NOTES ON

Matrix Analysis

Hao Su

Mainly based on NJU 0810Z1B001 Fundamentals of Matrix Theory.

July 30, 2026

Contents

Chapter 1

Algebraic and Analytic Structures

Definition 1.1. A *binary operation* on a set S is a map $S \times S \to S$.

Definition 1.2. A *monoid* is a set G with a binary operation that is associative, and having a unit element e (so that G is nonempty).

Note that e is unique.

Definition 1.3. Say that a monoid G is *commutative* or *abelian* if the binary operation is commutative.

Definition 1.4. A *group* G is a monoid, such that for every element $x \in G$ there exists an element $y \in G$ such that $xy = yx = e$. Such an element y is called an *inverse* for x.

Note that such an inverse is unique.

Example 1.5. Let k be a field and V a vector space over k. Let $\mathcal{GL}(V)$ (general linear) denote the set of invertible k-linear maps of V onto itself. Then $\mathcal{GL}(V)$ is a group under composition of mappings. Similarly, let k be a field and let $\mathcal{GL}(n, k)$ be the set of invertible n-by-n matrices with components in k. Then $\mathcal{GL}(n, k)$ is a group under the multiplication of matrices. For $n \geq 2$, this group is not abelian.

Definition 1.6. A *ring* A is a set, together with two binary operations called multiplication and addition respectively, and written as a product and as a sum respectively, satisfying the following conditions:

RI 1. A is a abelian group with respect to addition.

RI 2. A is a monoid with respect to multiplication.

RI 3. Distributivity.

A ring is commutative if the multiplication is.

Definition 1.7. A field $\mathbb{F}$ is a commutative ring, if every nonzero element $x \in \mathbb{F}$ has a multiplicative inverse $y \in \mathbb{F}$.

Algebraic structures refer to sets with operations. An algebra refers to a vector space equipped with a bilinear multiplication.

Example 1.8. Equivalence relations on sets of square matrices.

1. Matrices $A, B \in \mathbb{F}^{n,n}$ are said to be *similar*, denoted by $A \sim B$, if there is a nonsingular matrix $S \in \mathcal{GL}(n, \mathbb{F})$ such that $A = S^{-1}BS$.
2. Matrices $A, B \in \mathbb{F}^{n,n}$ are said to be *congruent*, denoted by $A \cong B$, if there is a nonsingular matrix $S \in \mathcal{GL}(n, \mathbb{F})$ such that $A = S^*BS$, where $*$ denotes matrix conjugate transposition.
3. Matrices $A, B \in \mathbb{R}^{n,n}$ are said to be *orthogonally similar*, denoted by $A \sim_O B$, if there is an orthogonal matrix $U \in \mathcal{O}(n)$ such that $A = U^{-1}BU$.
4. Matrices $A, B \in \mathbb{C}^{n,n}$ are said to be *unitarily similar*, denoted by $A \sim_U B$, if there is a unitary matrix $U \in \mathcal{U}(n)$ such that $A = U^{-1}BU$.
5. Matrices $A, B \in \mathbb{F}^{n,n}$ are said to be *similar by permutation*, denoted by $A \sim_P B$, if there is a permutation matrix $P \in \mathcal{P}(n)$ such that $A = P^{-1}BP$.

Definition 1.9. A relation on a set is a *partial order* if it satisfies reflexivity, anti-symmetry, and transitivity. A set with a partial order is called a *partially ordered set* or *poset*, denoted by $(S, \preceq)$.

Read $a \preceq b$ as *a precedes b* or *b supercedes a*.

Definition 1.10. For real symmetric matrices $A, B \in \mathcal{S}(n)$, define $A \leq B$ if $x^{\mathrm{T}}Ax \leq x^{\mathrm{T}}Bx$ for all $x \in \mathbb{R}^n$. For Hermitian matrices $A, B \in \mathcal{H}(n)$, define $A \leq B$ if $x^*Ax \leq x^*Bx$ for all $x \in \mathbb{C}^n$. $(\mathcal{S}(n), \leq)$ and $(\mathcal{H}(n), \leq)$ are both posets. This partial order is called the Löwner order.

Definition 1.11. For $x, y \in \mathbb{R}^n$, say that x is *majorized* by y, denoted by $x \prec y$, if

$$\begin{aligned}
\max_{1\le i\le n} x_i &\le \max_{1\le i\le n} y_i \\
\max_{1\le i_1<i_2\le n} (x_{i_1} + x_{i_2}) &\le \max_{1\le i_1<i_2\le n} (y_{i_1} + y_{i_2}) \\
&\vdots \\
\max_{1\le i_1<i_2<\cdots<i_{n-1}\le n} (x_{i_1} + x_{i_2} + \cdots + x_{i_{n-1}}) &\le \max_{1\le i_1<i_2<\cdots<i_{n-1}\le n} (y_{i_1} + y_{i_2} + \cdots + y_{i_{n-1}}) \\
x_1 + x_2 + \cdots + x_n &= y_1 + y_2 + \cdots + y_n.
\end{aligned}$$

Say that x is *weakly majorized* by y, denoted by $x \prec_w y$, if the last equality above is changed to inequality "$\le$".

For example, $(2,2,2)^{\mathrm{T}} \prec (1,2,3)^{\mathrm{T}}$ since $\prec$ arranges vectors according to their level of fluctuation when the average is the same. Also notice that $(1,2,3)^{\mathrm{T}} \prec (3,3,3)^{\mathrm{T}}$ which shows that $\prec_w$ orders the level of fluctuation and the average in a combined way.

Neither the majorization $\prec$ nor the weak majorization $\prec_w$ is a partial order in $\mathbb{R}^n$ since neither is anti-symmetric. Nevertheless, both $\prec$ and $\prec_w$ are partial orders in $\mathbb{R}^n/\simeq$, the set of bags of n real numbers. For simplicity, $\mathbb{R}^n/\simeq$ will be denoted by $\mathbb{R}_n$ in the sequel.

Definition 1.12. In a set S with two equivalence relations $\equiv_1$ and $\equiv_2$, say that $\equiv_1$ is finer than $\equiv_2$ (or $\equiv_2$ is coarser than $\equiv_1$), denoted by $\equiv_1 \preceq \equiv_2$, if $a \equiv_1 b$ implies $a \equiv_2 b$.

Definition 1.13. Let $(S, \preceq_S)$ and $(T, \preceq_T)$ be two posets. A map $f : S \to T$ is said to be *monotonically increasing* if $a \preceq_S b$ implies $f(a) \preceq_T f(b)$. It is said to be *monotonically decreasing* if $a \preceq_S b$ implies $f(b) \preceq_T f(a)$.

Example 1.14. The trace is a monotonically increasing function from poset $(\mathcal{S}(n), \le)$ or $(\mathcal{H}(n), \le)$ to poset $(\mathbb{R}, \le)$, i.e., if $A \le B$, then $\operatorname{tr} A \le \operatorname{tr} B$.

Definition 1.15. A monotonically increasing function from a subset of $(\mathbb{R}_n, \prec)$ to $(\mathbb{R}, \le)$ is also said to be *Schur-convex*. Likewise, a monotonically decreasing function from a subset of $(\mathbb{R}_n, \prec)$ to $(\mathbb{R}, \le)$ is also said to be *Schur-concave*.

A few definitions, followed by several relevant results, are as follows, mainly on majorization and Schur-convexity.

Definition 1.16. A *T-transform* is an n-by-n matrix T obtained from the identity matrix I_n by replacing a 2-by-2 principal submatrix corresponding to indices $i \neq j$ with

$$\begin{pmatrix} t & 1-t \\ 1-t & t \end{pmatrix}, \qquad t \in [0,1],$$

and leaving all other diagonal entries as 1 and all other off-diagonal entries as 0. That is, T acts on a vector $x \in \mathbb{R}^n$ by averaging two components x_i and x_j:

$$(Tx)_i = tx_i + (1-t)x_j, \qquad (Tx)_j = (1-t)x_i + tx_j,$$

and $(Tx)_k = x_k$ for all $k \neq i, j$.

Definition 1.17. A matrix $A = (a_{ij}) \in \mathbb{R}^{n,n}$ is said to be *doubly stochastic* if

$$a_{ij} \geq 0, \qquad \sum_{i=1}^{n} a_{ij} = 1, \qquad \sum_{j=1}^{n} a_{ij} = 1.$$

Theorem 1.18. *The following are equivalent for $x, y \in \mathbb{R}^n$:*

(i) There exists a finite sequence of T-transforms $T_1, \dots, T_m$ such that $x = T_m \cdots T_1\, y$.
(ii) There exists a doubly stochastic matrix A such that $x = Ay$.

Proof. $(i) \Rightarrow (ii)$: Each T-transform is doubly stochastic, since every row and column of its defining 2-by-2 block

$$\begin{pmatrix} t & 1-t \\ 1-t & t \end{pmatrix}, \qquad t \in [0,1],$$

sums to 1, and all other diagonal entries are 1 with zeros elsewhere. Because the product of doubly stochastic matrices is again doubly stochastic, the matrix $A := T_m \cdots T_1$ is doubly stochastic, and hence $x = Ay$.

$(ii) \Rightarrow (i)$: Let A be a doubly stochastic matrix with $x = Ay$. Eliminate off-diagonal mass of A iteratively using pairwise averaging operations supported on two coordinates at a time. Each step replaces a 2-by-2 subblock of A by a matrix of the form $\left(\begin{smallmatrix} t & 1-t \\ 1-t & t \end{smallmatrix}\right)$ without violating the row or column sum constraints. After finitely many steps, the product of these transformations equals A. Applying this factorization to y gives $x = T_m \cdots T_1\, y$ as required. ⊣

Theorem 1.19. ***Birkhoff-von Neumann**** *The set $\mathcal{D}_n$ of all n-byn doubly stochastic matrices is the convex hull of the set $\mathcal{P}_n$ of all n-by-n permutation matrices. In particular, every $A \in \mathcal{D}_n$ can be written as*

$$A = \sum_{\sigma \in S_n} \lambda_\sigma P_\sigma, \qquad \lambda_\sigma \geq 0, \quad \sum_{\sigma \in S_n} \lambda_\sigma = 1,$$

where P_σ is the permutation matrix corresponding to $\sigma \in S_n$.

Proof. The inclusion $\operatorname{conv}(\mathcal{P}_n) \subseteq \mathcal{D}_n$ is immediate since every permutation matrix is doubly stochastic and the class of doubly stochastic matrices is convex.

For the reverse inclusion, observe that $\mathcal{D}_n$ is a bounded convex polytope in $\mathbb{R}^{n,n}$ defined by the linear equations

$$A\mathbf{1} = \mathbf{1}, \qquad A^\top \mathbf{1} = \mathbf{1}, \qquad A \geq 0.$$

Hence $\mathcal{D}_n$ is the convex hull of its extreme points. It therefore suffices to show that the only extreme points are the permutation matrices.

Suppose $A = (a_{ij}) \in \mathcal{D}_n$ is not a permutation matrix. Then there exists at least one row, say the r-th, containing two distinct positive entries $a_{rj_1}, a_{rj_2} > 0$. Because each column sum is 1, there exist rows $s_1 \neq r$ and $s_2 \neq r$ with $a_{s_1 j_1}, a_{s_2 j_2} > 0$. Consider a perturbation E supported on the four positions

$$(r, j_1),\ (r, j_2),\ (s_1, j_1),\ (s_2, j_2)$$

with entries $\pm\varepsilon$ arranged so that every row and column sum of E is zero. Then $A \pm E$ remain nonnegative and doubly stochastic for sufficiently small $\varepsilon > 0$, and

$$A = \tfrac{1}{2}(A + E) + \tfrac{1}{2}(A - E),$$

with $A + E \neq A - E$. Hence A is not an extreme point. Therefore, every extreme point must have exactly one 1 in each row and column and zeros elsewhere; that is, it is a permutation matrix, as desired. ⊣

Theorem 1.20. ***Hardy-Littlewood-Pólya*** *For $x, y \in \mathbb{R}^n$ the following are equivalent:*

(1) x is majorized by y.

(2) There exists a doubly stochastic matrix A such that $x = Ay$.

(3) For every convex $\phi : \mathbb{R} \to \mathbb{R}$,

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

where if ϕ is strictly convex then equality holds iff x is a permutation of y.

Proof. (2) ⇒ (3): Write $x_i = \sum_j a_{ij} y_j$ with $A = (a_{ij})$ doubly stochastic. By Jensen's inequality rowwise, $\phi(x_i) = \phi(\sum_j a_{ij} y_j) \le \sum_j a_{ij}\phi(y_j)$. Summing over i and using $\sum_i a_{ij} = 1$ gives $\sum_i \phi(x_i) \le \sum_j \phi(y_j)$. If ϕ is strictly convex and $\sum_i \phi(x_i) = \sum_j \phi(y_j)$, then every Jensen step is tight, forcing each row of A to concentrate on a single column; with column sums $= 1$ this implies A is a permutation matrix, so x is a permutation of y.

(3) ⇒ (1): Taking $\phi(u) = u$ yields $\sum_i x_i = \sum_i y_i$. For any $t \in \mathbb{R}$, the function $\phi_t(u) := (u-t)_+ = \max\{u-t, 0\}$ is convex, hence $\sum_i (x_i - t)_+ \le \sum_i (y_i - t)_+$ for all t. Let $x^\downarrow, y^\downarrow$ be decreasing rearrangements. For $t = x_k^\downarrow$, $\sum_i (x_i^\downarrow - t)_+ = \sum_{i=1}^k (x_i^\downarrow - x_k^\downarrow) = \sum_{i=1}^k x_i^\downarrow - k x_k^\downarrow$, and similarly for $y^\downarrow$, giving $\sum_{i=1}^k x_i^\downarrow \le \sum_{i=1}^k y_i^\downarrow$ for all k. Together with the sum equality, this is $x \prec y$.

(1) ⇒ (2): Assume $x \prec y$. We shall construct a doubly stochastic matrix A such that $x = Ay$. Choose the smallest index k for which $\sum_{i=1}^k x_i^\downarrow < \sum_{i=1}^k y_i^\downarrow$. Since the total sums are equal, there exists some $l > k$ such that $x_l^\downarrow > y_l^\downarrow$. Replace the pair $(y_k^\downarrow, y_l^\downarrow)$ by

$$(y'_k, y'_l) = (t\, y_k^\downarrow + (1-t) y_l^\downarrow,\ (1-t) y_k^\downarrow + t\, y_l^\downarrow)$$

for a suitable $t \in (0,1)$ so that the new vector y' satisfies $\sum_{i=1}^k x_i^\downarrow = \sum_{i=1}^k y_i'^\downarrow$ and still $y' \succ x$. This operation is precisely a T-transform applied to $(y_k^\downarrow, y_l^\downarrow)$. Repeating this procedure finitely many times yields $x = T_m \cdots T_1\, y$. By 1.18 we are done. ⊣

Lemma 1.21. *Schur-Ostrowski* *Let $D \subseteq \mathbb{R}^n$ be convex and permutation-invariant, and $f : D \to \mathbb{R}$ be C^1 and symmetric. Assume that for all $x \in D$ and $i \ne j$,*

$$(x_i - x_j)\big(\partial_i f(x) - \partial_j f(x)\big) \ge 0. \tag{$*$}$$

Then f is Schur-convex on D.

Proof. By 1.20, it suffices to show that f does not increase under any single T-transform. Fix $x \in D$ and indices $i \ne j$, and set $a = x_i$, $b = x_j$. Consider the averaging path $z(t)$, $t \in [0,1]$, given by

$$z_i(t) = (1-t)a + tb, \qquad z_j(t) = ta + (1-t)b, \qquad z_k(t) = x_k\ (k \notin \{i,j\}).$$

Define $\phi(t) = f(z(t))$. By the chain rule,

$$\phi'(t) = \big(\partial_i f(z(t)) - \partial_j f(z(t))\big)\,(b-a).$$

At each t, $(*)$ applied at $z(t)$ gives

$$\big(z_i(t) - z_j(t)\big)\big(\partial_i f(z(t)) - \partial_j f(z(t))\big) \ge 0,$$

and since $z_i(t) - z_j(t) = (1-2t)(a-b)$, we get $\phi'(t) \le 0$ for $t \in [0, \frac{1}{2}]$ and $\phi'(t) \ge 0$ for $t \in [\frac{1}{2}, 1]$. By $f(z(0)) = f(z(1))$ we conclude that f does not increase along the averaging step. ⊣

Second Proof without 1.20. By symmetry of f, we may restrict to vectors arranged in nonincreasing order:

$$x_1 \geq \cdots \geq x_n, \qquad y_1 \geq \cdots \geq y_n.$$

Write the cumulative sums

$$s_k := \sum_{i=1}^{k} x_i \quad (k = 1, \dots, n), \qquad S_k := \sum_{i=1}^{k} y_i,$$

with the convention $s_0 = S_0 := 0$. Majorization $x \prec y$ is exactly the set of constraints

$$s_k \leq S_k \quad (k = 1, \dots, n-1), \qquad s_n = S_n. \tag{$*$}$$

Conversely, the variables $(s_1, \dots, s_{n-1})$ parametrize the feasible x via

$$x_k = s_k - s_{k-1} \quad (k = 1, \dots, n).$$

Consider $F(s_1, \dots, s_{n-1}) := f(x(s))$ with s_n fixed. By the chain rule and the dependence $x_k = s_k - s_{k-1}$, we have for $k = 1, \dots, n-1$:

$$\frac{\partial F}{\partial s_k} = \sum_{i=1}^{n} \partial_i f(x) \frac{\partial x_i}{\partial s_k} = \partial_k f(x) - \partial_{k+1} f(x). \tag{$**$}$$

Since x is sorted, $x_k \geq x_{k+1}$. Applying $(*)$ with $(i, j) = (k, k+1)$ yields

$$(x_k - x_{k+1})\big(\partial_k f(x) - \partial_{k+1} f(x)\big) \geq 0 \quad \Longrightarrow \quad \partial_k f(x) - \partial_{k+1} f(x) \geq 0.$$

By $(**)$, this gives

$$\frac{\partial F}{\partial s_k} \geq 0 \qquad (k = 1, \dots, n-1).$$

Hence F is coordinatewise nondecreasing in $(s_1, \dots, s_{n-1})$. Under the constraints $(*)$, F is therefore maximized when each s_k attains its upper bound S_k, i.e. at $s_k = S_k$ for $k = 1, \dots, n-1$. This choice recovers $x_k = S_k - S_{k-1} = y_k$, so the maximizer is $x = y$ and

$$\forall x \prec y: \quad f(x) = F(s) \leq F(S) = f(y).$$

Thus f is Schur-convex. ⊣

Theorem 1.22. *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and symmetric, then f is Schur-convex.*

Proof. Assume $x \prec y$. By 1.20 ((1) ⇒ (2)), there exists a doubly stochastic matrix A with $x = Ay$. By 1.19 and the convexity and symmetry of f,

$$f(x) = f(Ay) = f\left(\sum_{\sigma} \lambda_\sigma P_\sigma y\right) \leq \sum_{\sigma} \lambda_\sigma f(P_\sigma y) = \sum_{\sigma} \lambda_\sigma f(y) = f(y).$$ ⊣

Corollary 1.23. *The Hölder p-norms are Schur-convex. Cf. 2.5.*

This is trivial by 1.22. The following proof uses 1.21 though, which is a stronger result than 1.22.

Proof. For $g = \sum_{i=1}^n x_i^p$, we have $\partial_i g(x) = p\, x_i^{p-1}$ and, for $p \geq 1$, the map $t \mapsto t^{p-1}$ is increasing on $[0, \infty)$. Hence for all $i \neq j$,

$$(x_i - x_j)\big(\partial_i g(x) - \partial_j g(x)\big) = p\,(x_i - x_j)\big(x_i^{p-1} - x_j^{p-1}\big) \geq 0,$$

so g is Schur-convex by 1.21. Monotone composition preserves Schur-convexity, giving $\|\cdot\|_p$ for $p \geq 1$. Finally, $\max_i x_i$ is Schur-convex (e.g., as a pointwise maximum of linear functionals), giving the case $p = \infty$.

For an alternative definition of 1.15, where $\prec_w$ is used instead, the proof is trivial, for that p-norm is coordinatewise nondecreasing. ⊣

Definition 1.24. Let T be a subset of a poset $(S, \preceq)$. An element $a \in S$ is said to be a *lower bound* of T if $a \preceq b$ for all $b \in T$. Similarly, an element $a \in S$ is said to be an *upper bound* of T if $b \preceq a$ for all $b \in T$. A lower bound a of T is said to be the *greatest lower bound* or *infimum* of T, denoted by $\inf T$, if $c \preceq a$ for any other lower bound c of T. An upper bound a of T is said to be the *least upper bound* or *supremum* of T , denoted by $\sup T$, if $a \preceq c$ for any other upper bound c of T.
Let $a, b \in S$. Then $\inf\{a, b\}$ is also denoted by $a \wedge b$, called the *meet* of a, b; and $\sup\{a, b\}$ is also denoted by $a \vee b$, called the *join* of a, b.

Note that $\inf T$ and $\sup T$, if exist, are unique.

Definition 1.25. A poset $(S, \preceq)$ is called a *lattice* if $a \wedge b$ and $a \vee b$ exist for all $a, b \in S$.

Theorem 1.26. *A lattice is an algebraic structure $(S, \vee, \wedge)$, consisting of a set S and two binary, commutative and associative operations.*

Proof. We show that $\wedge$ is associative: $\inf\{a, b, c\} \preceq a$ and $\inf\{a, b, c\} \preceq b$ give that $\inf\{a, b, c\} \preceq a \wedge b$. Thus by $\inf\{a, b, c\} \preceq c$ we have that $\inf\{a, b, c\} \preceq (a \wedge b) \wedge c$. By transitivity we have also that $(a \wedge b) \wedge c \preceq a$ and $(a \wedge b) \wedge c \preceq b$. Obviously $(a \wedge b) \wedge c \preceq c$. Thus $(a \wedge b) \wedge c \preceq \inf\{a, b, c\}$. By anti-symmetry we have that $(a \wedge b) \wedge c = \inf\{a, b, c\}$. The rest is trivial. ⊣

Example 1.27. Examples on lattices.

1. $(\mathbb{N}, |)$ is a lattice. Here $a \wedge b$ is the greatest common divisor and $a \vee b$ is the least common multiple.
2. Let $\mathcal{P}$ be the set of all nonzero monic polynomials. Then $(\mathcal{P}, |)$ is also a lattice.
3. For a given set S, $(2^S, \subseteq)$ is a lattice. Here the meet and join are the intersection and union respectively.
4. For the set of real-valued continuous functions on $[0, 1]$, denoted by $C([0, 1])$, we say $f \leq g$ if $f(x) \leq g(x)$ for all $x \in [0, 1]$. $(C([0, 1]), \leq)$ is a lattice with
$$(f \wedge g)(x) = \min\{f(x), g(x)\}, \quad (f \vee g)(x) = \max\{f(x), g(x)\}.$$
5. $(\mathcal{S}(n), \leq)$ and $(\mathcal{H}(n), \leq)$ are not lattices. Cf. 1.3.
6. For $x, y \in \mathbb{R}_n$, if their entries have different total sums, then it is impossible to find their meet or join under partial order $\prec$. Hence $(\mathbb{R}_n, \prec)$ is not a lattice. On the other hand, $(\mathbb{R}_n, \prec_w)$ is a lattice. Cf. 1.4.

Definition 1.28. A lattice S is said to be *distributive* if for all $a, b, c \in S$,
$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c) \tag{i}$$
and
$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c) \tag{ii}$$
Cf. 1.29.

Theorem 1.29. *(i) and (ii) in 1.25 are equivalent.*

Proof. ⊣

Definition 1.30. A lattice S is *modular* if for all $a, b, c \in S$ such that $c \preceq a$,
$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c) = (a \wedge b) \vee c.$$

Definition 1.31. Let T be a subset of lattice $(S, \preceq)$. Then $(T, \preceq)$ is a poset, and is said to be a *sublattice* of $(S, \preceq)$ if it is a lattice and $x \vee_T y$ and $x \wedge_T y$ in T are the same as $x \vee_S y$ and $x \wedge_S y$ in S for all $x, y \in T$.

Definition 1.32. In a set S, a function $d : S \times S \to \mathbb{R}$ is a *metric* if for any $x, y, z \in S$,

1. $d(x, y) > 0$ if $x \neq y$ and $d(x, y) = 0$ if $x = y$. (Positivity)
2. $d(x, y) = d(y, x)$. (Symmetry)
3. $d(x, z) \leq d(x, y) + d(y, z)$. (Triangle Inequality)

A set with a metric is a *metric space*.

Definition 1.33. A sequence $\{x_k\}$ in a metric space (S, d) *converges* if there is a $x_\infty \in S$ such that for each $\epsilon > 0$ there is a $K > 0$ such that $d(x_k, x_\infty) < \epsilon$ for all $k > K$. In this case, say that $\{x_k\}$ has a limit x_∞ and write $\lim_{k\to\infty} x_k = x_\infty$. Say that $\{x_k\}$ is *Cauchy* if for each $\epsilon > 0$ there is a $K > 0$ such that $d(x_m, x_n) < \epsilon$ for all $m, n > K$.

Note that every convergent sequence is Cauchy.

Definition 1.34. A metric space is said to be *complete* if every Cauchy sequence converges. A metric space is said to be *compact* if every sequence has a convergent subsequence (with limit in it).

Exercises

Exercise 3. * Show that the poset of real symmetric matrices $(\mathcal{S}(n), \leq)$ (1.10) and Hermitian matrices $(\mathcal{H}(n), \leq)$ is not a lattice (1.25) when $n > 1$.

Pick two 2-by-2 real symmetric matrices $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. Note that A and B are incomparable. Now we try to find a supremum of $\{A, B\}$. Say that $C = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is. Intuitively we solve for $\det(C - A) = \det(C - B) = 0$, which gives

$$C^\pm = \begin{pmatrix} a & \pm\sqrt{(a-1)a} \\ \pm\sqrt{(a-1)a} & a \end{pmatrix},$$

where we require that $a > 1$.

Now we show that

$$C^+ = \begin{pmatrix} a & \sqrt{(a-1)a} \\ \sqrt{(a-1)a} & a \end{pmatrix} \text{ and } C^- = \begin{pmatrix} a & -\sqrt{(a-1)a} \\ -\sqrt{(a-1)a} & a \end{pmatrix}$$

are indeed the minimum upper bounds of $\{A, B\}$. Let $M_1^\pm = C^\pm - A$ and $M_2^\pm = C^\pm - B$.

Suppose $D \geq 0$ is nonzero. If u_1 is a nonzero vector in $\ker(M_1^\pm)$, then either

$$u_1^\top D u_1 > 0 \quad \Rightarrow \quad u_1^\top (M_1^\pm - D) u_1 < 0,$$

so $M_1^\pm - D$ is not PSD; or else $u_1^\top D u_1 = 0$, which forces $Du_1 = 0$. Similarly, let $u_2 \in \ker(M_2^\pm)$. If $u_2^\top D u_2 > 0$ then $M_2^\pm - D$ is not PSD; otherwise $Du_2 = 0$. But u_1, u_2 are linearly independent, hence if D kills both it must vanish identically, contradicting $D \neq 0$. Thus at least one of $M_1^\pm - D$ or $M_2^\pm - D$ ceases to be PSD, as desired.

C^+ and C^- are obviously incomparable. That is, the poset of real symmetric matrices $(\mathcal{S}(2), \leq)$ is not a lattice. The rest cases for Hermitian matrices or $n > 2$ are trivial generalizations.

Exercise 4. Show that $(\mathbb{R}_n, \prec_w)$ is a lattice by giving ways to compute the meet and join. Also answer the following questions:

1. Is this lattice distributive? Is this lattice modular?
2. Is $(\mathbb{N}_n, \prec_w)$ a sublattice of $(\mathbb{R}_n, \prec_w)$?

to do

Exercise 5. 1.29.

Exercise 7. * 1.23.

Exercise 10. * For $x, y \in [0, \infty)^n$, say that x is *log majorized by* y, denoted by $x \prec_{\log} y$, if $\log x \prec \log y$, where the function log is applied elementwise. Show that $x \prec_{\log} y$ implies $x \prec_w y$.

For each $k \in \{1, \dots, n\}$ define on $\mathbb{R}^k$ function $F_k(z) = \sum_{i=1}^k e^{z_i}$. Note that F_k is C^1 and symmetric. For all $x \in \mathbb{R}^k$ and $i \neq j$,

$$(x_i - x_j)\big(\partial_i f(x) - \partial_j f(x)\big) = (x_i - x_j)(e^{x_i} - e^{x_j}) \geq 0.$$

By 1.21, F_k is Schur convex. (Or immediately by 1.22.) By symmetry of F_k, we may restrict to vectors arranged in nonincreasing order:

$$x_1 \geq \dots \geq x_n, \qquad y_1 \geq \dots \geq y_n.$$

By $x \prec_{\log} y$ we have that $\log x^{(1:k)} \prec \log y^{(1:k)}$, and hence

$$\sum_{i=1}^{k} x_i = F_k(\log x^{(1:k)}) \leq F_k(\log y^{(1:k)}) = \sum_{i=1}^{k} y_i$$

for each K, as desired.

Chapter 2

Linear Spaces

2.1 Linear Spaces and Linear Transformation

Let $\mathcal{X}$ and $\mathcal{Y}$ be vector spaces over $\mathbb{F}$, $\mathcal{U}$ and $\mathcal{V}$ subspaces of $\mathcal{X}$, and $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$.

Proposition 2.1. *$(\mathcal{X}, \subseteq)$ is a lattice, where $\mathcal{U} \wedge \mathcal{V} = \mathcal{U} \cap \mathcal{V}$ and $\mathcal{U} \vee \mathcal{V} = \mathcal{U} + \mathcal{V}$. The set of A-invariant subspaces is a lattice, under the partial order $\subseteq$. Hence it is a sublattice of the lattice of all subspaces of $\mathcal{X}$.*

Definition 2.2. Denote

$$A\mathcal{U} = \{Au \,:\, u \in \mathcal{U}\}$$

and call it *the image of $\mathcal{U}$ under A*. Let $\mathcal{V}$ be a subspace of $\mathcal{Y}$. Denote

$$A^{-1}\mathcal{V} = \{x \in \mathcal{X} \,:\, Ax \in \mathcal{V}\}$$

and call it *the inverse image of $\mathcal{V}$*. A may not be invertible. We have that $\mathcal{R}(A) = A\mathcal{X}$ and $\mathcal{N}(A) = A^{-1}\{0\}$, where $\mathcal{R}$ denotes range and $\mathcal{N}$ denotes null.

Theorem 2.3. *Suppose that $\mathcal{U}_1, \mathcal{U}_2, \mathcal{V}_1$, and $\mathcal{V}_2$ are subspaces of $\mathcal{X}$. Then*

$$A(\mathcal{U}_1 + \mathcal{U}_2) = A\mathcal{U}_1 + A\mathcal{U}_2 \tag{1}$$

$$A(\mathcal{U}_1 \cap \mathcal{U}_2) \subseteq A\mathcal{U}_1 \cap A\mathcal{U}_2 \tag{2}$$

$$A^{-1}(\mathcal{V}_1 + \mathcal{V}_2) \supseteq A^{-1}\mathcal{V}_1 + A^{-1}\mathcal{V}_2 \tag{3}$$

$$A^{-1}(\mathcal{V}_1 \cap \mathcal{V}_2) = A^{-1}\mathcal{V}_1 \cap A^{-1}\mathcal{V}_2. \tag{4}$$

(2) becomes an equality if and only if

$$(*) \quad \mathcal{N}(A) \cap (\mathcal{U}_1 + \mathcal{U}_2) = \mathcal{N}(A) \cap \mathcal{U}_1 + \mathcal{N}(A) \cap \mathcal{U}_2,$$

and (3) becomes an equality if and only if

$$(**) \quad \mathcal{R}(A) \cap (\mathcal{V}_1 + \mathcal{V}_2) = \mathcal{R}(A) \cap \mathcal{V}_1 + \mathcal{R}(A) \cap \mathcal{V}_2.$$

Proof. Denote the equality version of (2) and (3) by (2∗) and (3∗). We only prove (2 ∗) ⇔ (∗) and (3 ∗) ⇔ (∗∗) since the rest are trivial.

(2 ∗) ⇒ (∗): Take $n \in \mathcal{N}(A) \cap (\mathcal{U}_1 + \mathcal{U}_2)$. Write $n = x_1 + x_2$ where $x_i \in \mathcal{U}_i$. Thus $A(x_1 + x_2) = 0$ and $Ax_1 = -Ax_2 \in A\mathcal{U}_1 \cap A\mathcal{U}_2$. By (2 ∗) there exists $u \in \mathcal{U}_1 \cap \mathcal{U}_2$ such that $Au = Ax_1 = -Ax_2$. That is, $\{x_1 - u, x_2 + u\} \subseteq \mathcal{N}(A)$. Then $n = x_1 - u + x_2 + u \in \mathcal{N}(A) \cap \mathcal{U}_1 + \mathcal{N}(A) \cap \mathcal{U}_2$. The other direction is omitted as trivial.

(∗) ⇒ (2 ∗): Take $n \in A\mathcal{U}_1 \cap A\mathcal{U}_2$. Write $n = Ax_1 = Ax_2$, where $Ax_i \in A\mathcal{U}_i$. Thus $x_1 - x_2 \in \mathcal{N}(A) \cap (\mathcal{U}_1 + \mathcal{U}_2)$. By (∗) there exist $y_1 \in \mathcal{N}(A) \cap \mathcal{U}_1$ and $y_2 \in \mathcal{N}(A) \cap \mathcal{U}_2$ such that $y_1 - y_2 = x_1 - x_2$. Thus $x_1 - y_1 = x_2 - y_2 \in \mathcal{U}_1 \cap \mathcal{U}_2$. Hence $n \in A(\mathcal{U}_1 \cap \mathcal{U}_2)$. The other direction is omitted as trivial.

(3 ∗) ⇒ (∗∗): Take $n \in \mathcal{R}(A) \cap (\mathcal{V}_1 + \mathcal{V}_2)$. Write $n = Av$, thus $v \in A^{-1}(\mathcal{V}_1 + \mathcal{V}_2)$. By (3 ∗) there exist $x_1 \in A^{-1}\mathcal{V}_1$ and $x_2 \in A^{-1}\mathcal{V}_2$ such that $v = x_1 + x_2$. Hence $n = Av = Ax_1 + Ax_2 \in \mathcal{R}(A) \cap \mathcal{V}_1 + \mathcal{R}(A) \cap \mathcal{V}_2$. The other direction is omitted as trivial.

(∗∗) ⇒ (3 ∗): Take $n \in A^{-1}(\mathcal{V}_1 + \mathcal{V}_2)$. Then $An \in \mathcal{R}(A) \cap (\mathcal{V}_1 + \mathcal{V}_2)$. By (∗∗) there exist $Ax_1 \in \mathcal{R}(A) \cap \mathcal{V}_1$ and $Ax_2 \in \mathcal{R}(A) \cap \mathcal{V}_2$ such that $An = Ax_1 + Ax_2$. Thus $A(n - x_1) = Ax_2 \in \mathcal{R}(A) \cap \mathcal{V}_2$. Then $n = x_1 + n - x_1$, where $x_1 \in A^{-1}\mathcal{V}_1$ and $n - x_1 \in A^{-1}\mathcal{V}_2$, as desired. The other direction is omitted as trivial. ⊣

2.2 Normed Linear Space

Definition 2.4. A *norm* on a vector space $\mathcal{X}$ is a function $\|\cdot\| : \mathcal{X} \to [0, \infty)$ such that the following properties hold for all $u, v \in \mathcal{X}$ and all $\alpha \in \mathbb{F}$:

(a) $\|u\| = 0$ if and only if $u = 0$;
(b) $\|\alpha u\| = |\alpha|\|u\|$;
(c) $\|u + v\| \le \|u\| + \|v\|$.

A function satisfying only (b) and (c) is called a *semi-norm*. Note that a normed vector space is a metric space (1.32), with the metric defined by $(x, y) \mapsto \|x - y\|$.

Definition 2.5. *Hölder p-norms*, where $p \ge 1$ or $p = \infty$, are defined on $\mathbb{F}^n$ by

$$\left\|\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \qquad \text{for } p \ge 1 \text{ and } \left\|\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right\|_\infty = \max_{1 \le i \le n} |x_i|.$$

The following result shows that Höleder p-norms are indeed norms.

Theorem 2.6. ***Minkowski*** *For $p \geq 1$ and $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n) \in \mathbb{R}^n$,*

$$\|x + y\|_p \;\leq\; \|x\|_p + \|y\|_p.$$

Proof. For $p \geq 1$, the scalar function $\phi(t) = |t|^p$ is convex on $\mathbb{R}$ because

$$\phi''(t) = p(p-1)|t|^{p-2} \geq 0.$$

Since the sum of convex functions is convex, it follows that

$$f(x) = \sum_{i=1}^{n} \phi(x_i)$$

is convex on $\mathbb{R}^n$. Define the unit sublevel set

$$B_p = \{x \in \mathbb{R}^n \,:\, f(x)^{1/p} \leq 1\} = \{x \,:\, f(x) \leq 1\}.$$

Because f is convex, each set $\{x \,:\, f(x) \leq c\}$ is convex. Hence B_p is convex.

For arbitrary x, y, let $a = \|x\|_p$ and $b = \|y\|_p$. Then $\frac{x}{a}, \frac{y}{b} \in B_p$, so

$$\frac{x+y}{a+b} = \frac{a}{a+b}\frac{x}{a} + \frac{b}{a+b}\frac{y}{b} \in B_p$$

by convexity of B_p. Therefore,

$$\left\|\frac{x+y}{a+b}\right\|_p \leq 1 \quad\implies\quad \|x+y\|_p \leq a + b = \|x\|_p + \|y\|_p. \qquad \dashv$$

Comment. This is why we require $p \geq 1$ in 2.5. Note that the case where $p = \infty$ is trivial. Note also that for any norm $\|\cdot\|$ with unit ball $B = \{x \,:\, \|x\| \leq 1\}$, the triangle inequality is equivalent to convexity of B. Sets like B_p that are convex and closed under scaling ($x \in S \to \lambda x \in S$ for $0 \leq \lambda \leq 1$) are called *balanced*.

Definition 2.7. For $1 \leq p \leq \infty$, define

$$\ell_p(\mathbb{Z}) = \left\{\; x = (x(k))_{k\in\mathbb{Z}} \;:\; \|x\|_p < \infty \;\right\},$$

where $\|\cdot\|_p$ is the Hölder p-norm on $\mathbb{F}^{\mathbb{Z}}$.

Definition 2.8. Ky Fan k-norms are defined on $\mathbb{F}^n$ by

$$\left\|\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right\|_{(k)} = \max_{1 \le i_1 < \cdots < i_k \le n} \left(|x_{i_1}| + \cdots + |x_{i_k}|\right).$$

Definition 2.9. Let $A = (a_{ij}) \in \mathbb{F}^{m,n}$ and let $p \in [1, \infty) \cup \{\infty\}$. The *entrywise Hölder p-norm* of A is defined by

$$\|A\|_{p,E} = \|\operatorname{vec} A\|_p,$$

where $\|\cdot\|_p$ denotes the Hölder p-norm on $\mathbb{R}^{mn}$.
In particular, for $p = 2$, the entrywise 2-norm coincides with the *Frobenius norm*:

$$\|A\|_F = \|A\|_{2,E} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}} = |\operatorname{Tr}(AA^*)|^{\frac{1}{2}}.$$

Definition 2.10. Let $A \in \mathbb{F}^{m,n}$ and $1 \le p \le \infty$. The *induced p-norm* (or *operator p-norm*) of A is defined by

$$\|A\|_p = \sup_{x \ne 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\|x\|_p = 1} \|Ax\|_p.$$

where $\|\cdot\|_p$ denotes the Hölder p-norm on $\mathbb{R}^m$.

Let $\sigma_1(A)$ denote the largest singular value of A, then

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|, \qquad \|A\|_2 = \sigma_1(A), \qquad \|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|.$$

Definition 2.11. Let $A \in \mathbb{F}^{m,n}$ and let $\sigma_1(A) \ge \sigma_2(A) \ge \cdots \ge \sigma_{\min\{m,n\}}(A) \ge 0$ denote the singular values of A. For $1 \le p \le \infty$, the *Schatten p-norm* of A is defined by

$$|A|_p = \|(\sigma_i(A))\|_p,$$

where $\|\cdot\|_p$ denotes the Hölder p-norm on $\mathbb{R}^{\min\{m,n\}}$.

$|A|_1$ is called the *trace norm*, $|A|_2$ coincides with the *Frobenius norm*, and $|A|_\infty$ is called the *spectral norm*.

We may define convergence of sequence over vector space and Cauchy sequence like 1.33, and completeness like 1.34, expect that the metric is defined by $d(x, y) = \|x - y\|$. The following results are stated without proof.

Theorem 2.12. *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on a finite-dimensional vector space $\mathcal{X}$. Then there exist finite positive constants C_m and C_M such that*

$$C_m\|x\|_\alpha \le \|x\|_\beta \le C_M\|x\|_\alpha, \qquad \forall x \in \mathcal{X}.$$

Definition 2.13. Say that two norms, denoted by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, are *equivalent* if for each sequence of vectors $\{x_k\}$ in $\mathcal{X}$,

$$\lim_{k\to\infty} x_k = x \text{ with respect to } \|\cdot\|_\alpha \quad \text{iff} \quad \lim_{k\to\infty} x_k = x \text{ with respect to } \|\cdot\|_\beta.$$

Theorem 2.14. *In a finite-dimensional vector space, all norms are equivalent.*

Theorem 2.15. *All finite-dimensional vector spaces are complete.*

2.3 Condition Numbers: Linear Equation

Definition 2.16. A matrix norm $\|\cdot\|$ on $\mathbb{F}^{m,n}$ is said to be *submultiplicative* if for all matrices $A \in \mathbb{F}^{m,n}$ and $B \in \mathbb{F}^{n,p}$,

$$\|AB\| \le \|A\|\,\|B\|.$$

Definition 2.17. For a nonsingular matrix $A \in \mathbb{F}^{n,n}$ and a given matrix norm $\|\cdot\|$ on $\mathbb{F}^{n,n}$ satisfying submultiplicativity, the *condition number* of A with respect to $\|\cdot\|$ is defined by

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|.$$

By convention, if A is singular, we set $\kappa(A) = \infty$.

Theorem 2.18. *Properties of condition number.*

(a) For any nonsingular $A \in \mathbb{F}^{n,n}$,

$$\kappa(A) \geq 1,$$

since

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \cdot \|A^{-1}\| = \kappa(A).$$

(b) For any nonsingular $A \in \mathbb{F}^{n,n}$ and scalar $\alpha \in F \setminus \{0\}$,

$$\kappa(\alpha A) = \|\alpha A\| \cdot \|(\alpha A)^{-1}\| = |\alpha|\,\|A\| \cdot |\alpha|^{-1}\|A^{-1}\| = \kappa(A).$$

Thus, $\kappa(A)$ is invariant under nonzero scalar multiplication.

For $A \in \mathbb{C}^{n,n}$, let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A) > 0$ denote the singular values of A (since A is nonsingular).

Definition 2.19. Suppose we choose 2.10 as the norm. Then the *spectral condition number* is defined by

$$\kappa_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1(A)}{\sigma_n(A)}.$$

The spectral condition number is the ratio of the largest to the smallest singular value of A, measuring the "stretching ratio" of A in the Euclidean space $\mathbb{C}^n$.

Definition 2.20. Suppose we choose $\|\cdot\|_F$ defined in 2.9 as the norm. Then the *Frobenius condition number* is defined by

$$\kappa_F(A) = \|A\|_F \cdot \|A^{-1}\|_F = \sqrt{\left(\sum_{i=1}^{n} \sigma_i^2(A)\right)\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2(A)}\right)}.$$

The Frobenius condition number is less common in error analysis than the spectral condition number, but is useful in numerical linear algebra for matrix approximation problems.

Definition 2.21. We like wise define κ_1 and κ_∞. For the 1-norm (column-sum norm),

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|, \qquad \kappa_1(A) = \|A\|_1 \cdot \|A^{-1}\|_1.$$

For the ∞-norm (row-sum norm),

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|, \qquad \kappa_\infty(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty.$$

Note that $\kappa_\infty(A) = \kappa_1(A^\top)$, since $\|A\|_\infty = \|A^\top\|_1$ and $\|A^{-1}\|_\infty = \|(A^\top)^{-1}\|_1$.

A central application of condition numbers is quantifying the sensitivity of solutions to linear systems

$$Ax = b,$$

where $A \in \mathbb{F}^{n,n}$ is nonsingular and $b \in F^n \setminus \{0\}$.

Suppose b is perturbed by $\Delta b \in F^n$, leading to a perturbed solution $x + \Delta x$ satisfying

$$A(x + \Delta x) = b + \Delta b.$$

Then

$$A\Delta x = \Delta b \quad \Rightarrow \quad \Delta x = A^{-1}\Delta b.$$

Taking vector norms (compatible with the matrix norm used for $\kappa(A)$),

$$\|b\| = \|Ax\| \le \|A\| \cdot \|x\| \;\Rightarrow\; \|x\| \ge \frac{\|b\|}{\|A\|},$$

$$\|\Delta x\| = \|A^{-1}\Delta b\| \le \|A^{-1}\| \cdot \|\Delta b\|.$$

Combining these gives the relative error bound

$$\frac{\|\Delta x\|}{\|x\|} \le \kappa(A)\frac{\|\Delta b\|}{\|b\|}.$$

Note that a large $\kappa(A)$ means small relative errors in b can cause large relative errors in x, making A *ill-conditioned.*

Suppose A is perturbed by $\Delta A \in \mathbb{F}^{n,n}$) such that $A + \Delta A$ is nonsingular, and the perturbed system

$$(A + \Delta A)(x + \Delta x) = b$$

holds. Expanding and neglecting the higher-order term $\Delta A\,\Delta x$ (valid for small $\|\Delta A\|$) gives

$$A\Delta x \approx -\Delta A x \quad \Rightarrow \quad \Delta x \approx -A^{-1}\Delta A x.$$

Taking norms,

$$\frac{\|\Delta x\|}{\|x\|} \le \kappa(A)\frac{\|\Delta A\|}{\|A\|}.$$

We may improve this bound for small perturbations with

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \frac{\kappa(A)\,\|\Delta A\|/\|A\|}{1 - \kappa(A)\,\|\Delta A\|/\|A\|},$$

which holds when $\kappa(A)\,\|\Delta A\|/\|A\| < 1$.

If both A and b are perturbed ($A \mapsto A + \Delta A$, $b \mapsto b + \Delta b$), then

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \frac{\kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right)}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}}.$$

valid for $\kappa(A)\,\|\Delta A\|/\|A\| < 1$.

Exercises

Exercise 2. 2.3.

Exercise 10. Prove that a norm satisfying the parallelogram equality comes from an inner product.

Proof. Real case. Define

$$\langle x, y\rangle := \tfrac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right).$$

Complex case. Define

$$\langle x, y\rangle := \tfrac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 + i\|x+iy\|^2 - i\|x-iy\|^2\right).$$

We prove only that the function in complex case is indeed an inner product. First note that $\langle x, x\rangle = \|x\|^2$ satisfies positivity and definiteness. For conjugate symmetry, note that

$$\begin{aligned}
\overline{\langle x, y\rangle} &= \tfrac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 - i\|x+iy\|^2 + i\|x-iy\|^2\right) \\
&= \tfrac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 - i\,|i|\,\|-ix+y\|^2 + i\,|-i|\,\|+ix+y\|^2\right) \\
&= \tfrac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 - i\|y+ix\|^2 + i\|y-ix\|^2\right) = \langle y, x\rangle.
\end{aligned}$$

Before we prove additivity and homogeneity in the first argument, set $q(x) := \|x\|^2$. The parallelogram identity is

$$q(u+v) + q(u-v) = 2q(u) + 2q(v) \qquad (\forall u, v \in V). \tag{P}$$

And define for $x, y \in V$

$$\langle x, y\rangle := \frac{1}{4}\Big(q(x+y) - q(x-y) + i\,q(x+iy) - i\,q(x-iy)\Big). \tag{C0}$$

Additivity. Fix $y \in V$ and $x_1, x_2 \in V$. First handle the real part. Apply (P) with $u = x_1 + y, v = x_2$:

$$q(x_1 + x_2 + y) + q(x_1 - x_2 + y) = 2q(x_1 + y) + 2q(x_2). \tag{A1}$$

Apply (P) with $u = x_1 - y, v = x_2$:

$$q(x_1 + x_2 - y) + q(x_1 - x_2 - y) = 2q(x_1 - y) + 2q(x_2). \tag{A2}$$

Subtract (A2) from (A1):

$$\begin{aligned}&\big(q(x_1 + x_2 + y) - q(x_1 + x_2 - y)\big) + \big(q(x_1 - x_2 + y) - q(x_1 - x_2 - y)\big)\\ &= 2\big(q(x_1 + y) - q(x_1 - y)\big).\end{aligned} \tag{A3}$$

Now swap x_1, x_2 in the same derivation. Using $u = x_2 + y, v = x_1$ and $u = x_2 - y, v = x_1$ we obtain

$$\begin{aligned}&\big(q(x_1 + x_2 + y) - q(x_1 + x_2 - y)\big) + \big(q(-x_1 + x_2 + y) - q(-x_1 + x_2 - y)\big)\\ &= 2\big(q(x_2 + y) - q(x_2 - y)\big).\end{aligned} \tag{A4}$$

Since q is even, $q(-x_1 + x_2 \pm y) = q(x_1 - x_2 \mp y)$. Substitute these into (A4) to get

$$\begin{aligned}&\big(q(x_1 + x_2 + y) - q(x_1 + x_2 - y)\big) + \big(q(x_1 - x_2 - y) - q(x_1 - x_2 + y)\big)\\ &= 2\big(q(x_2 + y) - q(x_2 - y)\big).\end{aligned} \tag{A5}$$

Adding (A3) and (A5) cancels the middle terms and yields

$$\begin{aligned}&2\big(q(x_1 + x_2 + y) - q(x_1 + x_2 - y)\big)\\ &= 2\Big(\big(q(x_1 + y) - q(x_1 - y)\big) + \big(q(x_2 + y) - q(x_2 - y)\big)\Big).\end{aligned} \tag{A6}$$

Divide by 8 to obtain

$$\operatorname{Re}\langle x_1 + x_2, y\rangle = \operatorname{Re}\langle x_1, y\rangle + \operatorname{Re}\langle x_2, y\rangle. \tag{A7}$$

For the imaginary part, replace y by iy in the previous argument. Repeating (A1)–(A6) with $y \mapsto iy$ gives

$$\operatorname{Re}\langle x_1 + x_2, iy\rangle = \operatorname{Re}\langle x_1, iy\rangle + \operatorname{Re}\langle x_2, iy\rangle. \tag{A8}$$

But by (C0),

$$\operatorname{Im}\langle x, y\rangle = \frac{1}{4}\big(q(x+iy) - q(x-iy)\big) = \operatorname{Re}\langle x, iy\rangle. \tag{A9}$$

Combining (A8) and (A9) gives

$$\operatorname{Im}\langle x_1 + x_2, y\rangle = \operatorname{Im}\langle x_1, y\rangle + \operatorname{Im}\langle x_2, y\rangle. \tag{A10}$$

Together, (A7) and (A10) imply

$$\langle x_1 + x_2, y\rangle = \langle x_1, y\rangle + \langle x_2, y\rangle. \tag{Add}$$

Homogeneity.

(a) Integer and rational real scalars. From (Add) with $x_1 = x_2 = \cdots = x$ (induction) and the identity $\langle -x, y\rangle = -\langle x, y\rangle$ (obvious from (C0)), we have

$$\langle nx, y\rangle = n\langle x, y\rangle \quad (\forall n \in \mathbb{Z}). \tag{H1}$$

Let $r = \frac{m}{n} \in \mathbb{Q}$ with $n \neq 0$. Then by (H1),

$$n\langle rx, y\rangle = \langle nrx, y\rangle = \langle mx, y\rangle = m\langle x, y\rangle,$$

hence

$$\langle rx, y\rangle = r\langle x, y\rangle \quad (\forall r \in \mathbb{Q}). \tag{H2}$$

(b) Real scalars by continuity. The map $t \mapsto \langle tx, y\rangle$ is continuous because it is a linear combination of continuous maps $t \mapsto q(tx \pm y)$. Since $\mathbb{Q}$ is dense, (H2) extends to all $t \in \mathbb{R}$:

$$\langle tx, y\rangle = t\langle x, y\rangle \quad (\forall t \in \mathbb{R}). \tag{H3}$$

(c) Multiplication by i. Using (C0) and the norm homogeneity $\|iz\| = \|z\|$,

$$\begin{aligned}
\langle ix, y\rangle &= \frac{1}{4}\Big(q(ix + y) - q(ix - y) + i\, q(ix + iy) - i\, q(ix - iy)\Big) \\
&= \frac{1}{4}\Big(q(x - iy) - q(x + iy) + i\, q(x + y) - i\, q(x - y)\Big) \\
&= -\frac{1}{4}\big(q(x + iy) - q(x - iy)\big) \;+\; i\,\frac{1}{4}\big(q(x + y) - q(x - y)\big) \\
&= -\operatorname{Im}\langle x, y\rangle \;+\; i\,\operatorname{Re}\langle x, y\rangle \;=\; i\Big(\operatorname{Re}\langle x, y\rangle + i\,\operatorname{Im}\langle x, y\rangle\Big) \;=\; i\,\langle x, y\rangle.
\end{aligned} \tag{H4}$$

(d) General complex scalars. For $\alpha = a + ib$ with $a, b \in \mathbb{R}$, using (H3), (H4), and (Add) with $x_1 = ax$, $x_2 = bx$,

$$\begin{aligned}
\langle \alpha x, y\rangle &= \langle ax + b(ix), y\rangle = \langle ax, y\rangle + \langle b(ix), y\rangle = a\langle x, y\rangle + b\,\langle ix, y\rangle \\
&= (a + ib)\langle x, y\rangle = \alpha\langle x, y\rangle.
\end{aligned} \tag{Hom}$$

(Add) and (Hom) gives linearity of $\langle \cdot, y\rangle$ in the first argument over $\mathbb{C}$. ⊣

Exercise 13. Prove that 2.5 and 2.8 are indeed norms in $\mathbb{F}^n$.

2.5 are indeed norms, due to 2.6.

to do

Chapter 3

Matrix Factorizations and Decompositions

Theorem 3.1. *Let $A \in \mathbb{C}^{n,n}$. If A admits a right eigenvector $v \neq 0$ associated with eigenvalue $\lambda \in \mathbb{C}$, i.e.*

$$Av = \lambda v,$$

then there exists a nonzero vector w such that

$$w^* A = \lambda w^*.$$

That is, a left eigenvector always exists corresponding to λ.

Proof. Since $Av = \lambda v$, we have $\det(A - \lambda I) = 0$. Taking conjugate transposes gives

$$\det(A - \lambda I) = \overline{\det(A - \lambda I)} = \det\left((A - \lambda I)^*\right) = \det(A^* - \overline{\lambda} I) = 0.$$

Hence $\overline{\lambda}$ is an eigenvalue of A^*. Therefore, there exists a nonzero vector w satisfying

$$A^* w = \overline{\lambda} w.$$

Taking conjugate transposes of both sides yields

$$w^* A = \lambda w^*.$$

Thus w^* is a left eigenvector of A associated with λ. ⊣

Definition 3.2. The set of eigenvalues of A, called the *spectrum* of A, is denoted by $\sigma(A)$. The *spectral radius* of $A \in \mathbb{F}^{n,n}$, denoted by $\rho(A)$, is defined as

$$\rho(A) := \max\{\, |\lambda| \, : \, \lambda \text{ is an eigenvalue of } A \,\}.$$

Definition 3.3. Let $A = (a_{ij}) \in \mathbb{C}^{n,n}$. For each pair of indices $1 \leq i, j \leq n$, the *algebraic* *cofactor* of a_{ij}, denoted by A_{ij}, is defined as

$$A_{ij} = (-1)^{i+j} \det(A'_{ij}),$$

where A'_{ij} is the $(n-1) \times (n-1)$ matrix obtained from A by deleting the i-th row and j-th column.

Definition 3.4. The *cofactor matrix* of A is the matrix

$$C(A) = (A_{ij})_{1 \le i,j \le n},$$

whose (i, j)-entry is the algebraic cofactor of a_{ij}.

Definition 3.5. The *adjugate* (or *classical adjoint*) of A, denoted by $\operatorname{adj}(A)$, is the transpose of its cofactor matrix:

$$\operatorname{adj}(A) = C(A)^\top.$$

It satisfies the fundamental identity

$$A \operatorname{adj}(A) = \operatorname{adj}(A)\, A = \det(A) I_n.$$

Definition 3.6. Let $A \in \mathbb{C}^{n,n}$ and $z \in \mathbb{C}$ such that $(zI - A)$ is invertible. The matrix function

$$R(z, A) = (zI - A)^{-1}$$

is called the *resolvent* of A.

Theorem 3.7. *From linear algebra, we know that*

$$(zI - A)^{-1} = \frac{\operatorname{adj}(zI - A)}{c_A(z)}$$

Theorem 3.8. ***Leverrier–Souriau–Faddeeva–Frame*** *Let*

$$c_A(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_n,$$
$$\operatorname{adj}(zI - A) = B_1 z^{n-1} + B_2 z^{n-2} + \cdots + B_n.$$

Then $a_1, \dots, a_n$ and $B_1, \dots, B_n$ can be computed by the following formulas:

$$\begin{aligned} B_1 &= I, \\ a_i &= -\frac{1}{i}\operatorname{tr}(B_i A), \quad i = 1, 2, \dots, n, \\ B_{i+1} &= B_i A + a_i I, \quad i = 1, 2, \dots, n-1. \end{aligned}$$

By 3.7 we have

$$c_A(z)I = \operatorname{adj}(zI - A)(zI - A),$$

we have

$$\begin{aligned} &(z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_n)I \\ &= B_1 z^n + (B_2 - B_1 A)z^{n-1} + (B_3 - B_2 A)z^{n-2} + \cdots - B_n A. \end{aligned}$$

This gives $B_1 = I$, $B_{i+1} = B_i A + a_i I$ for $i = 1, 2, \dots, n-1$, and $a_n I + B_n A = 0$, giving an algorithm to compute A^{-1}.

Theorem 3.9. ***Cayley–Hamilton*** $c_A(A) = 0$.

Proof. Using 3.8 we have

$$\begin{aligned} 0 &= B_n A + a_n I \\ &= (B_{n-1}A + a_{n-1}I)A + a_n I \\ &= B_{n-1}A^2 + a_{n-1}A + a_n I \\ &= (B_{n-2}A + a_{n-2}I)A^2 + a_{n-1}A + a_n I \\ &= B_{n-2}A^3 + a_{n-2}A^2 + a_{n-1}A + a_n I \\ &\ \ \vdots \\ &= (B_1 A^{n-1} + a_1 A^{n-2} + \cdots + a_{n-2}A^2 + a_{n-1}A) + a_n I \\ &= A^n + a_1 A^{n-1} + \cdots + a_{n-1}A + a_n I = c_A(A) \end{aligned}$$

⊣

Definition 3.10. Let $A \in \mathbb{C}^{n,n}$ with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ (counted without repetition). For each λ_i, denote by

$$\begin{aligned} m_i &= \text{algebraic multiplicity of } \lambda_i, \\ d_i &= \text{geometric multiplicity of } \lambda_i = \dim\ker(A - \lambda_i I). \end{aligned}$$

The matrix A is called *defective* if there exists at least one eigenvalue λ_i such that

$$d_i < m_i.$$

Equivalently, A does not possess a complete set of linearly independent eigenvectors and is therefore not diagonalizable.
The matrix A is called *nondefective* if

$$d_i = m_i \quad \text{for all } i,$$

that is, A has n linearly independent eigenvectors and is diagonalizable over $\mathbb{C}$.

We can diagonalize a matrix by finding n linearly independent eigenvectors. Cf. 1.8. Similarity transformation of a matrix corresponds to representing its underlying linear transformation on $\mathbb{C}^n$ in another basis.Hence easily, similar matrices share same eigenvalues and the same characteristic polynomial. Note that having the same eigenvalues does not imply similarity.

Definition 3.11. A matrix $A \in \mathbb{C}^{n,n}$ is said to be *cyclic* if there exists a vector $x \in \mathbb{C}^n$ such that

$$\{x,\ Ax,\ A^2x,\ \dots,\ A^{n-1}x\}$$

forms a basis of $\mathbb{C}^n$. Such a vector x is called a *cyclic vector* for A. Equivalently, A is cyclic if and only if its minimal polynomial has degree n.

Obviously, if A is cyclic, then A is similar to a Jordan block.

Lemma 3.12. *For linear maps U, V on an n-dimensional space,*

$$\dim \mathcal{N}(UV) \le \dim \mathcal{N}(U) + \dim \mathcal{N}(V).$$

Proof. By Sylvester's rank inequality,

$$\dim \mathcal{R}(UV) \ge \dim \mathcal{R}(U) + \dim \mathcal{R}(V) - n.$$

Hence

$$\begin{aligned}\dim \mathcal{N}(UV) &= n - \dim \mathcal{R}(UV) \\ &\le n - (\dim \mathcal{R}(U) + \dim \mathcal{R}(V) - n) \\ &= \dim \mathcal{N}(U) + \dim \mathcal{N}(V).\end{aligned}$$

⊣

Theorem 3.13. *If A has a single eigenvalue with geometric multiplicity 1, then $(A - \lambda I)$ is cyclic.*

Proof. Set $N := A - \lambda I$. Then N has the sole eigenvalue 0 and $\dim \mathcal{N}(N) = 1$. By the Cayley–Hamilton theorem applied to $A - \lambda I$, the characteristic polynomial of N is t^n, hence $N^n = 0$. Thus N is nilpotent, and its nilpotency index is some $s \le n$. We will show that $s = n$, i.e. $N^{n-1} \neq 0$, which suffices.

Apply 3.12 inductively with $U = N$ and $V = N^k$:

$$\dim \mathcal{N}(N^{k+1}) \le \dim \mathcal{N}(N) + \dim \mathcal{N}(N^k), \tag{$*$}$$

which gives

$$\dim \mathcal{N}(N^k) \le k \, \dim \mathcal{N}(N) = k$$

for all $k \ge 1$, since $\dim \mathcal{N}(N) = \dim \ker N = 1$.

Because $N^n = 0$, we have $\ker N^n = \mathbb{C}^n$, hence $\dim \mathcal{N}(N^n) = n$. But the inequality above says $\dim \mathcal{N}(N^n) \le n$, that is, each inequality in $(*)$ is an equality. Therefore the nilpotency index s equals n. In particular, $N^{n-1} \neq 0$, as desired. ⊣

Theorem 3.14. *For each $A \in \mathbb{C}^{n,n}$ there exists an invertible $S \in \mathcal{GL}(n, \mathbb{C})$ such that*

$$S^{-1} A S = \operatorname{diag}(J_1, \dots, J_\ell),$$

where each block $J_i \in \mathbb{C}^{n_i, n_i}$ is itself a block–diagonal matrix

$$J_i = \operatorname{diag}(J_{i1}, J_{i2}, \dots, J_{im_i}), \qquad \sum_{i=1}^{\ell} n_i = n,$$

and each J_{ij} is a Jordan block associated with the eigenvalue λ_i:

$$J_{ij} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix} = \lambda_i I_{k_{ij}} + N_{k_{ij}},$$

with $k_{ij} = \dim J_{ij}$ and $N_{k_{ij}}$ the nilpotent matrix having ones on the first superdiagonal and zeros elsewhere. ℓ is the number of distinct eigenvalues, m_i is the geometric multiplicity of λ_i and n_i the algebraic multiplicity of λ_i.

Definition 3.15. A matrix U is *co-isometry* if $UU^* = I$.

A unitary matrix is both isometry and co-isometry.

Exercises

Exercise 5. Show the following.

1. *Singular value decomposition*: For $A \in \mathbb{F}^{m,n}$, there exist unitary matrices U, V, and a positive diagonal matrix Σ such that
$$A = U\Sigma V^*.$$
2. *Polar decomposition*: For each $A \in \mathbb{F}^{n,n}$, there exist a positive semidefinite matrix P and a unitary matrix U such that
$$A = PU.$$
3. *Cholesky factorization*: A positive definite matrix A can be factorized as $A = T^*T$ where T is upper triangular.
4. *CS decomposition*: For a $2n$-by-$2n$ unitary matrix W, there exist n-by-n unitary matrices U_1, U_2, V_1, V_2 such that
$$W = \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} \begin{bmatrix} C & -S \\ S & C \end{bmatrix} \begin{bmatrix} V_1^* & 0 \\ 0 & V_2^* \end{bmatrix}$$
where
$$C = \begin{bmatrix} \cos\theta_1 & & \\ & \ddots & \\ & & \cos\theta_n \end{bmatrix}, \qquad S = \begin{bmatrix} \sin\theta_1 & & \\ & \ddots & \\ & & \sin\theta_n \end{bmatrix},$$
for some $\theta_i \in [0, \frac{\pi}{2}]$, $i = 1, \dots, n$.

1. Follows from spectral theorem, which follows from Schur decomposition theorem. Cf. Notes on Linear Algebra.
2. Follows from SVD. Cf. Notes on Linear Algebra.
3. Follows from QR factorization. Cf. Notes on Linear Algebra.
4. to do

Exercise 11. Show the following:

1. Let $A \in \mathbb{R}^{n,m}$ and $B \in \mathbb{R}^{n,p}$. Then there exists $C \in \mathbb{R}^{p,m}$ such that $BC = A$ if and only if $\mathcal{R}(A) \subseteq \mathcal{R}(B)$.
2. Let $A \in \mathbb{R}^{n,m}$ and $B \in \mathbb{R}^{p,m}$. Then there exists $C \in \mathbb{R}^{n,p}$ such that $CB = A$ if and only if $\mathcal{N}(B) \subseteq \mathcal{N}(A)$.

1. That $BC = A \Rightarrow \mathcal{R}(A) \subseteq \mathcal{R}(B)$ is trivial. We show the other direction. $\mathcal{R}(A) \subseteq \mathcal{R}(B)$ implies that $Ax_i \in \mathcal{R}(B)$, where $\{x_i\}_{i=1}^m$ is a basis of $\mathbb{R}^m$. Then there exist a list of vectors $\{y_i\}_{i=1}^m$ such that $By_i = Ax_i$. $\{y_i\}$ and $\{x_i\}$ give C.

2. That $CB = A \Rightarrow \mathcal{N}(B) \subseteq \mathcal{N}(A)$ is trivial. We show the other direction. Say that $\{x_i\}_{i=1}^k$ is a basis of $\mathcal{R}(B)$, and let $\{y_i\}_{i=1}^k$ be a list such that $By_i = x_i$. Extend $\{x_i\}_{i=1}^k$ to $\{x_i\}_{i=1}^p$ to be a basis of $\mathbb{R}^p$, and extend $\{y_i\}_{i=1}^k$ to $\{y_i\}_{i=1}^p$ where $y_j = 0$ for $j > k$. Then $\{x_i\}_{i=1}^p$ and $\{Ay_i\}_{i=1}^p$ give C. $\mathcal{N}(B) \subseteq \mathcal{N}(A)$ is necessary, implying that $\{y_i\}_{i=1}^k$ spans $\mathbb{R}^m \setminus \mathcal{N}(A)$.

Alternatively, by that $CB = A$ iff $B^*C^* = A^*$ and that $\mathcal{N}(B) = (\mathcal{R}(B))^\perp$ and hence $\mathcal{N}(B) \subseteq \mathcal{N}(A)$ iff $\mathcal{R}(A *) \subseteq \mathcal{R}(B^*)$, it follows directly from (1).

Exercise 14. Consider linear equation

$$Ax = b$$

where $A \in \mathbb{F}^{m,n}$, $b \in \mathbb{F}^m$, and $x \in \mathbb{F}^n$. If A is square and nonsingular, we can get the unique solution of the equation easily as $x = A^{-1}b$. However, we are often interested in the case when A is rectangular or singular.

1. Assume $\operatorname{rank} A = m$. Then there may be infinitely many solutions. Show that the minimum norm solution, i.e., the solution with the smallest 2-norm, is
$$x = A^*(AA^*)^{-1}b.$$
2. Assume $\operatorname{rank} A = n$. Then there may not be solutions. Show that the so-called least square solution, i.e., the x which minimizes $\|Ax - b\|_2$, is
$$x = (A^*A)^{-1}A^*b.$$
3. In general, we are interested in the minimum norm least square solution, which is the one with the smallest 2-norm among all x that minimizes $\|Ax - b\|_2$. Show that the minimum norm least square solution of the linear equation can be given in the form
$$x = A^\dagger b$$
for some matrix $A^\dagger \in \mathbb{F}^{n,m}$.

4. For the linear equation
$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} x = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$
find its minimum norm least square solution.

It is a known result (cf. Notes on Linear Algebra) that $x = A^\dagger b$ minimizes $\|Ax - b\|_2$ and, among all minimizers, has the smallest $\|x\|_2$. It then suffices for (1) and (2) to show that the expression given for x coincides with $A^\dagger x$ and for (3) to give the expression for $A^\dagger$.

Let $A \in \mathbb{F}^{m,n}$ have the SVD

$$A = U\Sigma V^*, \qquad U \in \mathbb{F}^{m,m}, \quad V \in \mathbb{F}^{n,n}, \quad \Sigma = \begin{bmatrix} \operatorname{diag}(\sigma_1, \dots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix},$$

where $r = \operatorname{rank} A$ and $\sigma_i > 0$. The Moore–Penrose pseudoinverse is

$$A^\dagger = V\Sigma^\dagger U^*, \qquad \Sigma^\dagger = \begin{bmatrix} \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}) & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus (3) is already solved.

For (1), $r = m$ and $\Sigma\Sigma^* = \operatorname{diag}(\sigma_1^2, \dots, \sigma_m^2)$ is invertible. We compute

$$AA^* = U\Sigma\Sigma^* U^*, \qquad (AA^*)^{-1} = U(\Sigma\Sigma^*)^{-1}U^*.$$

Then
$$A^*(AA^*)^{-1} = V\Sigma^* U^* \, U(\Sigma\Sigma^*)^{-1}U^* = V\big(\Sigma^*(\Sigma\Sigma^*)^{-1}\big)U^* = V\Sigma^\dagger U^* = A^\dagger.$$
Therefore,
$$x = A^*(AA^*)^{-1}b = A^\dagger b.$$
For (2), $r = n$ and $\Sigma^*\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$ is invertible. We compute

$$A^*A = V\Sigma^*\Sigma V^*, \qquad (A^*A)^{-1} = V(\Sigma^*\Sigma)^{-1}V^*.$$

Then
$$(A^*A)^{-1}A^* = V(\Sigma^*\Sigma)^{-1}V^* \, V\Sigma^* U^* = V\big((\Sigma^*\Sigma)^{-1}\Sigma^*\big)U^* = V\Sigma^\dagger U^* = A^\dagger.$$
Therefore,
$$x = (A^*A)^{-1}A^*b = A^\dagger b.$$
For (4), we can directly optimize $\|Ax - b\|$ or we can compute $A^\dagger$: let

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}, \qquad u = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Since $A = uv^\top$,

$$A = \|u\|\|v\|\ \hat{u}\hat{v}^\top$$

already give an SVD of A, with

$$\sigma_1 = \|u\|\,\|v\|.$$

Thus

$$A^\dagger = \frac{1}{\sigma_1}\hat{v}\hat{u}^\top = \frac{1}{\|u\|^2\|v\|^2}vu^\top = \frac{1}{50}\begin{bmatrix}1 & 3\\ 2 & 6\end{bmatrix}.$$

Exercise 18. Show the following

1. $\|A\|_p \geq \rho(A)$;
2. for each $\epsilon > 0$, there exists $S \in \mathcal{GL}(n, \mathbb{F})$ such that $\|S^{-1}AS\|_p < \rho(A) + \epsilon$;
3. $\lim_{k\to\infty} A^k = 0$ iff $\rho(A) < 1$;
4. $\lim_{k\to\infty} \|A^k\|_p^{1/k} = \rho(A)$.

to do

Chapter 4

Matrix Analysis

Definition 4.1. A matrix $A \in \mathbb{F}^{n,n}$ is *Hermitian* if $A = A^*$, denoted by $A \in \mathcal{H}(n, \mathbb{F})$.

Definition 4.2. A matrix $A \in \mathbb{F}^{n,n}$ is *skew Hermitian* if $A = -A^*$.

Theorem 4.3. *The following statements are equivalent:*

1. $A > 0$.
2. $\operatorname{eig}(A) > 0$.
3. $\det\begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix} > 0$ *for all* $i = 1, 2, \ldots, n$.

The determinants in statement (3) of the above theorem are called *leading principal minors*. Thus a Hermitian matrix is positive definite if and only if all its leading principal minors are positive.

Example 4.4. Matrix

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

is not positive semidefinite, but

$$\det(a_{11}) = 0 \quad \text{and} \quad \det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = 0.$$

Definition 4.5. Define the triple of the numbers of negative eigenvalues, zero eigenvalues, and positive eigenvalues, all counted with multiplicity, as the *inertia* of A, denoted by

$$\pi(A) = \{\pi_-(A),\ \pi_0(A),\ \pi_+(A)\},$$

where $\pi_-(A)$, $\pi_0(A)$, and $\pi_+(A)$ are the numbers of negative, zero, and positive eigenvalues of A, respectively. The *signature* of a matrix A is defined to be $\pi_+(A) - \pi_-(A)$.

For example, an n-by-n positive definite matrix A has $\pi(A) = \{0, 0, n\}$. Also note that the rank of A is $\pi_+(A) + \pi_-(A)$.

In the quadratic form x^*Ax, if we make a variable change $x = Tz$ where $T \in \mathcal{GL}(n, \mathbb{F})$, then the quadratic form becomes z^*T^*ATz. The corresponding matrix changes from A to T^*AT.

Definition 4.6. The transformation from A to T^*AT, where T is nonsingular, is called a *congruence transformation*. The transformation from A to $T^{-1}AT$, where T is nonsingular, is called a *similarity transformation*.

An immediate consequence of the spectral theorem is that for each $A \in \mathcal{H}(n, \mathbb{F})$, there exists a $T \in \mathcal{GL}(n, \mathbb{F})$ such that

$$T^*AT = \begin{bmatrix} I_{\pi_+(A)} & 0 & 0 \\ 0 & 0_{\pi_0(A)} & 0 \\ 0 & 0 & -I_{\pi_-(A)} \end{bmatrix}.$$

T^*AT here is called a *signature matrix*, whose trace gives the signature of A.

Theorem 4.7. ***Sylvester*** *For any Hermitian matrix A and any nonsingular matrix T, the inertia of A remains unchanged under a congruence transformation:*

$$\pi(T^*AT) = \pi(A).$$

Proof. Say that

$$S^*AS = \begin{bmatrix} I_{\pi_+(A)} & 0 & 0 \\ 0 & 0_{\pi_0(A)} & 0 \\ 0 & 0 & -I_{\pi_-(A)} \end{bmatrix}.$$

where $S \in \mathcal{GL}(n, \mathbb{F})$. Then

$$T^*AT = T^*(S^*)^{-1} \begin{bmatrix} I_{\pi_+(A)} & 0 & 0 \\ 0 & 0_{\pi_0(A)} & 0 \\ 0 & 0 & -I_{\pi_-(A)} \end{bmatrix} S^{-1}T.$$

Let

$$\begin{aligned} R &= T^*(S^*)^{-1}, \\ R_+ &= R_{1:\pi_+(A),1:\pi_+(A)}, \\ R_- &= R_{n-\pi_-(A)+1;n,n-\pi_-(A)+1;n}. \end{aligned}$$

Then

$$T^*AT = \begin{bmatrix} R_+R_+^* & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -R_-R_-^* \end{bmatrix}$$

$R_+R_+^*$ and $R_-R_-^*$ are clearly positive definite. Thus the inertia of T^*AT equals that of A, as desired. ⊣

Theorem 4.8. *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{bmatrix} \in H(n_1 + n_2, \mathbb{F}),$$

where A_{11} is invertible. Then

$$\pi(A) = \pi(A_{11}) + \pi(A/A_{11}).$$

In particular, $A > 0$ if and only if both $A_{11} > 0$ and $A/A_{11} > 0$.

Proof. Cf. 8.3.

$$\begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}^* \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{bmatrix} \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & A/A_{11} \end{bmatrix}.$$

⊣

Theorem 4.9. *Let $A \in H(n)$. Denote by $d(A) = [a_{11}, \dots, a_{nn}]$ the vector of diagonal entries and by $\lambda(A) = [\lambda_1, \dots, \lambda_n]$ the vector of eigenvalues. Then*

$$d(A) \prec \lambda(A).$$

Proof. By the spectral theorem, $A = U\Lambda U^*$, where U is unitary and

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n).$$

The diagonal entries of A are

$$d_i = a_{ii} = \sum_{j=1}^{n} \lambda_j |u_{ij}|^2.$$

The matrix $S = (|u_{ij}|^2)$ is doubly stochastic, since $\sum_j |u_{ij}|^2 = 1$ and $\sum_i |u_{ij}|^2 = 1$. Therefore $d(A)$ is majorized by $\lambda(A)$. ⊣

Theorem 4.10. *Let $A \geq 0$ be an n-by-n positive semidefinite matrix with diagonal elements $d_1 \geq d_2 \geq \dots \geq d_n$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Then*

$$d(A) \prec_{\log} \lambda(A).$$

In particular, for $k = 1$, we have the Hadamard inequality

$$\det A \leq \prod_{i=1}^{n} d_i.$$

Proof. Firstly, assume that H is positive definite, so the diagonal elements and eigenvalues are all positive. Note that if $f(t)$ is a convex function, then $x \prec y$ implies $[f(x_1), \dots, f(x_n)]^{\mathrm{T}} \prec_w [f(y_1), \dots, f(y_n)]^{\mathrm{T}}$. Therefore, applying the decreasing convex function $f(t) = -\ln t$ to the majorization $d \prec \lambda$, we obtain the desired result. For general positive semidefinite matrices, the statement still holds by a continuity argument. ⊣

Theorem 4.11. *Given $d, \lambda \in \mathbb{R}^n$ satisfying $d \prec \lambda$, there exists a real symmetric (hence Hermitian) matrix with diagonal d and eigenvalues λ.*

Definition 4.12. In $\mathbb{R}^{m,n}$, define a partial order $\trianglelefteq$ for $A, B \in \mathbb{R}^{m,n}$,

$$A \trianglelefteq B \quad \text{iff} \quad a_{ij} \leq b_{ij} \text{ for all } i, j.$$

A matrix $A \in \mathbb{R}^{m,n}$ is said to be *positive* if $0 \triangleleft A$. The set of all positive m-by-n matrices is denoted by $\mathbb{R}^{m,n}_{+}$. A matrix $A \in \mathbb{R}^{m,n}$ is said to be *nonnegative* if $0 \trianglelefteq A$. The set of all nonnegative m-by-n matrices is denoted by $\mathbb{R}^{m,n}_{0,+}$.

The relation is called a *partial order* because not every pair of matrices is comparable.

Lemma 4.13. *If $0 \trianglelefteq B \trianglelefteq A$, then $\rho(B) \leq \rho(A)$.*

Proof. For nonnegative matrices A and B, $B \trianglelefteq A$ implies $B^k \trianglelefteq A^k$, which implies

$$\|B^k\|_\infty^{1/k} \leq \|A^k\|_\infty^{1/k}.$$

Finally, by 3.18 we see

$$\rho(B) = \lim_{k\to\infty} \|B^k\|_\infty^{1/k} \leq \lim_{k\to\infty} \|A^k\|_\infty^{1/k} = \rho(A). \qquad ⊣$$

Hence the function ρ from the poset $(\mathbb{R}^{n,n}_{+}, \trianglelefteq)$ to the poset $(\mathbb{R}, \leq)$ is monotonically increasing.

Theorem 4.14. *For $A \in \mathbb{R}^{n,n}_{0,+}$,*

$$\min_{1\le i\le n} \sum_{j=1}^{n} a_{ij} \le \rho(A) \le \max_{1\le i\le n} \sum_{j=1}^{n} a_{ij},$$

and

$$\min_{1\le j\le n} \sum_{i=1}^{n} a_{ij} \le \rho(A) \le \max_{1\le j\le n} \sum_{i=1}^{n} a_{ij}.$$

Theorem 4.15. ***Perron*** *If $A \in \mathbb{R}^{n,n}$ is positive, then*

1. *$\rho(A) > 0$,*
2. *$\rho(A)$ is a simple eigenvalue of A whose algebraic multiplicity equals 1,*
3. *A has a positive eigenvector for $\rho(A)$,*
4. *all other eigenvalues have modulus less than $\rho(A)$.*

Definition 4.16. Consider $A \in \mathbb{R}^{n,n}_{0,+}$.

1. A is called *reducible* if there exists a permutation matrix P such that

$$P^\top AP = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

   where B and D are square (of sizes k and $n-k$ with $1 \le k \le n-1$).
2. A is called *irreducible* if it is not reducible.
3. A is called *primitive* iff $\exists k \in \mathbb{N}$ such that $A^k \in \mathbb{R}^{n,n}_{+}$.

Theorem 4.17. *When $A \trianglerighteq 0$, the following are equivalent.*

(i) A is irreducible.

(ii) For every nonempty proper subset $S \subseteq \{1, \dots, n\}$, there exist $i \in S$ and $j \notin S$ with $a_{ij} \neq 0$.

(iii) The directed graph $G(A)$ with vertices $\{1, \dots, n\}$ and an edge $i \to j$ whenever $a_{ij} > 0$ is strongly connected.

Theorem 4.18. *For $n \geq 1$,*

$$\mathbb{R}^{n,n}_{+} \subset \{\textit{primitive matrices}\} \subset \{\textit{irreducible matrices}\} \subset \mathbb{R}^{n,n}_{0,+}.$$

Proof. Suppose A is primitive, so $A^k > 0$ for some k. If A were reducible, there exists a permutation matrix P and a $1 \leq r \leq n-1$ with

$$P^\top AP = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

whence for all $m \geq 1$,

$$(P^\top AP)^m = \begin{bmatrix} B^m & * \\ 0 & D^m \end{bmatrix},$$

whose lower-left block is still 0. Conjugating back gives $(A^m)_{ij} = 0$ for all $i > r \geq j$, contradicting $A^k > 0$. Hence A is irreducible. ⊣

Equivalently, a nonnegative A is primitive iff it is irreducible and aperiodic (the gcd of all directed cycle lengths in $G(A)$ equals 1).

Theorem 4.19. ***Frobenius*** *If $A \in \mathbb{R}^{n,n}_{0,+}$, then $\rho(A)$ is an eigenvalue of A and there exists a nonnegative corresponding eigenvector. Moreover, if A is irreducible, then*

1. *$\rho(A) > 0$,*
2. *$\rho(A)$ is a simple eigenvalue,*
3. *A has a positive eigenvector for $\rho(A)$.*

If A is primitive, all other eigenvalues of A have modulus strictly less than $\rho(A)$.

For a nonnegative matrix A, its spectral radius is called the Perron-Frobenius eigenvalue, and a nonnegative corresponding eigenvector is called a Perron-Frobenius eigenvector.

A nonnegative matrix $S = [s_{ij}] \in \mathbb{R}^{n,n}$ is called *stochastic* if each row sum equals 1, i.e.

$$\sum_{j=1}^{n} s_{ij} = 1 \quad \text{for all } i.$$

If S and T are stochastic, then ST is also stochastic. The Perron–Frobenius eigenvalue of S is 1, and one Perron–Frobenius eigenvector is $e = [1, 1, \ldots, 1]^\mathrm{T}$.

A matrix $M \in \mathbb{C}^{n,n}$ whose diagonal entries are nonnegative and off-diagonal entries are nonpositive can always be expressed in the form

$$M = sI - A,$$

for some scalar $s > 0$ and some matrix $A \trianglerighteq 0$.

Definition 4.20. A matrix $M = sI - A$ with $A \trianglerighteq 0$ is called an *M-matrix* if $s \geq \rho(A)$.

Every M-matrix has all its eigenvalues lying in the closed right half of the complex plane, with the eigenvalue $s - \rho(A)$ being the one closest to the origin.

Theorem 4.21. *A matrix M is a nonsingular M-matrix if and only if its inverse M^{-1} is nonnegative.*

Proof. ($\Rightarrow$): Write $M = sI - B$, where $B \trianglerighteq 0$ and $s > \rho(B)$. Then

$$M^{-1} = (sI - B)^{-1} = \frac{1}{s} \sum_{k=0}^{\infty} \left(\frac{B}{s}\right)^k \trianglerighteq 0.$$

to do ⊣

Definition 4.22. An M-matrix is called a *Laplacian matrix* if the sum of the entries in each row equals zero.

Every Laplacian matrix has a zero eigenvalue, and all its other eigenvalues have positive real parts.

Definition 4.23. Given a simple graph G with n vertices $v_1, \ldots, v_n$, the associated Laplacian matrix $L \in \mathbb{R}^{n,n}$ is defined element-wise as

$$L_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j, \\ 0 & \text{otherwise.} \end{cases}$$

Chapter 5

Singular Values

Variational characterization is a field of mathematical analysis that uses variations to find maxima and minima of functionals: mappings from a set of functions to the real numbers.

Definition 5.1. For any $A \in \mathcal{H}(n)$ and nonzero $x \in \mathbb{C}^n$, the *Rayleigh* (or *Rayleigh–Ritz*) quotient $R_A : \mathbb{C}^n \setminus \{0\} \to \mathbb{R}$ is

$$R_A(x) = \frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{x^*Ax}{x^*x}.$$

The Rayleigh quotient of an eigenvector is the eigenvalue it corresponds to.

Theorem 5.2. *For any $A \in \mathcal{H}(n)$,*

$$\lambda_{\min}(A) \le R_A(x) \le \lambda_{\max}(A),$$

where

$$\lambda_{\max}(A) = \max_{\substack{x\in\mathbb{C}^n \\ \|x\|_2=1}} x^*Ax, \qquad \lambda_{\min}(A) = \min_{\substack{x\in\mathbb{C}^n \\ \|x\|_2=1}} x^*Ax.$$

Proof. Take eigendecomposition on A to get $A = U^*\Lambda U$. Let $y = Ux$, we have

$$x^*Ax = y^*\Lambda y = \sum_{i=1}^n \lambda_i |y_i|^2.$$

Since A is Hermitian, all its eigenvalues are real. We have the following inequality:

$$\lambda_{\min} \sum_{i=1}^n |y_i|^2 \le \sum_{i=1}^n \lambda_i |y_i|^2 \le \lambda_{\max} \sum_{i=1}^n |y_i|^2.$$

Since the 2-norm is unitarily invariant, dividing by $\|x\|_2^2$ completes the proof. ⊣

Hence the Rayleigh quotient may be seen as a weighted average of the eigenvalues. Easily we have that $\lambda_{\max}(A) = \max R_A(x)$ and $\lambda_{\min}(A) = \min R_A(x)$.

Remark 5.3. For any A and B,

$$\max_B f(A,B) \geq f(A,B) \quad \text{and} \quad f(A,B) \geq \min_A f(A,B).$$

Therefore, for any A and B,

$$\max_B f(A,B) \geq \min_A f(A,B).$$

Taking the minimum over A and the maximum over B on both sides gives

$$\min_A \max_B f(A,B) \geq \max_B \min_A f(A,B).$$

Theorem 5.4. ***Courant–Fischer*** *Let $A \in \mathcal{H}(n)$ with nonincreasing eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then for $1 \leq k \leq n$,*

$$\lambda_k = \max_{\dim \mathcal{U}=k} \min_{x \in \mathcal{U} \setminus \{0\}} R_A(x) = \min_{\dim \mathcal{U}=n-k+1} \max_{x \in \mathcal{U} \setminus \{0\}} R_A(x).$$

Corollary 5.5. *Let $A \in \mathbb{C}^{m,n}$ with nonincreasing singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$. Then for $1 \leq k \leq \min(m,n)$,*

$$\sigma_k = \max_{\dim \mathcal{U}=k} \min_{\substack{x \in \mathcal{U} \setminus \{0\} \\ \|x\|_2=1}} \|Ax\|_2 = \min_{\dim \mathcal{U}=n-k+1} \max_{\substack{x \in \mathcal{U} \setminus \{0\} \\ \|x\|_2=1}} \|Ax\|_2.$$

Theorem 5.6. *For $A \in \mathbb{C}^{n,n}$,*

$$|\lambda(A)| \prec_{\log} \sigma(A).$$

Proof. Let $A \in \mathbb{C}^{n,n}$. By the Schur decomposition there exist a unitary matrix Q and an upper triangular matrix T such that $A = QTQ^*$, and the diagonal entries of T are the eigenvalues of A, i.e. $T_{ii} = \lambda_i(A)$ for $i = 1, \ldots, n$. Since multiplication by unitary matrices does not change singular values, $\sigma_i(A) = \sigma_i(T)$ for all i. Fix k with $1 \leq k \leq n$, and let $T[k]$ denote the leading principal k-by-k submatrix of T. Because T is upper triangular, $T[k]$ is also upper triangular, hence

$$\prod_{i=1}^{k} |\lambda_i(A)| = |\det T[k]|.$$

Two standard facts are used:

1. (*Determinant–singular value bound*) For any k-by-k matrix B with singular values $\sigma_1(B) \geq$

$\cdots \geq \sigma_k(B) \geq 0$,

$$|\det B| \leq \prod_{i=1}^{k} \sigma_i(B).$$

2. (*Monotonicity of singular values for submatrices*) If B is any k-by-k principal submatrix of an n-by-n matrix M, then for $i = 1, \dots, k$,

$$\sigma_i(B) \leq \sigma_i(M).$$

Applying these to $T[k]$ yields

$$\prod_{i=1}^{k} |\lambda_i(A)| = |\det T[k]| \leq \prod_{i=1}^{k} \sigma_i(T[k]) \leq \prod_{i=1}^{k} \sigma_i(T) = \prod_{i=1}^{k} \sigma_i(A),$$

and equality holds for $k = n$. Therefore $|\lambda(A)| \prec_{\log} \sigma(A)$. ⊣

Theorem 5.7. *For $A, B \in \mathbb{C}^{n,n}$, the singular values of the product satisfy*

$$[\sigma_i(A)\sigma_{n\downarrow i+1}(B)] \prec_{\log} \sigma(AB),$$

or equivalently, for every $1 \leq i_1 < i_2 < \cdots < i_k \leq n$,

$$\prod_{j=1}^{k} \sigma_{i_j}(AB) \leq \prod_{j=1}^{k} \sigma_j(A)\sigma_{i_j}(B),$$

with equality when $k = n$.

Theorem 5.8. *For $A, B \in \mathbb{C}^{n,n}$,*

$$\sigma(AB) \prec_{\log} [\sigma_i(A)\sigma_i(B)],$$

or equivalently, for $1 \leq k \leq n$,

$$\prod_{i=1}^{k} \sigma_i(AB) \leq \prod_{i=1}^{k} \sigma_i(A)\sigma_i(B),$$

where equality holds for $k = n$.

Lemma 5.9. ***Ky Fan Maximum Principle*** *For $X \in \mathcal{H}(n)$ and $1 \leq k \leq n$,*

$$\max_{\substack{S \subseteq \mathbb{C}^n \\ \dim S = k}} \operatorname{Tr}(P_S X) = \sum_{j=1}^{k} \lambda_j^{\downarrow}(X),$$

where P_S denotes the orthogonal projection onto the subspace S.

Proof. Let X admit the spectral decomposition

$$X = \sum_{i=1}^{n} \lambda_i \, u_i u_i^*, \qquad \lambda_1 \geq \cdots \geq \lambda_n, \quad \{u_i\}_{i=1}^{n} \text{ orthonormal.}$$

For any k-dimensional subspace S with orthogonal projector P_S,

$$\operatorname{Tr}(P_S X) = \operatorname{Tr}\left(P_S \sum_{i=1}^{n} \lambda_i u_i u_i^* \right) = \sum_{i=1}^{n} \lambda_i \operatorname{Tr}(P_S u_i u_i^*) = \sum_{i=1}^{n} \lambda_i \langle P_S u_i, u_i \rangle.$$

Define $p_i := \langle P_S u_i, u_i \rangle = \|P_S u_i\|^2$. Since P_S is an orthogonal projection,

$$0 \leq p_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} p_i = \operatorname{Tr}(P_S) = k.$$

Hence

$$\operatorname{Tr}(P_S X) = \sum_{i=1}^{n} \lambda_i p_i = \sum_{i=1}^{k} \lambda_i - \sum_{i=1}^{k} \lambda_i q_i + \sum_{i=1}^{n-k} \lambda_{k+i} r_i,$$

where $0 \leq q_i \leq 1$, $0 \leq r_i \leq 1$, and $\sum q_i = \sum r_i$, thus

$$\operatorname{Tr}(P_S X) \leq \sum_{i=1}^{k} \lambda_i.$$

Equality is achieved by taking $S = \operatorname{span}\{u_1, \dots, u_k\}$, for which $P_S u_i = u_i$ if $i \leq k$ and $P_S u_i = 0$ otherwise. ⊣

Theorem 5.10. *For $A, B \in \mathcal{H}(n)$,*

$$\lambda(A + B) \prec \lambda^{\downarrow}(A) + \lambda^{\downarrow}(B).$$

Proof. For $X \in \mathcal{H}(n)$, let $s_k(X) := \sum_{i=1}^{k} \lambda_i^{\downarrow}(X)$.

Apply 5.9 to $X = A + B$:

$$s_k(A+B) = \max_{\dim S = k} \operatorname{Tr}\big(P_S(A+B)\big) = \max_{\dim S = k} \big(\operatorname{Tr}(P_S A) + \operatorname{Tr}(P_S B)\big).$$

For any fixed S, $\operatorname{Tr}(P_S A) \le \max_{\dim U = k} \operatorname{Tr}(P_U A) = s_k(A)$ and similarly $\operatorname{Tr}(P_S B) \le s_k(B)$. Hence

$$s_k(A+B) \;\le\; s_k(A) + s_k(B) \qquad (1 \le k \le n).$$

For $k = n$ we have equality because $s_n(X) = \operatorname{Tr} X$ is additive, as desired. ⊣

Theorem 5.11. *For $A, B \in \mathbb{H}(n)$,*

$$\lambda^{\downarrow}(A) + \lambda^{\uparrow}(B) \prec \lambda(A+B),$$

or equivalently,

$$\lambda^{\downarrow}(A+B) - \lambda^{\downarrow}(A) \prec \lambda(B).$$

Theorem 5.12. *For $A, B \in \mathbb{C}^{n,n}$,*

$$|\sigma^{\downarrow}(A) - \sigma^{\downarrow}(B)| \prec_w \sigma(A-B),$$

or equivalently, for any $1 \le i_1 < i_2 < \cdots < i_k \le n$,

$$\sum_{j=1}^{k} |\sigma_{i_j}(A) - \sigma_{i_j}(B)| \le \sum_{i=1}^{k} \sigma_i^{\downarrow}(A-B).$$

Corollary 5.13. *For $A, B \in \mathbb{C}^{n,n}$,*

$$\sigma(A+B) \prec_w \sigma^{\downarrow}(A) + \sigma^{\downarrow}(B).$$

Definition 5.14. A norm $\|\cdot\|$ on $\mathbb{C}^{m,n}$ is *unitarily invariant* if it satisfies

$$\|U^* A V\| = \|A\|$$

for all unitary U and V.

Definition 5.15. A function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is called a *symmetric gauge function* if it satisfies the following conditions:

1. $\Phi(x) \geq 0$ and $\Phi(x) = 0$ if and only if $x = 0$,
2. $\Phi(\lambda x) = |\lambda|\Phi(x)$ for any $\lambda \in \mathbb{R}$,
3. $\Phi(x + y) \leq \Phi(x) + \Phi(y)$,
4. $\Phi(Px) = \Phi(x)$ for any permutation matrix P,
5. $\Phi(|x|) = \Phi(x)$, where $|x| = (|x_1|, |x_2|, \dots, |x_n|)$.

Theorem 5.16. ***von Neumann*** *There is a one-to-one correspondence between unitarily invariant norms $\|\cdot\|$ on $\mathbb{C}^{n,n}$ and symmetric gauge functions Φ on $\mathbb{R}^n$, determined by*

$$\|A\| = \Phi(\sigma(A)),$$

where $\sigma(A)$ denotes the vector of singular values of A.

Theorem 5.17. *Let the singular value decompositions of $A, B \in \mathbb{C}^{m,n}$ be $A = U_A \Sigma_A V_A^*$ and $B = U_B \Sigma_B V_B^*$, respectively. Then, for all unitarily invariant norms,*

$$\|\Sigma_A - \Sigma_B\| \leq \|A - B\|.$$

Definition 5.18. A norm $\|\cdot\|$ over $\mathbb{C}^{n,n}$ is called *symmetric* if for any $A, B, C \in \mathbb{C}^{n,n}$,

$$\|BAC\| \leq \sigma_1(B)\sigma_1(C)\|A\|.$$

Theorem 5.19. *A norm is symmetric if and only if it is unitarily invariant.*

Theorem 5.20. *All unitarily invariant (that is, symmetric) norms are submultiplicative.*

Theorem 5.21. *The Frobenius norm is unitarily invariant; for any unitary $U \in \mathbb{C}^{m,m}$ and $V \in \mathbb{C}^{n,n}$,*

$$\|UAV\|_F = \|A\|_F.$$

Hence

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\operatorname{tr}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)}.$$

Theorem 5.22. *The Frobenius norm satisfies:*

1. *It is a matrix norm, i.e. nonnegativity, positive homogeneity, and triangle inequality hold.*
2. *It is consistent with the Euclidean norm on vectors.*
3. *It is submultiplicative (2.16).*

Corollary 5.23. *For any $A \in \mathbb{F}^{m,n}$,*

$$\|A\|_2 \le \|A\|_F \le \sqrt{r}\,\|A\|_2,$$

where $r = \min(m, n)$. Hence the spectral norm $\|\cdot\|_2$ and Frobenius norm $\|\cdot\|_F$ are equivalent.

Theorem 5.24. ***Schmidt–Mirsky–Eckart–Young*** *For both the induced 2-norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$, the truncated singular value decomposition (SVD) gives the best rank-k approximation:*

$$\|A - A_k\| \le \|A - X\|, \quad \text{for all rank-}k \text{ matrices } X.$$

Moreover, the approximation error is given by

$$\|A - A_k\| = \begin{cases} \sigma_{k+1}, & \text{for the induced } \|\cdot\|_2 \text{ norm,} \\ \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}, & \text{for the } \|\cdot\|_F \text{ norm.} \end{cases}$$

Exercises

Exercise 1. Show that for given matrices A, B, C of appropriate sizes, it holds that

$$\min_{X \in \mathbb{C}^{m,n}} \left\| \begin{bmatrix} A & B \\ C & X \end{bmatrix} \right\|_2 = \max\left\{ \left\| \begin{bmatrix} A & B \end{bmatrix} \right\|_2, \left\| \begin{bmatrix} A \\ C \end{bmatrix} \right\|_2 \right\}.$$

Find a minimizer.

Proof. Let

$$M(X) := \begin{bmatrix} A & B \\ C & X \end{bmatrix}, \qquad M_1 := \begin{bmatrix} A & B \end{bmatrix}, \qquad M_2 := \begin{bmatrix} A \\ C \end{bmatrix}.$$

It is trivial by 2.10 that

$$\|M(X)\|_2 \geq \max\{\|M_1\|_2, \|M_2\|_2\}.$$

I don't know how to show it but it seems that

$$X^* := CA^\dagger B$$

is a minimizer. ⊣

Exercise 5. 5.10

Chapter 6

Angles and Metrics between Subspace

Definition 6.1. The *Riemann sphere* is defined as $\mathbb{C} \cup \{\infty\}$ satisfying the following arithmetic operations:

1. $z + \infty = \infty$ for any $z \in \mathbb{C}$, and $\infty + \infty = \infty$.
2. $z \times \infty = \infty$ for any $z \neq 0 \in \mathbb{C}$, and $\infty \times \infty = \infty$.
3. $z/0 = \infty$ for any $z \neq 0 \in \mathbb{C}$, and $\infty/0 = \infty$.
4. $z/\infty = 0$ for any $z \in \mathbb{C}$.

One possible realization is to treat the complex plane as a tangent plane to the corresponding Riemann sphere at its bottom, given by the equation

$$x^2 + y^2 + \left(z - \tfrac{1}{2}\right)^2 = \left(\tfrac{1}{2}\right)^2,$$

which characterizes a unit diameter sphere S^2 in $\mathbb{R}^3$. Let $N = (0, 0, 1)$ be the north pole of the sphere.

Definition 6.2. The stereographic projection mapping $\rho : \mathbb{C} \to S^2 \setminus N$ is defined as

$$\rho(z) = \frac{(\operatorname{Re} z, \operatorname{Im} z, |z|^2)}{1 + |z|^2}.$$

Definition 6.3. The *chordal metric* $d : S^2 \times S^2 \to \mathbb{R}$ is defined as

$$d(\rho(z), \rho(w)) = \frac{2|z - w|}{\sqrt{1 + |z|^2}\,\sqrt{1 + |w|^2}}.$$

It measures the length of the chord (arc) connecting the stereographic projections of z, w onto the Riemann sphere.

Chapter 8

Special Topics

Definition 8.1. For a square matrix $X \in \mathbb{C}^{n,n}$, the *matrix exponential* is defined by the power series

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!} = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots,$$

where $X^0 = I$.

Theorem 8.2. *For any square matrix X and invertible Y of the same dimension:*

1. $e^0 = I$.
2. $\exp(X^\top) = (\exp X)^\top$ *and* $\exp(\overline{X}) = \overline{\exp X}$.
3. $e^{YXY^{-1}} = Ye^XY^{-1}$.
4. *If* $XY = YX$, *then* $e^Xe^Y = e^{X+Y}$.
5. *For scalars* $a, b \in \mathbb{C}$, $e^{aX}e^{bX} = e^{(a+b)X}$.
6. $e^Xe^{-X} = I$.

The matrix exponential naturally arises in systems of linear differential equations. For the scalar equation $y' = ay$, the solution is $y(t) = e^{at}y(0)$. For a system $y' = Ay$ with $A \in \mathbb{C}^{n,n}$, the solution is

$$y(t) = e^{At}y(0).$$

If A has n independent eigenvectors, then $A = V\Lambda V^{-1}$ for $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, and

$$e^{At} = Ve^{\Lambda t}V^{-1} = V \begin{bmatrix} e^{\lambda_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n t} \end{bmatrix} V^{-1}.$$

Thus the solution to $y' = Ay$ is

$$y(t) = e^{At}y(0) = c_1e^{\lambda_1 t}x_1 + \cdots + c_ne^{\lambda_n t}x_n,$$

where x_i are eigenvectors of A.

Definition 8.3. Let

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{F}^{(n_1+n_2),(m_1+m_2)}.$$

If A_{ij} is invertible, then the Schur complement of A_{ij} in A is denoted A/A_{ij}. In particular,

$$\begin{aligned} A/A_{11} &= A_{22} - A_{21}A_{11}^{-1}A_{12}, \\ A/A_{12} &= A_{21} - A_{22}A_{12}^{-1}A_{11}, \\ A/A_{21} &= A_{12} - A_{11}A_{21}^{-1}A_{22}, \\ A/A_{22} &= A_{11} - A_{12}A_{22}^{-1}A_{21}. \end{aligned}$$

Application to linear equations. Consider

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Assuming A is invertible, eliminate x via $x = A^{-1}(u - By)$ to obtain

$$(D - CA^{-1}B)y = v - CA^{-1}u.$$

Theorem 8.4. *Assume the required inverses exist. Then:*

1. $\det(A) = \det(A_{11})\det(A/A_{11})$.
2. $\det(A) = \det(A_{22})\det(A/A_{22})$.
3. $$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}(A/A_{11})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A/A_{11})^{-1} \\ -(A/A_{11})^{-1}A_{21}A_{11}^{-1} & (A/A_{11})^{-1} \end{bmatrix}.$$
4. $$A^{-1} = \begin{bmatrix} (A/A_{22})^{-1} & -(A/A_{22})^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}(A/A_{22})^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}(A/A_{22})^{-1}A_{12}A_{22}^{-1} \end{bmatrix}.$$
5. $$A^{-1} = \begin{bmatrix} (A/A_{22})^{-1} & (A/A_{12})^{-1} \\ (A/A_{21})^{-1} & (A/A_{11})^{-1} \end{bmatrix}.$$

Definition 8.5. Let $A \in \mathbb{F}^{m,n}$ and $B \in \mathbb{F}^{p,q}$. The *Kronecker product* of A and B is

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{F}^{mp,nq}.$$

Theorem 8.6. *Let $A \in \mathbb{F}^{m,n}$, $B \in \mathbb{F}^{p,q}$, $C \in \mathbb{F}^{n,\ell}$, and $D \in \mathbb{F}^{q,r}$.*

1. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.
2. *If A, B are invertible,* $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
3. *If A, B are square,* $\sigma(A \otimes B) = \sigma(A)\sigma(B)$.
4. *If A, B are positive semidefinite (or definite), so is $A \otimes B$.*
5. *If A, B are unitary, so is $A \otimes B$.*

Theorem 8.7. *For $A \in \mathbb{F}^{p,m}$, $X \in \mathbb{F}^{m,n}$, and $B \in \mathbb{F}^{n,q}$,*

$$\operatorname{vec}(AXB) = (B^{\mathsf{T}} \otimes A) \operatorname{vec}(X).$$

Definition 8.8. For square matrices $A \in \mathbb{F}^{a,a}$ and $B \in \mathbb{F}^{b,b}$, define *Kronecker sum* to be

$$A \oplus B = A \otimes I_b + I_a \otimes B.$$

Definition 8.9. For $A, B \in \mathbb{F}^{m,n}$, the *Hadamard product* is the entrywise product

$$A \circ B = [a_{ij} b_{ij}] \in \mathbb{F}^{m,n}.$$

Theorem 8.10. ***Schur Product Theorem*** *If A and B are positive semidefinite, then $A \circ B$ is also positive semidefinite.*

Theorem 8.11. *For matrices A, B, C of the same size and scalar k,*

$$\begin{aligned} A \circ B &= B \circ A, \\ A \circ (B \circ C) &= (A \circ B) \circ C, \\ A \circ (B + C) &= A \circ B + A \circ C, \\ (kA) \circ B &= A \circ (kB) = k(A \circ B), \\ A \circ 0 &= 0 \circ A = 0. \end{aligned}$$

Theorem 8.12. *If A, C have the same dimensions and B, D likewise, then*

$$(A \otimes B) \circ (C \otimes D) = (A \circ C) \otimes (B \circ D).$$

Appendix A

Homework

A.1

1. 1.3

2. 1.23

3. 1.10

4. Let $\mathbb{C}$ be equipped with the Euclidean metric and with the chordal metric (6.3). If $\{z_k\}$ is a sequence in $\mathbb{C}$ that is Cauchy (1.33) with respect to one of these metrics, is it necessarily Cauchy with respect to the other?

(1) If $\{z_k\}$ is Cauchy with respect to the Euclidean metric, then for each $\epsilon > 0$ there is a $K > 0$ such that

$$d(\rho(z_m), \rho(z_n)) = \frac{2|z_m - z_n|}{\sqrt{1 + |z_m|^2}\sqrt{1 + |z_n|^2}} \leq 2|z_m - z_n| \leq 2 \cdot \frac{1}{2}\epsilon = \epsilon$$

for all $m, n > K$. That is, $\{z_k\}$ is Cauchy with respect to the chordal metric.

(2) The other direction does not hold, considering a counterexample $\{z_k\}$ where $z_i = i \in \mathbb{N}$.

A.2

1. 2.3

2. Show that 2.9 and 2.10 are indeed norms for $p = 1, 2, \infty$.

We prove the triangle inequality for 2.9. Note that $\|A+B\|_{p,E} \le \|A\|_{p,E} + \|B\|_{p,E}$ iff $\|\operatorname{vec}(A+B)\|_p \le \|\operatorname{vec} A\|_p + \|\operatorname{vec} B\|_p$ iff $\|\operatorname{vec} A + \operatorname{vec} B\|_p \le \|\operatorname{vec} A\|_p + \|\operatorname{vec} B\|_p$, but $\|\cdot\|_p$ is a norm, so we are done.

For the triangle inequality for 2.10, again by that $\|\cdot\|_p$ is a norm we have that

$$\begin{aligned}
\|A+B\|_p &= \sup_{\|x\|_p=1} \|(A+B)x\|_p = \sup_{\|x\|_p=1} \|Ax+Bx\|_p \\
&\le \sup_{\|x\|_p=1} (\|Ax\|_p + \|Bx\|_p) \le \sup_{\|x\|_p=1} \|Ax\|_p + \sup_{\|x\|_p=1} \|Bx\|_p \\
&= \|A\|_p + \|B\|_p
\end{aligned}$$

> **3.** (This is the real homework. The preceding one was mistaken.) Verify whether 2.9 and 2.10 are submultiplicative for $p = 1, 2, \infty$.

Let $A \in \mathbb{F}^{m,r}$, $B \in \mathbb{F}^{r,n}$.

For 2.9, if p=1:

$$\|AB\|_{1,E} = \sum_{i,j} \Big| \sum_k a_{ik} b_{kj} \Big| \le \sum_{i,j,k} |a_{ik}||b_{kj}| = \sum_k \Big(\sum_i |a_{ik}| \Big) \Big(\sum_j |b_{kj}| \Big)$$

Since all terms are nonnegative,

$$\sum_k \Big(\sum_i |a_{ik}| \Big) \Big(\sum_j |b_{kj}| \Big) \le \Big(\sum_{i,k} |a_{ik}| \Big) \Big(\sum_{k,j} |b_{kj}| \Big) = \|A\|_{1,E} \, \|B\|_{1,E}.$$

Hence $\|AB\|_{1,E} \le \|A\|_{1,E} \|B\|_{1,E}$.

If p=2:

$$\|AB\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n \Big| \sum_{k=1}^r a_{ik} b_{kj} \Big|^2.$$

For each fixed pair (i,j), apply the Cauchy–Schwarz inequality to the vectors $\big(a_{ik}\big)_{k=1}^r$ and $\big(b_{kj}\big)_{k=1}^r$:

$$\Big| \sum_{k=1}^r a_{ik} b_{kj} \Big|^2 \le \Big(\sum_{k=1}^r |a_{ik}|^2 \Big) \Big(\sum_{k=1}^r |b_{kj}|^2 \Big).$$

Summing over i, j yields

$$\|AB\|_F^2 \le \sum_{i,j} \Big(\sum_k |a_{ik}|^2 \Big) \Big(\sum_k |b_{kj}|^2 \Big) = \Big(\sum_{i,k} |a_{ik}|^2 \Big) \Big(\sum_{k,j} |b_{kj}|^2 \Big) = \|A\|_F^2 \, \|B\|_F^2.$$

Taking square roots gives $\|AB\|_F \le \|A\|_F \, \|B\|_F$.

When $p = \infty$, the submultiplicativity does not hold for 2.9. Consider

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Then $\|A\|_{\infty,E} = \|B\|_{\infty,E} = 1$. But

$$AB = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

so

$$\|AB\|_{\infty,E} = 2 > 1 \cdot 1.$$

For 2.10,

$$\begin{aligned} \|AB\|_p &= \sup_{x \neq 0} \frac{\|ABx\|_p}{\|x\|_p} = \sup_{x \neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \frac{\|Bx\|_p}{\|x\|_p} \\ &\leq \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \sup_{x \neq 0} \frac{\|Bx\|_p}{\|x\|_p} = \|A\|_p \|B\|_p. \end{aligned}$$

4. 2.10

A.3

1. 3.5

2. 3.11

3. 3.14

A.4

1. What additional condition on a nonsingular M-matrix M ensures that $M^{-1} \triangleright 0$? Moreover, if $M = \varepsilon I + L$, where L is a graph Laplacian, what corresponding condition on the underlying graph ensures that $M^{-1} \triangleright 0$ for all $\varepsilon > 0$?

Write $M = sI - B$, where $B \trianglerighteq 0$ and $s > \rho(B)$. Then

$$M^{-1} = (sI - B)^{-1} = \frac{1}{s}\sum_{k=0}^{\infty}\left(\frac{B}{s}\right)^k.$$

Note that the (i, j)-th entry of B^k is positive if and only if there exists a directed walk of length k from vertex j to i in the digraph associated with B. Therefore, $M^{-1} \triangleright 0$ holds if and only if for all i, j there exists k with $(B^k)_{ij} > 0$, which is equivalent to the digraph of B being *strongly connected*. Hence:

$$M^{-1} \triangleright 0 \quad \iff \quad M \text{ is irreducible.}$$

For $M = \varepsilon I + L$ with $\varepsilon > 0$, where L is a graph Laplacian, write

$$M = D - A, \quad D = \operatorname{diag}(\varepsilon + d_i), \quad A = (w_{ij})_{i \neq j}.$$

Then $M = D(I - P)$ with $P = D^{-1}A \trianglerighteq 0$ satisfying $\rho(P) < 1$, and hence

$$M^{-1} = (I - P)^{-1}D^{-1} = \left(\sum_{k=0}^{\infty} P^k\right)D^{-1}.$$

The matrix P is irreducible if and only if the underlying graph is strongly connected. In that case, $\sum_{k\geq 0} P^k \triangleright 0$, yielding $M^{-1} \triangleright 0$. Conversely, if the graph is not strongly connected, M is reducible and some entries of M^{-1} must vanish. Thus the condition on the underlying graph is that it has to be strongly connected.

2. If $x(0) \trianglerighteq 0$ and $\dot{x}(t) = -Mx(t)$, where M is a given M-matrix, show that $x(t) \trianglerighteq 0$, for all $t \geq 0$.

Proof. Note that $A := -M$ has *nonnegative* off-diagonal entries, i.e. A is a Metzler matrix. We claim that for any Metzler matrix A, $e^{At} \trianglerighteq 0$ for all $t \geq 0$. Indeed, choose $\alpha > 0$ large enough so that $B := A + \alpha I \trianglerighteq 0$. Then

$$e^{At} = e^{-\alpha t}e^{(A+\alpha I)t} = e^{-\alpha t}\sum_{k=0}^{\infty}\frac{(Bt)^k}{k!}.$$

Since $B \trianglerighteq 0$, every power B^k is entrywise nonnegative, so the series has entrywise nonnegative partial sums, and hence $e^{(A+\alpha I)t} \trianglerighteq 0$. Multiplying by the positive scalar $e^{-\alpha t}$ preserves entrywise nonnegativity, proving $e^{At} \trianglerighteq 0$.

The solution of $\dot{x}(t) = Ax(t)$ with $A = -M$ is $x(t) = e^{At}x(0)$. If $x(0) \trianglerighteq 0$, then by the claim $e^{At} \trianglerighteq 0$, so

$$x(t) = e^{At}x(0) \trianglerighteq 0 \qquad \text{for all } t \geq 0.$$ ⊣

3. 4.7

4. Given an arbitrary undirected connected graph, let D be its degree matrix (a diagonal matrix where each diagonal element is the degree of the corresponding vertex) and L be its Laplacian matrix. Then

$$\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

is called the normalized Laplacian matrix. Prove that the eigenvalues of $\hat{L}$ lie in the interval $[0, 2]$.

Proof. Set $M := D^{-1/2} A D^{-1/2}$. Write the eigenvalues of M as $\mu_1, \dots, \mu_n$ and those of $\widehat{L}$ as $\lambda_1, \dots, \lambda_n$. Since $\widehat{L} = I - M$, they satisfy

$$\lambda_i = 1 - \mu_i \qquad (i = 1, \dots, n).$$

Hence it suffices to prove $\mu_i \in [-1, 1]$.

For any nonzero $x \in \mathbb{R}^n$, let $y = D^{-1/2} x$. Then $x^\top x = y^\top D y$ and

$$x^\top M x \;=\; y^\top A y \;=\; \sum_{i,j} A_{ij} y_i y_j \;=\; 2 \sum_{\{i,j\} \in E} y_i y_j.$$

Using $2|ab| \le a^2 + b^2$ edgewise,

$$|y^\top A y| \;=\; \left| 2 \sum_{\{i,j\} \in E} y_i y_j \right| \;\le\; \sum_{\{i,j\} \in E} (y_i^2 + y_j^2) \;=\; \sum_i d_i y_i^2 \;=\; y^\top D y.$$

Therefore the Rayleigh quotient of M obeys

$$\left| \frac{x^\top M x}{x^\top x} \right| = \left| \frac{y^\top A y}{y^\top D y} \right| \le 1.$$

Since M is symmetric, all its eigenvalues lie between the minimum and maximum Rayleigh quotients. Hence $\mu_i \in [-1, 1]$ for all i, as desired. ⊣

A.5

1. 5.1

2. 5.10

3. We do not assume A to be Hermitian and define

$$F(A) = \left\{ \frac{x^* A x}{x^* x} \; : \; x \in \mathbb{C}^n,\; x \neq 0 \right\}.$$

1. Show that

   (a)
   $$\min_{x \neq 0} \left| \frac{x^* A x}{x^* x} \right| \;\leq\; |\lambda_i| \;\leq\; \max_{x \neq 0} \left| \frac{x^* A x}{x^* x} \right|;$$

   (b)
   $$\sigma(A+B) \subseteq F(A+B) \subseteq F(A) + F(B),$$

   where $\sigma(\cdot)$ denotes the spectrum;

   (c) $F(A)$ is a convex subset of $\mathbb{C}$.

2. Compute
$$F\left(\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\right), \quad F\left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\right), \quad \text{and} \quad F\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right).$$

3. Is it true that
$$F(AB) \subseteq F(A)F(B)\,?$$
   Prove it or give a counterexample.

Note that

$$F(A) = \{\langle Ax, x\rangle \; : \; \|x\| = 1\}, \qquad \langle x, y\rangle := x^* y.$$

1. (a) For any eigenvalue λ_i of A choose a corresponding eigenvector $0 \neq v_i$ with $\|v_i\| = 1$. Then
$$\frac{\langle Av_i, v_i\rangle}{\langle v_i, v_i\rangle} = \langle Av_i, v_i\rangle = \langle \lambda_i v_i, v_i\rangle = \lambda_i.$$
   Hence $\lambda_i \in F(A)$. Since $F(A)$ is compact (continuous image of the compact unit sphere), the continuous map $z \mapsto |z|$ attains its minimum and maximum on $F(A)$; therefore
$$\min_{\|x\|=1} \big|\langle Ax, x\rangle\big| \;\leq\; |\lambda_i| \;\leq\; \max_{\|x\|=1} \big|\langle Ax, x\rangle\big|.$$

   (b) In finite dimensions the spectrum $\sigma(\cdot)$ consists of eigenvalues, and by the argument above $\sigma(A+B) \subseteq F(A+B)$. For the second inclusion, for any $\|x\| = 1$,
$$\langle (A+B)x, x\rangle = \langle Ax, x\rangle + \langle Bx, x\rangle \in F(A) + F(B),$$
   whence $F(A+B) \subseteq F(A) + F(B)$.

(c) (Toeplitz–Hausdorff) For $\theta \in [0, 2\pi)$ set

$$H_\theta := \mathrm{Re}\left(e^{-i\theta} A\right) = \frac{e^{-i\theta} A + e^{i\theta} A^*}{2},$$

which is Hermitian. For $\|x\| = 1$,

$$\mathrm{Re}\left(e^{-i\theta} \langle Ax, x\rangle\right) = \langle H_\theta x, x\rangle \leq \lambda_{\max}(H_\theta).$$

Hence

$$F(A) \subseteq \bigcap_{\theta \in [0,2\pi)} \left\{ z \in \mathbb{C} : \mathrm{Re}\left(e^{-i\theta} z\right) \leq \lambda_{\max}(H_\theta) \right\}.$$

Claim: The reverse inclusion holds and the display is an equality. The right-hand side is an intersection of half-planes, therefore convex. Thus $F(A)$ is convex. to do

2. (a) A is hermitian, hence

$$F(A) = [\lambda_{\min}(A), \lambda_{\max}(A)].$$

The eigenvalues solve $\det\left(\begin{smallmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{smallmatrix}\right) = \lambda^2 - \lambda - 1 = 0$, i.e. $\lambda_\pm = \frac{1 \pm \sqrt{5}}{2}$. Therefore

$$F\left(\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\right) = \left[\frac{1-\sqrt{5}}{2}, \frac{1+\sqrt{5}}{2}\right].$$

(b) Write $A = I + N$ where $N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. For $x = (e^{i\alpha} \cos t,\ e^{i\beta} \sin t)$,

$$\langle Ax, x\rangle = 1 + \langle Nx, x\rangle = 1 + e^{i(\beta-\alpha)} \cos t\ \sin t = 1 + \frac{1}{2} e^{i(\beta-\alpha)} \sin(2t).$$

As t varies, $\frac{1}{2}|\sin(2t)|$ ranges over $[0, \frac{1}{2}]$, and the phase $e^{i(\beta-\alpha)}$ is arbitrary; hence

$$F\left(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right) = \{ z \in \mathbb{C} : |z - 1| \leq \tfrac{1}{2} \},$$

the closed disk centered at 1 of radius $1/2$.

(c) A is Hermitian with eigenvalues 2 and 0. Thus

$$F\left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = [0, 2].$$

3. Take

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad AB = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Since A and B are Hermitian, their numerical ranges are real intervals

$$F(A) = \operatorname{conv} \sigma(A) = [-1, 1], \qquad F(B) = \operatorname{conv} \sigma(B) = [-1, 1],$$

hence

$$F(A)F(B) = \{st \ : \ \ s, t \in [-1, 1]\} = [-1, 1] \subseteq \mathbb{R}.$$

However AB is skew-Hermitian: $(AB)^* = -AB$. Take the unit vector $x = \frac{1}{\sqrt{2}}\binom{1}{i}$. Then

$$\langle ABx, x\rangle = x^* ABx = \frac{1}{2}\begin{pmatrix} 1 & -i \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ i \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & -i \end{pmatrix}\begin{pmatrix} i \\ -1 \end{pmatrix} = i.$$

Thus $i \in F(AB)$, which is not real. Consequently

$$i \notin F(A)F(B) = [-1, 1].$$

Therefore the inclusion $F(AB) \subseteq F(A)F(B)$ is *false* in general.