) < \infty$. Since the MacLaurin coefficients of g are eventually nonnegative, Theorem 7.12.4 yields that g must have a singular point $\xi > 0$ whose residue at ξ is positive. Since $\rho(A)$ is the only positive eigenvalue of A, all other positive residues of g, coming from $2f_{\mathrm{S}}$ are not located on positive numbers. Hence the residues of g at its poles located on the positive axes are negative integers. The above contradiction shows that $g = 0$, i.e. S is a Frobenius multiset. $\square$

Let $A \in \mathbb{C}^{n\times n}$ and assume that S(A) is the eigenvalue multiset of A. Then

$$f_{\mathrm{S(A)}}(z) = \operatorname{tr}(I - zA)^{-1} = \sum_{i=0}^{\infty} (\operatorname{tr} A^i) z^i. \tag{7.12.7}$$

Corollary 7.12.7 *Let $A \in \mathbb{C}^{n\times n}$. Denote by* S *be the multiset consisting of all eigenvalues of A, counted with multiplicities. Assume that the traces of $A^k, k \in \mathbb{N}$ are eventually nonnegative. Then the following conditions hold.*

1. *$\rho(A)$ is an eigenvalue of A.*

2. *Assume that the algebraic multiplicity of $\rho(A)$ is μ. Let λ be an eigenvalue of A of multiplicity $m(\lambda)$ satisfying $|\lambda| = \rho(A)$. Then $m(\lambda) \le \mu$.*

3. *Assume that $\rho(A) > 0$ and suppose that $\lambda_1 = \rho(A), \lambda_2, \ldots, \lambda_m$ are all the distinct eigenvalues of A satisfying the conditions $|\lambda_i| = \rho(A), m(\lambda_i) = \mu$ for $i = 1, \ldots, m$. Then $\frac{\lambda_i}{\rho(A)}, i = 1, \ldots, m$ are the m distinct roots of 1.*

4. *Let $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$. Then $\zeta\mathrm{S}(\rho(\mathrm{A})) = \mathrm{S}$.*

5. *If $\rho(A) > 0$ is an algebraically simple eigenvalue of A, and none of the other eigenvalues of A are positive, then* S *is a Frobenius multiset.*

Definition 7.12.8 *$A \in \mathbb{R}^{n\times n}$ is called eventually nonnegative if $A^k \ge 0$ for all integers $k > N$.*

Lemma 7.12.9 *Let $B \in \mathbb{R}^{n\times n}$. Then there exists a positive integer M with the following property. Assume that $L > M$ is a prime. Suppose that B^L is similar to a nonnegative matrix. Then the eigenvalue multiset of B is a union of Frobenius multisets.*

Proof. Associate with the eigenvalues of B the following set $\mathrm{T} \subset \mathrm{S}^1$. For $0 \ne \lambda \in \operatorname{spec} \mathrm{B}$ we let $\frac{\lambda}{|\lambda|} \in \mathrm{T}$. For $\lambda \ne \kappa \in \operatorname{spec}\mathrm{B}$ satisfying the conditions $|\lambda| = |\kappa| > 0$ we assume that $\frac{\lambda}{\kappa}, \frac{\kappa}{\lambda} \in \mathrm{T}$. Let $\mathrm{T}_1 \subset \mathrm{T}$ be the set of all roots of 1 that are in T. Recall that $\eta \in \mathrm{S}^1$ is called a primitive k-root of 1, if $\eta^k = 1$, and $\eta^{k'} \ne 1$ for all integers $k' \in [1, k)$. k is called the primitivity index of η. Let $L > k$ be a prime. Then η^L is a k-primitive root of 1. Furthermore, the map $\eta \mapsto \eta^L$ is an isomorphism of the group of of all $k-th$ roots of 1, which commutes with conjugation $\eta \mapsto \bar{\eta}$. Clearly, if $\eta \in \mathrm{S}$, and η is not a root of unity then η^L is not root of unity. Define $M \in \mathbb{N}$ to be the maximum over all primitivity indices of $\eta \in \mathrm{T}_1$. If $\mathrm{T}_1 = \emptyset$ then $M = 1$. Let $L > M$ be a prime. Assume that B^L is similar to $C \in \mathbb{R}_+^{n\times n}$. Apply Theorem 6.4.4 and the Perron-Frobenius theorem to each irreducible diagonal block of of the matrix in (6.4.3), to deduce that the eigenvalue multiset of $\mathrm{S(C)}$ is $\cup_{j=1}^{t+f}\mathrm{F_j}$ and each $\mathrm{F_j}$ a Frobenius multiset.

Clearly, $\overline{\mathrm{spec\,B}} = \mathrm{spec\,B}$ and $\mathrm{spec\,B^L} = \mathrm{spec\,C}$. Observe next that the condition that L is a prime satisfying $L > M$ implies that the map $z \mapsto z^L$ induces a $1-1$ and onto map $\phi : \mathrm{spec\,B} \to \mathrm{spec\,C}$. Moreover, $\phi^{-1}(r) > 0$ if and only of $r > 0$. Hence ϕ can be extended to a $1-1$ and onto map $\tilde{\phi} : \mathrm{S(B)} \to \mathrm{S(C)}$. Furthermore, $\phi^{-1}(\mathrm{F_j})$ is a Frobenius set, where the number of distinct points $\mathrm{F_j(r(F_j))}$ is equal to the number of points in $\phi^{-1}(\mathrm{F_j})(\mathrm{r}(\phi^{-1}(\mathrm{F_j})))$. Hence $\mathrm{S(B)} = \cup_{\mathrm{j=1}}^{\mathrm{t+f}}\phi^{-1}(\mathrm{F_j})$ is a decomposition of $\mathrm{S(B)}$ to a union of Frobenius multisets. □

Corollary 7.12.10 *Assume that a matrix $B \in \mathbb{R}^{n\times n}$ is similar to an eventually nonnegative matrix. Then the eigenvalue multiset* S(B) *of B is a union of Frobenius multisets.*

Theorem 7.12.11 *Assume that the eigenvalue multiset* S(B) *of $B \in \mathbb{R}^{n\times n}$ is a union of Frobenius multisets. Then there an eventually nonnegative $A \in \mathbb{R}^{n\times n}$, such that* $\mathrm{S(A)} = \mathrm{S(B)}$.

Proof. It is enough to show that for a given Frobenius multiset F there exists an eventually $A \in \mathbb{R}^{n\times n}$ such that $\mathrm{S(A)} = \mathrm{F}$. The claim is trivial if $\mathrm{F} = \{0\}$. Assume that $r(\mathrm{F}) > 0$. Without loss of generality we can assume that $r(\mathrm{F}) = 1$. Suppose first that $\mathrm{F} \cap \mathrm{S}^1 = \{1\}$. To each real point $\lambda \in \mathrm{F}$ of multiplicity $m(\lambda)$ we associate $m(\lambda)$ the diagonal matrix $G(\lambda) = \lambda I_{m(\lambda)} \in \mathbb{R}^{m(\lambda)\times m(\lambda)}$. For nonreal points $\lambda \in \mathrm{F}$ of multiplicity $m(\lambda)$ we associate the block diagonal matrix $H(\lambda) = I_{m(\lambda)} \otimes \begin{bmatrix} 2\Re(\lambda) & |\lambda|^2 \\ -1 & 0 \end{bmatrix}$. Note that $H(\bar{\lambda}) = H(\lambda)$. Let

$$C := [1] \oplus_{\lambda\in\mathrm{F}\cap\mathbb{R}\setminus\{1\}} G(\lambda) \oplus_{\lambda\in\mathrm{F},\Im\lambda>0} H(\lambda),$$
$$C = C_0 + C_1, \quad C_0 = [1] \oplus 0_{(n-1)\times(n-1)},$$
$$C_1 = [0] \oplus (\oplus_{\lambda\in\mathrm{F}\cap\mathbb{R}\setminus\{1\}} G(\lambda) \oplus_{\lambda\in\mathrm{F},\Im\lambda>0} H(\lambda)).$$

Clearly,

$$\mathrm{S(C) = F,\ C_0C_1 = C_1C_0 = 0,\ C_0^m = C_0},$$
$$C^m = C_0^m + C_1^m,\ \rho(C_1) < 1,\ \lim_{m\to\infty} C_1^m = 0.$$

Let $X \in \mathbf{GL}(n,\mathbb{R})$ be a matrix such $X\mathbf{1} = X^\top \mathbf{1} = \mathbf{e}_1$. Define

$$A_0 := X^{-1}C_0X = \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \quad A_1 = X^{-1}C_1X, \quad A := A_0 + A_1 = X^{-1}X.$$

So $\mathrm{S(A) = S(C) = F}$. Also

$$A^m = A_0^m + A_1^m,\ A_0^m = A_0,\ \lim_{m\to\infty} A_1^m \Rightarrow \lim_{m\to\infty} A^m = A_0.$$

So A is eventually positive.

Assume now that F is a Frobenius set with $r(\mathrm{F}) = 1$ such that $\mathrm{F} \cap \mathrm{S}^1$ consist of exactly $m > 1$ roots of unity. Let $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$. Recall that $\zeta\mathrm{F} = \mathrm{F}$. Let $\mathrm{F} = \mathrm{F}_1 \cup \mathrm{F}_0$, where $0 \notin \mathrm{F}_1$ and F_0 consists of $m(0)$ copies of 0. $(m(0) = 0 \iff \mathrm{F}_0 = \emptyset$. If $\mathrm{F}_0 \neq \emptyset$ then the zero matrix of order $m(0)$ has F_0 as its eigenvalue multiset. Thus it is enough to show that there exists an eventually nonnegative matrix B whose eigenvalue multiset is F_1. Clearly, F_1 is a Frobenius set satisfying $r(\mathrm{F}_1) = 1$ and $\mathrm{F} \cap \mathrm{S}^1$ consist of exactly $m > 1$ roots of unity. Assume that all the elements of F_1, counted with their mulitplicity are the coordinates of the vector $\mathbf{z} = (z_1, \ldots, z_N)^\top \in \mathbb{C}^N$. Let

be the the elementary symmetric polynomials in $z_1, \ldots, z_N$. Hence the multiset F_1 consists of the roots of $P(z) := z^N + \sum_{k=1}^N (-1)^k \sigma_k(\mathbf{z}) z^{N-k}$. Since $\bar{\mathrm{F}}_1 = \mathrm{F}_1$ it follows that each $\sigma_k(\mathbf{z})$ is real. As $\zeta\mathrm{F}_1 = \mathrm{F}_1$ we deduce that $N = mN'$ and $\sigma_k = 0$ if m does not divide k. Let F_2 be the root multiset $Q(z) := z^{N'} + \sum_{k=1}^{N'} (-1)^{mk} \sigma_{km}(\mathbf{z}) z^{N'-k}$. Clearly, $\bar{\mathrm{F}}_2 = \mathrm{F}_2$ Since $1 \in \mathrm{F}_1$ it follows that $1 \in \mathrm{F}_2$. Furthermore, $\mathrm{F}_1 = \phi^{-1}(\mathrm{F}_2)$, where $\phi(z) : \mathbb{C} \to \mathbb{C}$ is the map $z \mapsto z^m$. That is, if $z \in \mathrm{F}_2$ has multiplicity $m(z)$ then $\phi^{-1}(z)$ consists of m points, each of multiplicity $m(z)$ such that these m-points are all the solutions of $w^m = z$. Hence $\mathrm{F}_2 \cap \mathrm{S}^1 = \{1\}$. Therefore F_2 is a Frobenius set.

According to the previous case there exists an eventually nonnegative matrix $A \in \mathbb{R}^{N' \times N'}$ such that F_2 is its eigenvalue multiset. Let $P \in \mathcal{P}_m$ be a permutation matrix corresponding to the cyclic permutation on $\langle m \rangle$ $i \mapsto i+1$, for $i = 1, \ldots, m$, where $m+1 \equiv 1$. Consider the matrix $B = P \otimes A$. Then B is eventually nonnegative, and the eigenvalue multiset of B is F_1 □

7.13 Inverse eigenvalue problem for nonnegative matrices

The following problem is called the *inverse eigenvalue problem for nonnegative matrices*, abbreviated an IEPFNM:

Problem 7.13.1 *Let* $\mathrm{S} \subset \mathbb{C}$ *be a multiset consisting of* n *points, (counting with their multiplicities.) Find necessary and sufficient conditions such that there exists a nonnegative* $A \in \mathbb{R}^{n\times n}$ *whose eigenvalue multiset is* S.

Proposition 7.13.2 *Let* $A \in \mathbb{R}_+^{n\times n}$*. Then the eigenvalue multiset* S *satisfies the following conditions.*

1. S *is a union of Frobenius multisets.*

2. *All the moments of* $s_k(\mathrm{S}) \ge 0$.

3. $s_{kl}(\mathrm{S}) \ge \frac{s_k(\mathrm{S})^l}{n^{l-1}}$ *for each* $k, l \in \mathbb{N}$.

Proof. 1 Follows from Theorem 6.4.4 and the Perron-Frobenius theorem applied to each irreducible diagonal block of of the matrix in (6.4.3). Since $A^k \ge 0$ it follows that $\operatorname{tr} A^k \ge 0$. Hence 2 holds. Since $A^k \ge 0$ it is enough to show the inequality in 3 for $k = 1$. Decompose $A = [a_{ij}]$ as $D + A_0$, where $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ and $A_0 := A - D \ge 0$. So $A^l - D^l \ge A_0^l \ge 0$. Hence $\operatorname{tr} A^l \ge \operatorname{tr} D^l = \sum_{i=1}^l a_{ii}^l$. Hólder inequality for $p = l$ yield that $\sum_{i=1}^n a_{ii} \le (\sum_{i=1}^n a_{ii}^l)^{\frac{1}{l}} n^{\frac{l-1}{l}}$, which yields 3. □

The following result gives simple sufficient conditions for a mulitset S to be the eigenvalue multiset of a nonnegative matrix.

Proposition 7.13.3 *Let* $\mathrm{S} \subset \mathbb{C}$ *be a multiset containing* n *elements, counting with multiplicities. Assume that the elementary symmetric polynomials corresponding to* S *satisfy* $(-1)^{k-1}\sigma_k(\mathrm{S}) \ge 0$ *for* $k = 1, \ldots, n$*. Then there exists* $A \in \mathbb{R}_+^{n\times n}$ *such that* S *is the eigenvalue multiset of* A.

Proof. Note that the companion matrix to the polynomial $P(z) = z^n + \sum_{i=1}^n (-1)^i \sigma_i(\mathrm{S}) \mathrm{z}^{\mathrm{n-i}}$ is a nonnegative matrix. □

Recall the MacLaurin inequalities [HPL52, p' 52].

Proposition 7.13.4 *Let* $\mathbf{w} = (w_1, \ldots, w_{n-1})^\top \in \mathbb{R}_+^{n-1}$*. Then the sequence* $\big(\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\big)^{\frac{1}{k}}$ *nonincreasing for* $k = 1, \ldots, n-1$.

Proposition 7.13.5 *Let* S *be a multiset of real numbers, which contains exactly one positive number. Assume that the sum of all elements in* S *is nonnegative. Then* S *satisfies the conditions of Proposition 7.13.3. In particular, there exists $A \in \mathbb{R}_+^{n\times n}$ such that* S *is the eigenvalue multiset of A.*

Proof. Without loss of generality we may assume that $S = \{1, -w_1, \ldots, w_{n-1}\}$ where $w_i \ge 0$ for $i = 1, \ldots, n-1$ and $1 \ge \sum_{i=1}^{n-1} w_i$. Denote $\mathbf{z} = (1, -w_1, \ldots, -w_{n-1})^\top$ and $\mathbf{w} = (w_1, \ldots, w_{n-1})^\top$. Clearly, $\sigma_1(\mathbf{z}) \ge 0, (-1)^{n-1}\sigma_n(\mathbf{z}) = \sigma_{n-1}(\mathbf{w}) \ge 0$. Observe next that

$$\sigma_{k+1}(\mathbf{z}) = (-1)^k(\sigma_k(\mathbf{w}) - \sigma_{k+1}(\mathbf{w})) \text{ for } k = 1, \ldots, n-2.$$

Thus to prove that $(-1)^k\sigma_{k+1}(\mathbf{z}) \ge 0$ it is enough to show that that the sequence $\sigma_i(\mathbf{w}), i = 1, \ldots, n-1$ is nonincreasing.

Observe that $\sigma_1(\mathbf{w}) \le 1$. We now use Use Proposiiton 7.13.5. First observe that

$$\Big(\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\Big)^{\frac{1}{k}} \le \frac{\sigma_1(\mathbf{w})}{n-1} \le \frac{1}{n-1} \text{ for } k = 1, \ldots, n-1.$$

Next

$$\sigma_k(\mathbf{w}) - \sigma_{k+1}(\mathbf{w}) \ge \sigma_k(\mathbf{w}) - \Big[\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\Big]^{\frac{k+1}{k}}\binom{n-1}{k+1} =$$
$$\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\Big[\binom{n-1}{k} - \Big(\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\Big)^{\frac{1}{k}}\binom{n-1}{k+1}\Big] \ge$$
$$\frac{\sigma_k(\mathbf{w})}{\binom{n-1}{k}}\Big[\binom{n-1}{k} - \frac{1}{n-1}\binom{n-1}{k+1}\Big] \ge 0$$

Hence S satisfies the conditions of Proposition 7.13.3. The last part of Proposition 7.13.3 yields that there exists $A \in \mathbb{R}_+^{n\times n}$ such that S is the eigenvalue multiset of A. □

Example 7.13.6 *Let* $\mathrm{S} = \{\sqrt{2}, \sqrt{-1}, -\sqrt{-1}\}$. *Then* S *is a Frobenius set. Furthermore, $s_2(\mathrm{S}) = 0$ and all other moments of* S *are positive. Hence the condition 3 of Proposition 7.13.2 does not hold for $k = 1, l = 2$. In particular, there is an eventually nonnengative matrix $A \in \mathbb{R}_+^{3\times 3}$, which can not be nonnegative, whose eigenvalue multiset is* S.

Theorem 7.13.7 *Let* $\mathrm{S} = \{\lambda_1, \lambda_2, \lambda_3\}$ *be a multiset satisfying the following properties.*

1. $r(\mathrm{S}) \in \mathrm{S}$.
2. $\overline{\mathrm{S}} = \mathrm{S}$.
3. $s_1(\mathrm{S}) \geq 0$.
4. $(s_1(\mathrm{S}))^2 \leq 3\mathrm{s}_2(\mathrm{S})$.

Then there exist $A \in \mathbb{R}_+^{3\times 3}$ *such that* S *is the eigenvalue multiset of* A.

Proof. Suppose first that $\mathrm{S} \subset \mathbb{R}$. It is straightforward to show that S is a union of Frobenius multisets. In that case the theorem can be shown straightforward. See Problem 2. It is left to discuss the following renormalized case $\mathrm{S} = \{\mathrm{r}, \mathrm{e}^{\sqrt{-1}\theta}, \mathrm{e}^{-\sqrt{-1}\theta}\}$, where $r \geq 1, \theta \in (0, \pi)$. The condition $s_1(\mathrm{S}) \geq 0$ yields that

$$2\cos\theta + r \geq 0. \tag{7.13.1}$$

The condition $(s_1(\mathrm{S}))^2 \leq 3\mathrm{s}_2(\mathrm{S})$ boils down to

$$(r - 2\cos(\frac{\pi}{3} + \theta))(r - 2\cos(\frac{\pi}{3} - \theta)) \geq 0.$$

For $r \geq 1, \theta \in (0, \pi)$ we have $r - 2\cos(\frac{\pi}{3} + \theta) > 0$. Hence the condition 4 is equivalent to

$$r - 2\cos(\frac{\pi}{3} - \theta) \geq 0. \tag{7.13.2}$$

Let U be the orthogonal matrix $\frac{1}{6}\begin{bmatrix} \sqrt{2} & \sqrt{3} & -1 \\ \sqrt{2} & 0 & 2 \\ \sqrt{2} & -\sqrt{3} & -1 \end{bmatrix}$ and $J = \mathbf{1}_3\mathbf{1}_3^\top$. So $U^\top JU = \mathrm{diag}(3, 0, 0)$. S is the eigenvalue set of $B = \begin{bmatrix} r & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}$. Then $A := UBU^\top$ is the following matrix

$$\frac{r}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} - \frac{2}{3}\begin{bmatrix} -\cos\theta & \cos(\frac{\pi}{3} + \theta) & \cos(\frac{\pi}{3} - \theta) \\ \cos(\frac{\pi}{3} - \theta) & -\cos\theta & \cos(\frac{\pi}{3} + \theta) \\ \cos(\frac{\pi}{3} + \theta) & \cos(\frac{\pi}{3} - \theta) & -\cos\theta \end{bmatrix}.$$

The above inequalities show that $A \geq 0$. □

A weaker version of the solution of Problem 7.13.1 was given in [BoH91].

Theorem 7.13.8 *Let* $\mathrm{T} \subset \mathbb{C}\backslash\{0\}$ *be Frobenius mulitiset satisfying the following conditions.*

1. *$\mathrm{T}(\mathrm{r}(\mathrm{T})) = \{\mathrm{r}(\mathrm{T})\}$.*

2. *$s_k(\mathrm{T}) \geq 0$ for $k \in \mathbb{N}$.*

3. *If $s_k(\mathrm{T}) > 0$ then $s_{kl}(\mathrm{T}) > 0$ for all $l \in \mathbb{N}$.*

Then there exist a square nonnegative primitive matrix A, whose eigenvalues multiset is a union of T and $m_0 \geq 0$ copies of 0.

We prove the above theorem under the stronger assumption

$$s_k(\mathrm{T}) > 0 \text{ for k} \geq 2, \tag{7.13.3}$$

following the arguments of [Laf10].

Lemma 7.13.9 *Let $A_n \in \mathbb{C}^{n\times n}$ be the following lower Hessenberg matrix*

$$A_n = \begin{bmatrix} s_1 & 1 & 0 & & . & . & . & & & 0\\ s_2 & s_1 & 2 & 0 & & . & . & . & & 0\\ s_3 & s_2 & s_1 & 3 & 0 & & . & . & . & 0\\ . & s_3 & . & . & . & . & & & & \\ . & . & & . & . & . & & & & \\ . & & & & . & & & & & \\ & & & & & . & & & & \\ & & & & & & & & & \\ s_{n-1} & s_{n-2} & & . & . & . & & s_2 & s_1 & n-1\\ s_n & s_{n-1} & & & . & . & . & s_3 & s_2 & s_1 \end{bmatrix} \tag{7.13.4}$$

Let $\mathrm{S} = \{\lambda_1, \ldots, \lambda_\mathrm{n}\} \subset \mathbb{C}$ be the unique multiset such that $s_k = s_k(\mathrm{S})$ for $k = 1, \ldots, n$. Let $\sigma_1, \ldots, \sigma_n$ be the n-elementary symmetric polynomials corresponding to S. Then the characteristic polynomial of $A_n$is given by

$$\det\,(zI_n - A_n) = z^n + \sum_{i=1}^{n} (-1)^i i! \binom{n}{i} \sigma_i. \tag{7.13.5}$$

Proof. Recall the Newton identities.

$$s_1 = \sigma_1, \quad s_k = (-1)^{k-1} k\sigma_k + \sum_{i=1}^{k-1} (-1)^{i-1} \sigma_i s_{k-i} \text{ for } k = 2, \ldots, n.$$

Let $p(z)$ be the polynomial given by the right-hand side of (7.13.5). Denote by $\mathrm{C}(\mathrm{p}(\mathrm{z})) \in \mathbb{C}^{\mathrm{n}\times\mathrm{n}}$ the companion matrix corresponding to $p(z)$. Let $Q = [q_{ij}] \in \mathbb{C}^{n\times n}$ be the following lower triangular matrix.

$$q_{ij} = \frac{(-1)^{i-j}\sigma_{i-j}}{(j-1)!}, \quad j = 1, \ldots, i, \ i = 1, \ldots, n, \text{ where } \sigma_0 := 1.$$

Use the Newton identities to verify the equality $A_nQ = C(p(z))Q$. Hence A_n is similar to $C(p(z))$, and the characteristic polynomial of A_n is given by (7.13.5). □

Proof of Theorem 7.13.8 under the assumption 7.13.3. Let $\mathrm{T} = \{\lambda_1, \ldots, \lambda_\mathrm{n}\}$ be a multiset in $\mathbb{C}$. Denote by $\sigma_1, \ldots, \sigma_n$ the elementary symmetric polynomials corresponding to T. Let $p(z) = z^n + \sum_{i=1}^n (-1)^i \sigma_i z^{n-i}$ be the normalized polynomial whose zero set is T. For $m \in \mathbb{N}$ denote $\mathrm{S_m} := \mathrm{T} \cup \{\underbrace{0, \ldots, 0}_{\mathrm{m}}\}$. Let $\sigma_{i,m}$ be the $i-th$ elementary symmetric polynomial corresponding to $\mathrm{S_m}$ for $i = 1, \ldots, n+m$. Then $\sigma_{i,m} = \sigma_i$ for $i = 1, \ldots, n$ and $\sigma_{i,m} = 0$ for $i = n+1, \ldots, n+m$. The $s_k(\mathrm{T}) = \mathrm{s_k}(\mathrm{S_m})$ for all $k \in \mathbb{N}$. Denote by $A_{n+m} \in \mathbb{C}^{(n+m)\times(n+m)}$ the matrix (7.13.4), where $s_k = s_k(\mathrm{T})$ for $k = 1, \ldots, n+m$. Observe that

$$\det\,(zI_{n+m} - \frac{1}{n+m}A_{n+m}) = (z^n + \sum_{i=1}^n (\prod_{j=1}^i (1 - \frac{j-1}{n+m}))(-1)^i \sigma_i z^{n-i})z^m.$$

Let

$$p_m(z) := z^n + \sum_{i=1}^n (\prod_{j=1}^i (1 - \frac{j-1}{n+m})^{-1})(-1)^i \sigma_i z^{n-i}. \tag{7.13.6}$$

Denote by $T_m = \{\lambda_{1,m}, \ldots, \lambda_{n,m}\}$ the multiset formed the n zeros of p_m. Since $\lim_{m\to\infty} p_m(z) = p(z)$ we deduce that $\lim_{m\to\infty} \mathrm{pdist}(\mathrm{T_m}, \mathrm{T}) = 0$. That is, we can rename $\lambda_{1,m}, \ldots, \lambda_{n,m}, m \in \mathbb{N}$ such that $\lim_{m\to\infty} \lambda_{i,m} = \lambda_i$ for $i = 1, \ldots, n$. Let $B_{m+n} \in \mathbb{C}^{(m+n)\times(m+n)}$ be the matrix defined by (7.13.4), where $s_k, k = 1, \ldots, m+n$ are the moments corresponding to $\mathrm{T_m}$. Then $\det\,(zI_{n+m} - \frac{1}{n+m}B_{n+m}) = z^m p(z)$. Thus, if the first $n+m$ moments corresponding to $\mathrm{T_m}$ are nonnegative, it follows that that the multiset $\mathrm{S_m}$ is realized as an eigenvalue set of a nonnegative matrix.

We now show that the above condition holds for $m \ge N$, if T satisfies the assumption 1 of Theorem 7.13.8 and (7.13.3). It is enough to consider the case where $\mathrm{T} = \{\lambda_1 = 1, \lambda_2, \ldots, \lambda_\mathrm{n}\}$, where $1 > |\lambda_2| \ge \ldots \ge |\lambda_n|$. Let $\varepsilon := \frac{1-|\lambda_2|}{4}$. First we choose M big enough such that after renaming the elements of the multiset of T_m we have that $|\lambda_{i,m} - \lambda_i| \le \varepsilon$ for $i = 1, \ldots, n$ and $m \ge M$. Note that since $\bar{T}_m = T_m$ it follows that $\lambda_{1,m} \in \mathbb{R}$ and $\lambda_{1,m} \ge 1 - \varepsilon$ for $m \ge M$. Furthermore, $|\lambda_{i,m}| \le 1 - 3\varepsilon$ for $i = 2, \ldots, n$. Hence

$$\mathbf{s}_k(T_m) \ge (1-\varepsilon)^k (1 - (n-1)(\frac{1-3\varepsilon}{1-\varepsilon})^k).$$

Thus for $k > k(\varepsilon) := \frac{\log(n-1)}{\log(1-\varepsilon)-\log(1-3\varepsilon)}$ and $m \ge N$ we have that $s_k(T_m) > 0$. Clearly $s_1(T_m) = s_1(T)$ and $\lim_{m\to\infty} s_k(T_m) = s_k(T)$ for $k = 2, \ldots, \lceil k(\varepsilon) \rceil$. Since $s_k(T) > 0$ for $n > 1$ we deduce the positivity of all $s_k(T_m)$ for all $k > 1$ if $m \ge N \ge M$. □

It is straightforward to generalize this result to a general Frobenius multiset. See [Fri09].

Theorem 7.13.10 *Let* $\mathrm{T} \subset \mathbb{C}\backslash\{0\}$ *be a Frobenius mulitiset satisfying the following conditions.*

1. $\mathrm{T(r(T))} = \{\mathrm{r(T)}, \zeta\mathrm{r(T)}, \ldots, \zeta^{\mathrm{m}-1}\mathrm{r(T)}\}$ *for* $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$ *where* $m > 1$ *is an integer.*

2. $s_k(\mathrm{T}) \ge 0$ *for* $k \in \mathbb{N}$.

3. *If* $s_k(\mathrm{T}) > 0$ *then* $s_{kl}(\mathrm{T}) > 0$ *for all* $l \in \mathbb{N}$.

Then there exist a square nonnegative irreducible matrix A*, whose eigenvalues multiset is a union of* T *and* $m_0 \ge 0$ *copies of* 0.

Proof. Observe first that $s_k(\mathrm{T}) = 0$ if $m \not| k$. Let $\phi : \mathbb{C} \to \mathbb{C}$ be the map $z \mapsto z^m$. Since $\zeta\mathrm{T} = \mathrm{T}$ it follows that for $z \in \mathrm{T}$ with multiplicity $m(z)$ we obtain the multiplicity z^m in $\phi(\mathrm{T})$ is $mm(z)$. Hence $\phi(\mathrm{T})$ is a union of m copies of a Frobenius set T_1, where $r(\mathrm{T}_1) = \mathrm{r(T)^m}$ and $\mathrm{T}_1(\mathrm{r}(\mathrm{T}_1)) = \{\mathrm{r}(\mathrm{T}_1)\}$. Moreover $s_{km}(\mathrm{T}) = \mathrm{ms_k}(\mathrm{T}_1)$. Hence T_1 satisfies the assumptions of Theorem 7.13.8. Thus there exists a primitive matrix $B \in \mathbb{R}_+^{n\times n}$ whose nonzero eigenvalue multiset is T_1. Let $A = [A_{ij}]_{i=j=1}^m$ be the following nonnegative matrix of order mn.

$$A = \begin{bmatrix} 0_{n\times n} & I_n & 0_{n\times n} & 0_{n\times n} & \ldots & 0_{n\times n} \\ 0_{n\times n} & 0_{n\times n} & I_n & 0_{n\times n} & \ldots & 0_{n\times n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n\times n} & 0_{n\times n} & 0_{n\times n} & 0_{n\times n} & \ldots & I_n \\ B & 0_{n\times n} & 0_{n\times n} & 0_{n\times n} & \ldots & 0_{n\times n} \end{bmatrix}. \tag{7.13.7}$$

Then A is irreducible and the nonzero part of eigenvalue multiset if T. (See Problems 4 and 5.)

Problems

1. **Definition 7.13.11** *Let* $\mathrm{S} \subset \mathbb{C}$ *be a finite multiset.* S *is called a semi Frobenius multiset if either* S *has* m *elements all equal to* 0*, or the following conditions hold.*

   (a) $r(\mathrm{S}) > 0,\ \bar{\mathrm{S}} = \mathrm{S},\ \mathrm{r}(\mathrm{S}) \in \mathrm{S}$.

   (b) $m(z) \le \mu := m(r(\mathrm{S}))$ *for each* $z \in \mathrm{S}$ *such that* $|z| = r(\mathrm{S})$.

   (c) Assume that S *contains exactly* m *distinct points satisfying* $|z| = r(\mathrm{S}), \mathrm{m}(\mathrm{z}) = \mu$*. Then* $\zeta\mathrm{S} = \mathrm{S}$ *for* $\zeta = e^{\frac{2\pi\sqrt{-1}}{m}}$.

   S *is called an almost a Frobenius multiset if the the number points in* S*, counted with their mulitplicites, satisfying* $z \in \mathrm{S}, |\mathrm{z}| = \mathrm{r}(\mathrm{S}), \mathrm{m}(\mathrm{z}) < \mu$ *is strictly less than* $m\mu$*.*

   Let f_{S} by (7.12.6). Show

   (a) $\mathrm{S} = \{1, 1, \mathrm{z}, \bar{\mathrm{z}}\}$, with $|z| = 1, z \ne \pm 1$ is a semi Frobenius multiset, and f_{S} has nonnegative moments.

   (b) Let $\mathrm{S} = \cup_{\mathrm{i}=1}^{\mathrm{j}} \mathrm{S}_{\mathrm{i}}$, where each S_{i} is almost a Frobenius multiset. Then f_{S} has eventually nonnegative MacLaurin coeffients.

   (c) Assume that the MacLaurin coefficients f_{S} are eventually nonnegative. Then $r(\mathrm{S}) \in \mathrm{S}$. Suppose furthermore that $0 < \alpha < r(\mathrm{S})$ is the second largest positive number contained in S. Then $\mathrm{S} \cap \{\mathrm{z} \in \mathbb{C},\ \alpha < |\mathrm{z}| \le \mathrm{r}(\mathrm{S})\}$ is a semi Frobenius set.

   (d) Assume that the MacLaurin coefficients f_S are eventually nonnegative. Suppose that S contains only one positive number of mulitplicity one. Then S is semi Frobenius.

   (e) Assume that the MacLaurin coefficients f_S are eventually nonnegative. Suppose that S contains only two positive number of mulitplicity one each: $r(\mathrm{S}) > \alpha > 0$. Decompose S to $\mathrm{S}_1 \cup \mathrm{S}_2$, where S_1 is a maximal semi Frobenius set containing $\{z \in \mathbb{C},\ \alpha < |z| \le r(\mathrm{S})\}$. If $\alpha \in \mathrm{S}_1$ then $\mathrm{S}_2 = \emptyset$. Suppose that $\alpha \in \mathrm{S}_2$. Then $\mathrm{S}_3 := \mathrm{S}_2 \cap \{\mathrm{z} \in \mathbb{C},\ |\mathrm{z}| = \alpha\}$ is a set, i.e. $m(z) = 1$ for each $z \in \mathrm{S}_3$. Assume for simplicity of the exposition that $\alpha = 1$. Let $m' \in [1, l)$ be the greatest divisor of $m > 1$, entering in the definition of the Frobenius multiset S_1, such that all m' roots of 1 are in S_3. Let $m'' := \frac{m}{m'} > 1$. Then there exists $r \in \mathbb{N}$ coprime with m'' such that one of the following conditions hold.

      i. If m'' is even then $\mathrm{S}_3 = \mathrm{S}_4$, where S_4 consists of all $m'r$ roots of 1.

      ii. If m'' is odd then either $\mathrm{S}_3 = \mathrm{S}_4$ or $\mathrm{S}_3 = \mathrm{S}_4 \cup \mathrm{S}_5$, where

$$\mathrm{S}_5 = \cup_{\mathrm{q,k,j}=1}^{\mathrm{m}',\mathrm{r},\mathrm{m}''-1} \mathrm{e}^{2\pi\sqrt{-1}((\mathrm{q}+\frac{2\mathrm{k}-1}{2\mathrm{r}}+\frac{\mathrm{j}}{\mathrm{m}''})\frac{1}{\mathrm{m}'})}.$$

Hint: Use the function g in the proof of Theorem 7.12.6, or/and consult with [Fri78b, Thm 4] and its proof.

2. Let $\mathrm{S} = \{\lambda_1, \ldots, \lambda_\mathrm{n}\} \subset \mathbb{R}$ be a union of Frobenius multiset. Assume furthermore that $\sum_{i=1}^{n} \lambda_i \ge 0$. Show that if $n \le 4$ then there exists a nonnegative $n \times n$ matrix whose eigenvalue multiset if S. **Hint**: For $n = 3$ use Proposition 7.13.5. For $n = 4$ and the case where S contains exactly two negative numbers consult with the proof of [LoL78, Thm. 3].

3. Show that for $n \ge 4$ the multiset $\mathrm{S} := \{\sqrt{2}, \sqrt{2}, \sqrt{-1}, -\sqrt{-1}, 0, \ldots, 0\}$ satisfies all the conditions of Proposition 7.13.2. However there is no $A \in \mathbb{R}_+^{n\times n}$ with the eigenvalue set S.

4. Let $B \in \mathbb{R}_+^{n\times n}$ be a primitive matrix. Show that the matrix $A \in \mathbb{R}_+^{mn\times mn}$ defined (7.13.7) is irreducible for any integer $m > 1$.

5. Let $B \in \mathbb{C}^{n\times n}$. Assume that T is the eigenvalue multiset of B Assume that $A \in \mathbb{C}^{mn\times mn}$ is defined by (7.13.7). Let S be the eigenvalue multiset of A. Show that $w \in \mathrm{S}$ if and only if $w^m \in \mathrm{T}$. Furthermore the multiplicity of $0 \ne w \in \mathrm{S}$ equals to the multiplicity of w^m in T. The mulitplicity of $0 \in \mathrm{S}$ is m times the multiplicity of $0 \in \mathrm{T}$.

7.14 Cones

Let $\mathbf{V}$ be a vector space over $\mathbb{C}$. Then $\mathbf{V}$ is a vector space over $\mathbb{R}$, which we denote by $\mathbb{V}_\mathbb{R}$, or simply $\mathbf{V}$ when no ambiguity arises. See Problem 1.

Definition 7.14.1 *Let $\mathbf{V}$ be a finite dimensional vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. A set $\mathrm{K} \subset \mathbf{V}$ is called a cone if*

1. $\mathrm{K} + \mathrm{K} \subset \mathrm{K}$, *i.e.* $\mathbf{x} + \mathbf{y} \in \mathrm{K}$ *for each* $\mathbf{x}, \mathbf{y} \in \mathrm{K}$.

2. $\mathbb{R}_+\mathrm{K} \subset \mathrm{K}$, *i.e.* $a\mathbf{x} \in \mathrm{K}$ *for each* $a \in [0, \infty)$ *and* $\mathbf{x} \in \mathrm{K}$

Assume that $\mathrm{K} \subset \mathbf{V}$ is a cone. (Note that K is convex set.) Then

1. $\mathrm{ri}\, K, \dim K$, *is the relative interior and the dimension of K, viewed as a convex set.*

2. $\mathrm{K}^* := \{\mathbf{f} \in \mathbf{V}^*,\ \Re\mathbf{f}(\mathbf{x}) \ge o$ *for all* $\mathbf{x} \in \mathbf{V}\}$ *is called the conjugate cone, (in* $\mathbf{V}^*$*).*

3. K *is called pointed if* $\mathrm{K} \cap \mathrm{K} = \{\mathbf{0}\}$.

4. K *is called generating if* $\mathrm{K}-\mathrm{K}=\mathbf{V}$*, i.e. any* $\mathbf{z}\in\mathbf{V}$ *can be represented as* $\mathbf{x}-\mathbf{y}$ *for some* $\mathbf{x},\mathbf{y}\in\mathrm{K}$.

5. K *is called proper if* K *is closed, pointed and generating.*

6. *For* $\mathbf{x},\mathbf{y}\in\mathbf{V}$ *we denote:* $\mathbf{x}\ge^{\mathrm{K}}\mathbf{y}$ *if* $\mathbf{x}-\mathbf{y}\in\mathrm{K}$*;* $\mathbf{x}\gneq^{\mathrm{K}}\mathbf{y}$ *if* $\mathbf{x}\ge^{\mathrm{K}}\mathbf{y}$ *and* $\mathbf{x}\ne\mathbf{y}$*;* $\mathbf{x}>^{\mathrm{K}}\mathbf{y}$ *if* $\mathbf{x}-\mathbf{y}\in\mathrm{ri}\,\mathrm{K}$.

7. *For* $\mathbf{x}\in\mathbf{V}$ *we call:* $\mathbf{x}$ *nonnegative relative to* K *if* $\mathbf{x}\ge^{\mathrm{K}}\mathbf{0}$*;* $\mathbf{x}$ *is semipositive relative to* K *if* $\mathbf{x}\gneq^{\mathrm{K}}\mathbf{0}$*;* $\mathbf{x}$ *is positive relative to* K *if* $\mathbf{x}\in\mathrm{ri}\,\mathrm{K}$. *When there is no ambiguity about the cone* K *we drop the term relative to* K.

8. $\mathrm{K}_1\subset\mathrm{K}$ *is called a subcone of* K *if* K_1 *is a cone in* $\mathbf{V}$. $\mathrm{F}\subset\mathrm{K}$ *is called a face of* K*, if* F *is a subcone of* K*, and* $\mathbf{y}\in\mathrm{F}$ *if* $\mathbf{y}\in\mathrm{K}$ *and there exists* $\mathbf{x}\in\mathrm{F}$ *such that* $\mathbf{x}\ge^{K}\mathbf{y}$. $\dim\mathrm{F}$*, the dimension of* F*, is called the dimension of* F. $\mathrm{F}=\{0\},\mathrm{F}=\mathrm{K}$ *are called trivial faces, (*$\dim\{0\}=0$*).* $\mathbf{x}\gneq\mathbf{0}$ *is called an extreme ray if* $\mathbb{R}_+\mathbf{x}$ *is a face in* K*, (of dimension* 1*). For a set* $\mathrm{X}\subset\mathrm{K}$*, the face* $\mathrm{F}(\mathrm{X})$ *generated by* X*, is the intersections of all faces of* K *containing* X.

9. *Let* $T\in\mathrm{Hom}\,(\mathbf{V},\mathbf{V})$*. Then:* $T\ge^{\mathrm{K}}0$*, and* T *is called nonnegative with respect to* K*, if* $T\mathrm{K}\subset\mathrm{K}$*;* $T\gneq^{\mathrm{K}}0$*, and* T *is called semipositive with respect to* K*, if* $T\ge^{\mathrm{K}}0$ *and* $T\ne0$*;* $T>^{\mathrm{K}}0$*, and* T *is called positive with respect to* K*, if* $T(\mathrm{K}\backslash\{\mathbf{0}\})\subset\mathrm{ri}\,\mathrm{K}$. $T\ge^{\mathrm{K}}0$ *is called primitive with respect to* K*, if* F *is a face of* K *satisfying* $T\mathrm{F}\subset\mathrm{F}$*, i.e.* F *is* T *invariant, then* F *is a trivial face of* K. T *is called eventually positive with respect to* K *if* $T^l>^{\mathrm{K}}0$ *for all integers* $l\ge L(\ge1)$*. When there is no ambiguity about the cone* K *we drop the term relative to* K*. Denote by* $\mathrm{Hom}\,(\mathbf{V},\mathbf{V})^{\mathrm{K}}$ *the set of all* $T\ge^{\mathrm{K}}0$*. For* $T,S\in\mathrm{Hom}\,(\mathbf{V},\mathbf{V})$ *we denote:* $T\ge^{\mathrm{K}}S\iff T-S\ge^{\mathrm{K}}0$*,* $T\gneq^{\mathrm{K}}S\iff T-S\gneq^{\mathrm{K}}0$*,* $T>^{\mathrm{K}}S\iff T-S>^{\mathrm{K}}0$.

As pointed out in Problem 3, without loss of generality we can discuss only the cones in real vector spaces. Also, in most of the applications the cones of interest lie in the real vector spaces. Since most of the results we state hold for cones over complex vector spaces, we state our results for cones in real or complex vector spaces, and give a proof only for the real case, when possible.

Lemma 7.14.2 *Let* $\mathbf{V}$ *be a finite dimensional vector space over* $\mathbb{F}=\mathbb{R},\mathbb{C}$*. Let* K *be a cone in* $\mathbf{V}$*. Then* $\mathbf{V}=\mathrm{K}-\mathrm{K}$*, i.e.* K *is generating, if and only if the interior of* K *is nonempty, i.e.* $\dim\mathrm{K}=\dim_{\mathbb{R}}\mathbf{V}$.

Proof. It is enough to assume that $\mathbf{V}$ is an n dimensional vector space over $\mathbb{R}$. Let $k = \dim \mathrm{K}$. Then span K is a k dimensional vector space in $\mathbf{V}$. Assume that $\mathrm{K} - \mathrm{K} = \mathbf{V}$. Since $\mathrm{K} - \mathrm{K} \subset$ span K we deduce that $\dim \mathrm{K} = \mathrm{n}$. Hence K must have an interior.

Assume now that K has an interior. Hence it interior must contain n linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which form a basis in $\mathbf{V}$. So $\sum_{i=1}^n a_i \mathbf{x}_i \in \mathrm{K}$ for any $a_1, \ldots, a_n \geq 0$. Since any $\mathbf{z} \in \mathbf{V}$ is of the form $\sum_{i=1}^n z_i \mathbf{x}_i$ is of the form $\sum_{z_i \geq 0} z_i x_i - \sum_{z_i < 0} (-z_i) \mathbf{x}_i$ we deduce that $\mathrm{K} - \mathrm{K} = \mathbf{V}$. □

Theorem 7.14.3 *Let* $\mathrm{K} \subset \mathbf{V}$ *be a proper cone over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$, *where* $\dim \mathbf{V} \in [1, \infty)$. *Then the following conditions holds.*

1. *There exists* $\mathbf{f} \in \mathrm{K}^*$ *which is strictly positive, i.e.* $\Re \mathbf{f}(\mathbf{x}) > 0$ *if* $\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0}$.

2. *Every* $\mathbf{x} \gneq \mathbf{0}$ *is a nonnegative linear combination of at most* $\dim_{\mathbb{R}} \mathbf{U}$ *extreme rays of* K.

3. *The conjugate cone* $\mathrm{K}^* \subset \mathbf{V}^*$ *is proper.*

Proof. It is enough to assume that $\mathbf{V}$ is a vector space over $\mathbb{R}$. Observe that for any $\mathbf{u} \gneq^{\mathrm{K}} \mathbf{0}$ the set $\mathrm{I}(\mathbf{u}) := \{\mathbf{x} \in \mathbf{V},\ \mathbf{u} \geq^{\mathrm{K}} \mathbf{x} \geq^{\mathrm{K}} -\mathbf{u}\}$ is a compact set. Clearly, $\mathrm{I}(\mathbf{u})$ is closed. It is left to show that that $\mathrm{I}(\mathbf{u})$ is bounded. Fix a norm $\|\cdot\|$ on $\mathbf{V}$. Assume to the contrary that there exists a sequence $\mathbf{0} \neq \mathbf{x}_m \in \mathrm{C}$ such that $\lim_{m\to\infty} \|\mathbf{x}_m\| = \infty$. Let $\mathbf{y}_m = \frac{1}{\|\mathbf{x}_m\|}\mathbf{x}_m, m \in \mathbb{N}$. Since $\|\mathbf{y}_m\| = 1, m \in \mathbb{N}$ it follows that there exists a subsequence $m_k, k \in \mathbb{N}$ such that $\lim_{k\to\infty} \mathbf{y}_{m_k} = \mathbf{y}, \|\mathbf{y}\| = 1$. Since $\mathbf{y}_m \in \mathrm{I}(\frac{1}{\|\mathbf{x}_m\|}\mathbf{u})$ it follows that $\mathbf{y} \in \mathrm{I}(\mathbf{0})$. So $\mathbf{y} \in \mathrm{K} \cap -\mathrm{K} = \{\mathbf{0}\}$ which is impossible. Hence $\mathrm{I}(\mathbf{u})$ is compact.

Choose $\mathbf{u} \in$ ri K. We claim that $\mathbf{u}$ is an isolated extreme point of $\mathrm{I}(\mathbf{u})$. Since $\mathbf{u} \in$ ri K it follows that there exist $r > 0$ so that $\mathbf{u} + \mathbf{x} \in \mathrm{K}$ for each $\|\mathbf{x}\| \leq r$. Suppose that there exist $\mathbf{v}, \mathbf{w} \in \mathrm{I}(\mathbf{u})$ such that $t\mathbf{v} + (1-t)\mathbf{w} = \mathbf{u}$ for some $t \in (0,1)$. So $\mathbf{v} = \mathbf{u} - \mathbf{v}_1, \mathbf{w} = \mathbf{u} - \mathbf{v}_2$ for some $\mathbf{v}_1, \mathbf{w}_1 \geq^{\mathrm{K}} \mathbf{0}$. The equality $\mathbf{u} = (1-t)\mathbf{v} + t\mathbf{w}$ yields $\mathbf{0} = (1-t)\mathbf{v}_1 + t\mathbf{w}_1 \geq (1-t)\mathbf{v}_1 \geq \mathbf{0}$. Hence $\mathbf{v}_1 = 0$. (See Problem 2). Similarly, $\mathbf{w}_1 = \mathbf{0}$. Hence $\mathbf{u}$ is an extreme point.

We now show that for any $\mathbf{x} \in \mathbf{U}$ such that $\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0}, \|\mathbf{x}\| \leq r$ the point $\mathbf{u} - \mathbf{x}$ is not an extreme point of $\mathrm{I}(\mathbf{u})$. Indeed, $\mathbf{u} - \frac{3}{2}\mathbf{x}, \mathbf{u} - \frac{1}{2}\mathbf{x} \in \mathrm{I}(\mathbf{u})$ and $\mathbf{u} - \mathbf{x} = \frac{1}{2}(\mathbf{u} - \frac{3}{2}\mathbf{x}) + \frac{1}{2}(\mathbf{u} - \frac{1}{2})\mathbf{x}$. Since $\mathbf{u}$ is an isolated extreme point, Corollary 7.1.10 yields that $\mathbf{u}$ is exposed. Hence there exists $\mathbf{f} \in \mathbf{U}^*$ such that $\mathbf{f}(\mathbf{u}) > \mathbf{f}(\mathbf{u} - \mathbf{x})$ for any $\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0}$ satisfying $\|\mathbf{x}\| < r$. So $\mathbf{f}(\mathbf{x}) > 0$ for any $\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0}$ satisfying $\|\mathbf{x}\| < r$. Hence $\mathbf{f}(\mathbf{y}) = \frac{\|\mathbf{y}\|}{r} f(\frac{r}{\|\mathbf{y}\|}\mathbf{y}) > 0$ for any $\mathbf{y} \gneq^{\mathrm{K}} \mathbf{0}$. This proves the part 1 of the theorem.

Let $\mathrm{C} = \{\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0},\ \mathbf{f}(\mathbf{x}) = 1\}$. Since K is closed, it follows that C is a convex closed set. We claim that C is compact, i.e. bounded. Fix a norm $\|\cdot\|$ on $\mathbf{U}$. Assume to the contrary that there exists a sequence $\mathbf{x}_m \in \mathrm{C}$ such that $\lim_{m\to\infty} \|\mathbf{x}_m\| = \infty$. Let $\mathbf{y}_m = \frac{1}{\|\mathbf{x}_m\|}\mathbf{x}_m, m \in \mathbb{N}$. Since $\|\mathbf{y}_m\| = 1, m \in \mathbb{N}$ it follows that there exists a subsequence $m_k, k \in \mathbb{N}$ such that $\lim_{k\to\infty} \mathbf{y}_{m_k} = \mathbf{y}, \|\mathbf{y}\| = 1$. Since K is closed it follows that $\mathbf{y} \gneq^{\mathrm{K}} \mathbf{0}$. Note that

$$\mathbf{f}(\mathbf{y}) = \lim_{k\to\infty} f(\mathbf{y}_{m_k}) = \lim_{k\to\infty} \frac{f(\mathbf{x}_{m_k})}{\|\mathbf{x}_k\|} = \lim_{k\to\infty} \frac{1}{\|\mathbf{x}_{m_k}\|} = 0.$$

This contradicts the assumption that $\mathbf{f}$ is strictly positive on K. Thus C is a convex compact set. We next observe that dim C = n − 1. First observe that $\mathbf{f}(\mathrm{C} - \mathbf{x}) = 0$ for any $\mathbf{x} \in \mathbb{C}$. Hence dim C $\leq$ n − 1. Observe next that if $\mathbf{f}(\mathbf{z}) = 0$ and $\|\mathbf{z}\| \leq r$ then $\mathbf{u} + \mathbf{z} \in \mathrm{C}$. Hence dim C = n − 1. Let $\mathbf{w} \gneq^{\mathrm{K}} \mathbf{0}$. Define $\mathbf{w}_1 = \frac{1}{\mathbf{f}(\mathbf{w})}\mathbf{w} \in \mathrm{C}$. Caratheodory theorem claims that $\mathbf{w}_1$ is a convex combination of at most n extreme points of C. This proves the part 2 of the theorem.

Let $R := \{\max \|\mathbf{x}\|,\ \mathbf{x} \in \mathrm{C}\}$. Let $\mathbf{g} \in \mathbf{U}^*, \|\mathbf{g}\|^* \leq \frac{1}{R}$. Then for $\mathbf{x} \in \mathrm{C}$ $|\mathbf{g}(\mathbf{x})| \leq 1$. Hence $(\mathbf{f} + \mathbf{g})(\mathbf{x}) \geq 1 - |\mathbf{g}(\mathbf{x})| \geq 0$. Thus $\mathbf{f} + \mathbf{g} \in \mathrm{K}^*$. So $\mathbf{f}$ is an interior point of K^*. Clearly K^* is a closed and a pointed cone. Hence part 3 of the theorem hold. □

Theorem 7.14.4 *Let* $\mathbf{V}$ *be a vector space over* $\mathbb{F} = \mathbb{R}, \mathbb{C}$. *Assume that* $\mathrm{K} \subset \mathbf{V}$ *be a proper cone. Assume that* $T \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})^{\mathrm{K}}$. *Let* $\mathrm{S(T)} \subset \mathbb{C}$ *be the eigenvalue multiset of* T, *(i.e. the root set of the polynomial* $\det\,(zI - T)$.*) Then*

1. $\rho(T) \in \mathrm{S(T)}$.
2. *Let* $\lambda \in \mathrm{S(T)}(\rho(\mathrm{T}))$. *Then* $\mathrm{index}\,(\lambda, T)) \leq k := \mathrm{index}\,(\rho(T), T)$.
3. *There exists* $\mathbf{x} \gneq^{\mathrm{K}} \mathbf{0}$ *such that* $T\mathbf{x} = \rho(T)\mathbf{x}$, *and* $\mathbf{x} \in (\rho(T)I - T)^{k-1}\mathbf{V}$.
4. *If* $T >^{\mathrm{K}} 0$ *then* $\rho(T) > 0, k = 1, \mathrm{S(T)}(\rho(\mathrm{T})) = \{\rho(\mathrm{T})\}$ *and* $\rho(A)$ *is a simple root of the characteristic polynomial of* T. *(This statement can hold only if* $\mathbb{F} = \mathbb{R}$.*)*

Assume in addition that $\rho(T) = 1$. *Let* $P \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})$ *be the spectral projection, associated with* T, *on the generalized eigenspace corresponding to* 1. *Then*

$$\lim_{m\to\infty} \frac{k!}{m^k} \sum_{i=0}^{m-1} T^i = (T - I)^{k-1} P \gneq^{\mathrm{K}} 0. \tag{7.14.1}$$

Assume finally that $\mathbb{F} = \mathbb{R}$, $\lambda \neq 1, |\lambda| = 1$ *is an eigenvalue of* T *of index* k. *Let* $P(\lambda) \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})$ *be the spectral projection, associated with* T, *on the generalized eigenspace corresponding to* λ. *View* $P(\lambda) = P_1 + \sqrt{-1}P_2$, *where* $P_1, P_2 \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})$. *Then*

(7.14.2) $|\mathbf{f}((T-\lambda I)^{k-1}P(\lambda)\mathbf{y})| \le \mathbf{f}((T-I)^{k-1}P\mathbf{y})$ *for any* $\mathbf{y} \in \mathrm{K}, \mathbf{f} \in \mathrm{K}^*$.

Proof. Suppose first that $\rho(T) = 0$, i.e. T is nilpotent. Then parts 1-2 are trivial. Choose $\mathbf{y} \gneq^{\mathrm{K}} \mathbf{0}$. Then there exists an integer $j \in [0, k-1]$ so that $T^j\mathbf{y} \gneq^{\mathrm{K}} \mathbf{0}$ and $T^{j+1}\mathbf{y} = \mathbf{0}$. Then $\mathbf{x} := T^j\mathbf{y}$ is an eigenvector of A which lies in K. Since $A\mathbf{x} = \mathbf{0}$ it follows that A can not be positive.

From now on we assume that $\rho(T) > 0$, and without loss of generality we assume that $\rho(T) = 1$. In particular, $\dim \mathbf{V} \ge 1$. Choose a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ in $\mathbf{V}$. Assume first that $\mathbb{F} = \mathbb{R}$. Then T represented in the basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ by $A = [a_{ij}] \in \mathbb{R}^{n\times n}$. Consider the matrix $B(z) = (I - zA)^{-1} = [b_{ij}]_{i=j=1}^n \mathbb{C}(z)^{n\times n}$. Using the Jordan canonical form of A we deduce that all the singular points of all $b_{ij}(z)$ are of the form $\mu := \frac{1}{\lambda}$ where λ is a nonzero eigenvalue of A. Furthermore, if $0 \neq \lambda \in \mathrm{spec}\ (\mathrm{A})$, and λ has index $l = l(\lambda)$. Then for each i, j, $b_{ij}(z)$ may have a pole at μ of order l at most, and there is at least one entry $b_{ij}(z)$, where $i = i(\lambda), j = j(\lambda)$, such that $b_{ij}(z)$ has a pole of order l exactly. In particular, for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ the rational function $\mathbf{y}^\top B(z)\mathbf{x}$ may have a pole of order l at most μ. Furthermore, there exists $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} = \mathbf{x}(\lambda), \mathbf{y} = \mathbf{y}(\lambda)$ such that $\mathbf{y}^\top B(z)\mathbf{x}$ has a pole at μ of order l. (See Problem 7.)

Let $\hat{\mathrm{K}} \in \mathbb{R}^n$ denote the induced cone by $\mathrm{K} \subset \mathbf{V}$. Then $\hat{\mathrm{K}}$ is a proper cone. Denote by

$$\hat{\mathrm{K}}^* := \{\mathbf{y} \in \mathbb{R}^n, \mathbf{y}^\top\mathbf{x} \ge 0 \text{ for all } \mathbf{x} \in \hat{\mathrm{K}}^*\}.$$

Theorem 7.14.3 implies that $\hat{\mathrm{K}}^*$ is a proper cone. Observe next that $A\hat{\mathrm{K}} \subset \hat{\mathrm{K}}$. Clearly, we have the following MacLaurin expansion

$$B(z) = (I - zA)^{-1} = \sum_{i=0}^{\infty} z^i A^i, \text{ for } |z| < \frac{1}{\rho(A)}, \tag{7.14.3}$$

$$\mathbf{y}^\top B(z)\mathbf{x} = \sum_{i=0}^{\infty} (\mathbf{y}^\top A^i\mathbf{x})z^i, \text{ for } |z| < \frac{1}{\rho(A)}. \tag{7.14.4}$$

Note that $\mathbf{y}^\top B(z)\mathbf{x}$ is a rational function. Denote by $r(\mathbf{x}, \mathbf{y}) \in (0, \infty]$ the convergence radius of $\mathbf{y}^\top B(z)\mathbf{x}$. So $r(\mathbf{x}, \mathbf{y}) = \infty$ if and only if $\mathbf{y}^\top B(z)\mathbf{x}$ is polynomial. Assume first that $\mathbf{x} \in \hat{\mathrm{K}}, \mathbf{y} \in \hat{\mathrm{K}}^*$. Then the MacLaurin coefficients of $\mathbf{y}^\top B(z)\mathbf{x}$ are nonnegative. Hence we can apply the Vivanti-Pringsheim theorem 7.12.2, i.e. $r(\mathbf{x}, \mathbf{y})$ is a singular point of $\mathbf{y}^\top B(z)\mathbf{x}$.

Hence $\frac{1}{r(\mathbf{x},\mathbf{y})} \in$ spec (A) if $r(\mathbf{x}, y) < \infty$. Suppose now that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Since $\hat{\mathrm{K}}$ and $\hat{\mathrm{K}}^*$ are generating it follows $\mathbf{x} = \mathbf{x}_+ - \mathbf{x}_-$, $\mathbf{y} = \mathbf{y}_+ - \mathbf{y}_-$ for some $\mathbf{x}_+, \mathbf{x}_- \in \hat{\mathrm{K}}$, $\mathbf{y}_+, \mathbf{y}_- \in \hat{\mathrm{K}}^*$. So

$$\mathbf{y}^\top B(z)\mathbf{x} = \mathbf{y}_+^\top B(z)\mathbf{x}_+ + \mathbf{y}_-^\top B(z)\mathbf{x}_- - \mathbf{y}_-^\top B(z)\mathbf{x}_+ - \mathbf{y}_+^\top B(z)\mathbf{x}_-.$$

Hence
(7.14.5)
$r(\mathbf{x}, \mathbf{y}) \geq r(\mathbf{x}_+, \mathbf{x}_-, \mathbf{y}_+, \mathbf{y}_-) := \min(r(\mathbf{x}_+, \mathbf{y}_+), r(\mathbf{x}_-, \mathbf{y}_-), r(\mathbf{x}_+, \mathbf{y}_-), r(\mathbf{x}_-, \mathbf{y}_+)).$

Let $\lambda \in$ spec (A), $|\lambda| = \rho$(A), and assume that $l =$ index (λ). Choose $\mathbf{x}, \mathbf{y}$ such that $\mu = \frac{1}{\lambda}$ is a pole of $\mathbf{y}^\top B(z)\mathbf{x}$ of order l. Hence we must have equality in (7.14.5). More presicely, there exists $\mathbf{x}_1 \in \{\mathbf{x}_+, \mathbf{x}_-\}, \mathbf{y}_1 \in \{\mathbf{y}_+, \mathbf{y}_-\}$ such that $r(\mathbf{x}, \mathbf{y}) = r(\mathbf{x}_1, \mathbf{y}_1)$ and $\mathbf{y}_1^\top B(z)\mathbf{x}_1$ has a pole of at μ of order l. Vivanti-Pringsheim theorem yields that $r(\mathbf{x}_1, \mathbf{y}_1)$ is pole of order $k' \geq l$ of $\mathbf{y}_1^\top B(z)\mathbf{x}_1$. Hence $\rho(A) \in$ spec (A) and index $(\rho(A)) \geq k' \geq l =$ index (λ). This proves parts 1-2. Choose $\lambda = \rho(A)$ that satisfies the above assumptions. Hence $B(z)\mathbf{x}_1$ must have at least one coordinate with a pole at $\rho(A)^{-1}$ of order $k =$ index $(\rho(A))$. Problem 7 yields that $\lim_{t \nearrow 1} B(t\rho(A)^{-1}\mathbf{x}_1 = \mathbf{u} \neq \mathbf{0}$ such that $A\mathbf{u} = \rho(A)\mathbf{u}$, and $\mathbf{u} \in (\rho(A)I - A)^{k-1}\mathbb{R}^n$. Use the fact that for $z = t\rho(A)^{-1}, t \in (0, 1)$ we have the equality (7.14.3). So $(1-t)^k B(t\rho(A)^{-1})\mathbf{x}_1 \in \hat{\mathrm{K}}$ for each $t \in (0, 1)$. Since $\hat{\mathrm{K}}$ is closed we deduce that $\mathbf{u} \in \hat{\mathrm{K}}$. This proves part 3.

The equality (7.14.1) follows from the Tauberian theorem 8 and it application to the series (7.14.3).

Assume now that $A >^{\hat{\mathrm{K}}} 0$. Observe first that the eigenvector $\mathbf{x} \gneq^{\hat{\mathrm{K}}} \mathbf{0}, A\mathbf{x} = \mathbf{x}$ satisfies $\mathbf{x} >^{\hat{\mathrm{K}}} \mathbf{0}$, i.e. $\mathbf{x} \in$ ri $\hat{\mathrm{K}}$. Next we claim that the dimension of the eigenspace $\{\mathbf{y},\ (A - I)\mathbf{y} = \mathbf{0}\}$ is 1. Assume to the contrary that $A\mathbf{y} = \mathbf{y}$ and $\mathbf{x}, \mathbf{y}$ are linearly independent. Since $6.7.limparexpbzx(s_0) = A\mathbf{x}(s_0) >^{\hat{\mathrm{K}}} \mathbf{0}$ we obtain a contradiction. Hence $\mathbf{x}$ is a geometrically simple eigenvalue.

Nest we claim that $k =$ index $(1) = 1$. Assume to the contrary that $k > 1$. Recall that $\mathbf{x} \in (A - I)^{k-1}\mathbb{R}^n$. So $\mathbf{x} = (A - I)\mathbf{y}$. Hence $\mathbf{x} = (A - I)(\mathbf{y} + t\mathbf{x})$. Choose $t > 0$ big enough so that $\mathbf{z} = \mathbf{y} + t\mathbf{x} = t(\frac{1}{t}\mathbf{y} + \mathbf{x}) >^{\hat{\mathrm{K}}} \mathbf{0}$. Since $\mathbf{x} >^{\hat{\mathrm{K}}} \mathbf{0}$ it follows that there exists $s > 0$ such that $(A-I)\mathbf{z} - r\mathbf{z} = \mathbf{x} - r\mathbf{z} >^{\hat{\mathrm{K}}} \mathbf{0}$. That is $A\mathbf{z} >^{\hat{\mathrm{K}}} (1 + r)\mathbf{z}$. Hence $A^m\mathbf{z} \geq (1 + r)^m\mathbf{z} \Rightarrow (\frac{1}{1+r}A)^m\mathbf{z} \geq^{\hat{\mathrm{K}}} \mathbf{z}$. Since $\rho(\frac{1}{1+r}A) = \frac{1}{1+r} < 1$ it follows that $\mathbf{0} = \lim_{m\to\infty}(\frac{1}{1+r}A)^m\mathbf{z} \geq^{\hat{\mathrm{K}}} \mathbf{z}^{\hat{\mathrm{K}}} >^{\hat{\mathrm{K}}} \mathbf{0}$, which is impossible. Hence index $(1) = 1$.

We now show that if $\lambda \in$ spec (A) and $|\lambda| = 1$ then $\lambda = 1$. Let J $:= \{\mathbf{y} \in \hat{\mathrm{K}}, \|\mathbf{y}\|_2 = 1\}$. Since $\hat{\mathrm{K}}$ is closed it follows that J is compact set. Since $A >^{\hat{\mathrm{K}}} 0$ it follows that $AJ \in$ ri $\hat{\mathrm{K}}$. Hence there exists $s \in (0, 1)$ such that

$A\mathbf{y}-s\mathbf{z} >^{\hat{\mathrm{K}}} \mathbf{0}$ for any $\mathbf{y}\in \mathrm{J}, \mathbf{z}\in\mathbb{R}^n, \|\mathbf{y}\|_2 = 1$. In particular, $A\mathbf{y}-s\mathbf{y} >^{\hat{\mathrm{K}}} \mathbf{0}$ for each $\mathbf{y}\in\mathrm{J}$. That is $(A-sI)\mathrm{J}\in \mathrm{ri}\,\hat{\mathrm{K}}$. Hence $(A-sI) >^{\hat{\mathrm{K}}} 0$. Note that $(A-sI)\mathbf{x} = (1-s)\mathbf{x} >^{\hat{\mathrm{K}}} \mathbf{0}$. So $\rho(A-sI) = 1-s$. Each eigenvalue of $A-sI$ is $\lambda - s$, where $\lambda\in \mathrm{spec}\,(\mathrm{A})$. Apply part 1 of the theorem to deduce that $|\lambda - s|\le 1-s$. Since for any $\zeta\in \mathrm{S}\backslash\{1\}$ we must have that $|\zeta - s| > |\zeta| - s = 1-s$, we obtain that $\mathrm{S(A)}(1)=\{1\}$, which concludes the proof of part 4 for $\mathbb{F}=\mathbb{R}$.

Assume next that $\rho(T)=\rho(A)=1$, and $k=\mathrm{index}\,(1)$. (7.14.11) of Problem 9 yields the equality in (7.14.1). Since any sum in the left-hand side of (7.14.1) is nonnegative with respect to the cone K it follows that $(A-I)^{k-1}P\ge^{\mathrm{K}} 0$. Let $\lambda_1 = 1$ so $s_1 = k$, see notation of Problem 7. Recall that $(A-I)^{k-1}P$ in the basis $\mathbf{b}_1,\ldots,\mathbf{b}_n$ is represented by the component $Z_{1(k-1)}\ne 0$. Hence $(A-I)^{k-1}P \gneq^{\mathrm{K}} 0$.

Assume finally that $\lambda\in\mathrm{spec}\,(\mathrm{T}), |\lambda| = 1, \lambda\ne 1, \mathrm{index}\,(\lambda) = \mathrm{k}$. Let $P(\lambda)$ be the spectral projection on the eigenvalue λ. Then (7.14.11) yields

$$\lim_{m\to\infty}\frac{k!}{m^k}\sum_{r=0}^{m-1}\bar{\lambda}^r T^r = \bar{\lambda}^{k-1}(T-\lambda I)^{k-1}P(\lambda). \tag{7.14.6}$$

Let $\mathbf{y}\in\mathrm{K}, \mathbf{f}\in\mathrm{K}$. Since $\mathbf{f}(T^r\mathbf{y})\ge 0$ and $|\lambda|=1$ we obtain $|\mathbf{f}(\bar{\lambda}^rT^r\mathbf{y})| = \mathbf{f}(T^r\mathbf{y})$. The triangle inequality

$$|\frac{k!}{m^k}\sum_{r=0}^{m-1}\mathbf{f}(\bar{\lambda}^rT^r\mathbf{y})| \le \frac{k!}{m^k}\sum_{r=0}^{m-1}\mathbf{f}(T^r\mathbf{y})$$

Let $m\to\infty$ and use the equalities (7.14.6) and (7.14.1) to deduce (7.14.2).

We now point out why our results hold for a vector space $\mathbf{V}$ over $\mathbb{C}$. Let $T\in\mathrm{Hom}\,(\mathbf{V},\mathbf{V})$ and assume that $\mathbf{b}_1,\ldots,\mathbf{b}_n$ is a basis $\mathbf{V}$. Then $\mathbf{V}_\mathbb{R}$ has a basis $\mathbf{b}_1,\ldots,\mathbf{b}_n,\sqrt{-1}\mathbf{b}_1,\ldots,\sqrt{-1}\mathbf{b}_n$. Clearly T induces an operator $\tilde{T}\in\mathrm{Hom}\,(\mathbf{V}_\mathbb{R},\mathbf{V}_\mathbb{R})$. Let $A\in\mathbb{C}^{n\times n}$ represents T in the basis $\mathbf{b}_1,\ldots,\mathbf{b}_n$. Observe that $A = B+\sqrt{-1}C$, where $B, C\in\mathbb{R}^{n\times n}$. Then $\tilde{T}$ is presented by the matrix $\tilde{A} = \begin{bmatrix} B & -C\\ C & B\end{bmatrix}$ in the basis $\mathbf{b}_1,\ldots,\mathbf{b}_n,\sqrt{-1}\mathbf{b}_1,\ldots,\sqrt{-1}\mathbf{b}_n$. See Problem 10.

Problems

1. Let $\mathbf{V}$ be a vector space of dimension n over $\mathbb{C}$, with a basis $\mathbf{z}_1,\ldots,\mathbf{z}_n$. Show.

(a) $\mathbf{V}$ is a vector space over $\mathbb{R}$ of dimension $2n$, with a basis $\mathbf{z}_1, \sqrt{-1}\mathbf{z}_1, \ldots, \mathbf{z}_n, \sqrt{-1}\mathbf{z}_n$. We denote this real vector space by $\mathbf{V}_{\mathbb{R}}$ and its dimension $\dim_{\mathbb{R}}\mathbf{V}$.

(b) Let the assumptions of 1a hold. Then $\mathbf{V}^*$ can be identified with $(\mathbf{V}_{\mathbb{R}})^*$ as follows. Each $\mathbf{f} \in \mathbf{V}^*$ gives rise to $\hat{\mathbf{f}} \in (\mathbf{V}_{\mathbb{R}})^*$ by the formula $\hat{\mathbf{f}}(\mathbf{z}) = \Re\mathbf{f}(\mathbf{z})$. In particular, if $\mathbf{f}_1, \ldots, \mathbf{f}_n$ form a basis in $\mathbf{V}^*$ then $\hat{\mathbf{f}}_1, \widehat{\sqrt{-1}\mathbf{f}_1}, \ldots, \hat{\mathbf{f}}_n, \widehat{\sqrt{-1}\mathbf{f}_n}$ is a basis in $(\mathbf{V}_{\mathbb{R}})^*$.

2. Let K be a cone. Show that K is pointed if and only the two inequalities $\mathbf{x} \ge^{\mathrm{K}} \mathbf{y}, \mathbf{y} \ge^{\mathrm{K}} \mathbf{x}$ imply that $\mathbf{x} = \mathbf{y}$.

3. Let the assumptions of Problem 1 hold. Assume that $\mathrm{K} \subset \mathbf{V}$ is a cone. Denote by $\mathrm{K}_{\mathbb{R}}$ the induced cone in $\mathbf{V}_{\mathbb{R}}$. Show

   (a) K is closed if and only if $\mathrm{K}_{\mathbb{R}}$ is closed.

   (b) K is pointed if and only if $\mathrm{K}_{\mathbb{R}}$ is pointed.

   (c) K is generating if and only if $\mathrm{K}_{\mathbb{R}}$ is generating.

   (d) K is pointed if and only if $\mathrm{K}_{\mathbb{R}}$ is pointed.

4. Let $\mathbf{U}$ be a real vector space. Denote by $\mathbf{U}_{\mathbb{C}}$ as in Proposition 4.1.2. Assume that $\mathrm{K} \subset \mathbf{U}$ is a cone. Let $\mathrm{K}_{\mathbb{C}} := \{(\mathbf{x}, \mathbf{y}),\ \mathbf{x}, \mathbf{y} \in \mathrm{K}\}$. Show

   (a) $\mathrm{K}_{\mathbb{C}}$ is a cone in $\mathbf{U}_{\mathbb{C}}$.

   (b) K is closed if and only if $\mathrm{K}_{\mathbb{C}}$ is closed.

   (c) K is pointed if and only if $\mathrm{K}_{\mathbb{C}}$ is pointed.

   (d) K is generating if and only if $\mathrm{K}_{\mathbb{C}}$ is generating.

   (e) K is proper if and only if $\mathrm{K}_{\mathbb{C}}$ is proper.

5. Let the assumptions of Problem 4 hold. Assume that $A \in \mathrm{Hom}\,(\mathbf{U}, \mathbf{U})^{\mathrm{K}}$. Define $\hat{A} : \mathbf{U}_{\mathbb{C}} \to \mathbf{U}_{\mathbb{C}}$ by $\hat{A}(\mathbf{x}, \mathbf{y}) = (A\mathbf{x}, A\mathbf{y})$. Show

   (a) $\hat{A} \in \mathrm{Hom}\,(\mathbf{U}_{\mathbb{C}}, \mathbf{U}_{\mathbb{C}})^{\mathrm{K}_{\mathbb{C}}}$.

   (b) $\det\,(zI - A) = \det\,(zI - \hat{A})$.

   (c) $\hat{A}$ is not positive with respect to $\mathrm{K}_{\mathbb{C}}$.

6. Let $\mathbf{V}$ be a vector space over $\mathbb{F} = \mathbb{R}, \mathbb{C}$. Assume that $\mathrm{K} \subset \mathbf{V}$ and $A \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})^{\mathrm{K}}$. Then $A^* \in \mathrm{Hom}\,(\mathbf{V}^*, \mathbf{V}^*)^{\mathrm{K}^*}$.

7. Let $A \in \mathbb{C}^{n\times n}$. Assume that $\mathrm{S(A)} = \{\lambda_1, \ldots, \lambda_{\mathrm{n}}\}$ is the eigenvalue multiset of A. Consider the matrix

   $B(z) = (I - zA)^{-1} = [b_{ij}]_{i=j=1}^n \in \mathbb{C}(z)^{n\times n}$. Show

(a) Let $Z_{i1}, \ldots, Z_{i(s_i-0)}, i = 1, \ldots, \ell$ be all the matrix components of A as in §3.1. Then

i. Z_{i0} is the spectral projection of A on λ_i. (See §3.4.) Furthermore

$$(7.14.7) \qquad (A - \lambda_i I)^{s_i-1} Z_{i0} = Z_{i(s_i-1)} \text{ for } i = 1, \ldots, \ell.$$

(Use (3.1.6).) So $Z_{i(s_i-1)}\mathbb{C}^n$ is a subspace of eigenvectors of A corresponding to all Jordan blocks of A of order s_i and eigenvalue λ_i.

ii.

$$(7.14.8)\, (I - zA)^{-1} = \sum_{i=1}^{\ell} \sum_{j=0}^{s_i-1} \frac{z^j}{(1-\lambda_i z)^{j+1}} Z_{ij},$$

$$(7.14.9) \lim_{t \nearrow 1} (1-t)^{s_i} (I - \frac{t}{\lambda_i} A)^{-1} = \frac{1}{\lambda_i^{s_i}} Z_{i(s_i-1)}.$$

(**Hint:** To show the first equality use (3.4.1) by letting $\lambda = \frac{1}{z}$ and divide (3.4.1) by z.)

(b) All the singular points of all $b_{ij}(z)$ are of the form $\mu := \frac{1}{\lambda}$ where λ is a nonzero eigenvalue of A. Furthermore, if $0 \ne \lambda \in \text{spec (A)}$, and λ has index $l = l(\lambda)$. Then for each i, j $b_{ij}(z)$ may have a pole at μ of order l at most, and there is at least one entry $b_{ij}(z)$, where $i = i(\lambda), j = j(\lambda)$, such that $b_{ij}(z)$ has a pole of order l exactly. Furthermore Suppose furthermore, that for $\mathbf{x} \in \mathbb{C}^n$, at least one of the entries of $B(z)\mathbf{x}$ has a pole of order l at μ. Then $\lim_{t \nearrow 1} (1-t)^l B(t\mu)\mathbf{x} = \mathbf{y} \ne \mathbf{0}$, $A\mathbf{y} = \lambda\mathbf{y}$ and $\mathbf{y} \in (\lambda I - A)^{l-1}\mathbb{C}^n$.

(c) Let $\mathbf{e}_i = (\delta_{i1}, \ldots, \delta_{in})^\top, i \in \langle n \rangle$. For each $0 \ne \lambda \in \text{spec (A)}$ of index $l = l(\lambda)$ there exists $\mathbf{e}_i, \mathbf{e}_j$, $i = i(\lambda), j = j(\lambda)$ such that $\mathbf{e}_i^\top B(z)\mathbf{e}_j$ has a pole at $\frac{1}{\lambda}$ of order l exactly.

8. Let $k, l \in \mathbb{N}$ and consider the rational function

$$f(z) = \frac{1}{(1-\mu z)^l} = \sum_{i=0}^{\infty} (-1)^i \binom{-l}{i} \mu^i z^i.$$

Show

(a)

$$\frac{k!}{m^k} \sum_{i=0}^{m-1} (-1)^i \binom{-k}{i} = \frac{k!}{m^k} \sum_{i=0}^{m-1} |\binom{-k}{i}| = 1.$$

Hint: Use the Riemann sums for the integral $\int_0^1 x^{k-1}dx$ to show

$$\lim_{m\to\infty} \frac{k}{n^k}\sum_{i=1}^{m} i^{k-1} = 1.$$

(b)

$$\frac{k!}{m^k}\sum_{i=0}^{m-1}(-1)^i\binom{-l}{i}\mu^i = 0.$$

Under the following assumptions

i. $|\mu| = 1, \mu \neq 1$ and $l = k$. **Hint**: Recall the identity

$$s_m(z) = \sum_{i=0}^{m-1}(-1)^i\binom{-1}{i}z^i = \sum_{i=0}^{m-1} z^i = \frac{1-z^m}{1-z}.$$

Show that

$$\frac{1}{(k-1)!}\frac{d^{k-1}}{dz^{k-1}}s_{m+k-1}(z) = \sum_{i=0}^{m-1}(-1)^i\binom{-k}{i}z^i.$$

ii. $|\mu| = 1$ and $l < k$. **Hint:** Sum the absolute values of the corresponding terms and use part 8a.

iii. $|\mu| < 1$. **Hint:** Use the Cauchy-Hadamard formula to show that $\sum_{i=0}^{m-1} |\binom{-l}{i}|\,|\mu|^i < \infty$.

9. Let the assumptions of Problem 7 hold.

(a) For $m \geq \max(s_1,\ldots,s_\ell)$

(7.14.10)
$$\sum_{r=0}^{m-1} z^r A^r = \sum_{i=1}^{\ell}\sum_{j=0}^{s_i-1} z^j\Big(\sum_{r=0}^{m-1-j}(-1)^r\binom{-(j+1)}{r}\lambda_i^r z^r Z_{ij}\Big).$$

(Use the first m terms of MacLaurin expansion of both sides of (7.14.8).)

(b) Assume furthermore that $\rho(A) = 1$ and k is the maximal index of all eigenvalues λ_i satisfying $|\lambda_i| = 1$. Assume that $\lambda = \lambda_1, |\lambda| = 1$, and $k = \text{index}\,(\lambda_1) = s_1$. Let $P(\lambda) = Z_{10}$ be the spectral projection on λ and $Z_{1(s_1-1)} = (A-\lambda)^{k-1}P(\lambda_1)$. Then (7.14.10) and Problem 8 implies.

$$\lim_{m\to\infty}\frac{k!}{m^k}\sum_{r=0}^{m-1}\bar{\lambda}^r A^r = \bar{\lambda}^{k-1}(A-\lambda I)^{k-1}P(\lambda). \tag{7.14.11}$$

10. Let $\mathbf{V}$ be a vector space over $\mathbb{C}$ with a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$. Then $\mathbf{V}_\mathbb{R}$ has a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n, \sqrt{-1}\mathbf{b}_1, \ldots, \sqrt{-1}\mathbf{b}_n$. Let $T \in \mathrm{Hom}\,(\mathbf{V}, \mathbf{V})$. Show Clearly T induces an operator $\tilde{T} \in \mathrm{Hom}\,(\mathbf{V}_\mathbb{R}, \mathbf{V}_\mathbb{R})$. Let $A \in \mathbb{C}^{n\times n}$ represents T in the basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$. Observe that $A = B + \sqrt{-1}C$, where $B, C \in \mathbb{R}^{n\times n}$. Then $\tilde{T}$ is presented by the matrix $\tilde{A} = \begin{bmatrix} B & -C \\ C & B \end{bmatrix}$ in the basis $\mathbf{b}_1, \ldots, \mathbf{b}_n, \sqrt{-1}\mathbf{b}_1, \ldots, \sqrt{-1}\mathbf{b}_n$

Bibliography

[Arn71] V.I. Arnold, On matrices depending on paramemters, *Usp. Math. Nauk* 26 (1971), 101-114.

[Avr03] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Dover Publishing, 2003.

[BR97] R.B. Bapat and T.E.S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, Cambridge, UK, 1997.

[BeZ06] I. Bengtsson and K. Zyckowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge U. Press, 2006.

[BeP79] A. Berman and R.J. Plemmons, *Nonnegative Matrices in Mathematical Sciences*, Academic Press, New York 1979.

[Bha84] R. Bhatia, Variation of symmetric tensors and permanents, *Linear Algebra Appl.* 62 (1984), 268–276.

[BhE90] R. Bhatia and L. Elsner, On the variation of permanents, *Linear Multilin. Algebra* 27 (1990), 105-109.

[BhF81] R. Bhatia and S. Friedland, Variation of Grassman powers and spectra, *Linear Algebra Appl.* 40 (1981), 1-18.

[Big96] N. Biggs, *Algebraic Graph Theory*, Cambridge University Press, Cambridge, second edition, 1996.

[BoH91] M. Boyle and D. Handelman, The spectra of nonnegative matrices via symbolic dynamics, *Annals of Math.* 133 (1991), 249–316.

[BPS66] R.A. Brualdi, S.V. Parter and H. Schneider, The diagonal equivalence of a nonnegative matrix to a stochastic matrix, *J. Math. Anal. Appl.* 16 (1966), 31–50.

[BR91] R.A. Brualdi and H.J. Ryser, *Combinatorial Matrix Theory*, Cambridge Univ. Press, Cambridge, 1991.

[Coh79] J.E. Cohen, Random evolutions and the spectral radius of a nonnegative matrix, *Mat. Proc. Camb. Phil. Soc* 86 (1979), 345-350.

[CuR62] C.W. Curtis and I. Reiner, *Representation Theory of Finite Groups and Associative Algebras*, Interscience, New York, 1962.

[Die69] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York-London, 1969.

[DV75] M.D. Donsker and S.R.S. Varadhan, On a variational formula for the principal eigenvalue for operators with maximum principle, *Proc. Nat. Acad. Sci. U.S.A.* 72 (1975), 780–783.

[DDG51] M.A. Drazin, J.W. Dungey and K.W. Gruenberg, Some theorems on commutative matrices, *J. London Math. Soc.*, 26 (1951), 221-228.

[Dan63] G. B. Dantzig, *Linear Programming and Extensions*, Princeton, NJ: Princeton University Press, 1963.

[DM64] A.L. Dulmage and N.S. Mendelsohn, Gaps in the exponent of primitive matrices, *Illinois J. Math.* 8 (1964), 642-656.

[Els82] L. Elsner, On the variation of spectra of matrices, *Linear Algebra Appl.* 47 (1982), 127–132.

[Els88] L. Elsner, A note on the variation of permanents, *Linear Algebra Appl.* 109 (1988), 37-39.

[EJRS83] L. Elsner, C. Johnson, J. Ross, and J. Schönheim, On a generalized matching problem arising in estimating the eigenvalue variation of two matrices, *European J. Combin.* 4 (1983), 133–136.

[Fad66] D.K. Fadeev, On the equivalence of systems of integral matrices, *Izv. Akad. Nauk SSSR*, Ser. Mat. 30 (1966), 449-454.

[Fri73] S. Friedland, Extremal eigenvalue problems for convex sets of symmetric matrices and operators, *Israel J. Math.* 15 (1973), 311-331.

[Fri77] S. Friedland, Inverse eigenvalue problems, *Linear Algebra Appl.* 17 (1977), 15-51.

[Fri78] S. Friedland, Extremal eigenvalue problems, *Bull. Brazilian Math. Soc.* 9 (1978), 13-40.

[Fri78b] S. Friedland, On inverse problem for nonnegative and eventually nonnegative matrices, *Israel J. Math.* 29 (1978), 43-60.

[Fri80a] S. Friedland, On pointwise and analytic similarity of matrices, *Israel J. Math.* 35 (1980), 89-108.

[Fri80b] S. Friedland, Analytic similarities of matrices, *Lectures in Applied Math.*, Amer. Math. Soc. 18 (1980), 43-85 (edited by C.I. Byrnes and C.F. Martin).

[Fri81] S. Friedland, A generalization of the Kreiss matrix theorem, *SIAM J. Math. Anal.* 12 (1981), 826-832.

[Fri81a] S. Friedland, Stable convex sets of matrices, *Linear Multilin. Algebra* 9 (1981), 299-316.

[Fri82] S. Friedland, Variation of tensor powers and spectra, *Linear Multilin. Algebra* 12 (1982), 81–98.

[Fri83] S. Friedland, Simultaneous similarity of matrices, *Adv. in Math.* 50 (1983), 189-265.

[Fri84] S. Friedland, Convex spectral functions, *Linear Multilin. Algebra* 16 (1984), 285-294.

[Fri06] S. Friedland, Convergence of products of matrices in projective spaces, *Linear Alg. Appl* 413(2006), 247-263.

[Fri08] S. Friedland, On the graph isomorphism problem,

[Fri09] S. Friedland, A note on the nonzero spectrum of irreducible matrices, 5 pages, 2009, arXiv:0910.3415v1.

[FrK75] S. Friedland and S. Karlin, Some inequalities for the spectral radius of non-negative matrices and applications, *Duke Mathematical Journal* 42 (3), 459-490, 1975.

[FMMN08] S. Friedland, V. Mehrmann, A. Miedlar and M. Nkengla,Fast low rank approximations of matrices and tensors, prepint 2008.

[FrS80] S. Friedland and H. Schneider, The growth of powers of nonnegative matrix, *SIAM J. Algebraic Discrete Methods* 1 (1980), 185-200.

[FrT84] S. Friedland and E. Tadmor, Optimality of the Lax-Wendroff condition, *Linear Alg. Appl* 56 (1984), 121–129.

[FrT08] S. Friedland and C. W. Tan, Maximizing sum rates in Gaussian interference-limited channels, 24 pp., *submitted*, arXiv:0806.2860.

[FriT07] S. Friedland and A. Torokhti. Generalized rank-constrained matrix approximations, *SIAM J. Matrix Anal. Appl.* 29 (2007), 656 – 659.

[FrZ84] S. Friedland and C. Zenger, All spectral dominant norms are stable, *Linear Algebra Appl.* 58 (1984), 97–107.

[Gan59] F.R. Gantmacher, *The Theory of Matrices*, Vol. I and II, Chelsea Publ. Co., New York 1959.

[GaJ79] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, 1979.

[GaB77] M.A. Gauger and C.I. Byrnes, Characteristic free, improved decidability criteria for the similarity problem, *Lin. Multilin. Alg.* 5 (1977), 153-158.

[Gin78] H. Gingold, A method for global block diagonalization for matrix-valued functions, *SIAM J. Math. Anal.* 9 (1978), 1076-1082.

[GolV96] G.H. Golub and C.F. Van Loan. Matrix Computation, *John Hopkins Univ. Press, 3rd Ed.*, Baltimore, 1996.

[GorT01] S.A. Goreinov and E.E. Tyrtyshnikov, The maximum-volume concept in approximation by low-rank matrices, *Contemporary Mathematics* 280 (2001), 47 – 51.

[GorTZ95] S.A. Goreinov, E.E. Tyrtyshnikov, and N.L. Zamarashkin, Pseudo-skeleton approximations of matrices, *Reports of the Russian Academy of Sciences* 343(2) (1995), 151 – 152.

[GorTZ97] S.A. Goreinov, E.E. Tyrtyshnikov, and N.L. Zamarashkin, A theory of pseudo-skeleton approximations of matrices, *Linear Algebra Appl.* 261 (1997), 1 – 21.

[GrH78] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, J. Wiley, New York, 1978.

[GuR65] R. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice-Hall, New Jersey, 1965.

[Gur80] R.M. Guralnick, A note on the local-global prinicple for similarity of matrices, *Lin. ALg. Appl.* 30 (1980), 241–245.

[Gur81] R.M. Guralnick, Similarity of matrices over local rings, *Linear Algebra Appl.* 41 (1981), 161-174.

[Gur03] L. Gurvits, Classical deterministic complexity of Edmond's problem and quantum entanglement, In: *Proc. 35th STOC* 2003, pp. 10-19.

[Hal35] P. Hall, On representation of subsets, *J. London Math. Soc.* 10 (1935), 26-30.

[Har49] G.H. Hardy, *Divergent Series*, Oxford, Clarendon Press, 1949.

[HPL52] G.H. Hardy, J.E. Littlewood and G. Pólya, *Inequalities*, Cambridge Univ. Press, Second edition, 1952.

[HJ88] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, New York 1988.

[Hel40] O. Helmer, Divisibility properties of integral functions, *Duke. Math. J.* 6 (1940), 345-356.

[Hel43] O. Helmer, The elementary divisor theorems for certain rings without chain conditions, *Bull. Amer. Math. Soc.* 49 (1943), 225-236.

[Hor96] M. Horodecki, P. Horodecki, R. Horodecki, Separability of Mixed States: Necessary and Sufficient Conditions, *Physics Letters A* 223 (1996), 1–8.

[Kap49] I. Kaplansky, Elementary divisors and modules, *Trans. Amer. Math. Soc.* 66 (1949), 464-491.

[Kar84] N.K. Karmakar, A new polynomial agorithm for linear programming, *Combinatorica* 4 (1984), 373-395.

[Kat80] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, 2nd ed., New York 1980.

[Kat82] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, 2nd ed., New York 1982.

[Kha79] L.G. Khachian, A Polynomial Algorithm in Linear Programming, *Dokl. Akad. Nauk SSSR* 244 (1979), 1093-1096. English translation in *Soviet Math. Dokl.* 20, 191-194, 1979.

[Knu98] D.E. Knuth, *The Art of Computer Programming, vol. 3: Sorting and Searching*, Second Edition, Addison-Wesley, 1998.

[Kre62] H.O. Kreiss, Über die stabilitätsdefinition für differenzgleichungen die partielle differentiel gleichungen approximieren, *Nordisk Tidskr., Informations-Behandling* 2 (1962), 153-181.

[Kro90] L. Kronecker, Algebraische reduction der schaaren bilinear formen, *S-B Akad. Berlin*, 1890, 763-778.

[Laf10] T.J. Laffey, A constructive version of the Boyle-Handelman theorem on the spectra of nonnegative matrices, arXiv:1005.0929.

[Lan58] S. Lang, *Introduction to Algebraic Geometry*, Interscience, New York, 1958.

[Lan67] S. Lang, *Algebra*, Addison-Wesley, Mass., 1967.

[Lea48] W.G. Leavitt, A normal form for matrices whose elements are holomorphic functions, *Duke Math. J.* 15 (1948), 463-472.

[LoL78] R. Loewy and D. London, A note on an inverse problem for nonnegative matrices, *Linear Multilin. Algebra* 6 (1978) 83–90.

[Lov86] L. Lovász, *An ALgorithmic Theory of Numbers, Graphs and Convexity*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 50, SIAM, Philadelphia, PA, 1986.

[Luo08] Z.-Q. Luo and Z. Zhang, Dynamic Spectrum Management: Complexity and Duality, *IEEE Journal on Selected Areas in Signal Processing*, 2 (1), 57-73, 2008.

[McD33] C.C. MacDuffee, *The Theory of Matrices*, Springer, Berline, 1933.

[MaM64] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Prindle, Weber & Schmidt, Boston, 1964.

[MoF80] N. Moiseyev and S. Friedland, The association of resonance states with incomplete spectrum of finite complex scaled Hamiltonian matrices, *Phys. Rev.* A 22 (1980), 619-624.

[MoT52] T.S. Motzkin and O. Taussky, Pairs of matrices with property L, *Trans. Amer. Math. Soc.* 73 (1952), 108-114.

[MoT55] T.S. Motzkin and O. Taussky, Pairs of matrices with property L. II, *Trans. Amer. Math. Soc.* 80 (1955), 387-401.

[Neu77] C. Neumann, Untersuchungen über das logarithmische Newtonsche Potential, Teubner Leipzig, 1877.

[NoW99] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 1999.

[Olv74] F.W.J. Olver, *Asymptotic and Special Functions*, Academic Press, New York, 1974.

[Ost66] A. Ostrowski, *Solutions of Equations and Systems of Equations*, 2nd ed., Academic Press, New York 1966.

[Pea66] C. Pearcy, An elementary proof of the power inequality for the numerical radius, *Michigan Math. J.* 13 (1966), 289–291.

[Per96] A. Peres, Separability Criterion for Density Matrices, *Phys. Rev. Lett.* 77 (1996), 1413–1415.

[Pin09] M.A. Pinsky, *Introduction to Fourier analysis and wavelets*, American Mathematical Society, 2009.

[Pri94] A. Pringsheim, Über Functionen, welche in gewissen Punkten endliche differentialquotienten jeder endlichen Ordnung, aber keine Taylor'sche Reihenentwickeling besitzen, *Math. Annalen* 44 (1984), 41–56.

[Rem98] R. Remmert, *Theory of Complex Functions*, Springer-Verlag, 4th printing, 1998.

[Roc70] R.T. Rockafeller, *Convex Analysis*, Princeton Univ. Press, 1970.

[Rot81] U. Rothblum, Expansions of sums of matrix powers, *SIAM Rev.* 23 (1981), 143-164.

[Rot52] W.E. Roth, On the equaiton $AX - YB = C$ and $AX - XB = C$ in matrices, *Proc. Amer. Math. Soc.* 3 (1952), 392-396.

[Rud74] W. Rudin, *Real and Complex Analysis*, McGraw Hill, New York, 1974.

[Sch50] R. Schatten, *A Theory of Cross Spaces*, Ann. Math. Studies, no. 26, 1950.

[Sed59] O. Sedláček, On incedenčnich matiach orientirovaných grafò, *Časop. Pěst. Mat.* 84 (1959), 303-316.

[Sen81] E. Seneta, *Non-Negative Matrices and Markov Chains*, 2nd ed. Springer, New York, 1981.

[Sri03] R. Srikant, *The Mathematics of Internet Congestion Control*, Birkhauser, Boston, 2003.

[Str35] S. Straszewicz, Über exponierte punkte abgeschlossener punktmengen, *Fund. Math.* 24 (1935), 139-143.

[Tad81] E. Tadmor, The equivalence of L_2-stability, the resolvent condition, and strict H-stability, *Linear Alg. Appl.* 41 (1981), 151-159.

[TeD07] F. de Terán and F.M. Dopico, Low rank perturbation of Kronecker structure without full rank, *SIAM J. Matrix Anal. Appl.*, 29 (2007), 496-529.

[Val79] L.G. Valiant, The complexity of computing the permanent, *Theoretical Computer Science* 8 (1979), 189-201.

[vdW59] B.L. Van der Waerden, *Algebra*, Vol. I and II, Springer, Berlin, 1959.

[Var63] R. S. Varga, Matrix iterative analysis, *Prentice Hall*, 1963.

[Viv93] G. Vivanti, Sulle serie di potenze, *Rivista di Matematica* 3 (1893), 111–114.

[vNe37] J. von Neumann, Some matrix-inequalities and metrization of matrix-space, *Tomsk Univ. Rev.* 1 (1937), 286-300.

[Was63] W. Wasow, On holomorphically similar matrices, *J. Math. Anal. Appl.* 4 (1963), 202-206.

[Was77] W. Wasow, Arnold's canonical matrices and asymptotic simplifcation of ordinary differential equations, *Lin. Alg. Appl.* 18 (1977), 163-170.

[Was78] W. Wasow, *Topics in Theory of Linear Differential Equations Having Singularities with respect to a Parameter*, IRMA, Univ. L. Paster, Strassbourg, 1978.

[Wei67] K. Weierstrass, Zur theorie der bilinearen un quadratischen formen, *Monatsch. Akad. Wiss. Berlin*, 310-338, 1867.

[Wie50] H. Wielandt, Unzerlegbare nicht-negative Matrizen, *Math. Z.* 52 (1950), 642-648.

[Wie67] H. Wielandt, *Topics in Analytic Theory of Matrices*, Lecture Notes, Univ. Wisconsin, Madison, 1967.