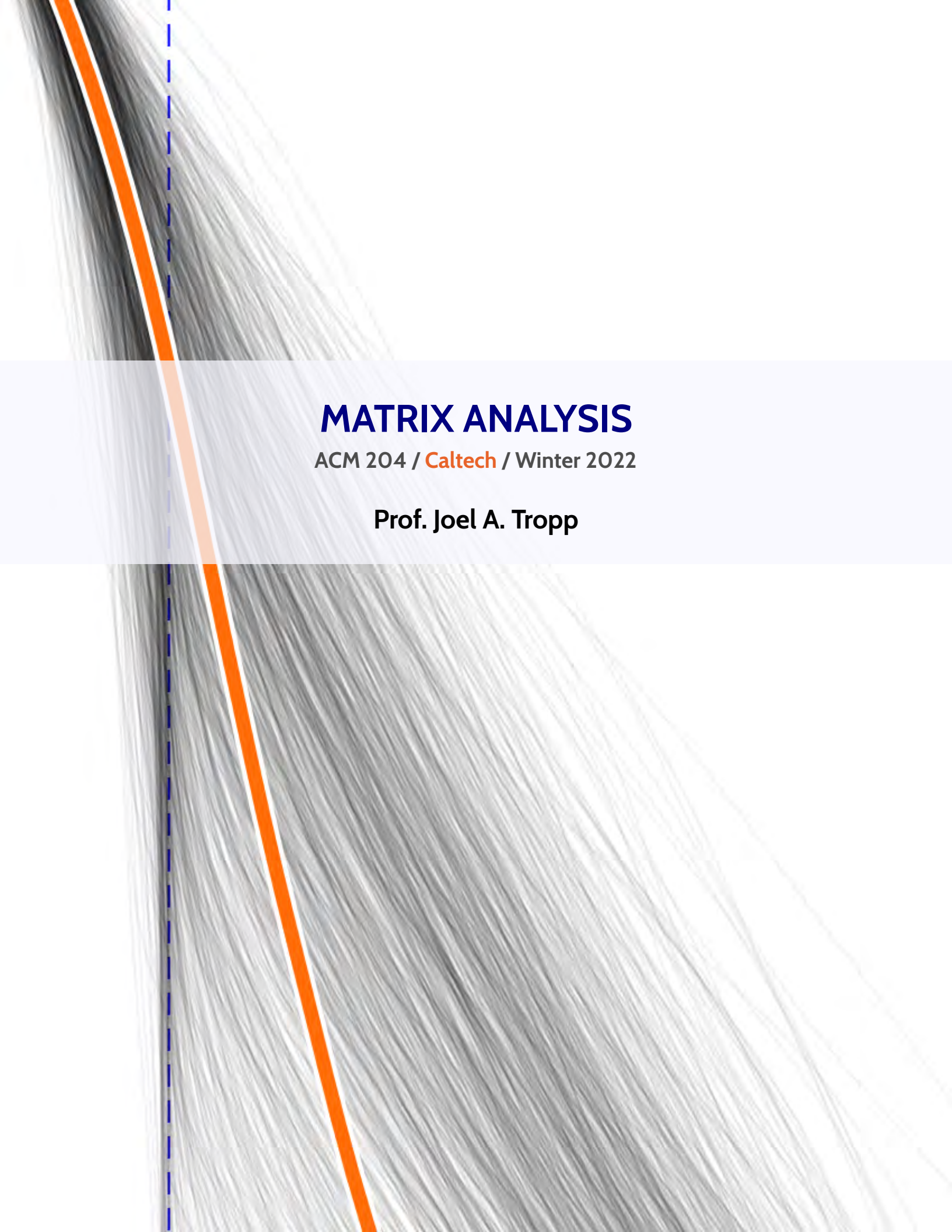

MATRIX ANALYSIS

ACM 204 / Caltech / Winter 2022

Prof. Joel A. Tropp

Typeset on August 22, 2022

Cite as:
Joel A. Tropp, *ACM 204: Matrix Analysis*, Caltech CMS Lecture Notes 2022-01, Pasadena, Winter 2022. doi:10.7907/nwsv-df59.

Cover image: Sample paths of a randomized block Krylov method for estimating the largest eigenvalue of a symmetric matrix.

Lecture images: Falling text in the style of *The Matrix* was created by Jamie Zawinski, ©1999–2003.

Contents

Preface

"The Matrix is everywhere. It is all around us. Even now in this very room."

—Morpheus, *The Matrix*, 1999

Matrices are a foundational tool in the mathematical sciences, in statistics, in engineering, and in computer science. The purpose of this course is to develop a deeper understanding of matrices, their structure, and function using tools from linear algebra, convexity theory, and analysis.

Course overview

The topics of this course vary from term to term, depending on the audience. This term, we covered the following material:

- Basics of multilinear algebra
- Majorization and doubly stochastic matrices
- Symmetric and unitarily invariant norms
- Uniform smoothness of matrix spaces
- Complex interpolation methods for matrix inequalities
- Variational principles for eigenvalues
- Perturbation theory for eigenvalues
- Angles between subspaces and perturbation theory for eigenspaces
- Tensor products and matrix equations
- Positive and completely positive linear maps
- Matrix monotonicity and convexity
- Differentiation of standard matrix functions
- Loewner's theorems on matrix monotone functions
- Matrix means
- Convexity of matrix trace functions
- Positive-definite functions and Bochner's theorem
- Entrywise positivity preservers and Schoenberg's theorem

The course had three optional problem sets, which helped to cement some of the foundational material. The problem sets are attached to the notes.

The primary assignment was a project, where each student read some classic or modern papers in matrix analysis and wrote a synthetic treatment of the material. A selection of the projects is attached to the lecture notes.

Prerequisites

ACM 204 is designed for G2 and G3 students in the mathematical sciences. The prerequisites for this course are differential and integral calculus (e.g., Caltech Ma 1ac), ordinary differential equations (e.g., Ma 2), and intermediate linear algebra (e.g., Ma 1b and ACM 104). Exposure to linear analysis (e.g., CMS 107), functional analysis (e.g., ACM 105), and optimization theory (e.g., CMS 122) is also valuable.

Supplemental textbooks

There is no required textbook for the course. Some recent books that cover related material include

- [Bha97] Bhatia, *Matrix Analysis*, Springer, 1997.
- [Bha07a] Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, SIAM, 2007.
- [Bha07b] Bhatia, *Positive-Definite Matrices*, Princeton, 2007.
- [Car10] Carlen, *Trace Inequalities and Quantum Entropy*, 2010.
- [Hia10] Hiai, *Matrix Analysis: Matrix Monotone Functions, Matrix Means, and Majorization*, 2010.
- [HP14] Hiai & Petz, *Introduction to Matrix Analysis and Applications*, Springer, 2014.
- [Hig08] Higham, *Functions of Matrices*, SIAM, 2008.
- [HJ13] Horn & Johnson, *Matrix Analysis*, 2nd ed., Cambridge, 2013.
- [HJ94] Horn & Johnson, *Topics in Matrix Analysis*, Cambridge, 1994.
- [Kat95] Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, 1995.
- [MOA11] Marshall et al., *Inequalities: Theory of Majorization and Its Applications*, 2nd ed., Springer 2011.

Bhatia's books [Bha97; Bha07b] are the primary sources for this course.

These notes

These lecture notes document ACM 204 as taught in Winter 2022, and they are primarily intended as a reference for students who have taken the class. The notes are prepared by student scribes with feedback from the instructor. The notes have been edited by the instructor to try to correct his own failures of presentation. *All remaining errors and omissions are the fault of the instructor.*

Please be aware that these notes reflect material presented in a classroom, rather than a formal scholarly publication. In some places, *the notes may lack appropriate citations* to the literature. There is no claim that the arrangement or presentation of the material is primarily due to the instructor.

The notes also contain the projects of students who wished to share their work. They received feedback and made revisions, but the projects have not been edited. They represent the students' individual work.

Acknowledgements

These notes were transcribed by students taking the course in Winter 2022. They are Eray Atay, Jag Boddapati, Edoardo Calvello, Ruizhi Cao, Anthony (Chi-Fang) Chen, Nicolas Christianson, Matthieu Darcy, Rohit Dilip, Ethan Epperly, Salvador Gomez, Taylan Kargin, Eitan Levin, Elvira Moreno, Nicholas Nelson, Roy (Yixuan) Wang, and Jing Yu. Without their care and attention, we would not have such an excellent record.

Joel A. Tropp
Steele Family Professor of Applied & Computational Mathematics
California Institute of Technology

jtropp@cms.caltech.edu
http://users.cms.caltech.edu/~jtropp

Pasadena, California
March 2022

Notation

I have selected notation that is common in the linear algebra and probability literature. I have tried to been consistent in using the symbols that are presented below. There are some minor variations in different lectures, including the letter that indicates the dimension of a matrix and the indexing of sums. *Scribes are expected to use this same notation!*

Set theory

The Pascal notation $:=$ and $=:$ generates a definition. Sets without any particular internal structure are denoted with sans serif capitals: $\mathsf{A}, \mathsf{B}, \mathsf{E}$. Collections of sets are written in a calligraphic font: $\mathscr{A}, \mathscr{B}, \mathscr{F}$. The power set (that is, the collection containing all subsets) of a set E is written as $\mathscr{P}(\mathsf{E})$.

The symbol $\emptyset$ is reserved for the empty set. We use braces to denote a set. The character $\in$ (or, rarely, $\ni$) is the member-of relation. The set-builder notation

$$\{x \in \mathsf{A} : P(x)\}$$

carves out the (unique) set of elements that belong to a set A and that satisfy the predicate P. Basic set operations include union ($\cup$), intersection ($\cap$), symmetric difference ($\triangle$), set difference ($\setminus$), and the complement ($^{\mathsf{c}}$) with respect to a fixed set. The relations $\subseteq$ and $\supseteq$ indicate set containment.

The natural numbers $\mathbb{N} := \{1, 2, 3, \dots\}$. Ordered tuples and sequences are written with parentheses, e.g.,

$$(a_1, a_2, a_3, \dots, a_n) \quad \text{or} \quad (a_1, a_2, a_3, \dots)$$

Alternative notations include things like $(a_i : i \in \mathbb{N})$ or $(a_i)_{i \in \mathbb{N}}$ or simply (a_i).

Real analysis

We mainly work in the field $\mathbb{R}$ of real numbers, equipped with the absolute value $|\cdot|$. The extended real numbers $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ are defined with the usual rules of arithmetic and order. In particular, we instate the conventions that $0/0 = 0$ and $0 \cdot \pm\infty = 0$. We use the standard (American) notation for open and closed intervals; e.g.,

$$(a, b) := \{x \in \overline{\mathbb{R}} : a < x < b\} \quad \text{and} \quad [a, b] := \{x \in \overline{\mathbb{R}} : a \le x \le b\}.$$

Occasionally, we may visit the rational field $\mathbb{Q}$, and we very commonly use the complex field $\mathbb{C}$. The imaginary unit, i, is written in an upright font.

We use modern conventions for words describing order; these may be slightly different from what you are used to. In this course, we enforce the definition that *positive* means ≥ 0 and *negative* means ≤ 0. For example, the positive integers compose the set $\mathbb{Z}_+ := \{0, 1, 2, 3, \dots\}$ and the positive reals compose the set $\mathbb{R}_+ := \{x \in \mathbb{R} : x \ge 0\}$. When required, we may deploy the phrase *strictly positive* to mean > 0 and *strictly negative* to mean < 0. Similarly, *increasing* means "never going down" and *decreasing* means "never going up."

Warning: Positive means ≥ 0! ■

Linear algebra

We work in a real or complex linear space. The letters d and n (and occasionally others) are used to denote the dimension of this space, which is always finite. For example, we write $\mathbb{R}^d$ or $\mathbb{C}^n$. We may write $\mathbb{F}$ to refer to either field, or we may omit the field entirely if it is not important.

We use the delta notation for standard basis vectors: $\boldsymbol{\delta}_i$ has a one in the ith coordinate and zeros elsewhere. The vector $\mathbf{1}$ has ones in each entry. The dimension of these vectors is determined by context.

The symbol * denotes the (conjugate) transpose of a vector or a matrix. In particular, z^* is the complex conjugate of a complex number $z \in \mathbb{C}$. We may also write $^\mathsf{T}$ for the ordinary transpose to emphasize that no conjugation is performed.

We equip $\mathbb{F}^d$ with the standard inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^* \boldsymbol{y}$. The inner product generates the Euclidean norm $\|\boldsymbol{x}\|^2 := \langle \boldsymbol{x}, \boldsymbol{x} \rangle$.

We write $\mathbb{H}_d(\mathbb{F})$ for the real-linear space of $d \times d$ self-adjoint matrices with entries in the field $\mathbb{F}$. Recall that a matrix is self-adjoint when $\boldsymbol{A} = \boldsymbol{A}^*$. The symbols $\mathbf{0}$ and $\mathbf{I}$ denote zero matrix and the identity matrix; their dimensions are determined by context or by an explicit subscript.

We equip the space $\mathbb{H}_d$ with the trace inner product $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle := \operatorname{tr}(\boldsymbol{XY})$, which generates the Frobenius norm $\|\boldsymbol{X}\|_{\mathrm{F}}^2 := \langle \boldsymbol{X}, \boldsymbol{X} \rangle$. The map $\operatorname{tr}(\cdot)$ returns the trace of a square matrix; the parentheses are often omitted. We instate the convention that nonlinear functions bind before the trace.

The spectral theorem states that every self-adjoint matrix $\boldsymbol{A} \in \mathbb{H}_n$ admits a *spectral resolution*:

$$\boldsymbol{A} = \sum\nolimits_{i=1}^{m} \lambda_i \boldsymbol{P}_i \quad \text{where} \quad \sum\nolimits_{i=1}^{m} \boldsymbol{P}_i = \mathbf{I}_n \quad \text{and} \quad \boldsymbol{P}_i \boldsymbol{P}_j = \delta_{ij} \boldsymbol{P}_i.$$

Here, $\lambda_1, \dots, \lambda_m$ are the distinct (real) eigenvalues of $\boldsymbol{A}$. The range of the orthogonal projector $\boldsymbol{P}_i$ is the invariant subspace associated with λ_i. In this context, δ_{ij} is the Kronecker delta.

The maps $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ return the minimum and maximum eigenvalues of a self-adjoint matrix. The ℓ_2 operator norm $\|\cdot\|$ of a self-adjoint matrix satisfies the relation

$$\|\boldsymbol{A}\| := \max\left\{ |\lambda_{\max}(\boldsymbol{A})|, |\lambda_{\min}(\boldsymbol{A})| \right\} \quad \text{for } \boldsymbol{A} \in \mathbb{H}_n.$$

A self-adjoint matrix is *positive semidefinite (psd)* if its eigenvalues are nonnegative; a self-adjoint matrix is *positive definite (pd)* if its eigenvalues are positive. The symbol $\preccurlyeq$ refers to the psd order: $\boldsymbol{A} \preccurlyeq \boldsymbol{H}$ if and only if $\boldsymbol{H} - \boldsymbol{A}$ is psd.

We can define a *standard matrix function* for a self-adjoint matrix using the spectral resolution. For an interval $\mathsf{I} \subseteq \mathbb{R}$ and for a function $f : \mathsf{I} \to \mathbb{R}$,

$$\boldsymbol{A} = \sum\nolimits_{i=1}^{m} \lambda_i \, \boldsymbol{P}_i \quad \text{implies} \quad f(\boldsymbol{A}) = \sum\nolimits_{i=1}^{m} f(\lambda_i) \, \boldsymbol{P}_i.$$

Implicitly, we assume that the eigenvalues of the matrix $\boldsymbol{A}$ lie within the domain I of the function f. When we apply a real function to a self-adjoint matrix, we are always referring to the associated standard matrix function. In particular, we often encounter powers, exponentials, and logarithms.

We write $\mathbb{M}_n(\mathbb{F})$ for the linear space of $n \times n$ matrices over the field $\mathbb{F}$. We also define the linear space $\mathbb{M}^{m \times n}(\mathbb{F})$ of $m \times n$ matrices over the field $\mathbb{F}$. We can extend the trace inner-product and Frobenius norm to this setting:

$$\langle \boldsymbol{B}, \boldsymbol{C} \rangle := \operatorname{tr}(\boldsymbol{B}^* \boldsymbol{C}) \quad \text{and} \quad \|\boldsymbol{B}\|_{\mathrm{F}}^2 := \langle \boldsymbol{B}, \boldsymbol{B} \rangle \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}^{m \times n}.$$

A square matrix $\boldsymbol{Q} \in \mathbb{M}_n$ that satisfies $\boldsymbol{Q}^*\boldsymbol{Q} = \mathbf{I}_n$ is called *orthogonal* (resp. *unitary*) in the real (resp. complex) case. A tall, rectangular matrix $\boldsymbol{B} \in \mathbb{M}^{m\times n}$ with $n \leq m$ that satisfies $\boldsymbol{B}^*\boldsymbol{B} = \mathbf{I}_n$ is called *orthonormal*; this terminology is common in the numerical literature. More generally, a rectangular matrix $\boldsymbol{B} \in \mathbb{M}^{m\times n}$ is called a *partial isometry* if $\boldsymbol{B}^*\boldsymbol{B}$ is an orthogonal projector.

Every matrix $\boldsymbol{B} \in \mathbb{M}^{m\times n}(\mathbb{F})$ admits a singular value decomposition:

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \quad \text{where } \boldsymbol{U} \in \mathbb{F}^{m\times m} \text{ and } \boldsymbol{V} \in \mathbb{F}^{n\times n}.$$

The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal (or unitary). The rectangular matrix $\boldsymbol{\Sigma} \in \mathbb{F}^{m\times n}$ is diagonal, in the sense that the entries $(\boldsymbol{\Sigma})_{ij} = 0$ whenever $i \neq j$. The diagonal entries of $\boldsymbol{\Sigma}$ are called *singular values*. They are conventionally arranged in decreasing order and written with the following notation.

$$\sigma_{\max}(\boldsymbol{B}) \coloneqq \sigma_1(\boldsymbol{B}) \geq \sigma_2(\boldsymbol{B}) \geq \cdots \geq \sigma_r(\boldsymbol{B}) =: \sigma_{\min}(\boldsymbol{B}) \quad \text{where } r \coloneqq \min\{m, n\}.$$

The symbol $\|\cdot\|$ always refers to the ℓ_2 operator norm; it returns the maximum singular value of its argument.

We write $\operatorname{lin}(\cdot)$ for the linear hull of a family of vectors. The operators $\operatorname{range}(\cdot)$ and $\operatorname{null}(\cdot)$ extract the range and null space of a matrix. The operator † extracts the pseudoinverse.

Probability

The map $\mathbb{P}\{\cdot\}$ returns the probability of an event. The operator $\mathbb{E}[\cdot]$ returns the expectation of a random variable taking values in a linear space. We only include the brackets when it is necessary for clarity, and we impose the convention that nonlinear functions bind before the expectation.

The symbol $\sim$ means "has the distribution." We abbreviate (statistically) *independent and identically distributed (iid)*. Named distributions, such as NORMAL and UNIFORM, are written with small capitals.

We say that a random vector $\boldsymbol{x} \in \mathbb{F}^n$ is *centered* when $\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$. A random vector is *isotropic* when $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^*] = \mathbf{I}_n$. A random vector that is both centered and isotropic is *standardized*.

An important property of the standard normal distribution, which we use heavily, is the fact that it is rotationally invariant. If $\boldsymbol{x} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I})$, then $\boldsymbol{Q}\boldsymbol{x}$ is also standard normal for every matrix $\boldsymbol{Q}$ that is orthogonal (in the real case) or unitary (in the complex case).

Order notation

We sometimes use the familiar order notation from computer science. The symbol $\Theta(\cdot)$ refers to asymptotic equality. The symbol $O(\cdot)$ refers to an asymptotic upper bound.

I. lectures

1. Tensor Products

Date: 4 January 2022 **Scribe:** Rohit Dilip

In this lecture, we develop the theory of tensor products, which provides us with a way to "multiply" vector spaces. We begin by discussing the axioms that such a construction should satisfy; then we develop a rigorous way to implement these axioms. We show that the tensor product of two Hilbert spaces is itself a Hilbert space, equipped with an induced inner product. By developing an isomorphism with bilinear forms, we show how to regard tensor products as matrices. Finally, we show that we can construct linear operators on a tensor product space in a very natural way, and we develop their spectral theory in a transparent way.

Agenda:
1. Tensor product: Motivation
2. Tensor product spaces
3. Bilinear forms
4. Theory of linear operators

1.1 Tensor product: Motivation

We begin with some background and motivation.

1.1.1 Setting

Throughout this chapter, we will assume H is an n-dimensional Hilbert space over a field $\mathbb{F}$ (either $\mathbb{R}$ or $\mathbb{C}$). A Hilbert space is endowed with an inner product denoted by $\langle \cdot, \cdot \rangle$. By convention, we assume the inner product is *conjugate linear in the first coordinate* and linear in the second coordinate. We fix an orthonormal basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$. That is, $\langle \boldsymbol{e}_j, \boldsymbol{e}_k \rangle = \delta_{jk}$. Finally, we denote the space of linear operators acting on H by $\mathscr{L}(\mathsf{H})$. Since H is finite-dimensional, we can regard every element of $\mathscr{L}(\mathsf{H})$ is as a matrix with dimension $\dim(\mathsf{H})$.

The Kronecker delta $\delta_{jj} = 1$ for all j, while $\delta_{jk} = 0$ when $j \neq k$.

1.1.2 Axioms for the tensor product

Vectors spaces admit scaling and addition, but we do not usually talk about how to "multiply" two vectors together. Intuitively, this should be a fundamental binary operation that distributes over addition, and it should return an object living in a larger vector space. The latter condition can be understood via example: if we multiply two sides of a rectangle to find its area, we understand that the product lives in a different space because it has different units. This space is also larger in some sense than either of the constituent spaces (i.e., 2D instead of 1D).

In this section, we enumerate the axioms that a reasonable interpretation of vector "multiplication" might satisfy. We call this operation a *tensor product*.

Definition 1.1 (Tensor product: Axioms). The tensor product operation $\otimes$ maps a pair of vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$ to an object called the tensor product $\boldsymbol{x} \otimes \boldsymbol{y}$. We can add tensors and scale them. The product should satisfy the following properties:

1. **Additivity.** The product should distribute across the addition operation. For

all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathsf{H}$, we require the following two equalities to hold.

$$(\boldsymbol{x} + \boldsymbol{y}) \otimes \boldsymbol{z} = \boldsymbol{x} \otimes \boldsymbol{z} + \boldsymbol{y} \otimes \boldsymbol{z};$$
$$\boldsymbol{x} \otimes (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \otimes \boldsymbol{y} + \boldsymbol{x} \otimes \boldsymbol{z}.$$

2. **Homogeneity.** We would also like the product operation to behave well with scalar multiplication in the field $\mathbb{F}$ of the Hilbert space. In particular, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$,

$$\alpha(\boldsymbol{x} \otimes \boldsymbol{y}) = (\alpha \boldsymbol{x}) \otimes \boldsymbol{y} = \boldsymbol{x} \otimes (\alpha \boldsymbol{y}) \quad \text{for all } \alpha \in \mathbb{F}.$$

The homogeneity property is independent of the argument index in the product; that is, there is no conjugation of the field element α.

3. **Interaction with the zero vector.** The zero vector $\mathbf{0}$ is the identity operator over addition; i.e., $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathsf{H}$. We require that the tensor product with zero to be absorbing. Thus, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$,

$$\boldsymbol{x} \otimes \mathbf{0} = \mathbf{0} \otimes \mathbf{0} = \mathbf{0} \otimes \boldsymbol{y}.$$

4. **Faithfulness.** Finally, the tensor product should be faithful, i.e., multiplying two nonzero vectors produces a nonzero vector. Put differently, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$, if $\boldsymbol{x} \otimes \boldsymbol{y} = \mathbf{0} \otimes \mathbf{0}$, then either $\boldsymbol{x} = \mathbf{0}$ or $\boldsymbol{y} = \mathbf{0}$.

Intuitively, faithfulness guarantees that for a fixed vector $\boldsymbol{y}$, the mapping $\boldsymbol{x} \mapsto \boldsymbol{x} \otimes \boldsymbol{y}$ is one-to-one.

In the next section, we will show that there is a construction that is consistent with the tensor product axioms. First, let us consider some familiar binary vector operations to see whether these satisfy our desired axioms of a tensor product.

Example 1.2 (Inner product). The inner product does not generally satisfy the axioms of the tensor product. Consider the case where $\mathsf{H} = \mathbb{R}^2$. Then the inner product for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x}^\mathsf{T} \boldsymbol{y}.$$

However, it is not faithful, since one can easily find vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ that are not equal to $\mathbf{0}$, but where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. For instance, $\boldsymbol{x} = (1, 1)^\mathsf{T}$ and $\boldsymbol{y} = (1, -1)^\mathsf{T}$. ∎

Example 1.3 (Schur product). The Schur product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ is denoted by $\boldsymbol{x} \odot \boldsymbol{y}$ and is defined by elementwise multiplication of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. That is,

$$\odot : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n;$$
$$\odot : (\boldsymbol{x}, \boldsymbol{y}) \mapsto (x_i y_i : i = 1, \ldots, n).$$

Similar to the inner product, the Schur product does not satisfy faithfulness. For instance, when $\mathsf{H} = \mathbb{R}^2$, one can pick $\boldsymbol{x} = (1, 0)^\mathsf{T}$ and $\boldsymbol{y} = (0, 1)^\mathsf{T}$. Their Hadamard product is $\boldsymbol{x} \odot \boldsymbol{y} = (0, 0)^\mathsf{T}$, but both vectors are nonzero. ∎

Example 1.4 (Outer product). We will denote the outer product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{C}^n$ by $\boldsymbol{x} \otimes \boldsymbol{y}$, for reasons that will be clear in the following discussion. It is defined by

$$\otimes : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^{n \times n};$$
$$\otimes : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x}\boldsymbol{y}^\mathsf{T} = (x_i y_j)_{i=1,j=1}^{n}.$$

We can prove this by checking all four axioms for the tensor product.

1. **Additivity.** Since the outer product is a particular case of matrix multiplication (between an $n \times 1$ matrix and a $1 \times n$ matrix) and matrix multiplication is additive, the outer product is also additive.

2. **Homogeneity.** Given $\alpha \in \mathbb{C}$, homogeneity follows from explicitly constructing the outer product using the indices:

$$\begin{aligned} \boldsymbol{x}(\alpha \boldsymbol{y})^\mathsf{T} &= (x_i\, \alpha y_j)_{i=1,j=1}^n = (\alpha x_i y_j)_{i=1,j=1}^n; \\ (\alpha \boldsymbol{x})\boldsymbol{y}^\mathsf{T} &= (\alpha x_i y_j)_{i=1,j=1}^n; \\ \alpha(\boldsymbol{x}\boldsymbol{y}^\mathsf{T}) &= \alpha(x_i, y_j)_{i=1,j=1}^n = (\alpha x_i y_j)_{i=1,j=1}^n. \end{aligned}$$

Note that if we had used the conjugate transpose in our definition of the outer product, the outer product would not satisfy homogeneity, since $(\alpha \boldsymbol{x})\boldsymbol{y}^* \neq \boldsymbol{x}(\alpha \boldsymbol{y})^*$

3. **Interaction with zero.** If either $\boldsymbol{x}$ or $\boldsymbol{y}$ is $\boldsymbol{0}$, then either $x_i = 0$ for all i or $y_j = 0$ for all j. Then it must hold that

$$\boldsymbol{x}\boldsymbol{y}^\mathsf{T} = (x_i y_j)_{i,j=1}^n = (0)_{i,j=1}^n = \boldsymbol{0}\boldsymbol{0}^\mathsf{T}.$$

4. **Faithfulness.** If $\boldsymbol{x}\boldsymbol{y}^\mathsf{T} = \boldsymbol{0}$ (the matrix composed entirely of 0 elements), then $x_i y_j = 0$ for all i, j. Fix the index $i = 1$; then, either $x_1 = 0$ or $x_1 \neq 0$.
   1. If $x_1 \neq 0$, then $y_j = 0$ for all j, and $\boldsymbol{y} = \boldsymbol{0}$. This satisfies faithfulness.
   2. If $x_1 = 0$, then proceed to $i = 2$ and repeat this analysis.

   After proceeding in this way, either $\boldsymbol{y} = \boldsymbol{0}$, or $x_i = 0$ for all i. In the latter case, $\boldsymbol{x} = \boldsymbol{0}$, and faithfulness is satisfied.

The outer product thus satisfies all four axioms of the tensor product. ∎

Exercise 1.5 (Extension to different spaces). Let H_1 and H_2 be finite-dimensional Hilbert spaces, perhaps with different dimensions. Describe axioms for a tensor product operation for vectors $\boldsymbol{x} \in \mathsf{H}_1$ and $\boldsymbol{y} \in \mathsf{H}_2$.

Exercise 1.6 (Extension to multiple arguments). Generalize the tensor product to a product of k vectors from H. **Hint:** Each of the axioms in Definition 1.1 needs to be adjusted slightly, but we should still be able to draw on our intuition from multiplying numbers in $\mathbb{C}$. For instance, when multiplying together a series of numbers $x_1, x_2, \ldots, x_k$ in $\mathbb{C}$, we expect homogeneity in every argument so that for $\alpha \in \mathbb{C}$, the product $x_1 \times \cdots \times \alpha x_j \times \cdots \times x_k = \alpha(x_1 \times \cdots \times x_j \times \cdots \times x_k)$. How should we adjust the other axioms so that we preserve our intuitions from the $\mathsf{H} = \mathbb{C}$ case?

1.2 The space of tensor products

Having defined the tensor product operator $\otimes$ as a binary operator following the axioms in Definition 1.1, we will now show how to define the tensor product space, a linear space that contains the tensor products. We know this is a reasonable object to examine because the outer product is a valid tensor product, so there is at least one case where this object exists.

We first define an *elementary tensor* to be an object of the form $\boldsymbol{x} \otimes \boldsymbol{y}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are in a vector space H. We then construct the linear space $\mathsf{H} \otimes \mathsf{H}$ by taking all linear combinations of elementary tensors. We agree that two tensors are the same if we can reduce their difference to zero by repeatedly applying the axioms.

Mathematically, an element $\mathsf{T} \in \mathsf{H} \otimes \mathsf{H}$ can be expressed as a sum of $r \in \mathbb{N}$ elementary tensors weighted by values $\alpha_i \in \mathbb{F}$. That is,

$$\mathsf{T} = \sum_{i=1}^r \alpha_i\, \boldsymbol{x}_i \otimes \boldsymbol{y}_i$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are elements of H. This space is evidently closed under linear

combinations, since given tensors $\mathsf{T}_1 = \sum_{i=1}^{r} \alpha_i \boldsymbol{x}_i \otimes \boldsymbol{y}_i$ and $\mathsf{T}_2 = \sum_{i=1}^{p} \beta_i \boldsymbol{x}_i \otimes \boldsymbol{y}_i$,

$$\begin{aligned}\gamma_1 \mathsf{T}_1 + \gamma_2 \mathsf{T}_2 &= \sum_{i=1}^{r} \gamma_1 \alpha_i \boldsymbol{x}_i \otimes \boldsymbol{y}_i + \sum_{j=1}^{p} \gamma_2 \beta_j \boldsymbol{x}_j \otimes \boldsymbol{y}_j \\ &= \sum_{i=1}^{\max\{r,p\}} \lambda_i \boldsymbol{x}_i \otimes \boldsymbol{y}_i,\end{aligned}$$

where we have concatenated the two sums. Some intuition about non-elementary tensors can be obtained by considering the specific case of an outer product, as in the following example.

One can view the tensor product as a generalization of the outer product. For the particular case of vectors in $\mathbb{C}^n$, the tensor product produces matrices in $\mathbb{C}^{n\times n}$.

Example 1.7 (Elementary tensors). As a concrete example of a non-elementary tensor, consider the standard orthonormal basis $\{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2\}$ in $\mathbb{R}^2$. Two elementary tensors are

$$\boldsymbol{\delta}_1 \otimes \boldsymbol{\delta}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\delta}_2 \otimes \boldsymbol{\delta}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

where $\otimes$ specifically references the outer product. Their sum is

$$\boldsymbol{\delta}_1 \otimes \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 \otimes \boldsymbol{\delta}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this way, we can view an elementary tensor as a rank-1 matrix, while a non-elementary tensor is any higher rank matrix. ▪

However, clearly tensor decompositions are not unique. To impose uniqueness, we define equivalence classes where two vectors in $\mathsf{H} \otimes \mathsf{H}$ are equivalent if and only if they are related by the axioms of vector multiplication, as can be seen through the following example.

Example 1.8 (Non-uniqueness of representation). Take the vectors $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, and $\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2$ in $\mathbb{R}^2$. Then the linear combination

$$(\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2) \otimes (\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2) - \boldsymbol{\delta}_1 \otimes \boldsymbol{\delta}_2 - \boldsymbol{\delta}_2 \otimes \boldsymbol{\delta}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.1}$$

is equal to the sum in the previous example, but has an ostensibly different representation. We can show that these representations are equivalent by using the axioms in Definition 1.1 to distribute the sums in the first term of Equation (1.1) and simplify. ▪

Once these equivalence classes have been established, we can rigorously define a tensor product space as follows.

Definition 1.9 (Tensor product space). Let H be a finite Hilbert space over a field $\mathbb{F}$. The tensor product space $\mathsf{H} \otimes \mathsf{H}$ is defined by all linear combinations of expressions $\boldsymbol{x} \otimes \boldsymbol{y}$ with $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$ modulo the axioms presented in Section 1.1.2.

In summary: $\boldsymbol{x} \otimes \boldsymbol{y}$ is an expression formed from two vectors; taking linear combinations of such expressions forms a new space, which is well defined if we impose equivalence defined by the axioms in Section 1.1.2.

Exercise 1.10 (Extension to different spaces). Explain how to construct the tensor product of two different finite-dimensional Hilbert spaces H_1 and H_2.

1.3 Isomorphism to bilinear forms

Although we have completely defined a tensor product space, the preceding discussions are somewhat abstract. In this section, we will present an alternate way to view a

tensor product space by describing an isomorphism between tensor product spaces and bilinear forms. This is particularly useful because bilinear forms are isomorphic to matrices, so tensor products can be identified with matrices.

1.3.1 Bilinear forms

First, we need the machinery of bilinear forms.

Definition 1.11 (Bilinear forms). A *bilinear form* on a Hilbert space H is a scalar-valued function $B : \mathsf{H} \times \mathsf{H} \to \mathbb{F}$ that maps two arguments from H to a field $\mathbb{F}$ and satisfies the following properties.

1. For all $\boldsymbol{x} \in \mathsf{H}$, the map $B(\boldsymbol{x}, \cdot)$ is a linear functional on H.
2. For all $\boldsymbol{y} \in \mathsf{H}$, the map $B(\cdot, \boldsymbol{y})$ is a linear functional on H.

We will denote the space of bilinear forms acting on $\mathsf{H} \times \mathsf{H}$ by $\mathrm{Bil}(\mathsf{H})$.

Bilinear forms can be viewed more concretely by noting their correspondence with matrices.

Proposition 1.12 (Bilinear forms are isomorphic to matrices). The linear space $\mathrm{Bil}(\mathsf{H})$ is isomorphic to $\mathbb{M}_n(\mathbb{F})$.

$\mathbb{M}_n(\mathbb{F})$ is the linear space of $n \times n$ matrices over the field $\mathbb{F}$.

Proof. Every matrix $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{F})$ has an associated bilinear form defined by

$$B_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i,j=1}^{n} x_i a_{ij} y_j = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{y}.$$

Conversely, every bilinear form B on H has an associated matrix $\boldsymbol{A}$ defined by $a_{ij} = B(\boldsymbol{e}_i, \boldsymbol{e}_j)$ for $i, j = 1, \ldots, n$. A bilinear form can thus be viewed as equivalent to a matrix. ∎

This observation echoes a connection to tensor product spaces; the outer product of two vectors is, after all, a matrix. This observation also provides a straightforward way to see that the linear space $\mathrm{Bil}(\mathsf{H})$ of bilinear forms on H has dimension $(\dim \mathsf{H})^2$, since this is the dimension of $\mathbb{M}_n(\mathbb{F})$.

1.3.2 Connection to tensor product spaces

Having defined bilinear forms, we can present an alternate definition for tensor product spaces.

Definition 1.13 (Tensor product space). The tensor product space $\mathsf{H} \otimes \mathsf{H}$ is the algebraic dual space of the space $\mathrm{Bil}(\mathsf{H})$ of bilinear forms.

Recall that the algebraic dual space of a linear space V is the space of all linear functionals on V.

Concretely, we identify each elementary tensor in $\mathsf{H} \otimes \mathsf{H}$ as a linear functional on $\mathrm{Bil}(\mathsf{H})$ via the following mapping:

$$\boldsymbol{x} \otimes \boldsymbol{y} : B \mapsto B(\boldsymbol{x}, \boldsymbol{y}).$$

This can be easily extended to all tensors by linearity as follows:

$$\sum_i \alpha_i \, \boldsymbol{x}_i \otimes \boldsymbol{y}_i : B \mapsto \sum_i \alpha_i \, B(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Now, we must show that this construction satisfies the axioms in Definition 1.1.

In the following discussion,[1] let us consider bilinear forms B with underlying matrix representations $\boldsymbol{B}$.

[1] For brevity, we do not check every case for each axiom, but all the other cases follow similarly.

1. **Additivity.** Additivity follows from the linearity of matrix–vector multiplication. In particular, the mapping defined by $(\boldsymbol{x}+\boldsymbol{y})\otimes\boldsymbol{z}$ acts on B to output
$$\begin{aligned} B(\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}) &= (\boldsymbol{x}^\mathsf{T}+\boldsymbol{y}^\mathsf{T})\boldsymbol{B}\boldsymbol{z} \\ &= \boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{z}+\boldsymbol{y}^\mathsf{T}\boldsymbol{B}\boldsymbol{z} \\ &= B(\boldsymbol{x},\boldsymbol{z})+B(\boldsymbol{y},\boldsymbol{z}), \end{aligned}$$
which corresponds to the mapping defined by $\boldsymbol{x}\otimes\boldsymbol{z}+\boldsymbol{y}\otimes\boldsymbol{z}$
2. **Homogeneity.** Homogeneity follows similarly to additivity. In particular, given a mapping defined by $\alpha\boldsymbol{x}\otimes\boldsymbol{y}$
$$\begin{aligned} B(\alpha\boldsymbol{x},\boldsymbol{y}) &= \alpha\boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{y} \\ &= \boldsymbol{x}^\mathsf{T}\boldsymbol{B}(\alpha\boldsymbol{y}) \\ &= B(\boldsymbol{x},\alpha\boldsymbol{y}), \end{aligned}$$
which is the mapping defined by $\boldsymbol{x}\otimes\alpha\boldsymbol{y}$.
3. **Interaction with zero.** A zero vector in either argument will make terms like $\boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{y}$ equal to $\boldsymbol{0}$, so
$$B(\boldsymbol{x},\boldsymbol{0}) = B(\boldsymbol{0},\boldsymbol{y}) = B(\boldsymbol{0},\boldsymbol{0}) = \boldsymbol{0},$$
which implies that the mappings defined by $\boldsymbol{x}\otimes\boldsymbol{0}$, and $\boldsymbol{0}\otimes\boldsymbol{y}$, and $\boldsymbol{0}\otimes\boldsymbol{0}$ are all identical.
4. **Faithfulness.** If $\boldsymbol{x}\otimes\boldsymbol{y}=\boldsymbol{0}\otimes\boldsymbol{0}$, then every bilinear form is mapped to 0, since $\boldsymbol{0}^\mathsf{T}\boldsymbol{B}\boldsymbol{0}=0$. The only way for every matrix to be mapped to 0 is if for $\boldsymbol{x}=\boldsymbol{0}$ or for $\boldsymbol{y}=\boldsymbol{0}$; if $\boldsymbol{x}$ and $\boldsymbol{y}$ were both nonzero, I could always consider $\boldsymbol{B}=\boldsymbol{x}\boldsymbol{y}^\mathsf{T}$; then
$$\begin{aligned} \boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{y} &= \boldsymbol{x}^\mathsf{T}\boldsymbol{x}\boldsymbol{y}^\mathsf{T}\boldsymbol{y} \\ &= \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 > 0, \end{aligned}$$
which would violate the assumption that the tensor maps all bilinear forms to zero.

Example 1.14 (Matrix associations in $\mathbb{F}^n$). If $\mathsf{H}\simeq\mathbb{F}^n$, then the space of bilinear forms $\mathrm{Bil}(\mathsf{H})\simeq\mathbb{M}_n(\mathbb{F})$, and $\mathsf{H}\otimes\mathsf{H}=\mathbb{M}_n(\mathbb{F})$. We can then identify $\boldsymbol{x}\otimes\boldsymbol{y}$ with $\boldsymbol{x}\boldsymbol{y}^\mathsf{T}$. Because bilinear forms are isomorphic to matrices and tensors are isomorphic to the dual of the space of bilinear forms, tensors are also isomorphic to matrices. ▪

1.4 Inner products

Since the constituent Hilbert spaces have an inner product, it is natural to want an inner product that acts on vectors from the tensor product space and that interacts well with the inner products of the underlying Hilbert spaces. We define such an inner product as follows.

Definition 1.15 (Tensor inner product). Given two elementary tensors $\boldsymbol{x}_1\otimes\boldsymbol{x}_2$ and $\boldsymbol{y}_1\otimes\boldsymbol{y}_2$, let the inner product between the two elementary tensors be given by
$$\langle\boldsymbol{x}_1\otimes\boldsymbol{x}_2,\boldsymbol{y}_1\otimes\boldsymbol{y}_2\rangle := \langle\boldsymbol{x}_1,\boldsymbol{y}_1\rangle\langle\boldsymbol{x}_2,\boldsymbol{y}_2\rangle.$$
Extend to all tensors by (semi)linearity.

Concretely, the inner product between two general tensors in $\mathsf{H} \otimes \mathsf{H}$ is required to obey

$$\begin{aligned}\left\langle \sum_{i=1}^{r} \alpha_i \, \boldsymbol{x}_1^{(i)} \otimes \boldsymbol{x}_2^{(i)}, \sum_{j=1}^{p} \beta_j \boldsymbol{y}_1^{(j)} \otimes \boldsymbol{y}_2^{(j)} \right\rangle \\ = \sum_{i=1}^{r} \sum_{j=1}^{p} \alpha_i^* \beta_j \langle \boldsymbol{x}_1 \otimes \boldsymbol{x}_2, \boldsymbol{y}_1 \otimes \boldsymbol{y}_2 \rangle \\ = \sum_{i=1}^{r} \sum_{j=1}^{p} \alpha_i^* \beta_j \langle \boldsymbol{x}_1, \boldsymbol{y}_1 \rangle \langle \boldsymbol{x}_2, \boldsymbol{y}_2 \rangle.\end{aligned}$$

Exercise 1.16 (The tensor inner product is well-defined). Show that the tensor inner product is well-defined. In this case, this means that two different representatives of the same equivalence class (i.e., they are related by the axioms in Definition 1.1) should lead to the same inner product.

Proposition 1.17 (Tensor orthonormal basis). Given an orthonormal basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ for H, the basis $\{\boldsymbol{e}_i \otimes \boldsymbol{e}_j : i, j = 1, \ldots, n\}$ is an orthonormal basis for $\mathsf{H} \otimes \mathsf{H}$. In particular, the dimension of $\mathsf{H} \otimes \mathsf{H}$ is $(\dim \mathsf{H})^2$.

Proof. Using the tensor inner product, we note that

$$\langle \boldsymbol{e}_i \otimes \boldsymbol{e}_j, \boldsymbol{e}_i \otimes \boldsymbol{e}_j \rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}_i \rangle \langle \boldsymbol{e}_j, \boldsymbol{e}_j \rangle = 1.$$

However, for indices $k \neq i$ and $l \neq j$,

$$\langle \boldsymbol{e}_i \otimes \boldsymbol{e}_j, \boldsymbol{e}_k \otimes \boldsymbol{e}_l \rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}_k \rangle \langle \boldsymbol{e}_j, \boldsymbol{e}_l \rangle = 0,$$

using the orthonormality of the basis on H.

It remains to be shown that this basis is total; that is, there is no tensor linearly independent from the span of basis vectors. For the finite-dimensional case, this is straightforward. Any vector in $\mathsf{H} \otimes \mathsf{H}$ can definitionally be expressed as a linear combination of elementary tensors, which can then be decomposed into the underlying bases. If $\alpha, \beta, \gamma \in \mathbb{F}$ and $r \in \mathbb{N}$ then

$$\begin{aligned}\sum_{i}^{r} \alpha_i \, \boldsymbol{x}_j \otimes \boldsymbol{y}_j &= \sum_{i}^{r} \alpha_i \left(\sum_{j}^{n} \beta_j \boldsymbol{e}_j \right) \otimes \left(\sum_{k}^{n} \gamma_k \boldsymbol{e}_k \right) \\ &= \sum_{i}^{r} \sum_{j,k}^{n} \lambda_{ijk} \, \boldsymbol{e}_j \otimes \boldsymbol{e}_k,\end{aligned}$$

where we have applied Definition 1.1 to reduce all the terms to a linear combination of elements from the tensor orthonormal basis. This shows that $\mathsf{H} \otimes \mathsf{H}$ is in the span of the set of elementary tensors. Every linear combination of elementary tensors is trivially in $\mathsf{H} \otimes \mathsf{H}$, which completes the proof. ∎

Exercise 1.18 (The tensor orthonormal basis is complete). Repeat the previous proof of completeness using the equivalence of tensor product spaces with bilinear forms. **Hint:** Remember the correspondence between bilinear forms and matrices.

We will often choose to order the orthonormal basis lexicographically; that is, we say $\boldsymbol{e}_j \otimes \boldsymbol{e}_k$ precedes $\boldsymbol{e}_m \otimes \boldsymbol{e}_n$ if and only if $j < m$ or $j = m$ and $k < n$. In other words, we sort by the first index, then the second).

1.5 Theory of linear operators

Having found that $\mathsf{H} \otimes \mathsf{H}$ is a vector space in its own right, a natural further step is to define linear operators acting on $\mathsf{H} \otimes \mathsf{H}$.

Definition 1.19 (Elementary tensor operators). Given linear operators $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{L}(\mathsf{H})$ acting on H and vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$, define a linear operator acting on $\mathsf{H} \otimes \mathsf{H}$ by the following action on the elementary tensor $\boldsymbol{x} \otimes \boldsymbol{y}$:

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{x} \otimes \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}) \otimes (\boldsymbol{B}\boldsymbol{y}).$$

We extend the action to a general tensor in $\mathsf{H} \otimes \mathsf{H}$ by linearity; that is

$$(\boldsymbol{A} \otimes \boldsymbol{B}) \left(\sum_{i=1}^{r} \alpha_i \, \boldsymbol{x}_i \otimes \boldsymbol{y}_i \right) = \sum_{i=1}^{r} \alpha_i \, (\boldsymbol{A}\boldsymbol{x}) \otimes (\boldsymbol{B}\boldsymbol{y})$$

For the finite dimensional case, every linear operator can of course be described by a matrix.

A general linear operator can be constructed as a sum of elementary linear operators. This construction of linear operators interacts smoothly with familiar operations from linear algebra. The following example on compositions of linear operators is quite useful to derive further properties of tensored linear operators using properties of the constituent operators.

Proposition 1.20 (Composition). Given linear operators $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ acting on a Hilbert space H and vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$, the composition of the linear operators $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{C} \otimes \boldsymbol{D}$ is given by $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D})$.

Proof. By considering the action of linear operators on elementary tensors, we have

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D})(\boldsymbol{x} \otimes \boldsymbol{y}) = (\boldsymbol{A} \otimes \boldsymbol{B})((\boldsymbol{C}\boldsymbol{x}) \otimes (\boldsymbol{D}\boldsymbol{y})) = (\boldsymbol{A}\boldsymbol{C}\boldsymbol{x}) \otimes (\boldsymbol{B}\boldsymbol{D}\boldsymbol{y}),$$

which implies that $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D})$. ∎

Corollary 1.21 (Identity and inverses). The identity operator on $\mathsf{H} \times \mathsf{H}$ is given by $\mathbf{I} \otimes \mathbf{I}$, where $\mathbf{I}$ is the identity operator on the Hilbert space H. Furthermore, the inverse of an elementary operator is given by $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = (\boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1})$ provided that both $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible.

Proof. Use Proposition 1.20. ∎

We can also consider other operations familiar from linear algebra.

Proposition 1.22 (Adjoint). The adjoint of an elementary linear operator $\boldsymbol{A} \otimes \boldsymbol{B}$ is given by $(\boldsymbol{A} \otimes \boldsymbol{B})^* = \boldsymbol{A}^* \otimes \boldsymbol{B}^*$.

Recall that the adjoint of a linear map $\boldsymbol{A}$ on a Hilbert space H is the map A^* such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$, the following is true:

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{A}^* \boldsymbol{y} \rangle.$$

Proof. Given vectors $\boldsymbol{x}_i, \boldsymbol{y}_j \in \mathsf{H}$, we can express the condition defining the adjoint as

$$\begin{aligned} \langle (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{x}_1 \otimes \boldsymbol{y}_1), \boldsymbol{x}_2 \otimes \boldsymbol{y}_2 \rangle &= \langle \boldsymbol{A}\boldsymbol{x}_1 \otimes \boldsymbol{B}\boldsymbol{y}_1, \boldsymbol{x}_2 \otimes \boldsymbol{y}_2 \rangle \\ &= \langle \boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \langle \boldsymbol{B}\boldsymbol{y}_1, \boldsymbol{y}_2 \rangle \\ &= \langle \boldsymbol{x}_1, \boldsymbol{A}^* \boldsymbol{x}_2 \rangle \langle \boldsymbol{y}_1, \boldsymbol{B}^* \boldsymbol{y}_2 \rangle \\ &= \langle \boldsymbol{x}_1 \otimes \boldsymbol{y}_1, (\boldsymbol{A}^* \otimes \boldsymbol{B}^*)(\boldsymbol{x}_2 \otimes \boldsymbol{y}_2) \rangle. \end{aligned}$$

Comparing the first and final lines, $(\boldsymbol{A} \otimes \boldsymbol{B})^* = \boldsymbol{A}^* \otimes \boldsymbol{B}^*$. ∎

As seen above, this construction of linear operators makes it obvious when a tensored linear operator preserves properties of linear operators on the constituent spaces. A particularly important property is persistence.

Proposition 1.23 (Persistence). For all $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{L}(\mathsf{H})$, the following statements hold.

1. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are self-adjoint, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is self-adjoint.
2. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are unitary, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is unitary.

3. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are normal, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is normal.
4. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is positive semidefinite.

Recall that a matrix $\boldsymbol{A}$ is normal if it commutes with its conjugate transpose; i.e., $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^*$.

The converses do not necessarily hold.

Proof. As an example, let us prove the case for normal operators. If $\boldsymbol{A}, \boldsymbol{B}$ are normal, then

$$\begin{aligned}(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{A} \otimes \boldsymbol{B})^* &= (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{A}^* \otimes \boldsymbol{B}^*) \\ &= (\boldsymbol{A}\boldsymbol{A}^* \otimes \boldsymbol{B}\boldsymbol{B}^*) \\ &= (\boldsymbol{A}^*\boldsymbol{A}) \otimes (\boldsymbol{B}^*\boldsymbol{B}) \\ &= (\boldsymbol{A}^* \otimes \boldsymbol{B}^*)(\boldsymbol{A} \otimes \boldsymbol{B}).\end{aligned}$$

Therefore, $\boldsymbol{A} \otimes \boldsymbol{B}$ is also normal. ∎

Example 1.24 (Quantum computing). Quantum computers, devices that use properties of quantum states to perform calculations, are studied using a circuit model where each quantum gate acts on two qubits (the quantum equivalent of a bit). Formally, the Hilbert space of a quantum circuit is the tensor product of the Hilbert spaces of all the qubits. For physical reasons, only unitary operations are permitted in quantum circuits. Because each two-qubit gate is a unitary operation, the unitary case of persistence shows that the action on the entire space is also unitary. For further references, see [NCoo]. ▪

1.6 Spectral theory

Results describing the spectra of linear operators are fundamentally important in linear algebra, so it makes sense to ask about the eigenvalues of a linear operator acting on a tensor product space. We can use the construction developed in the previous section to easily answer this question.

Proposition 1.25 (Spectral Theory). If $\boldsymbol{A} \in \mathcal{L}(\mathsf{H})$ has eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and associated eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathsf{H}$, then $\boldsymbol{A} \otimes \boldsymbol{A}$ has eigenvalues $\{\lambda_i\lambda_j : i, j = 1, \ldots, n\}$ with associated eigenvectors $\{\boldsymbol{u}_i \otimes \boldsymbol{u}_j :, i, j = 1, \ldots, n\}$.

Proof. We prove the expression for the eigenvalues using the Schur decomposition.

$$\begin{aligned}\boldsymbol{A} \otimes \boldsymbol{A} &= (\boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^*) \otimes (\boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^*) \\ &= (\boldsymbol{Q} \otimes \boldsymbol{Q})(\boldsymbol{T} \otimes \boldsymbol{T})(\boldsymbol{Q} \otimes \boldsymbol{Q})^*.\end{aligned}$$

The Schur decomposition expresses an arbitrary complex square matrix $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^*$, where $\boldsymbol{Q}$ is unitary and $\boldsymbol{T}$ is an upper triangular matrix whose diagonal entries are the eigenvalues of $\boldsymbol{A}$.

In moving from the first to the second line, we have applied the Schur decomposition. The center term is an upper triangular matrix with diagonal elements $\{\lambda_i\lambda_j : i, j = 1, \ldots, n\}$ whose ordering is given by the lexicographic ordering of the orthonormal basis for the tensor product space. Reading off these entries proves the result.

For the eigenvectors, we can simply note that

$$(\boldsymbol{A} \otimes \boldsymbol{A})(\boldsymbol{u}_j \otimes \boldsymbol{u}_k) = (\boldsymbol{A}\boldsymbol{u}_j) \otimes (\boldsymbol{A}\boldsymbol{u}_k) = \lambda_j\lambda_k(\boldsymbol{u}_j \otimes \boldsymbol{u}_k),$$

which proves the result. ∎

A similar construction holds for singular values.

Proposition 1.26 (Singular value decomposition). If a linear operator $\boldsymbol{A} \in \mathcal{L}(\mathsf{H})$ has a singular value decomposition (SVD) given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, then the singular values of $\boldsymbol{A} \otimes \boldsymbol{A}$ are $\{\sigma_i\sigma_j : i, j = 1, \ldots, n\}$.

Proof. The tensor operator $\boldsymbol{A} \otimes \boldsymbol{A}$ can be written as

$$\begin{aligned}\boldsymbol{A} \otimes \boldsymbol{A} &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*) \otimes (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*) \\ &= (\boldsymbol{U} \otimes \boldsymbol{U})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})(\boldsymbol{V} \otimes \boldsymbol{V})^*,\end{aligned}$$

which has the same form of the SVD. As the center term is diagonal, we can read off the entries in lexicographic order to prove the result. ∎

The spectral decomposition and singular value decomposition therefore provide a method to construct tensor operators based on relations to the underlying linear operators with particular eigenvalues and eigenvectors.

Notes

Much of the material in the section is discussed briefly in [Bha97, Chap. I]. The book [Hal74] contains a more mathematical description of bilinear forms and tensor products.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Hal74] P. R. Halmos. *Finite-dimensional vector spaces*. 2nd ed. Springer-Verlag, New York-Heidelberg, 1974.

[NC00] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

2. Multilinear Algebra

Date: 6 January 2022 **Scribe:** Edoardo Calvello

Agenda:
1. Multivariate tensor product
2. Permutations
3. Wedge products
4. Wedge operators
5. Spectral theory of wedge operators
6. Determinants
7. Symmetric tensor product

In the first lecture, we considered an arbitrary n-dimensional Hilbert space H equipped with the semilinear inner product $\langle \cdot, \cdot \rangle$ and a distinguished orthonormal basis $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$. We defined the bivariate tensor product space $\mathsf{H} \otimes \mathsf{H}$ as the algebraic dual of $\mathrm{Bil}(\mathsf{H})$, the space of bilinear forms taking $\mathsf{H} \times \mathsf{H}$ to $\mathbb{F}$, and we characterized the general form of its elements, called tensors. Extending the definition of $\mathscr{L}(\mathsf{H})$ to linear operators on $\mathsf{H} \otimes \mathsf{H}$ leads to the space $\mathscr{L}(\mathsf{H} \otimes \mathsf{H})$ of linear operators on tensors. We developed a spectral theory for elementary tensor operators.

It is thus natural to ask whether these constructions can be extended to the multivariate case. In this lecture, we introduce the k-variate multilinear functionals as a mechanism to construct the k-fold tensor product space $\otimes^k \mathsf{H}$ and develop the relevant theory. Starting from the observation that the tensor product does not commute, we consider the idea of summing over all permutations of the factors of the product. After recalling some background material on permutations, we define the antisymmetric tensor product, more briefly called the wedge product. It holds that the associated wedge product space $\wedge^k \mathsf{H}$ forms a linear subspace of $\otimes^k \mathsf{H}$. Drawing from our discussion on tensor product operators, this leads us to the theory of wedge operators and their relationship to the determinant function. Finally, we mention the symmetric tensor product and investigate its relation to the permanent.

2.1 Multivariate tensor product

In this section, we aim to extend the discussion on bivariate tensor products from the previous chapter to the multivariate setting and develop the relevant theory.

2.1.1 Construction of multivariate tensor product

To extend the concept of a bilinear form to a higher-order product space, we appeal to the following definition.

Definition 2.1 (k-variate multilinear functional). For $k \in \mathbb{N}$, a function $B : \mathsf{H}^k \to \mathbb{F}$ that is linear in each coordinate is called a *multilinear functional*. We denote by $\mathrm{ML}_k(\mathsf{H})$ the space of k-variate multilinear functionals.

Starting from the algebraic dual of the space $\mathrm{ML}_k(\mathsf{H})$, denoted by $(\mathrm{ML}_k(\mathsf{H}))^*$, we define the k-fold tensor product space $\otimes^k \mathsf{H}$.

Definition 2.2 (Multivariate tensor product space). Given any Hilbert space H, we define the k-fold tensor product space $\otimes^k \mathsf{H}$ as the algebraic dual of the space $\mathrm{ML}_k(\mathsf{H})$. Indeed, we identify $\otimes^k \mathsf{H} \cong (\mathrm{ML}_k(\mathsf{H}))^*$.

Let us next define a k-variate elementary tensor.

Definition 2.3 **(k-variate elementary tensor).** For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathsf{H}$, we call $\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k$ a k-variate elementary tensor.

Following this definition, we can identify each elementary tensor with a linear functional on the space $\mathrm{ML}_k(\mathsf{H})$. Indeed, for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathsf{H}$, by defining the action of $\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k$ on $\mathrm{ML}_k(\mathsf{H})$ as

$$(\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k)(B) = B(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \quad \text{for any } B \in \mathrm{ML}_k(\mathsf{H}).$$

Thus, the elementary tensor $\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k$ determines a linear functional on $\mathrm{ML}_k(\mathsf{H})$. For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathsf{H}$, we have that $\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k \in \otimes^k \mathsf{H}$. Now, since any linear combination of linear functionals is a linear functional, it holds that any linear combination of elementary tensors belongs to the space $\otimes^k \mathsf{H}$. The following proposition completes the description of the k-fold tensor product space $\otimes^k \mathsf{H}$.

Proposition 2.4 **(Multivariate tensors).** Any element $x \in \otimes^k \mathsf{H}$ is a linear combination of elementary tensors. We call an arbitrary element of $\otimes^k \mathsf{H}$ a k-variate tensor.

Exercise 2.5 **(Multivariate tensors).** Provide a proof for Proposition 2.4.

The space of multilinear functionals satisfies $\dim(\mathrm{ML}_k(\mathsf{H})) = (\dim(\mathsf{H}))^k$. In the finite-dimensional setting, the dual space has the same dimension. We may conclude that $\dim\left(\otimes^k \mathsf{H}\right) = (\dim(\mathsf{H}))^k$.

2.1.2 Inner product

Since H is a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$, we can equip the k-fold tensor product space $\otimes^k \mathsf{H}$ with its own inner product $\langle \cdot, \cdot \rangle_{\otimes^k}$.

Definition 2.6 **(Inner product for multivariate tensor space).** The inner product $\langle \cdot, \cdot \rangle_{\otimes^k}$ is defined on elementary tensors in $\otimes^k \mathsf{H}$ by

$$\langle \boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k, \boldsymbol{y}_1 \otimes \cdots \otimes \boldsymbol{y}_k \rangle_{\otimes^k} := \prod_{i=1}^{k} \langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle,$$

for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \mathsf{H}$.

Exercise 2.7 **(Tensor inner product).** Verify that the function $\langle \cdot, \cdot \rangle_{\otimes^k}$ as introduced in Definition 2.6 is an inner product.

Indeed, this inner product can be extended to all tensors in $\otimes^k \mathsf{H}$ by linearity. Given the inner product $\langle \cdot, \cdot \rangle_{\otimes^k}$, it is possible to find an orthonormal basis for the space $\otimes^k \mathsf{H}$.

Proposition 2.8 **(Orthonormal basis for $\otimes^k \mathsf{H}$).** The set

$$\left(\boldsymbol{e}_{i_1} \otimes \cdots \otimes \boldsymbol{e}_{i_k} : i_j = 1, \ldots, n;\ j = 1, \ldots, k\right)$$

forms an orthonormal basis for the space $\otimes^k \mathsf{H}$.

Exercise 2.9 **(Tensor orthonormal basis).** Using the fact that $(\dim(\mathsf{H}))^k = n^k$, prove Proposition 2.8.

2.1.3 Linear operators on multivariate tensor product

We can also introduce a k-fold tensor product operator as follows.

Definition 2.10 **(k-fold tensor product operator).** Given $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, we define the (elementary) tensor product operator $\otimes^k \boldsymbol{A} : \otimes^k \mathsf{H} \to \otimes^k \mathsf{H}$ via

$$\left(\otimes^k \boldsymbol{A}\right)(\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_k) = (\boldsymbol{A}\boldsymbol{x}_1) \otimes \cdots \otimes (\boldsymbol{A}\boldsymbol{x}_k),$$

for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathsf{H}$. We extend this definition to all tensors via linearity.

Note that it is sufficient to define the action of the tensor product operator on elementary tensors, as the definition can be extended to general tensors by linearity.

Similarly to the bivariate case, the k-fold tensor product operator $\otimes^k \boldsymbol{A}$ possesses the following important properties.

Proposition 2.11 **(Tensor operator: Properties).** Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$. Let $\otimes^k \boldsymbol{A} : \otimes^k \mathsf{H} \to \otimes^k \mathsf{H}$ be the k-fold tensor product operator associated to $\boldsymbol{A}$.

1. **Composition.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{L}(\mathsf{H})$, it holds that $\otimes^k(\boldsymbol{A}\boldsymbol{B}) = \left(\otimes^k \boldsymbol{A}\right)\left(\otimes^k \boldsymbol{B}\right)$.
2. **Inverses.** For any $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, we have $\left(\otimes^k \boldsymbol{A}\right)^{-1} = \otimes^k \boldsymbol{A}^{-1}$.
3. **Adjoint.** For any $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, it holds that $\left(\otimes^k \boldsymbol{A}\right)^* = \otimes^k \boldsymbol{A}^*$.
4. **Persistence.** If $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$ is self-adjoint, unitary, normal, or positive semidefinite, then $\otimes^k \boldsymbol{A}$ inherits the respective property.

Exercise 2.12 **(Tensor product operators).** Provide a proof for Proposition 2.11.

2.1.4 Spectral theory

As in the previous chapter, it is possible to generalize to a spectral theory of k-fold tensor product operators. We thus formulate the following two propositions.

Proposition 2.13 **(Spectrum of k-fold tensor product operator).** Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, and let $\otimes^k \boldsymbol{A} : \otimes^k \mathsf{H} \to \otimes^k \mathsf{H}$ be the k-fold tensor product operator associated to $\boldsymbol{A}$. The operator $\otimes^k \boldsymbol{A}$ has eigenvalues $\left(\lambda_{i_1} \cdots \lambda_{i_k} : 1 \leq i_j \leq n,\ j = 1, \ldots, k\right)$, where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of the linear operator $\boldsymbol{A}$.

Proof. As before, the proof of Proposition 2.13 follows directly from the Schur decomposition of $\boldsymbol{A}$. ∎

Proposition 2.14 **(Singular values of k-fold tensor product operator).** Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, and let $\otimes^k \boldsymbol{A} : \otimes^k \mathsf{H} \to \otimes^k \mathsf{H}$ be the k-fold tensor product operator associated to $\boldsymbol{A}$. The operator $\otimes^k \boldsymbol{A}$ has singular values $\left(\sigma_{i_1} \cdots \sigma_{i_k} : 1 \leq i_j \leq n,\ j = 1, \ldots, k\right)$, where $\sigma_1, \ldots, \sigma_n$ are singular values of the linear operator $\boldsymbol{A}$.

Proof. The proof of Proposition 2.14 again follows directly from the singular value decomposition of $\boldsymbol{A}$. ∎

Exercise 2.15 **(Tensor spectral theory).** Elaborate on the details of the proofs of Propositions 2.13 and 2.14 by generalizing the arguments of Lecture 1 to the k-fold tensor product case.

2.2 Permutations

Before continuing our discussion on multilinear algebra, we first outline some background material on permutations which will be of use when defining antisymmetric and symmetric tensor products. We present the definition of a permutation as follows.

Definition 2.16 (Permutation on k symbols). A bijective function $\pi : \{1, \ldots, k\} \to \{1, \ldots, k\}$ is called a permutation on k symbols.

Exercise 2.17 (Counting permutations). Argue that there are $k!$ permutations on k symbols.

We next provide two informative examples of permutations.

Example 2.18 (Some permutations). Both of the following functions $\pi_1, \pi_2 : \{1, 2, 3\} \to \{1, 2, 3\}$ represent permutations.

1. Let $\pi_1 : \{1, 2, 3\} \to \{1, 2, 3\}$ be defined by the tableau

$$\begin{array}{ccc} 1 & 2 & 3 \\ \hline 1 & 3 & 2 \end{array}.$$

2. Let $\pi_2 : \{1, 2, 3\} \to \{1, 2, 3\}$ be defined by

$$\begin{array}{ccc} 1 & 2 & 3 \\ \hline 2 & 3 & 1 \end{array}.$$

Each map π_1, π_2 is a permutation on 3 symbols. ▪

We can define a particular type of permutation, known as a transposition.

Definition 2.19 (Transposition). A transposition is a permutation that interchanges two symbols and maps all other symbols to themselves.

Example 2.20 (Transposition). Referring back to Example 2.18, we can observe that the permutation π_1 is a transposition, while π_2 is not. ▪

We proceed by illustrating a well-known result regarding transpositions.

Theorem 2.21 (Representation of permutations by transpositions). Any permutation can be expressed as a composition of transpositions.

Exercise 2.22 (*Transpositions). Sketch a proof for Theorem 2.21. **Hint:** For a more intuitive argument, appeal to a "sorting" scheme. For a more group-theoretic argument, appeal to the "cycle representation." First, show that any permutation can be written as a product of disjoint cycles, and then show that a product of disjoint cycles can be expressed as a product of 2-cycles.

Furthermore, let us recall the following result.

Theorem 2.23 (Parity of a permutation). Fix a permutation π. Either every representation of π as a composition of transpositions contains an even number of transpositions or else every representation of π as a composition of transpositions contains an odd number of transpositions.

Exercise 2.24 (Parity). Provide a proof for Theorem 2.23.

A permutation is *even* when it is represented by an even number of transpositions, and it is *odd* when it is represented by an odd number of transpositions. As a consequence of Theorem 2.23, each permutation is either even or odd. Finally, we define the signature of a permutation.

Definition 2.25 (Signature of a permutation). The signature ε_π of the a permutation π

is defined by

$$\varepsilon_\pi = \begin{cases} +1 & \text{if } \pi \text{ is even,} \\ -1 & \text{if } \pi \text{ is odd.} \end{cases}$$

It is important to note that Theorem 2.23 ensures that the signature of a permutation is well-defined.

2.3 Wedge products

In general, $\boldsymbol{x} \otimes \boldsymbol{y} \neq \boldsymbol{y} \otimes \boldsymbol{x}$!

The fact that the tensor product does not commute leads to the observation that it may be of interest to associate tensors whose factors are the same up to order. This can be accomplished by summing over all permutations of the factors of the product. In addition, it turns out to be fruitful to weight each term of the sum by the signature of the corresponding permutation. This weighting confers to this so-defined product the significant property of antisymmetry. Considering this idea of a weighted sum over permutations leads to the definition of the wedge product.

Definition 2.26 (Wedge product). The wedge product $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k$ of vectors $\boldsymbol{x}_1, \ldots \boldsymbol{x}_k \in \mathsf{H}$ is defined as

$$\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k := \frac{1}{\sqrt{k!}} \sum\nolimits_{\pi \in \mathsf{S}_k} \varepsilon_\pi \cdot \boldsymbol{x}_{\pi(1)} \otimes \cdots \otimes \boldsymbol{x}_{\pi(k)},$$

where S_k is the symmetric group consisting of all permutations on k symbols and where ε_π denotes the signature of π. As such, $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k$ is clearly an element of the space $\otimes^k \mathsf{H}$.

The purpose of the normalizing the wedge product by $(k!)^{-1/2}$ is to ensure that the wedge product of orthonormal vectors yields a unit vector. We will turn to this matter later when we introduce the inner product.

Example 2.27 (Wedge product for $k = 2$). When $n = 1$, the wedge product $\boldsymbol{x} \wedge \boldsymbol{y} = \boldsymbol{0}$ always. In case $k = 2$, the wedge product of $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}$ is the tensor

$$\boldsymbol{x} \wedge \boldsymbol{y} = \frac{1}{\sqrt{2}} \left(\boldsymbol{x} \otimes \boldsymbol{y} - \boldsymbol{y} \otimes \boldsymbol{x} \right).$$

When $n = 3$ and $\mathbb{F} = \mathbb{R}$, the wedge product is equivalent with the well-known cross-product. For general n and $k = 2$, one can find an analogy between the wedge product and the skew part of a matrix. (Why?) ▪

We observe that the signature of the permutation in the summand makes the wedge product antisymmetric. Indeed, for any $i, j \in \{1, \ldots, k\}$ we have that

$$\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_i \wedge \cdots \wedge \boldsymbol{x}_j \wedge \cdots \wedge \boldsymbol{x}_k = -\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_j \wedge \cdots \wedge \boldsymbol{x}_i \wedge \cdots \wedge \boldsymbol{x}_k.$$

Indeed, exchanging any two vectors introduces an additional transposition into each permutation in Definition 2.26, which flips the signature. As a consequence, the wedge product is also known as the *antisymmetric tensor product*.

The important observation above leads to the conclusion that if $\boldsymbol{x}_i = \boldsymbol{x}_j$ for some $i \neq j$ in $\{1, \ldots, k\}$, then $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k = \boldsymbol{0}$. Exchanging $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ does not change the product, but flips its sign. Therefore, since this product is an element of a vector space, $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k = \boldsymbol{0}$.

This statement has a remarkable generalization to the case in which $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)$ is a linearly dependent set.

Proposition 2.28 (Linear dependence and wedge product). The set $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)$ is linearly dependent if and only if $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k = \boldsymbol{0}$.

Proof. Without loss of generality, by the linear dependence of the set, it is possible to write $\boldsymbol{x}_1 = \sum_{j=2}^{k} \alpha_j \boldsymbol{x}_j$ for some $\alpha_2, \dots, \alpha_k \in \mathbb{F}$. This means in particular that by linearity,

$$\begin{aligned} \boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k &= \left(\sum\nolimits_{j=2}^{k} \alpha_j \boldsymbol{x}_j \right) \wedge \boldsymbol{x}_2 \wedge \cdots \wedge \boldsymbol{x}_k \\ &= \sum\nolimits_{j=2}^{k} \alpha_j \left(\boldsymbol{x}_j \wedge \boldsymbol{x}_2 \wedge \cdots \wedge \boldsymbol{x}_k \right) \\ &= \boldsymbol{0}. \end{aligned}$$

The converse is true by a straightforward induction argument. ∎

Exercise 2.29 (Wedge: Linear dependence). Recalling that any permutation can be written as a composition of transpositions, complete the induction argument in the proof of Proposition 2.28.

We can now proceed by constructing the wedge product space.

Definition 2.30 (Wedge product space). We define the wedge product space $\wedge^k \mathsf{H}$ by

$$\wedge^k \mathsf{H} := \operatorname{lin} \{ \boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k : \boldsymbol{x}_i \in \mathsf{H} \text{ and } i = 1, \dots, k \}.$$

As such, $\wedge^k \mathsf{H}$ is a linear subspace of the tensor product space $\otimes^k \mathsf{H}$.

As a linear subspace of $\otimes^k \mathsf{H}$, the space $\wedge^k \mathsf{H}$ thus inherits the inner product $\langle \cdot, \cdot \rangle_{\otimes^k}$, from which it is possible to construct an orthonormal basis.

Proposition 2.31 (Orthonormal basis for $\wedge^k \mathsf{H}$). The set

$$\left(\boldsymbol{e}_{i_1} \wedge \cdots \wedge \boldsymbol{e}_{i_k} : 1 \leq i_1 < \cdots < i_k \leq n \right)$$

forms an orthonormal basis for the space $\wedge^k \mathsf{H}$.

Exercise 2.32 (Wedge: Orthonormal basis). Provide a proof for Proposition 2.31.

Since the orthonormal basis for the linear subspace $\wedge^k \mathsf{H}$ contains $\binom{n}{k}$ elements, we conclude that $\dim\left(\wedge^k \mathsf{H}\right) = \binom{n}{k}$. In particular, we have $\dim\left(\wedge^n \mathsf{H}\right) = 1$.

2.4 Wedge operators

We continue our discussion on wedge product spaces by introducing the wedge operator.

Definition 2.33 (Wedge operator). Given $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, we define the (elementary) wedge operator $\wedge^k \boldsymbol{A} : \wedge^k \mathsf{H} \to \wedge^k \mathsf{H}$ via

$$\left(\wedge^k \boldsymbol{A} \right) (\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k) = (\boldsymbol{A}\boldsymbol{x}_1) \wedge \cdots \wedge (\boldsymbol{A}\boldsymbol{x}_k),$$

for any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathsf{H}$. We extend to the wedge space by linearity.

As such, we can identify the wedge operator $\wedge^k \boldsymbol{A}$ as the restriction of the tensor operator $\otimes^k \boldsymbol{A}$ to the wedge product space $\wedge^k \mathsf{H}$. Indeed, appealing to Definition 2.26 one can conclude that for any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathsf{H}$,

$$\begin{aligned}\big(\otimes^k \boldsymbol{A}\big)(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) &= \frac{1}{\sqrt{k!}} \sum\nolimits_{\pi \in S_k} \varepsilon_\pi (\boldsymbol{A}\boldsymbol{x}_{\pi(1)}) \otimes \cdots \otimes (\boldsymbol{A}\boldsymbol{x}_{\pi(k)}) \\ &= (\boldsymbol{A}\boldsymbol{x}_1) \wedge \cdots \wedge (\boldsymbol{A}\boldsymbol{x}_k) \\ &= \big(\wedge^k \boldsymbol{A}\big)(\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k),\end{aligned}$$

where the last equality follows from Definition 2.33. The wedge product space $\wedge^k \mathsf{H}$ is thus an invariant subspace of the operator $\otimes^k \boldsymbol{A}$. As a consequence, the wedge operator $\wedge^k \boldsymbol{A}$ inherits all the properties of $\otimes^k \boldsymbol{A}$.

Proposition 2.34 (Wedge operator: Properties). Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, and let $\wedge^k \boldsymbol{A} : \wedge^k \mathsf{H} \to \wedge^k \mathsf{H}$ be the wedge operator associated to $\boldsymbol{A}$.

1. **Composition.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathscr{L}(\mathsf{H})$, it holds that $\wedge^k(\boldsymbol{A}\boldsymbol{B}) = \big(\wedge^k \boldsymbol{A}\big)\big(\wedge^k \boldsymbol{B}\big)$.
2. **Inverses.** For any $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, we have $\big(\wedge^k \boldsymbol{A}\big)^{-1} = \wedge^k \boldsymbol{A}^{-1}$.
3. **Adjoint.** For any $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, it holds that $\big(\wedge^k \boldsymbol{A}\big)^* = \wedge^k \boldsymbol{A}^*$.
4. **Persistence.** If $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$ is self-adjoint, unitary, normal, or positive semidefinite, then $\wedge^k \boldsymbol{A}$ inherits the respective property.

Proof. Proposition 2.34 is a straightforward consequence of the wedge space $\wedge^k \mathsf{H}$ being an invariant subspace of the operator $\otimes^k \boldsymbol{A}$. ∎

2.5 Spectral theory of wedge operators

Much like in the case of the tensor product operator, it is possible to develop a spectral theory of wedge product operators. We thus formulate the following two propositions.

Proposition 2.35 (Spectrum of wedge operators). Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, and let $\wedge^k \boldsymbol{A} : \wedge^k \mathsf{H} \to \wedge^k \mathsf{H}$ be the wedge operator associated to $\boldsymbol{A}$. The operator $\wedge^k \boldsymbol{A}$ has eigenvalues $\big(\lambda_{i_1} \cdots \lambda_{i_k} : 1 \leq i_1 < \cdots < i_k \leq n\big)$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the linear operator $\boldsymbol{A}$.

Proof. As in the case of the tensor product operator, the proof of Proposition 2.35 follows directly from the Schur decomposition of $\boldsymbol{A}$. This time, we use the fact that the orthonormal basis for the wedge space consists of tensors $\boldsymbol{e}_{i_1} \wedge \cdots \wedge \boldsymbol{e}_{i_k}$ with no repeated indices. ∎

Proposition 2.36 (Singular values of wedge operators). Let $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, and let $\wedge^k \boldsymbol{A} : \wedge^k \mathsf{H} \to \wedge^k \mathsf{H}$ be the wedge operator associated to $\boldsymbol{A}$. The operator $\wedge^k \boldsymbol{A}$ has singular values $\big(\sigma_{i_1} \cdots \sigma_{i_k} : 1 \leq i_1 < \cdots < i_k \leq n\big)$, where $\sigma_1, \dots, \sigma_n$ are the singular values of the linear operator $\boldsymbol{A}$.

Proof. As in the case of the tensor product operator, the proof the proof of Proposition 2.36 follows directly from the singular value decomposition of $\boldsymbol{A}$. ∎

2.6 Determinants

In this section, we develop the theory of determinants of matrices as an instant consequence of the theory of wedge products.

Suppose that $\dim \mathsf{H} = n$. Then we may identify the wedge operator $\wedge^n \boldsymbol{A}$ with the determinant of $\boldsymbol{A}$. Indeed, the operator $\wedge^n \boldsymbol{A}$ acts on the space $\wedge^n \mathsf{H} = \operatorname{lin}\{\boldsymbol{e}_1 \wedge \cdots \wedge \boldsymbol{e}_n\}$, which is a one-dimensional space. Recall that a linear operator on a one-dimensional space acts by scalar multiplication. In particular, the operator $\wedge^n \boldsymbol{A}$ scales by its own single eigenvalue, $\lambda_1 \cdots \lambda_n$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\boldsymbol{A}$. These considerations lead us to the following definition of the determinant.

Definition 2.37 (Determinant). For any linear operator $\boldsymbol{A} \in \mathscr{L}(\mathsf{H})$, the determinant of $\boldsymbol{A}$ is defined as

$$\det(\boldsymbol{A}) := \wedge^n \boldsymbol{A} = \lambda_1 \cdots \lambda_n,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\boldsymbol{A}$.

This definition is a little abusive because the wedge operator $\wedge^n \boldsymbol{A}$ is actually a scalar operator $\alpha \mathbf{I}$ acting on a one-dimensional subspace of $\otimes^n \mathsf{H}$.

With this definition, all properties of the determinant of a matrix follow. In particular, considering $\mathscr{L}(\mathsf{H}) = \mathbb{M}_n(\mathbb{F})$, we have the following important result.

Proposition 2.38 (Properties of the determinant of a matrix). Given the space $\mathbb{M}_n(\mathbb{F})$, the determinant has the following properties.

1. **Multiplicativity.** For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{F})$,

$$\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B}).$$

2. **Multilinearity.** The determinant is multilinear. Namely, letting $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, for any $\alpha \in \mathbb{F}$ and $\boldsymbol{b}_j \in \mathbb{F}^n$,

$$\begin{aligned}\det\left([\boldsymbol{a}_1, \ldots, \alpha \cdot \boldsymbol{a}_j + \boldsymbol{b}_j, \ldots, \boldsymbol{a}_n]\right) = \\ \alpha \cdot \det\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_j, \ldots, \boldsymbol{a}_n]\right) + \det\left([\boldsymbol{a}_1, \ldots, \boldsymbol{b}_j, \ldots, \boldsymbol{a}_n]\right),\end{aligned}$$

   for all $j = 1, \ldots, n$.
3. **Antisymmetry.** The determinant reverses the sign if two columns of its argument are swapped. Namely, letting $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, for any $i, j \in \{1, \ldots, n\}$,

$$\det\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_j, \ldots, \boldsymbol{a}_n]\right) = -\det\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_j, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_n]\right).$$

4. **Normalization.** For the identity $\mathbf{I} \in \mathbb{M}_n(\mathbb{F})$, we have $\det(\mathbf{I}) = 1$.

Exercise 2.39 (Determinant). Using Definition 2.37, provide a proof for Proposition 2.38.

Finally, we note the following theorem.

Theorem 2.40 (Uniqueness of the determinant). The determinant is the unique function from $\mathbb{M}_n(\mathbb{F})$ to $\mathbb{F}$ with the above properties of multiplicativity, multilinearity, antisymmetry and normalization.

Proof. For a proof of this statement see the first chapter of [Art11]. ∎

2.7 Symmetric tensor product

We recall that the antisymmetry of the wedge product is given by weighing the sum over all permutations by each permutation's signature. We omit the signature to define the symmetric tensor product.

Definition 2.41 (Symmetric tensor product). The symmetric tensor product $\boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_k$ of vectors $\boldsymbol{x}_1, \ldots \boldsymbol{x}_k \in \mathsf{H}$ is defined as

$$\boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_k := \frac{1}{\sqrt{k!}} \sum\nolimits_{\pi \in \mathsf{S}_k} \boldsymbol{x}_{\pi(1)} \otimes \cdots \otimes \boldsymbol{x}_{\pi(k)},$$

where S_k is the symmetric group consisting of all permutations on k symbols. As such, clearly $\boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_k$ is an element of the space $\otimes^k \mathsf{H}$.

Indeed, for any $i, j \in \{1, \ldots, k\}$ we have that

$$\boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_i \vee \cdots \vee \boldsymbol{x}_j \vee \cdots \vee \boldsymbol{x}_k = \boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_j \vee \cdots \vee \boldsymbol{x}_i \vee \cdots \vee \boldsymbol{x}_k,$$

hence the symmetry.

Similarly as for the wedge product, we can define the symmetric tensor product space.

Exercise 2.42 (Symmetric tensor product space). By using Definition 2.41, define the symmetric tensor product space $\vee^k \mathsf{H}$. Establish that its dimension is $\binom{n+k-1}{k}$. How quickly does the dimension grow as $k \to \infty$ for fixed n?

Furthermore, it is possible to define the symmetric tensor product operator and develop the relevant theory.

Exercise 2.43 (Symmetric tensor product operator). By using Definition 2.41 and the notion of symmetric tensor product space $\vee^k \mathsf{H}$, define and develop the theory for the symmetric tensor product operator.

Let us next define the permanent of a matrix.

Definition 2.44 (Permanent of a matrix). Let $\boldsymbol{A} = [a_{ij}]$ be any matrix in $\mathbb{M}_n(\mathbb{F})$. The permanent of $\boldsymbol{A}$ is defined as

$$\operatorname{perm}(\boldsymbol{A}) := \sum\nolimits_{\pi \in \mathsf{S}_n} a_{1\pi(1)} \cdots a_{n\pi(n)},$$

where S_n is the symmetric group consisting of all permutations on n symbols.

We refer to [Bha97] for an in depth discussion on the connections between the symmetric tensor product operator and the permanent of a matrix.

Notes

The material of this chapter can largely be found in [Bha97], albeit to varying degrees of detail. In particular, the above discussions represent a more in-depth exploration of the foundational elements of multilinear algebra. The interested reader can find a further exploration of these topics in [Bha97]. For a refresher on topics relating to linear algebra, we refer to [Art11].

Lecture bibliography

[Art11] M. Artin. *Algebra*. Pearson Prentice Hall, 2011.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

3. Majorization

Date: 11 January 2022 **Scribe:** Matthieu Darcy

This lecture introduces the concept of majorization of vectors, a way of comparing how "spiky" they are. Next, we explore the space of doubly stochastic matrices and a particular class of doubly-stochastic matrices, the T-transforms. We relate both concepts by characterizing majorization via doubly stochastic matrices. We conclude with an application: the Schur–Horn theorem, which relates the diagonal entries and the eigenvalues of an Hermitian matrix.

Agenda:
1. Majorization
2. Doubly stochastic matrices
3. Characterization of majorization
4. Schur–Horn theorem

3.1 Majorization

In this section, we introduce the concept of majorization for vectors. Informally speaking, a vector $\boldsymbol{x}$ is majorized by a vector $\boldsymbol{y}$, written $\boldsymbol{x} \prec \boldsymbol{y}$, if $\boldsymbol{y}$ is "spikier" (or more concentrated) than $\boldsymbol{x}$.

3.1.1 Setting

In this lecture, we will work in $\mathbb{R}^n$ equipped with the standard basis $(\boldsymbol{\delta}_i : i = 1, 2, \ldots, n)$ and the canonical inner product $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^\mathsf{T}\boldsymbol{b}$. We define the vector of ones:

$$\mathbf{1} := (1, 1, \ldots 1)^\mathsf{T} \in \mathbb{R}^n.$$

The *trace* of a vector is

$$\mathrm{tr}(\boldsymbol{x}) := \langle \mathbf{1}, \boldsymbol{x} \rangle = \sum\nolimits_{i=1}^{n} x_i.$$

Much later, when we study positive linear maps, we will see that the trace of a vector and the trace of a matrix are analogous.

3.1.2 Rearrangements

We now define rearrangements, a key concept in analysis.

Definition 3.1 (Rearrangement). Let $\boldsymbol{x} \in \mathbb{R}^n$. The decreasing rearrangement $\boldsymbol{x}^\downarrow \in \mathbb{R}^n$ of $\boldsymbol{x}$ is a vector with the same entries as $\boldsymbol{x}$, but placed in weakly decreasing order:

$$x_1^\downarrow \geq x_2^\downarrow \geq \cdots \geq x_n^\downarrow.$$

Likewise, an increasing rearrangement $\boldsymbol{x} \in \mathbb{R}^n$ has the same entries as $\boldsymbol{x}$, placed in weakly increasing order:

$$x_1^\uparrow \leq x_2^\uparrow \leq \cdots \leq x_n^\uparrow.$$

In situations where the order of the entries of a vector is unimportant, it may be useful to replace the vector by its decreasing rearrangement. More formally, we can define an equivalence relation between vectors whose decreasing rearrangements are the same and work with the equivalence classes.

Example 3.2 (Rearrangements). If $\boldsymbol{x} = (1, 3, 2) \in \mathbb{R}^3$, then its decreasing and increasing rearrangements are

$$\boldsymbol{x}^{\downarrow} = (3, 2, 1);$$
$$\boldsymbol{x}^{\uparrow} = (1, 2, 3).$$

Observe that the vector $\boldsymbol{y} = (2, 3, 1)$ has the same rearrangements. ▪

The next proposition provides us with an upper bound and a lower bound on the inner product of two vectors from the inner products of their rearrangements.

Proposition 3.3 (Chebyshev rearrangement). For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\langle \boldsymbol{x}^{\downarrow}, \boldsymbol{y}^{\uparrow} \rangle \leq \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \langle \boldsymbol{x}^{\downarrow}, \boldsymbol{y}^{\downarrow} \rangle;$$
$$\langle \boldsymbol{x}^{\uparrow}, \boldsymbol{y}^{\downarrow} \rangle \leq \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \langle \boldsymbol{x}^{\uparrow}, \boldsymbol{y}^{\uparrow} \rangle.$$

Proof. The proof is left as an exercise to the reader. **Hint:** Assume that $\boldsymbol{x}, \boldsymbol{y} \geq \mathbf{0}$, and apply summation by parts. Otherwise, see [HLP88, p. 261]. ■

3.1.3 Majorization order

Majorization compares two vectors by considering their *decreasing* rearrangements.

Definition 3.4 (Majorization order). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. We say that $\boldsymbol{y}$ majorizes $\boldsymbol{x}$, written $\boldsymbol{x} \prec \boldsymbol{y}$, when

$$\sum_{i=1}^{k} x_i^{\downarrow} \leq \sum_{i=1}^{k} y_i^{\downarrow} \quad \text{for each } k = 1, \dots n, \text{ and}$$
$$\sum_{i=1}^{n} x_i^{\downarrow} = \sum_{i=1}^{n} y_i^{\downarrow}.$$

The second condition can be stated as $\operatorname{tr}(\boldsymbol{x}) = \operatorname{tr}(\boldsymbol{y})$.

Alternatively, the majorization order can be stated using the *increasing* rearrangements. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if

$$\sum_{i=1}^{k} x_i^{\uparrow} \geq \sum_{i=1}^{k} y_i^{\uparrow} \quad \text{for each } k = 1, \dots n, \text{ and}$$
$$\sum_{i=1}^{n} x_i^{\uparrow} = \sum_{i=1}^{n} y_i^{\uparrow}.$$

Intuitively, a vector $\boldsymbol{x}$ is majorized by $\boldsymbol{y}$ if $\boldsymbol{x}$ is "flatter": its mass is more uniformly distributed over its entries than $\boldsymbol{y}$.

Example 3.5 (Finite probability distribution). Let $\boldsymbol{x} \in \mathbb{R}^n_+$ with $\operatorname{tr}(\boldsymbol{x}) = 1$. We can interpret any such vector as a discrete probability distribution on a finite sample space. The following relations hold:

$$\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) \prec \boldsymbol{x} \prec (1, 0, \dots, 0).$$

We can interpret $\left(\frac{1}{n}, \dots, \frac{1}{n}\right)$ as the maximally flat vector: its mass is uniformly distributed and therefore it takes the longest to accumulate to the total. On the other hand, $(1, 0, \dots, 0)$ is the spikiest vector: it is maximally concentrated and therefore accumulates to the total as quickly as possible. These examples are illustrated in Figure 3.1 with increasing rearrangements. ▪

We also have the following equivalent definition of majorization:

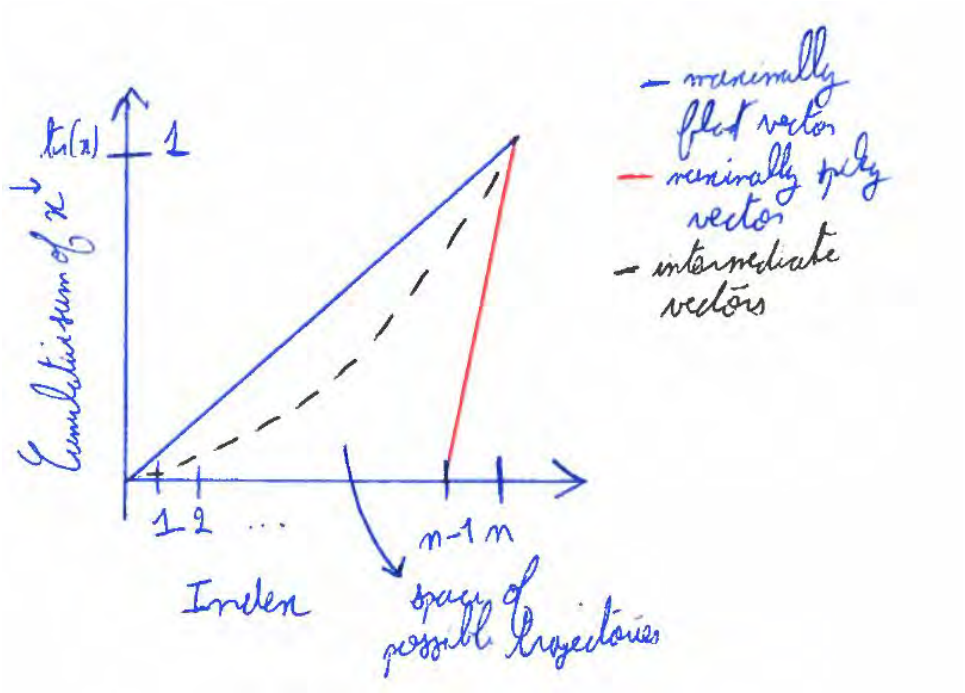

Figure 3.1 (Majorization for probability distributions). Illustration of majorization for *increasing* rearrangements of probability vectors on $\mathbb{R}^n$.

Proposition 3.6 (Majorization without rearrangement). For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if

$$\operatorname{tr}(|\boldsymbol{x} - t\mathbf{1}|) \leq \operatorname{tr}(|\boldsymbol{y} - t\mathbf{1}|) \quad \text{for all } t \in \mathbb{R}. \tag{3.1}$$

The symbol $|\cdot|$ denotes the entrywise absolute value. Note that (3.1) is exactly the statement that

$$\|\boldsymbol{x} - t\mathbf{1}\|_{\ell_1} \leq \|\boldsymbol{y} - t\mathbf{1}\|_{\ell_1} \quad \text{for all } t \in \mathbb{R}.$$

Proof. See Problem Set 1. ∎

Remark 3.7 (Partial order). The relation $\prec$ does not define a partial order because the relations $\boldsymbol{y} \prec \boldsymbol{x}$ and $\boldsymbol{x} \prec \boldsymbol{y}$ do not imply that $\boldsymbol{x} = \boldsymbol{y}$. It does however define a partial order on the set of decreasing rearrangements; that is the set of equivalence classes of vectors where $\boldsymbol{x} \sim \boldsymbol{y}$ if and only if $\boldsymbol{x}^{\downarrow} = \boldsymbol{y}^{\downarrow}$.

Remark 3.8 (Notation and conventions). Majorization and rearrangements arise in other fields, where one may encounter differing notations and conventions:

1. In statistics, rearrangements are called *order statistics*. They are usually denoted as $x_{[i]} = x_i^{\downarrow}$ and $x_{(i)} = x_i^{\uparrow}$.
2. In analysis, rearrangements are sometimes written as $x_i^* = x_i^{\downarrow}$.
3. In some fields, especially physics, $\boldsymbol{y} \succ \boldsymbol{x}$ means that $\boldsymbol{y}$ is more *chaotic* than $\boldsymbol{x}$. Hence uniform vectors are the highest in the partial order, which is the opposite of our definition.

When reading papers that use majorization, one must take care to note which conventions are being used!

3.2 Doubly stochastic matrices

In this section, we introduce the class of doubly stochastic matrices. We will see that these matrices enact the majorization relation.

Definition 3.9 (Doubly stochastic matrices). A matrix $\boldsymbol{S} \in \mathbb{M}_n(\mathbb{R})$ with entries $\boldsymbol{S} = [s_{ij}]$ is *doubly stochastic* if it has the following three properties:

1. **Positivity.** The entries $s_{ij} \geq 0$ for all $i, j = 1, 2, \dots, n$.
2. **Rows.** The row sums $\sum_{j=1}^{n} s_{ij} = 1$ for all $i = 1, 2, \dots, n$.
3. **Columns.** The column sums $\sum_{i=1}^{n} s_{ij} = 1$ for all $j = 1, 2, ..., n$.

The set DS_n collects all of the $n \times n$ doubly stochastic matrices.

Example 3.10 (Doubly stochastic matrices). The following are examples of doubly stochastic matrices:

1. The $n \times n$ identity matrix $\mathbf{I}_n$ is a doubly stochastic matrix.
2. The $n \times n$ matrix $\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ with constant entries is doubly stochastic.
3. Permutation matrices are doubly stochastic. Recall that $\boldsymbol{P}$ is a permutation matrix if it is a square matrix with 0–1 entries, and there is exactly one 1 per row and per column.
4. Orthostochastic matrices are doubly stochastic. A matrix $\boldsymbol{S}$ is *orthostochastic* when its entries $s_{ij} = |u_{ij}|^2$ for a unitary matrix $\boldsymbol{U}$.

Doubly stochastic matrices naturally arise in a variety of applications, such as the study of reversible Markov chains. ▪

Intuitively, we think of doubly stochastic matrices acting on vectors by averaging. Interpreting each row (and each column) of a doubly stochastic matrix $\boldsymbol{T}$ as a probability distribution, we can view the entries of $\boldsymbol{T}\boldsymbol{v}$ as the expectation of the (discrete) function $\boldsymbol{v}$ with respect to that probability distribution.

The next two propositions present some important properties of doubly stochastic matrices.

Proposition 3.11 (Doubly stochastic matrices: Characterization). Each matrix $\boldsymbol{S} \in \mathsf{DS}_n$ has the following properties:

1. **Positive.** For all $\boldsymbol{x} \in \mathbb{R}^n$, the relation $\boldsymbol{x} \geq \mathbf{0}$ implies that $\boldsymbol{S}\boldsymbol{x} \geq \mathbf{0}$.
2. **Trace preserving.** For all $\boldsymbol{x} \in \mathbb{R}^n$, we have $\operatorname{tr}(\boldsymbol{S}\boldsymbol{x}) = \operatorname{tr}(\boldsymbol{x})$.
3. **Unital.** We have $\boldsymbol{S}\mathbf{1} = \mathbf{1}$.

In fact, a matrix is doubly stochastic if and only if it satisfies these three properties.

Proof. The positivity property follows from the positivity of each coordinate $(\boldsymbol{S}\boldsymbol{x})_i = \sum_{j=1}^{n} s_{ij} x_j \geq 0$ for $i = 1, \dots n$.

The unital property follows easily from the observation that $(\boldsymbol{S}\mathbf{1})_i = \sum_{j=1}^{n} s_{ij} = 1$.

To prove the trace preservation property, note that $\boldsymbol{S}$ is doubly stochastic if and only if $\boldsymbol{S}^{\mathsf{T}}$ is also doubly stochastic. Therefore

$$\operatorname{tr}(\boldsymbol{S}\boldsymbol{x}) = \langle \mathbf{1},\ \boldsymbol{S}\boldsymbol{x} \rangle = \langle \boldsymbol{S}^{\mathsf{T}}\mathbf{1},\ \boldsymbol{x} \rangle = \langle \mathbf{1},\ \boldsymbol{x} \rangle = \operatorname{tr}(\boldsymbol{x}),$$

where we have used properties of the adjoint and the unital property of $\boldsymbol{S}^{\mathsf{T}}$. ∎

The next result describes structural properties of the full set DS_n of doubly stochastic matrices.

Proposition 3.12 (Properties of DS_n). The set DS_n has the following properties:

1. **Convexity.** The set DS_n is a compact, convex set of $\mathbb{R}^{n\times n}$.
2. **Composition.** If $\boldsymbol{S}, \boldsymbol{T} \in \mathsf{DS}_n$ then $\boldsymbol{ST} \in \mathsf{DS}_n$.

Proof. First, observe that the set DS_n is bounded. Indeed, for all $\boldsymbol{S} \in \mathsf{DS}_n$, we have $\|\boldsymbol{S}\|_\infty = \max_{ij} |s_{ij}| \leq 1$. To show that DS_n is closed and convex, simply observe that each row and each column of a doubly stochastic matrix can be defined as a vector belonging to the intersection of closed halfspaces.

In detail, we can write $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n]$ where $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ are the columns of $\boldsymbol{S}$. Then, for each index $i = 1, \ldots, n$, the condition $\langle \mathbf{1},\ \boldsymbol{s}_i\rangle = 1$ can be expressed as

$$\boldsymbol{s}_i \in \{\boldsymbol{x} \in \mathbb{R}^n : \langle \mathbf{1},\ \boldsymbol{x}\rangle \geq 1\} \cap \{\boldsymbol{x} \in \mathbb{R}^n : \langle \mathbf{1},\ \boldsymbol{x}\rangle \leq 1\}.$$

Likewise, for each index $i = 1, \ldots, n$, the condition $\boldsymbol{s}_i \geq \mathbf{0}$ can be expressed as

$$\boldsymbol{s}_i \in \{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{\delta}_j,\ \boldsymbol{x}\rangle \geq 0\} \quad \text{for } j = 1, \ldots, n.$$

The same holds for the rows of $\boldsymbol{S}$. The set DS_n is therefore the intersection of (a finite number of) closed and convex sets and hence is itself closed and convex.

To prove the composition property, consider two stochastic matrices $\boldsymbol{S}$, $\boldsymbol{T}$, and their product $\boldsymbol{ST}$. The entries of the product are clearly positive:

$$(\boldsymbol{ST})_{ij} = \sum_{k=1}^n s_{ik}t_{kj} \geq 0.$$

The row and column sum are also preserved in the product. For each row and column index,

$$\sum_{j=1}^n (st)_{ij} = \sum_{j=1}^n \sum_{k=1}^n s_{ik}t_{kj} = \sum_{k=1}^n \left(s_{ik} \sum_{j=1}^n t_{kj}\right) = 1.$$
$$\sum_{i=1}^n (st)_{ij} = \sum_{i=1}^n \sum_{k=1}^n s_{ik}t_{kj} = \sum_{k=1}^n \left(s_{kj} \sum_{i=1}^n t_{ik}\right) = 1.$$

Therefore, $\boldsymbol{ST}$ is also stochastic. ∎

Corollary 3.13 (Convex combinations of permutation matrices). Any convex combination of permutation matrices is a doubly stochastic matrix.

Proof. As was seen in Example 3.10, permutation matrices are doubly stochastic matrices and, by the previous theorem, the set DS_n is a convex set. ∎

The converse of the corollary above is also true: any doubly stochastic matrix is a convex combination of permutation matrices. This is the content of the Birkhoff–von Neumann theorem.

3.3 T-transforms

We now introduce a special class of of doubly stochastic matrices, the T-transforms. T-transforms allow us to take convex combinations of two entries of a vector while leaving all other coordinates unchanged. In other words, it averages two coordinates. These matrices will allow us to relate majorization to the space DS_n of doubly stochastic matrices.

Definition 3.14 (T-transform). A T-*transform* is a doubly stochastic matrix $\boldsymbol{T}$ of the form

$$\boldsymbol{T} = \tau \mathbf{I} + (1-\tau)\boldsymbol{P}$$

where $\tau \in [0,1]$ and $\boldsymbol{P}$ is a permutation that transposes two coordinates. For example, if the transposition acts on the coordinates (k, ℓ), then

$$(\boldsymbol{P}\boldsymbol{v})_i = \begin{cases} v_k & \text{if } i = \ell, \\ v_\ell & \text{if } i = k, \\ v_i & \text{otherwise.} \end{cases}$$

Thus, $\boldsymbol{T}$ only acts nontrivially on the coordinates (k, ℓ).

By Proposition 3.12, each T-transform is a doubly stochastic matrix because it is a convex combination of two permutation matrices. The next example shows that T-transforms and majorization are equivalent in $\mathbb{R}^2$.

Example 3.15 (T-transforms and majorization in $\mathbb{R}^2$). In $\mathbb{R}^2$, T-transforms take a particularly simple form because they act on both coordinates:

$$\boldsymbol{T} = \tau \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (1-\tau) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{for a parameter } \tau \in [0,1].$$

For vectors in $\mathbb{R}^2$, we will argue that $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y}$ for some choice of τ.

Indeed, let $\boldsymbol{x} = (x_1, x_2)$, and $\boldsymbol{y} = (y_1, y_2)$. The majorization relation $\boldsymbol{x} \prec \boldsymbol{y}$ states that

$$x_1 \le y_1 \quad \text{and} \quad x_1 + x_2 = y_1 + y_2. \tag{3.2}$$

Hence, $y_2 \le x_1 \le y_1$. As a consequence, there is a $\tau \in [0,1]$ such that

$$x_1 = \tau y_1 + (1-\tau) y_2.$$

Plugging the last display back in to (3.2), we obtain

$$x_2 = (1-\tau) y_1 + \tau y_2.$$

We conclude that $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y}$ for the T-transform with this distinguished parameter τ.

Conversely, we can reverse this chain of argument to see that $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y}$ implies that $\boldsymbol{x} \prec \boldsymbol{y}$ for vectors in $\mathbb{R}^2$. ∎

Example 3.16 (T-transforms in $\mathbb{R}^3$). In $\mathbb{R}^3$, we consider the T-transform that acts on the pair $(1, 3)$ of coordinates. Then the transposition matrix has the form

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

The T-transform becomes

$$\boldsymbol{T} = \begin{bmatrix} \tau & 0 & (1-\tau) \\ 0 & 1 & 0 \\ (1-\tau) & 0 & \tau \end{bmatrix} \quad \text{for some } \tau \in [0,1].$$

Observe that $(\boldsymbol{T}\boldsymbol{x})_2 = x_2$ so that the second coordinate is unchanged by this transformation.

This observation generalizes to $\mathbb{R}^n$ for $n \in \mathbb{N}$ and an arbitrary choice (ℓ, k) of coordinates. If $\boldsymbol{T}$ is a T-transform acting nontrivially on coordinates (k, ℓ), then $(\boldsymbol{T}\boldsymbol{x})_i = x_i$ for each $i \ne k, \ell$. ∎

Observe that the action of a T-transform induces a majorization relation.

Exercise 3.17 (T-transforms). Let $\boldsymbol{y} \in \mathbb{R}^n$ be a vector, and let $\boldsymbol{T}$ be a T-transform on $\mathbb{R}^n$. Then $\boldsymbol{T}\boldsymbol{y} \prec \boldsymbol{y}$.

3.4 Majorization and doubly stochastic matrices

The main theorem of this lecture relates majorization to the set of doubly stochastic matrices. Recall that a vector $\boldsymbol{x}$ is majorized by a vector $\boldsymbol{y}$ if the vector $\boldsymbol{x}$ is "more average." Likewise, doubly stochastic matrices act on a vector by averaging its entries with respect to discrete probability distributions defined by its rows. The next result shows that these concepts are equivalent: $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if $\boldsymbol{x}$ is obtained through a transformation of $\boldsymbol{y}$ by a doubly stochastic matrix.

Theorem 3.18 (Majorization and DS_n). Fix two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The following statements are equivalent:

1. $\boldsymbol{x} \prec \boldsymbol{y}$
2. $\boldsymbol{x} = \boldsymbol{T}_n \cdots \boldsymbol{T}_1 \boldsymbol{y}$ for certain T-transforms $\boldsymbol{T}_i$.
3. $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$ for some $\boldsymbol{S} \in \mathsf{DS}_n$.
4. $\boldsymbol{x} \in \operatorname{conv}\{\boldsymbol{P}\boldsymbol{y} : \boldsymbol{P} \text{ is a permutation matrix on } \mathbb{R}^n\}$.

Since each matrix $\boldsymbol{T}_1, \boldsymbol{T}_2, \dots, \boldsymbol{T}_n$ is a T-transform, observe that $\boldsymbol{T}_{n-1} \cdots \boldsymbol{T}_1 \boldsymbol{y}$ and $\boldsymbol{T}_n \cdots \boldsymbol{T}_1 \boldsymbol{y}$ only differ in two coordinates.

Proof. **(3 $\Leftrightarrow$ 4).** Corollary 3.13 gives the reverse implication. The forward implication follows from the Birkhoff–von Neumann theorem, which we will establish in Lecture 5. It can also be established directly; see Exercise 3.19.

(2 $\Rightarrow$ 3). Since $\boldsymbol{x} = \boldsymbol{T}_n \cdots \boldsymbol{T}_1 \boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}$ it suffices to show that $\boldsymbol{T}_n \cdots \boldsymbol{T}_1$ is doubly stochastic. This point holds because T-transforms are doubly stochastic and the set of doubly stochastic matrices is stable under products (Proposition 3.12).

(3 $\Rightarrow$ 1). Let $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$. For each $t \in \mathbb{R}$, we may calculate that

$$|\boldsymbol{x} - t\mathbf{1}| = |\boldsymbol{S}\boldsymbol{y} - t\mathbf{1}| = |\boldsymbol{S}(\boldsymbol{y} - t\mathbf{1})| \le \boldsymbol{S}(|\boldsymbol{y} - t\mathbf{1}|).$$

We have used the unital property and Jensen's inequality. Taking the trace,

$$\operatorname{tr}(|\boldsymbol{x} - t\mathbf{1}|) \le \operatorname{tr}(\boldsymbol{S}(|\boldsymbol{y} - t\mathbf{1}|)) = \operatorname{tr}(|\boldsymbol{y} - t\mathbf{1}|),$$

since doubly stochastic matrices are trace-preserving. This is the equivalent formulation of majorization from Proposition 3.6.

(1 $\Rightarrow$ 2). This is the hard part. Assume that $\boldsymbol{x} \prec \boldsymbol{y}$. We will construct a sequence of T-transforms $\boldsymbol{T}_1, \boldsymbol{T}_2, \dots, \boldsymbol{T}_n$ such that

$$\begin{aligned} &\boldsymbol{x} = \boldsymbol{T}_n \boldsymbol{T}_{n-1} \cdots \boldsymbol{T}_1 \boldsymbol{y} \quad \text{and} \\ &\boldsymbol{T}_{k+1} \boldsymbol{T}_k \dots \boldsymbol{T}_1 y \prec \boldsymbol{T}_k \dots \boldsymbol{T}_1 \boldsymbol{y} \quad \text{for all } k = 0, \dots, n-1. \end{aligned}$$

Without loss of generality, we may assume that $\boldsymbol{x} = \boldsymbol{x}^{\downarrow}$ and $\boldsymbol{y} = \boldsymbol{y}^{\downarrow}$ because reordering the vectors does not affect the majorization relation.

We proceed by induction on the dimension n. For the base case $n = 2$, see Example 3.15. For the induction step, observe that the two conditions

$$x_1 \le y_1 \quad \text{and} \quad \sum\nolimits_{i=1}^{n} x_i = \sum\nolimits_{i=1}^{n} y_i$$

imply that there is an index $k \in \{1, \dots, n-1\}$ such that $y_k \le x_1 \le y_{k-1}$. Thus,

$$x_1 = \tau y_1 + (1-\tau) y_k.$$

We can find a T-transform $\boldsymbol{T}_1$ on the set of coordinates $(1, k)$ such that $(\boldsymbol{T}_1 \boldsymbol{y})_1 = x_1$, leaving all other coordinates of $\boldsymbol{y}$ unchanged. By the implication **(3 $\Rightarrow$ 1)** or Exercise 3.17,

$$(x_1, \boldsymbol{w}) \coloneqq \boldsymbol{T}_1 \boldsymbol{y} \prec \boldsymbol{y},$$

where $\boldsymbol{w}$ denotes the last $n-1$ coordinates of $\boldsymbol{T}_1 \boldsymbol{y}$.

It remains to check that $(x_2, \dots, x_n) \prec \boldsymbol{w}$. Observe that

$$w_i = \begin{cases} y_i & \text{if } i \neq k \\ \tau y_i + (1-\tau) t y_1 & \text{if } i = k. \end{cases}$$

where we use the convention that the indices of $\boldsymbol{w}$ range from 2 to n. For $n < k$, since $x_1 < y_i$ for $i = 1, \dots, n$, we have that

$$\sum_{i=2}^{n} x_i \le \sum_{i=2}^{n} y_i = \sum_{i=2}^{n} w_i.$$

For $n \ge k$,

$$\sum_{i=1}^{n} x_i \le \sum_{i=1}^{n} y_i.$$

Therefore,

$$\sum_{i=2}^{n} x_i \le \sum_{i=1}^{n} y_i - x_1 = \sum_{i=2}^{n} w_i$$

because $w_k = y_1 + y_k - x_1$ and $w_i = y_i$ for $i \neq k$ by definition. Finally,

$$\sum_{i=2}^{n} w_i = \sum_{i=1}^{n} y_i - x_1 = \sum_{i=2}^{n} x_i.$$

As required, $(x_2, x_3, \dots, x_n) \prec \boldsymbol{w}$.

By our induction hypothesis, there is a T-transform on $\boldsymbol{w}$ satisfying the requirements. This transformation can be extended to a transformation $\boldsymbol{T}_2$ on $(x_1, \boldsymbol{w})$ by leaving the first coordinate unchanged, completing the induction step.

■

The next exercise completes the above proof without resorting to the Birkhoff–von Neumann theorem.

Exercise 3.19 (Direct proof of majorization and DS$_n$). Prove that **(2 $\Rightarrow$ 4)** in Theorem 3.18 by expanding the product of T-transforms.

3.5 The Schur–Horn theorem

We now prove the important Schur–Horn theorem on Hermitian matrices. On the one hand, Schur's theorem states that the diagonal entries of any Hermitian matrix are majorized by its eigenvalues. On the other, Horn's theorem tells us that we can go in the other direction: from a pair of vectors with a majorization relation there is a matrix with the corresponding diagonal and eigenvalues.

Theorem 3.20 (Schur–Horn). Let $\boldsymbol{A} \in \mathbb{H}_n$ be a Hermitian matrix. The following two claims hold.

1. **Schur.** We have $\operatorname{diag}(\boldsymbol{A}) \prec \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A})$.
2. **Horn.** If $\boldsymbol{u} \prec \boldsymbol{v}$, then there exists a Hermitian matrix $\boldsymbol{B}$ such that $\operatorname{diag}(\boldsymbol{B}) = \boldsymbol{u}$ and $\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{B}) = \boldsymbol{v}^{\downarrow}$.

The map $\boldsymbol{\lambda}^{\downarrow} : \mathbb{H}_n \mapsto \mathbb{R}^n$ returns the vector of eigenvalues arranged in decreasing order.

Proof. We present a proof of Schur's theorem only. For a proof of Horn's theorem, see [MOA11, p.302].

By the spectral theorem, we can write $\boldsymbol{A} = \boldsymbol{U}^* \boldsymbol{\Lambda} \boldsymbol{U}$ where $\boldsymbol{U}$ is unitary and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal entries are listed in $\boldsymbol{\lambda}_i^{\downarrow}(\boldsymbol{A})$. Write $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n]$ where $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n \in \mathbb{C}^n$ are the columns of $\boldsymbol{U}$. Then

$$a_{ii} = \boldsymbol{\delta}_i^* \boldsymbol{A} \boldsymbol{\delta}_i = \boldsymbol{\delta}_i^* \boldsymbol{U}^* \boldsymbol{\Lambda} \boldsymbol{U} \boldsymbol{\delta}_i = (\boldsymbol{U}\boldsymbol{\delta})^* \boldsymbol{\Lambda} (\boldsymbol{U}\boldsymbol{\delta}) = \boldsymbol{u}_i^* \boldsymbol{\Lambda} \boldsymbol{u}_i = \sum\nolimits_{j=1}^{n} |u_{ij}|^2 \lambda_j^{\downarrow}(\boldsymbol{A}).$$

Define the orthostochastic matrix $\boldsymbol{S}$ with entries $s_{ij} = |u_{ij}|^2$. The last display tells us that

$$a_{ii} = (\boldsymbol{S} \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}))_i.$$

In other words, $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{S} \boldsymbol{\lambda}^{\downarrow}(A)$ where $\boldsymbol{S}$ is doubly stochastic. By Theorem 3.18, we may conclude that $\operatorname{diag}(\boldsymbol{A}) \prec \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A})$. ∎

Notes

Majorization is a foundational topic in analysis. It plays a key role in the classic book *Inequalities* by Hardy, Littlewood, and Pólya [HLP88]. Majorization is also the core topic in the well-known book *Inequalities* by Marshall & Olkin, updated by Arnold [MOA11]. The presentation in this lecture is adapted from Bhatia's book [Bha97, Chap. II], which owes a heavy debt to Ando's arrangement of the material [And89]. Indeed, Ando understood that similar ideas are also at the heart of the theory of positive linear maps, which we will explore in the second half of the course.

Lecture bibliography

[And89] T. Ando. "Majorization, doubly stochastic matrices, and comparison of eigenvalues". In: *Linear Algebra and its Applications* 118 (1989), pages 163–248.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[HLP88] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Reprint of the 1952 edition. Cambridge University Press, Cambridge, 1988.

[MOA11] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: theory of majorization and its applications*. Second. Springer, New York, 2011. DOI: 10.1007/978-0-387-68276-1.

4. Isotone Functions

Date: 13 January 2022 **Scribe:** Anthony (Chi-Fang) Chen

We continue our discussion on majorization. A function $f : \mathbb{R} \to \mathbb{R}$ that respects the ordering on the real line is called a *monotone* function. In the same spirit, functions that respect the majorization order on vectors are called *isotone functions*. These functions provide convenient ways to transfer majorization properties from one space to another. We can identify isotone functions by checking simple properties, such as permutation invariance and convexity, and we will see plenty of examples.

Agenda:
1. Weyl majorant theorem
2. Isotonicity
3. Sufficient conditions and examples
4. Schur "convexity"

4.1 Recap

For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the majorization relation $\boldsymbol{x} \prec \boldsymbol{y}$ is defined by the following conditions:

$$\begin{aligned} \sum_{i=1}^{k} x_i^{\downarrow} &\le \sum_{i=1}^{k} y_i^{\downarrow} \quad \text{for each } k = 1, \dots, n; \\ \operatorname{tr}[\boldsymbol{x}] := \sum_{i=1}^{n} x_i^{\downarrow} &= \sum_{i=1}^{n} y_i^{\downarrow} =: \operatorname{tr}[\boldsymbol{y}]. \end{aligned}$$

In other words, majorization consists of $n - 1$ inequalities and one equality among the (sorted) entries of the vectors. Intuitively, this captures the idea that vector $\boldsymbol{y}$ is "spikier" than vector $\boldsymbol{x}$, or vector $\boldsymbol{x}$ is "flatter" than vector $\boldsymbol{y}$. Majorization plays a basic role in matrix analysis. For example, we saw an important theorem of Schur that relates the diagonal entries of a matrix to its decreasingly ordered eigenvalues.

Theorem 4.1 (Schur). Consider a self-adjoint matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathbb{C})$ written in the standard coordinate basis. The vector of diagonal entries of the matrix is majorized by the vector of eigenvalues:

$$\operatorname{diag}(\boldsymbol{A}) \prec \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}).$$

The function $\boldsymbol{\lambda}^{\downarrow} : \mathbb{H}_n \to \mathbb{R}^n$ reports the (real) eigenvalues in decreasing order.

4.2 Weyl majorant theorem

Today, we begin with another important theorem that establishes a majorization relationship between the eigenvalues and singular values of a square matrix. To state this result, we define a function $\boldsymbol{\lambda}^{\downarrow} : \mathbb{M}_n(\mathbb{C}) \to \mathbb{R}^n$ that lists the complex eigenvalues in decreasing order of magnitude, followed by increasing order of phase (modulo 2π). The function $\boldsymbol{\sigma} : \mathbb{M}_n(\mathbb{C}) \to \mathbb{R}^n$ reports the singular values in decreasing order.

Theorem 4.2 (Weyl majorant theorem). For every matrix $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$, we have

$$\prod_{i=1}^{k} |\lambda_i^{\downarrow}(\boldsymbol{A})| \le \prod_{i=1}^{k} \sigma_i^{\downarrow}(\boldsymbol{A}) \quad \text{for each } k = 1, \dots, n;$$
$$\prod_{i=1}^{n} |\lambda_i(\boldsymbol{A})| = \prod_{i=1}^{n} \sigma_i(\boldsymbol{A}).$$

We can express these relations with the shorthand

$$\log(\boldsymbol{\lambda}^{\downarrow}) \prec \log(\boldsymbol{\sigma}).$$

When there are zero eigenvalues and zero singular values, the product formulas give a precise meaning to the log-majorization.

The proof requires two lemmas.

Lemma 4.3 (Products of eigenvalues and singular values). For every matrix $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$, we have

$$\prod_{i=1}^{n} |\lambda_i^{\downarrow}(\boldsymbol{A})| = \prod_{i=1}^{n} \sigma_i(\boldsymbol{A}).$$

Proof. Let us express the determinant in terms of both the eigenvalues and the singular values.

$$\begin{aligned} |\det(\boldsymbol{A})| = |\det(\boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^*)| &= |\det(\boldsymbol{Q})| \cdot |\det(\boldsymbol{T})| \cdot |\det(\boldsymbol{Q}^*)| \\ &= \prod_{i=1}^{n} |\lambda_i^{\downarrow}(\boldsymbol{A})|; \qquad (4.1) \\ |\det(\boldsymbol{A})| = |\det(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*)| &= |\det(\boldsymbol{U})| \cdot |\det(\boldsymbol{\Sigma})| \cdot |\det(\boldsymbol{V}^*)| \\ &= \prod_{i=1}^{n} \sigma_i(\boldsymbol{A}). \qquad (4.2) \end{aligned}$$

In equation (4.1), we use Schur decomposition and the multiplicativity of the determinant, and then we evaluate the determinant of the upper triangular matrix $\boldsymbol{T}$. Note the determinant of any unitary matrix has magnitude one. Equation (4.2) is analogous but uses the singular value decomposition instead. ∎

Lemma 4.4 (Spectral radius and spectral norm). For every matrix $\boldsymbol{M} \in \mathbb{M}_n(\mathbb{C})$, the maximal singular value bounds the magnitude of each eigenvalue:

$$|\lambda_i(\boldsymbol{M})| \le \sigma_{\max}(\boldsymbol{M}) \quad \text{for each index } i.$$

Proof. Consider any eigenvalue $\lambda_i(\boldsymbol{M})$ of the matrix with unit-norm eigenvector $\boldsymbol{u}_i$. Then

$$|\lambda_i(\boldsymbol{M})| = \left|\boldsymbol{u}_i^* \boldsymbol{M} \boldsymbol{u}_i\right| \le \max\{|\boldsymbol{u}^* \boldsymbol{M} \boldsymbol{v}| : \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1\} = \sigma_{\max}(\boldsymbol{M}).$$

This is the advertised result. ∎

We are now prepared to establish Theorem 4.2.

Proof of Weyl majorant theorem. The equality is Lemma 4.3; for the inequalities, we cleverly use multilinear algebra. For any $k = 1, \dots, n$, consider the anti-symmetric subspace $\bigwedge^k \mathbb{C}^n$ and the induced operator $\bigwedge^k \boldsymbol{A} =: \boldsymbol{M}$. In Lecture 2, we showed that the eigenvalues of operator $\boldsymbol{M}$ are products of k eigenvalues of $\boldsymbol{A}$ with no repeated indices. By Lemma 4.4, we can bound the eigenvalues of $\boldsymbol{M}$ above by the maximal singular value of $\boldsymbol{M}$. Indeed,

Without invoking multilinear algebra, the proof gets horrible.

$$\left|\prod_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A})\right| = |\lambda_1^{\downarrow}(\boldsymbol{M})| \le \sigma_{\max}(\boldsymbol{M}) = \prod_{i=1}^{k} \sigma_i(\boldsymbol{A}).$$

This is the advertised result. ∎

4.3 Isotonicity

Today, we move on to study isotone functions that respect the majorization order. We will introduce easy-to-check conditions for isotonicity and see that many existing measures of "inequality" fit into this framework.

4.3.1 Motivation

Let us begin with an economics example. Suppose we wish to summarize how equal or unequal a society is. Consider a vector $\boldsymbol{x} \in \mathbb{R}^n_+$ with normalization $\mathrm{tr}[\boldsymbol{x}] = 1$ that describes the distribution of wealth of each individual in some society $\boldsymbol{x}$. As we have already seen, the majorization relation between societies $\boldsymbol{x} \prec \boldsymbol{y}$ is a way to quantify that "society $\boldsymbol{y}$ is more unequal than society $\boldsymbol{x}$".

Motivated by the above, can we find an even simpler summary of "inequality"? For example, can we reduce the distribution to a single number? Formally, we may ask if there is a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{implies} \quad \Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y}).$$

In fact, we will see that many familiar functions qualify.

Example 4.5 (Isotone functions). The following functions respect the majorization order:

- **k-max.** The sum of the largest k entries obviously preserves the majorization order:
$$\Phi(\boldsymbol{x}) = \sum\nolimits_{i=1}^{k} x_i^{\downarrow},$$
where $1 \le k \le n$.
- **Negative entropy.** The negative entropy reflects the amount of randomness or uniformity in a distribution:
$$\Phi(\boldsymbol{x}) = \mathrm{negent}(\boldsymbol{x}) = \sum\nolimits_{i=1}^{n} x_i \log(x_i).$$
- **Variance.** The variance is another measure of dispersion of a distribution:
$$\Phi(\boldsymbol{x}) = \mathrm{Var}[\boldsymbol{x}] = \frac{1}{n} \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2,$$
where $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$ is the mean.

By the end of this lecture, we will develop new tools to confirm that the negative entropy and the variance also respect the majorization order. ∎

4.3.2 Definitions

To prepare for our treatment of isotone functions, we will need some additional definitions.

Definition 4.6 (Weak majorization). For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we say that $\boldsymbol{y}$ *submajorizes* $\boldsymbol{x}$ and write $\boldsymbol{x} \prec_w \boldsymbol{y}$ when

$$\sum\nolimits_{i=1}^{k} x_i^{\downarrow} \le \sum\nolimits_{i=1}^{k} y_i^{\downarrow} \quad \text{for each } k = 1, \dots, n. \tag{4.3}$$

Similarly, we say that $\boldsymbol{y}$ *supermajorizes* $\boldsymbol{x}$ and write $\boldsymbol{x} \prec^{w} \boldsymbol{y}$ when

$$\sum_{i=1}^{k} x_i^{\uparrow} \geq \sum_{i=1}^{k} y_i^{\uparrow} \quad \text{for each } k = 1, \dots, n. \tag{4.4}$$

In comparison with majorization, the submajorization and supermajorization conditions both lack the trace equality constraint. This grants us more flexibility, but it also renders the two conditions (4.3) and (4.4) inequivalent.

Exercise 4.7 (Majorization and weak majorization). Check the following equivalence.

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{if and only if} \quad \boldsymbol{x} \prec^{w} \boldsymbol{y} \text{ and } \boldsymbol{x} \prec_{w} \boldsymbol{y}.$$

We now introduce the main concept of this lecture.

Definition 4.8 (Isotone function). We say a vector-valued function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is *isotone* if for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom}(\Phi)$,

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{implies} \quad \Phi(\boldsymbol{x}) \prec_{w} \Phi(\boldsymbol{y}).$$

In the definition, we also allow for functions defined only on a convex subset of vectors (e.g., those with positive entries).

One may wonder: Why not demand the full majorization relation $\Phi(\boldsymbol{x}) \prec \Phi(\boldsymbol{y})$? This condition turns out to be too limiting. For example, in the case of scalar-valued functions ($m = 1$), the trace constraint in majorization would force the function to be a constant.

Exercise 4.9 (Majorization on $\mathbb{R}$). For real numbers a, b, verify that

$$a \prec b \quad \text{if and only if} \quad a = b,$$
$$a \prec_{w} b \quad \text{if and only if} \quad a \leq b.$$

In other words, the majorization relation for numbers is very rigid.

4.3.3 Sufficient conditions

To study which functions are isotone, we first introduce some definitions that will be useful to formulate the sufficient conditions.

Definition 4.10 (Convex function: Vector-valued case). A vector-valued function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is *convex* when

$$\Phi\left((1-\tau)\boldsymbol{x} + \tau\boldsymbol{y}\right) \leq (1-\tau)\,\Phi(\boldsymbol{x}) + \tau\,\Phi(\boldsymbol{y}) \quad \text{for all } \tau \in [0,1]$$

and for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom}(\Phi)$. The inequality $\leq$ acts entrywise.

Definition 4.11 (Permutation covariance). A function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is *permutation covariant* if for each vector $\boldsymbol{x} \in \mathbb{R}^n$ and each permutation $\boldsymbol{P}$ on $\mathbb{R}^n$, there is a permutation $\boldsymbol{P}'$ on $\mathbb{R}^m$ such that

$$\Phi(\boldsymbol{P}\boldsymbol{x}) = \boldsymbol{P}'\Phi(\boldsymbol{x}).$$

The permutation $\boldsymbol{P}'$ may depend on both $\boldsymbol{x}$ and $\boldsymbol{P}$.

Example 4.12 (The real case). When the output dimension $m = 1$, the preceding definitions simplify. Permutation covariant functions must be permutation *invariant*:

$$\Phi(\boldsymbol{P}\boldsymbol{x}) = \Phi(\boldsymbol{x}) \quad \text{for each permutation } \boldsymbol{P}.$$

Convexity reduces to the familiar notion for real-valued functions. ∎

Convexity and permutation covariance give sufficient (but not necessary) conditions for isotonicity.

Theorem 4.13 (Isotonicity: Sufficient condition). If a function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is convex and permutation covariant, then it is isotone.

Proof. Fix vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ that satisfy the majorization relation $\boldsymbol{x} \prec \boldsymbol{y}$. In Lecture 3, we showed that there exists a doubly stochastic matrix $\boldsymbol{S}$ such that

$$\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y} = \sum\nolimits_{i=1}^{r} \alpha_i \boldsymbol{P}_i \boldsymbol{y}.$$

The second relation expresses the doubly stochastic matrix as a convex combination of permutations $\boldsymbol{P}_i$ with $\alpha_i \geq 0$ and $\sum_i^r \alpha_i = 1$. By convexity of the function Φ,

$$\begin{aligned} \Phi(\boldsymbol{x}) = \Phi\Big(\sum\nolimits_{i=1}^{r} \alpha_i \boldsymbol{P}_i \boldsymbol{y} \Big) &\leq \sum\nolimits_{i=1}^{r} \alpha_i \Phi(\boldsymbol{P}_i \boldsymbol{y}) \\ &= \sum\nolimits_{i=1}^{r} \alpha_i \boldsymbol{P}_i' \Phi(\boldsymbol{y}) =: \boldsymbol{z} \end{aligned}$$

where the inequality acts entrywise. Since the vector $\boldsymbol{z}$ is a convex combination of permutations of vector $\boldsymbol{y}_i$, we arrive at the majorization relation $\boldsymbol{z} \prec \Phi(\boldsymbol{y})$. The combined inequalities $\Phi(\boldsymbol{x}) \leq \boldsymbol{z} \prec \Phi(\boldsymbol{y})$ imply the submajorization relation $\Phi(\boldsymbol{x}) \prec_w \Phi(\boldsymbol{y})$, which is the advertised result. ∎

Exercise 4.14 (Submajorization deduction). Check that the following implication holds.

$$\Phi(\boldsymbol{x}) \leq \boldsymbol{z} \prec \Phi(\boldsymbol{y}) \quad \text{implies} \quad \Phi(\boldsymbol{x}) \prec_w \Phi(\boldsymbol{y}).$$

4.3.4 Isotone functions: Vector examples

First, let us give some examples of vector-valued functions that are isotone because of the condition in Theorem 4.13.

Exercise 4.15 (Isotonicity: Vectorized scalar functions). Consider a scalar convex function $\varphi : \mathbb{R} \to \mathbb{R}$, and extend it to vectors $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ by applying the scalar function φ entrywise:

$$\Phi(x_1, \cdots, x_n) := (\varphi(x_1), \cdots, \varphi(x_n)). \tag{4.5}$$

The constructed function Φ is convex and permutation covariant. Therefore, it is isotone.

Example 4.16 (Vectorized scalar functions). Here are some particular cases of the construction in Exercise 4.15:

- **Absolute powers.** Convex power functions, applied entrywise, are isotone. For example,

$$\Phi(\boldsymbol{x}) := (|x_1|^p, \cdots, |x_n|^p), \quad \text{for } p \geq 1.$$

- **Positive parts.** The positive part of a vector is an isotone function:

$$\Phi(\boldsymbol{x}) := ((x_1)_+, \cdots, (x_n)_+).$$

- **Exponential.** The entrywise exponential of a vector is an isotone function:

$$\Phi(\boldsymbol{x}) := (\mathrm{e}^{x_1}, \cdots, \mathrm{e}^{x_n}). \tag{4.6}$$

Note that each of these scalar functions is already convex. ▪

Corollary 4.17 (Log majorization). The log-majorization $\log(\boldsymbol{x}) \prec \log(\boldsymbol{y})$ implies the submajorization $\boldsymbol{x} \prec_w \boldsymbol{y}$.

Proof. This result follows immediately from the exponential example (4.6). ■

This corollary connects nicely with the Weyl majorant theorem (Theorem 4.2). We maintain the same notation for the decreasingly ordered eigenvalues and singular values.

Corollary 4.18 (Eigenvalue and singular value majorization). For every matrix $\boldsymbol{A} \in \mathbb{M}_n$, we have the submajorization relation $\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}) \prec_w \boldsymbol{\sigma}(\boldsymbol{A})$.

4.3.5 Isotone functions: Scalar examples

Next, let us consider the application of Theorem 4.13 in the special case when the function is scalar-valued ($m = 1$).

Corollary 4.19 (Scalar-valued isotone functions: Sufficient conditions). Suppose that the function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is convex and permutation invariant. Then it is isotone. That is,

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{implies} \quad \Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y}).$$

Example 4.20 (Isotonicity: Coordinate sums). If a function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then the function $\Phi(\boldsymbol{x}) = \sum_{i=1}^n \varphi(x_i)$ is isotone. Here are some examples.

- **Negative entropy.** This construction allows us to check that the negative entropy function is isotone.

  $$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{implies} \quad \operatorname{negent}(\boldsymbol{x}) \le \operatorname{negent}(\boldsymbol{y}),$$

  where $\operatorname{negent}(\boldsymbol{x}) := \sum_i x_i \log(x_i)$ is the negative entropy.
- **ℓ_p-norms.** Each ℓ_p norm ($p \ge 1$) is an isotone function.

  $$\boldsymbol{x} \prec \boldsymbol{y} \quad \text{implies} \quad \|\boldsymbol{x}\|_p \le \|\boldsymbol{y}\|_p.$$

These functions can also be viewed as tracing over the entries of the construction (4.5). Indeed, convexity and permutation covariance remain after taking the trace. ▪

Unfortunately, all the above examples rely on convexity. In the next section, we will see that convexity is not required to attain isotonicity.

4.4 Schur "convexity"

We present the necessary and sufficient conditions for isotonicity for a real-valued, differentiable function.

Definition 4.21 (Schur "convex" function). Consider an isotone real-valued function $\Phi : \mathbb{R}^n \to \mathbb{R}$. That is, the majorization relation $\boldsymbol{x} \prec \boldsymbol{y}$ implies the inequality $\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y})$. Then we say the function Φ is *Schur convex* or *S-convex*.

Warning: Despite the name, Schur "convex" functions need not be convex! ▪

Theorem 4.22 (Schur "convexity": Characterization). Suppose that the function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is differentiable on its convex domain. Then the following statements are equivalent.

1. The function Φ is isotone (Schur-convex).

2. The function Φ is permutation invariant. In addition, for all pairs (i, j) of coordinate indices and all vectors $\boldsymbol{x} \in \operatorname{dom}(\Phi)$,

$$(x_i - x_j)\left(\partial_i \Phi(\boldsymbol{x}) - \partial_j \Phi(\boldsymbol{x})\right) \geq 0. \tag{4.7}$$

For smooth functions, isotonicity is characterized locally by the derivative. This theorem provides more examples of isotone functions.

Example 4.23 (Schur "convex" functions). The following real-valued functions $\Phi : \mathbb{R}^n \to \mathbb{R}$ are isotone:

- **Product.** The negative product of entries of a positive vector is isotone.

$$\Phi(\boldsymbol{x}) = -\prod\nolimits_{i=1}^{n} x_i.$$

- **Variance.** The variance function is isotone.

$$\Phi(\boldsymbol{x}) = \frac{1}{n}\sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2,$$

  where $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean.

These examples are harder to verify directly from the definition of isotonicity, but the theorem makes short work of them. ∎

Let us prove the theorem.

Proof of Theorem 4.22. **(2 ⇒ 1).** We begin with the reverse implication. Consider vectors $\boldsymbol{u}, \boldsymbol{x}$ that satisfy the majorization relation $\boldsymbol{u} \prec \boldsymbol{x}$. To check isotonicity of the function Φ, we want to prove the inequality $\Phi(\boldsymbol{u}) \leq \Phi(\boldsymbol{x})$. In Lecture 3, we showed that majorization can be expressed using a sequence of T-transforms:

$$\boldsymbol{u} = \boldsymbol{T}_n \dots \boldsymbol{T}_1 \boldsymbol{x}.$$

Recall that a T-transform is a convex combination of the identity and a transposition:

$$\boldsymbol{T} = \tau \boldsymbol{I} + (1 - \tau)\boldsymbol{Q} \quad \text{for } \tau \in [0, 1] \text{ and } \boldsymbol{Q} \text{ a transposition.}$$

Therefore, it suffices to obtain the inequality $\Phi(\boldsymbol{u}) \leq \Phi(\boldsymbol{x})$ for a single transition $\boldsymbol{u} = \boldsymbol{T}\boldsymbol{x}$.

Without loss of generality, permutation invariance allows us to assume that the T-transform $\boldsymbol{T}$ averages the first two coordinates x_1, x_2 only. We can write explicitly

$$\boldsymbol{u} = \Big((1 - s)x_1 + sx_2, sx_1 + (1 - s)x_2, \dots, x_n\Big) \quad \text{for } s \in [0, 0.5]. \tag{4.8}$$

The key idea is to interpolate from the vector $\boldsymbol{x}$ to the vector $\boldsymbol{u}$. Define the function

$$\boldsymbol{x}(\tau) = \Big((1 - \tau)x_1 + \tau x_2, \tau x_1 + (1 - \tau)x_2, \dots, x_n\Big) \quad \text{for } \tau \in [0, s].$$

Note that $\boldsymbol{x}(0) = \boldsymbol{x}$ and $\boldsymbol{x}(s) = \boldsymbol{u}$. By the fundamental theorem of calculus (and the assumption that the function Φ is differentiable),

$$\begin{aligned}
\Phi(\boldsymbol{u}) - \Phi(\boldsymbol{x}) &= \int_0^s \frac{\mathrm{d}}{\mathrm{d}\tau}\left[\Phi(\boldsymbol{x}(\tau))\right] \mathrm{d}\tau \\
&= \int_0^s (x_2 - x_1)\Big(\partial_1 \Phi\big(\boldsymbol{x}(\tau)\big) - \partial_2 \Phi\big(\boldsymbol{x}(\tau)\big)\Big) \,\mathrm{d}\tau \\
&= -\int_0^s \frac{(x_2(\tau) - x_1(\tau))}{1 - 2\tau}\Big(\partial_1 \Phi\big(\boldsymbol{x}(\tau)\big) - \partial_2 \Phi\big(\boldsymbol{x}(\tau)\big)\Big) \,\mathrm{d}\tau \\
&\leq 0.
\end{aligned}$$

The second equality is the chain rule, and the last inequality is an assumption.

(1 ⇒ 2). Now we confirm the forward implication. Fix any vector $\boldsymbol{x}$. Observe that $\boldsymbol{x}$ is majorized by any permutation $\boldsymbol{P}\boldsymbol{x}$ because they have the identical sorted entries. Therefore, we can write down the two-sided majorization relation

$$\boldsymbol{x} \prec \boldsymbol{P}\boldsymbol{x} \prec \boldsymbol{P}^{-1}(\boldsymbol{P}\boldsymbol{x}) = \boldsymbol{x}.$$

Apply Schur convexity to obtain permutation invariance of the function Φ:

$$\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{P}\boldsymbol{x}) \le \Phi(\boldsymbol{x}).$$

That is, $\Phi(\boldsymbol{P}\boldsymbol{x}) = \Phi(\boldsymbol{x})$.

To establish the differential condition (4.7), for every vector in the domain $\boldsymbol{x} \in \operatorname{dom}(\Phi)$, we consider the vector

$$\boldsymbol{u}_{ij}(s) := \boldsymbol{T}_{ij}\boldsymbol{x} = \Big(\dots, (1-s)x_i + sx_j, \dots, sx_i + (1-s)x_j, \dots \Big).$$

This is analogous to (4.8), but the T-transform $\boldsymbol{T}_{ij}$ acts on the (i, j) pair of indices. Take the derivative as $s \to 0$ to obtain

$$\begin{aligned} 0 \ge \lim_{s \to 0} \frac{\Phi(\boldsymbol{u}_{ij}(s)) - \Phi(\boldsymbol{x})}{s} &= \frac{\mathrm{d}}{\mathrm{d}s}[\Phi(\boldsymbol{u}_{ij}(s))] \\ &= (x_j - x_i)\Big(\partial_i \Phi(\boldsymbol{x}) - \partial_j \Phi(\boldsymbol{x})\Big). \end{aligned}$$

The first inequality holds because of the majorization relation between the vectors and the Schur convexity of the function Φ. Rearrange to obtain the advertised result. ∎

Notes

The material in this lecture is adapted from Bhatia [Bha97, Chap. II], which is based on Ando's vision [And89].

Lecture bibliography

[And89] T. Ando. "Majorization, doubly stochastic matrices, and comparison of eigenvalues". In: *Linear Algebra and its Applications* 118 (1989), pages 163–248.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

5. Birkhoff and von Neumann

Date: 18 January 2022 **Scribe:** Jagannadh Boddapati

In this lecture, we study the geometry of DS_n, the set of $n \times n$ doubly stochastic matrices. We establish the classic result of Birkhoff & von Neumann, which states that the set of doubly stochastic matrices can be expressed as the convex hull of the permutation matrices. We use this result to prove the von Neumann trace theorem, which plays a basic role in understanding unitary invariant norms and convex trace functions.

Agenda:
1. Doubly stochastic matrices
2. The Birkhoff–von Neumann theorem
3. Minkowski theorem
4. Proof of Birkhoff theorem
5. The von Neumann trace theorem

5.1 Doubly stochastic matrices

In Lecture 3, we introduced the doubly stochastic matrices. We saw that doubly stochastic matrices act on a vector by averaging its entries. We also saw that doubly stochastic matrices arise in connection with majorization.

Definition 5.1 (Doubly stochastic matrices). A matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ is called *doubly stochastic* if it satisfies the following properties.

1. **Positive.** For all $\boldsymbol{x} \in \mathbb{R}^n$, the relation $\boldsymbol{x} \geq \boldsymbol{0}$ implies $\boldsymbol{S}\boldsymbol{x} \geq \boldsymbol{0}$. It is equivalent to say that all the entries of the matrix are positive: $s_{ij} \geq 0$ for all $i, j = 1, \dots, n$.
2. **Trace preserving.** For all $\boldsymbol{x} \in \mathbb{R}^n$, we have $\operatorname{tr}(\boldsymbol{S}\boldsymbol{x}) = \operatorname{tr}(\boldsymbol{x})$. This is equivalent to say that each column adds up to one.
3. **Unital.** $\boldsymbol{S}\boldsymbol{1} = \boldsymbol{1}$. It is equivalent to say that each row adds up to one.

Definition 5.2 (Birkhoff polytope). The set DS_n collects all of the $n \times n$ doubly stochastic matrices:

$$\mathsf{DS}_n := \left\{ \boldsymbol{S} \in \mathbb{R}^{n \times n} : \boldsymbol{S} \text{ is doubly stochastic} \right\}.$$

As we will discuss, the set DS_n is a convex polytope known as the *Birkhoff polytope*.

The Birkhoff polytope arises in the study of majorization.

Theorem 5.3 (Doubly stochastic matrices: Characterization). If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then $\boldsymbol{x} \prec \boldsymbol{y}$ holds if and only if $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$ for some $\boldsymbol{S} \in \mathsf{DS}_n$.

That is, when $\boldsymbol{x} \prec \boldsymbol{y}$, the entries of $\boldsymbol{x}$ are "more average" than the entries of $\boldsymbol{y}$.

5.1.1 Properties of doubly stochastic matrices

In this section, we are going to explore the geometry of the set of doubly stochastic matrices DS_n in more detail. We make the following observations.

1. **Polyhedron.** The set DS_n is a closed polyhedron, i.e., a finite intersection of closed halfspaces. Indeed, the positivity constraints restrict each entry to be

positive, which is a halfspace constraint. Each equality constraint in Definition 5.1 can be written as two halfspace constraints. Being a finite intersection of closed halfspaces, the set DS_n is convex and closed. An example of a polyhedron appears in Figure 5.1.

2. **Bounded.** All the entries of a doubly stochastic matrix lie between 0 and 1. Hence the set DS_n is bounded. In particular, DS_n is compact.
3. **Polytope.** The set DS_n is a polytope, i.e., is the convex hull of finitely many points. This claim follows from a major result in convex geometry known as the Weyl–Minkowski theorem, which states that every *bounded* polyhedron is a polytope. See Figure 5.2 for an example of a polytope. Please refer to [Bha97] for more details on doubly stochastic matrices and the Weyl–Minkowski theorem.

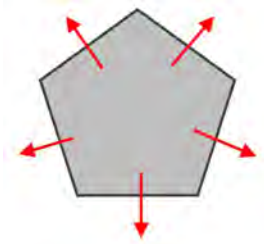

Figure 5.1 Example of a polyhedron formed by finite intersection of closed halfspaces.

Figure 5.2 Example of a polytope constructed as a convex-hull of five points shown in red.

Note that the Birkhoff polytope DS_n contains all the $n \times n$ permutation matrices. This observation leads us to explore what role the permutation matrices play in the structure of DS_n.

Exercise 5.4 (The convex hull of permutations). Deduce that the convex hull of the permutation matrices is a subset of the doubly stochastic matrices:

$$\operatorname{conv}\left\{\boldsymbol{P} \in \mathbb{R}^{n\times n} : \boldsymbol{P} \text{ is a permutation matrix}\right\} \subseteq \mathsf{DS}_n.$$

We will prove that this inclusion can be upgraded to a set equality.

5.2 The Birkhoff–von Neumann theorem

In this section, we will discuss the Birkhoff–von Neumann theorem which relates convex hull of permutation matrices to the set of doubly stochastic matrices DS_n. To prepare for our treatment, we will need some additional definitions.

Definition 5.5 (Extreme point). Let $\mathsf{K} \subseteq \mathbb{R}^d$ be a (nonempty) convex set. An *extreme point* of K is a point $\boldsymbol{x} \in \mathsf{K}$ such that $\boldsymbol{y}, \boldsymbol{z} \in \mathsf{K}$ and $\boldsymbol{x} = \frac{1}{2}(\boldsymbol{y} + \boldsymbol{z})$ together imply that $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{z}$.

In other words, we cannot represent an extreme point $\boldsymbol{x}$ as an average of *distinct* points in K. See Figure 5.3 for examples of extreme points.

Now, we state the Birkhoff–von Neumann theorem, which originally appeared in the paper [Bir46].

Theorem 5.6 (Birkhoff 1946; von Neumann 1953). The extreme points (i.e., vertices) of DS_n are precisely the permutation matrices. In particular, DS_n can be written as convex hull of permutation matrices,

$$\mathsf{DS}_n = \operatorname{conv}\left\{\boldsymbol{P} \in \mathbb{R}^{n\times n} : \boldsymbol{P} \text{ is a permutation matrix}\right\}$$

To prove this theorem, we must demonstrate that the permutation matrices compose the full set of extreme points of the Birkhoff polytope. First, we will present an important result from convex geometry that shows that the extreme points play a key role in the structure of convex sets.

Exercise 5.7 (Number of permutations). How many permutations suffice to express a doubly stochastic matrix in DS_n? **Hint**: Use the Carathéodory theorem.

A nice corollary of Theorem 5.6 is an independent proof of the geometric characterization of majorization from the last lecture.

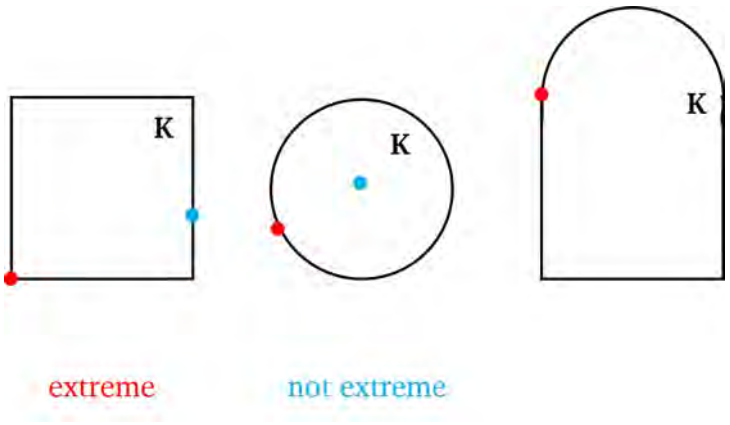

Figure 5.3 (Examples of extreme points). Points shown in red are extreme while the points shown in blue are not extreme.

Corollary 5.8 (Permutation matrix). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The condition $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{y}$ for $\boldsymbol{S} \in \mathsf{DS}_n$ is equivalent to, $\boldsymbol{x} \in \operatorname{conv}\{\boldsymbol{P}\boldsymbol{y} : \boldsymbol{P} \text{ is a permutation matrix}\}$.

To understand the meaning of this corollary, it is fruitful to define a new class of convex polytopes.

Definition 5.9 (Permutahedron). Let $\boldsymbol{y} \in \mathbb{R}^n$. A *permutahedron* is a polytope of the form

$$\operatorname{conv}\{\boldsymbol{P}\boldsymbol{y} : \boldsymbol{P} \in \mathbb{R}^{n \times n} \text{ a permutation matrix}\}.$$

As a consequence, we recognize that the set $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \prec \boldsymbol{y}\}$ of vectors that are majorized by $\boldsymbol{y}$ forms a convex polytope. This fact assigns a geometric meaning to the majorization relation.

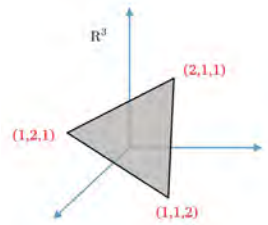

Figure 5.4 Example of a permutahedron in $\mathbb{R}^3$.

Exercise 5.10 (Permutahedra). By construction, the permutahedron is a polytope. By direct argument, show that a permutahedron is also a bounded polyhedron.

In addition to their role in majorization, permutahedra also arise from matching problems, network science, transportation problems, and convex optimization.

5.3 The Minkowski theorem on extreme points

We now present the Minkowski theorem, which explains how to represent a compact, convex set in terms of its extreme points. We will use this result to prove the Birkhoff theorem in the next section.

Theorem 5.11 (Minkowski). Let $\mathsf{K} \subseteq \mathbb{R}^d$ be a nonempty, compact, and convex set. Then K is the convex hull of the set of its extreme points:

$$\mathsf{K} = \operatorname{conv}(\operatorname{ext}(\mathsf{K})),$$

where $\operatorname{ext}(\mathsf{K})$ is the set of extreme of points of K.

Proof. We shall prove this theorem by induction on the dimension of K. If $\dim(\mathsf{K}) = 0$, then $\mathsf{K} = \{\boldsymbol{x}\}$ is a singleton and the result follows. Assume that the result holds true

The dimension of a convex set K is defined as $\dim \mathsf{K} := \dim \operatorname{aff}(\mathsf{K})$, the dimension of the smallest affine space containing the set K.

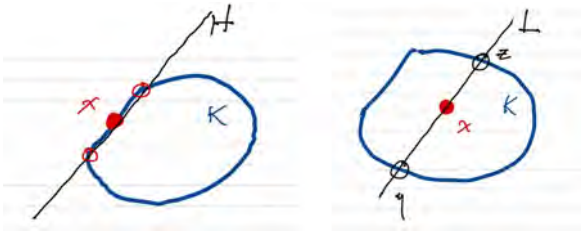

Figure 5.5 (Two cases for proof on Minkowski theorem by induction). The left figure demonstrates the case when $\boldsymbol{x}$ belongs to boundary(K). The right figure demonstrates the case when $\boldsymbol{x}$ does not belong to boundary(K).

for all nonempty, compact, convex sets with dimension $d - 1$.

Now, let us consider a nonempty, compact, convex set K with dimension d. We must show that every point $\boldsymbol{x} \in \mathsf{K}$ can be represented as a convex combination of extreme points of K. There are two cases as shown in Figure 5.5.

First, consider the case when $\boldsymbol{x} \in \text{boundary}(\mathsf{K})$. Then we can separate $\boldsymbol{x}$ weakly from K via a hyperplane H. Then the intersection of K and H is a nonempty, compact, and convex set. Hence, $\dim(\mathsf{K} \cap \mathsf{H})$ is less than or equal to $d - 1$, so induction applies. We deduce that

$$\boldsymbol{x} \in \mathsf{K} \cap \mathsf{H} = \operatorname{conv} \operatorname{ext}(\mathsf{K} \cap \mathsf{H}) \subseteq \operatorname{conv} \operatorname{ext}(\mathsf{K}).$$

We have used the fact that the extreme points of a face of a convex set are also extreme points of the set.

Second, consider the case when $\boldsymbol{x} \notin \text{boundary}(\mathsf{K})$. Choose a line L containing $\boldsymbol{x}$. Since K is compact and convex, this line hits the boundary at two points, say $\boldsymbol{y}, \boldsymbol{z}$. Therefore, we can write $\boldsymbol{x}$ as a convex combination of $\boldsymbol{y}, \boldsymbol{z}$:

$$\boldsymbol{x} = \tau \boldsymbol{y} + \bar{\tau} \boldsymbol{z} \quad \text{where } \tau \in [0, 1] \text{ and } \bar{\tau} = 1 - \tau.$$

Thus,

$$\boldsymbol{x} \in \operatorname{conv}\{\boldsymbol{y}, \boldsymbol{z}\} \subseteq \operatorname{conv} \operatorname{ext}(\mathsf{K}).$$

Indeed, by the induction, each of $\boldsymbol{y}, \boldsymbol{z} \in \operatorname{conv} \operatorname{ext}(\mathsf{K})$. ∎

Please refer to [Sch14] for more details on this theorem and its context. Let us discuss a few consequences and generalization of the Minkowski theorem.

Corollary 5.12 (Extreme point). Every nonempty, convex, compact set has an extreme point.

Exercise 5.13 (Bauer maximum principle). Show that every linear functional on a convex, compact set attains its maximum at an extreme point. In particular, if the maximizer is *unique*, it must be an extreme point of the set.

The Minkowski theorem has a generalization to infinite-dimensional spaces, a result that has far-reaching implications in mathematical analysis.

Fact 5.14 (Krein–Milman). Suppose X is a locally convex topological vector space (for example, a normed space). Suppose that K is a compact and convex subset of X. Then K is equal to the *closed* convex hull of its extreme points:

$$\mathsf{K} = \overline{\operatorname{conv}}(\operatorname{ext}(\mathsf{K})).$$

Moreover, if $\mathsf{B} \subseteq \mathsf{K}$, then K is equal to the closed convex hull of B if and only if $\mathrm{ext}(\mathsf{K}) \subseteq \mathrm{cl}(\mathsf{B})$, where $\mathrm{cl}(\mathsf{B})$ is the closure of B. ∎

5.4 Proof of Birkhoff theorem

In this section, we give a proof of the Birkhoff theorem, which states that the extreme points (that is, the vertices) of DS_n are precisely the permutation matrices. The first step is an exercise.

Exercise 5.15 (Permutations and DS_n). Show that each permutation matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ is an extreme point of DS_n. **Hint:** Argue that there is a linear functional that achieves a *unique* maximum on DS_n at $\boldsymbol{P}$. Invoke the Bauer maximum principle.

The following notion will be helpful in the proof that the permutation matrices are the only possible extreme points of the Birkhoff polytope.

Definition 5.16 (Perturbation). Let $\boldsymbol{S} \in \mathsf{DS}_n$ belong to the set of doubly stochastic matrices. A *perturbation* $\boldsymbol{E} \in \mathbb{R}^{n \times n}$ of the matrix $\boldsymbol{S}$ is a matrix with the property that $\boldsymbol{S} \pm \boldsymbol{E} \in \mathsf{DS}_n$.

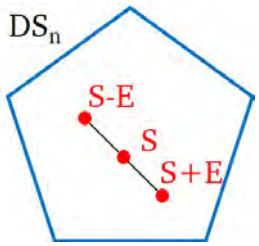

Figure 5.6 Example of perturbation.

Exercise 5.17 (Extreme point and perturbation). Show that $\boldsymbol{S}$ is an extreme point of DS_n if and only if $\boldsymbol{S}$ admits no perturbation $\boldsymbol{E}$, except the zero matrix. In particular, if $\boldsymbol{S}$ admits a nontrivial perturbation, then $\boldsymbol{S}$ is not an extreme point.

Proof of Theorem 5.6. We already know that the permutation matrices are extreme points of the Birkhoff polytope. We will argue that the permutation matrices compose the full set of extreme points. An application of Minkowski's theorem completes the argument.

To that end, let us suppose that $\boldsymbol{S} \in \mathsf{DS}_n$ is not a permutation. We will produce a nonzero perturbation $\boldsymbol{E}$. Therefore, $\boldsymbol{S}$ is not an extreme point of DS_n.

Observe that every doubly stochastic matrix that is not a permutation contains at least one entry that is not an integer. We will find a "cycle" consisting of nonintegral entries. First, find an entry in row i_1 and column j_1 such that $s_{i_1 j_1}$ lies between 0 and 1. Now, select another entry in *row* i_1 such that $s_{i_1 j_2}$ lies strictly between 0 and 1. We can do this because the entries in the row have to add up to 1. Find another entry in *column* j_2 such that s_{i_2, j_2} lies strictly between 0 and 1. We keep moving horizontally along the rows and vertically along the columns in sequence until we encounter an index pair that we have already seen. Figure 5.8 illustrates this process. This process have to terminate because there are only finite number of positions in the matrix.

Among all such sequences, we choose the one with fewest steps. This sequence must be a cycle:

$$(i_1, j_1) = (i_r, j_r).$$

Note that this cycle must have even number of steps. Indeed, in order to complete a cycle, we need to move from a row and a column in sequence. An odd number of steps in a given row or in a given column can be combined to form a shorter connection.

Given the indices in the cycle, we can find $\varepsilon > 0$ such that

$$s_{i_a j_a} \pm \varepsilon \in (0, 1) \quad \text{for each index } a.$$

We can construct a nontrivial perturbation $\boldsymbol{E}$ whose nonzero entries are

$$e_{i_1 j_1} = \varepsilon; \quad e_{i_1 j_2} = -\varepsilon; \quad e_{i_2 j_2} = \varepsilon; \quad e_{i_3 j_2} = -\varepsilon; \quad \text{etc.}$$

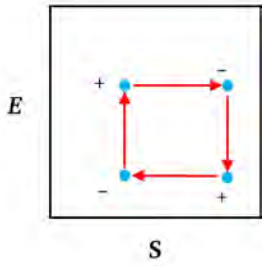

Figure 5.7 Perturbation matrix $\boldsymbol{E}$ used in proving Birkhoff theorem.

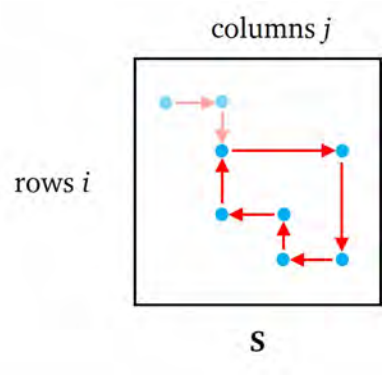

Figure 5.8 (Reordering indices to prove Birkhoff theorem). We move from one index where the entry is nonintegral to other index where the entry is nonintegral along a row and a column alternatively. We repeat the process until we end up at the same index. We consider only the shortest path and discard any indices that attach to the loop but will not close the loop (indicated in transparent colors).

Then the row sums and column sums of $\boldsymbol{E}$ are equal to zero. We can conclude that

$$\boldsymbol{S} \pm \boldsymbol{E} \in \mathsf{DS}_n.$$

We can conclude that $\boldsymbol{S}$ is not an extreme point. ∎

5.5 The Richter trace theorem

We may now prove the following theorem of Richter on a maximization problem that arises in the study of matrix traces. This is a close relative of an older result due to von Neumann (Exercise 5.20). The proof is due to Mirsky [Mir59].

Theorem 5.18 (Richter trace theorem). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$. Then

$$\max\{\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) : \boldsymbol{U} \in \mathbb{M}_n \text{ is unitary}\} = \sum\nolimits_{i=1}^{n} \lambda_i^{\downarrow}(\boldsymbol{A})\, \lambda_i^{\downarrow}(\boldsymbol{B}).$$

In other words, the theorem solves a matching problem. What is the best way we can rotate $\boldsymbol{A}$ to align it with $\boldsymbol{B}$? Part of the assertion is that the maximum is attained.

Proof. We prove the theorem by establishing upper and lower bounds on the trace. First, let us introduce the eigenvalue decompositions of $\boldsymbol{A}$ and $\boldsymbol{B}$.

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{Q}_{\boldsymbol{A}} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Q}_{\boldsymbol{A}}^* \qquad \text{where } \boldsymbol{\lambda} = \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}); \\ \boldsymbol{B} &= \boldsymbol{Q}_{\boldsymbol{B}} \operatorname{diag}(\boldsymbol{\mu}) \boldsymbol{Q}_{\boldsymbol{B}}^* \qquad \text{where } \boldsymbol{\mu} = \boldsymbol{\mu}^{\downarrow}(\boldsymbol{B}). \end{aligned}$$

We can compute the lower bound of the trace by choosing $\boldsymbol{U}_0 = \boldsymbol{Q}_{\boldsymbol{A}} \boldsymbol{Q}_{\boldsymbol{B}}^*$. Indeed,

$$\begin{aligned} \max\nolimits_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) &\geq \operatorname{tr}(\boldsymbol{U}_0^*\boldsymbol{A}\boldsymbol{U}_0\boldsymbol{B}) \\ &= \operatorname{tr}(\operatorname{diag}(\boldsymbol{\lambda}) \operatorname{diag}(\boldsymbol{\mu})) = \sum\nolimits_{i}^{n} \lambda_i \mu_i. \end{aligned}$$

It remains to show that this lower bound is indeed the largest possible value.

We can compute the upper bound of the trace as follows.

Recall that $\boldsymbol{\delta}_i$ is the ith standard basis vector.

$$\begin{aligned}\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) &= \max_{\boldsymbol{U}} \operatorname{tr}\left(\boldsymbol{U}^* \operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{U} \operatorname{diag}(\boldsymbol{\mu})\right) \\ &= \max_{\boldsymbol{U}} \sum\nolimits_{i=1}^{n} \boldsymbol{\delta}_i^*\boldsymbol{U}^* \operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{U} \operatorname{diag}(\boldsymbol{\mu})\boldsymbol{\delta}_i \\ &= \max_{\boldsymbol{U}} \sum\nolimits_{i}^{n} (\boldsymbol{U}\boldsymbol{\delta}_i)^* \operatorname{diag}(\boldsymbol{\lambda})(\boldsymbol{U}\boldsymbol{\delta}_i)\mu_i\end{aligned}$$

Here, the first equality is obtained by absorbing $\boldsymbol{Q}_{\boldsymbol{A}}, \boldsymbol{Q}_{\boldsymbol{B}}$ into $\boldsymbol{U}$. The second equality comes from the fact that the trace is the sum of diagonal entries. Next, we represent $\operatorname{diag}(\boldsymbol{\lambda})$ as a sum of rank-one matrices:

$$\operatorname{diag}(\boldsymbol{\lambda}) = \sum\nolimits_j \lambda_j \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*.$$

Combine the last two displays and simplify:

$$\begin{aligned}\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) &= \max_{\boldsymbol{U}} \sum\nolimits_{i,j=1}^{n} \mu_i\lambda_j (\boldsymbol{U}\boldsymbol{\delta}_i)^* \boldsymbol{\delta}_j \boldsymbol{\delta}_j^* (\boldsymbol{U}\boldsymbol{\delta}_i) \\ &= \max_{\boldsymbol{U}} \sum\nolimits_{i,j=1}^{n} \mu_i\lambda_j |\boldsymbol{\delta}_j^*\boldsymbol{U}\boldsymbol{\delta}_i|^2.\end{aligned}$$

Note that we have exposed the squared magnitudes of the entries of the unitary matrix, which compose an orthostochastic matrix $\boldsymbol{S} \in \mathsf{DS}_n$ with entries $s_{ij} = |u_{ij}|^2$.

We see that the maximum only increases if we pass to the full set of doubly stochastic matrices:

$$\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) \le \max_{\boldsymbol{S}\in\mathsf{DS}_n} \sum\nolimits_{i,j=1}^{n} \mu_i\lambda_j s_{ij}.$$

Invoke Bauer's maximum principle to see that this linear function attains its maximum at an extreme point of the set. Then use Birkhoff's theorem to recognize that the extreme points of DS_n are precisely the permutation matrices $\boldsymbol{P} \in \mathbb{R}^{n\times n}$. Thus,

$$\begin{aligned}\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) &\le \max_{\boldsymbol{P}} \sum\nolimits_{i,j=1}^{n} \mu_i\lambda_j p_{ij} \\ &= \max_{\pi} \sum\nolimits_{i=1}^{n} \lambda_{\pi(i)}\mu_i = \sum\nolimits_{i}^{n} \lambda_i\mu_i.\end{aligned}$$

We have passed from the permutation matrix to the associated permutation: $\pi(i) = j$ if and only if $p_{ij} = 1$. Last, we invoke the Chebyshev rearrangement inequality to see that the maximum is attained when the two vectors are both arranged in decreasing order.

Thus, both upper bound and lower bound on the trace are identical, and we have established the result. ∎

Here are some related results that often prove useful.

Exercise 5.19 (Richter trace theorem). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$. Prove that

$$\min\{\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{B}) : \boldsymbol{U} \in \mathbb{M}_n \text{ is unitary}\} = \sum\nolimits_{i=1}^{n} \lambda_i^{\downarrow}(\boldsymbol{A})\, \lambda_i^{\uparrow}(\boldsymbol{B}).$$

Exercise 5.20 (von Neumann trace theorem). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$. Prove that

$$\max\{\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}) : \boldsymbol{U}, \boldsymbol{V} \in \mathbb{M}_n \text{ are unitary}\} = \sum\nolimits_{i=1}^{n} \sigma_i^{\downarrow}(\boldsymbol{A})\, \sigma_i^{\downarrow}(\boldsymbol{B}).$$

Notes

Birkhoff and von Neumann independently established the result that shares their name. There is an influential geometric proof of the result due to Hoffmann & Wielandt [HW53], which also connects the result with perturbation theory for eigenvalues of a normal matrix. The material on convex geometry is adapted from Barvinok's book [Bar02] and from Schneider's treatise [Sch14]. We have extracted this direct proof of Birkhoff's theorem from a note by Glenn Hurlbert, which appears in [Hur10]. The proof of Richter's trace theorem is drawn from Mirsky's work [Mir59].

Lecture bibliography

[Bar02] A. Barvinok. *A course in convexity*. American Mathematical Society, Providence, RI, 2002. DOI: 10.1090/gsm/054.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bir46] G. Birkhoff. "Three observations on linear algebra". In: *Univ. Nac. Tacuman, Rev. Ser. A* 5 (1946), pages 147–151.

[HW53] A. J. Hoffman and H. W. Wielandt. "The variation of the spectrum of a normal matrix". In: *Duke J. Math.* 20 (1953), pages 37–39.

[Hur10] G. H. Hurlbert. *Linear optimization*. The simplex workbook. Springer, New York, 2010. DOI: 10.1007/978-0-387-79148-7.

[Mir59] L. Mirsky. "On the trace of matrix products". In: *Mathematische Nachrichten* 20.3-6 (1959), pages 171–174. DOI: 10.1002/mana.19590200306.

[Sch14] R. Schneider. *Convex bodies: the Brunn–Minkowski theory*. 151. Cambridge university press, 2014. DOI: 10.1017/CBO9781139003858.

6. Unitarily Invariant Norms

Date: 20 January 2022 **Scribe:** Elvira Moreno

In this lecture, we first introduce symmetric gauge functions, an important family of norms over $\mathbb{R}^n$ that have the property of being sign and permutation invariant. As part of our analysis, we develop duality theory for symmetric gauge functions and show that they satisfy some widely used norm inequalities. We then discuss a closely related family of matrix norms that are invariant under coordinate changes. By building on their connection to symmetric gauge functions, we develop duality theory for this important family of norms and provide a generalization of the well known Hölder inequality.

Agenda:
1. Symmetric gauge functions
2. Duality for symmetric gauge functions
3. Unitarily invariant norms
4. Duality for unitarily invariant norms

6.1 Symmetric gauge functions

We begin by introducing an important class of norms on $\mathbb{R}^n$.

Definition 6.1 (Symmetric gauge function). A *symmetric gauge function* is a map $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ that satisfies four properties:

1. **Norm:** Φ is a norm.
2. **Permutation invariance:** The equation $\Phi(\boldsymbol{P}\boldsymbol{x}) = \Phi(\boldsymbol{x})$ holds for any vector $\boldsymbol{x} \in \mathbb{R}^n$ and any permutation matrix $\boldsymbol{P} \in \mathbb{M}_n$.
3. **Sign invariance:** The equation $\Phi(\boldsymbol{D}\boldsymbol{x}) = \Phi(\boldsymbol{x})$ holds for any $\boldsymbol{x} \in \mathbb{R}^n$ and for any $n \times n$ diagonal matrix of the form $\boldsymbol{D} = \operatorname{diag}(\pm 1, \dots, \pm 1)$.
4. **Normalization:** Φ is normalized as $\Phi((1, 0, \dots, 0)) = 1$.

Recall that a norm on $\mathbb{R}^n$ is a function $\Phi : \mathbb{R}^n \to R$ satisfying the following three properties:

1. **Positive definiteness.** $\Phi(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, and $\Phi(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = 0$.
2. **Positive homogeneity.** $\Phi(\alpha\boldsymbol{x}) = |\alpha|\Phi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.
3. **Triangle inequality.** $\Phi(\boldsymbol{x} + \boldsymbol{y}) \leq \Phi(\boldsymbol{x}) + \Phi(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

In lecture 4, we defined the family of isotone functions on $\mathbb{R}^n$ as maps that respect the majorization preorder on $\mathbb{R}^n$, and we proved that convexity and permutation invariance are sufficient conditions for isotonicity. It follows from properties (1) and (2) that symmetric gauge functions are isotone, so they constitute a family of norms on $\mathbb{R}^n$ that preserve the majorization preorder.

Definition 6.2 (Symmetric gauge function on $\mathbb{C}^n$). We can extend the concept of a symmetric gauge function on $\mathbb{R}^n$ to the complex vector space $\mathbb{C}^n$ via

$$\Phi(\boldsymbol{z}) := \Phi(|\boldsymbol{z}|) \text{ for } \boldsymbol{z} \in \mathbb{C}^n,$$

where $|\boldsymbol{z}| := (|z_1|, \dots, |z_n|)$ is the entry-wise modulus of the vector $\boldsymbol{z}$.

Let us consider some examples of symmetric gauge functions.

Example 6.3 (ℓ_p norms). For a vector $x \in \mathbb{R}^n$, define

$$\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \text{ for each } p \in [1, \infty);$$
$$\|\boldsymbol{x}\|_\infty := \max_i |x_i|.$$

The functions $\|\cdot\|_p$ for $p \in [1, \infty]$ define norms on $\mathbb{R}^n$, commonly known as ℓ_p norms. It is easy to check that ℓ_p norms are symmetric gauge functions. ▪

Example 6.4 (Ky Fan norm). Fix $k \in \{1, \dots, n\}$. For each vector $x \in \mathbb{R}^n$, define

$$\|\boldsymbol{x}\|_{(k)} := \max_{|\mathsf{I}|=k} \sum\nolimits_{i \in \mathsf{I}} |x_i|.$$

That is, $\|\boldsymbol{x}\|_{(k)}$ is the sum of the k largest entries of the vector $|\boldsymbol{x}|$. The functions $\|\cdot\|_{(k)}$ are norms on $\mathbb{R}^n$, which are known by the name of *Ky Fan norms*. It is also easy to check that these norms are symmetric gauge functions. ▪

There are many other norms on $\mathbb{R}^n$ that are symmetric gauge functions. For instance, one could form combinations of ℓ_p and Ky Fan norms, or one could form weighted sums of the ordered entries of the vector.

Proposition 6.5 (Monotonicity). If $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is a symmetric gauge function, then $|\boldsymbol{x}| \le |\boldsymbol{y}|$ implies that $\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Recall that $\boldsymbol{x} \le \boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is interpreted entrywise and that $|\boldsymbol{x}| := (|x_1|, \dots, |x_n|)$ denotes the entrywise modulus.

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be such that $|\boldsymbol{x}| \le |\boldsymbol{y}|$. By sign invariance of Φ, we can assume that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are positive and that $0 \le x_i = t_i y_i$ for some values $t_i \in [0, 1]$ for each $i = 1, \dots, n$. By permutation invariance and iteration, it suffices to check the case where $t_2 = t_3 = \cdots = t_n = 1$. Indeed, if the result holds for the case were $\boldsymbol{x}$ and $\boldsymbol{y}$ differ by a single entry, it can be obtained for the general case by applying it to one coordinate at a time. Write $x_1 = t y_1$ for $t \in [0, 1]$. Then

$$\begin{aligned}
\Phi((x_1, x_2, \dots, x_n)) &= \Phi((t y_1, y_2, \dots, y_n)) \\
&= \Phi\left(\left(\frac{1+t}{2} y_1 - \frac{1-t}{2} y_1, \frac{1+t}{2} y_2 + \frac{1-t}{2} y_2, \dots, \frac{1+t}{2} y_n + \frac{1-t}{2} y_n\right)\right) \\
&= \Phi\left(\frac{1+t}{2}(y_1, y_2, \dots, y_n) + \frac{1-t}{2}(-y_1, y_2, \dots, y_n)\right) \\
&\le \frac{1+t}{2}\Phi((y_1, y_2, \dots, y_n)) + \frac{1-t}{2}\Phi((-y_1, y_2, \dots, y_n)) \\
&= \Phi((y_1, y_2, \dots, y_n)).
\end{aligned}$$

The inequality follows from convexity of Φ, while the last equality uses the fact that Φ is sign invariant. ∎

Next, we prove a theorem that demonstrates that the Ky Fan norms play an essential role in the theory of symmetric gauge functions.

Theorem 6.6 (Fan dominance: Vector case). Fix $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The following statements are equivalent:

1. $\|\boldsymbol{x}\|_{(k)} \le \|\boldsymbol{y}\|_{(k)}$ for each $k = 1, \dots, n$.
2. $\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y})$ for every symmetric gauge function Φ on $\mathbb{R}^n$.

Proof. Statement 1 follows immediately from 2, as Ky Fan norms are symmetric gauge functions. For the other implication, note that 1 is equivalent to the condition that $|\boldsymbol{x}| \prec_\omega |\boldsymbol{y}|$. As a consequence, there exists a vector $\boldsymbol{u} \in \mathbb{R}^n$ such that $|\boldsymbol{x}| \le \boldsymbol{u} \prec \boldsymbol{y}$. Therefore,

$$\Phi(\boldsymbol{x}) \le \Phi(|\boldsymbol{x}|) \le \Phi(\boldsymbol{u}) \le \Phi(|\boldsymbol{y}|) \le \Phi(\boldsymbol{y}).$$

The first and last inequalities follow from sign invariance of Φ. Monotonicity of Φ (Proposition 6.5) yields the second inequality. Finally, the third inequality follows from the fact that Φ is convex and permutation invariant, hence isotone. ∎

The Fan dominance theorem has a striking meaning. Given vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, we can check $\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y})$ for every symmetric gauge function Φ, by checking the inequality for only n functions, namely the Ky Fan k-norms for $k = 1, \dots, n$. In this sense, the theorem reduces a problem involving an infinite number of inequalities to the verification of a finite number.

Exercise 6.7 (Submajorization). Given $|\boldsymbol{x}| \prec_\omega |\boldsymbol{y}|$, explain how to construct a vector $\boldsymbol{u} \in \mathbb{R}^n$ such that $|\boldsymbol{x}| \le \boldsymbol{u} \prec \boldsymbol{y}$.

6.2 Duality for symmetric gauge functions

In this section, we develop the duality theory for symmetric gauge functions. More specifically, we define the dual norm of a symmetric gauge function and establish a generalization of the Hölder inequality.

Definition 6.8 (Duality). The dual norm Φ^* of a symmetric gauge function Φ is given by

$$\Phi^*(y) := \max\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle : \Phi(\boldsymbol{x}) \le 1\}.$$

Exercise 6.9 (Involution). Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a symmetric gauge function. Prove that $(\Phi^*)^* = \Phi$.

Exercise 6.10 (Dual symmetric gauge function). Prove that Φ^* is a symmetric gauge function if and only if Φ is a symmetric gauge function.

Let us consider the duality pairings for the examples studied in Section 6.1.

Example 6.11 (ℓ_p duality pairs). Let $p, q \in [1, \infty]$ be Hölder conjugates. Then

$$\|\boldsymbol{y}\|_p^* = \|\boldsymbol{y}\|_q \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^n.$$

In particular, ℓ_2 is self-dual, while ℓ_1 and ℓ_∞ form a dual pair. ▪

Recall that two real numbers $p, q \in [1, \infty)$ are said to be *Hölder conjugates* if $\frac{1}{p} + \frac{1}{q} = 1$. Also, recall that the Hölder conjugate of $p = 1$ is $q = \infty$.

Example 6.12 (Ky Fan duality pair). For each $k = 1 \dots n$, the dual norm of the Ky Fan k-norm is given by

$$\|\boldsymbol{x}\|_{(k)}^* = \max\left\{\|\boldsymbol{x}\|_{(1)}, \frac{1}{k}\|\boldsymbol{x}\|_{(n)}\right\} \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n.$$

Note that the Ky Fan 1-norm $\|\cdot\|_{(1)}$ corresponds to the ℓ_∞ norm, while $\|\cdot\|_{(n)}$ corresponds to the ℓ_1 norm. ▪

We now continue with two norm inequalities that hold for all symmetric gauge functions and specialize to familiar inequalities that frequently appear in analysis.

Proposition 6.13 (Dual norm inequality). For each symmetric gauge function Φ,

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \le \Phi^*(\boldsymbol{x})\,\Phi(\boldsymbol{y}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Proof. This follows immediately from the definition of the dual norm. ■

Notice that when $\Phi = \|\cdot\|_2$, the inequality in Proposition 6.13 corresponds to the Cauchy–Schwarz inequality.

Theorem 6.14 (Generalized Hölder inequality for symmetric gauge functions). Let $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ be a symmetric gauge function. Then,

$$\Phi(|\boldsymbol{x} \odot \boldsymbol{y}|) \le \left[\Phi(|\boldsymbol{x}|^p)\right]^{1/p} \left[\Phi(|\boldsymbol{y}|^q)\right]^{1/q},$$

Recall that $\odot$ denotes the entrywise product, also known as *Hadamard product* or *Schur product*. Also recall the notation $|\boldsymbol{x}| := (|x_1|, \dots, |x_n|)$.

where $p > 1$ and q is Hölder conjugate to p.

Proof. Refer to [Bha97, Thm. IV.1.6] for a proof of this theorem. ∎

When Φ is the ℓ_1 norm, the inequality reduces to Hölder's inequality

$$\sum_{i=1}^n |x_i\, y_i| \le \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}.$$

The theorem has interesting implications when applied to other symmetric gauge functions.

6.3 Unitarily invariant norms

In this section, we introduce an important family of matrix norms, called unitarily invariant norms, which are intimately related to symmetric gauge functions.

Definition 6.15 (Unitarily invariant norm). A norm $|||\cdot|||$ on $\mathbb{M}_n$ is *unitarily invariant* (UI) if

$$|||\boldsymbol{UAV}||| = |||\boldsymbol{A}||| \quad \text{for all } \boldsymbol{A}, \boldsymbol{U}, \boldsymbol{V} \in \mathbb{M}_n \text{ with } \boldsymbol{U}, \boldsymbol{V} \text{ unitary.}$$

We also insist that unitarily invariant norms be normalized in an appropriate fashion: $|||\mathrm{diag}(1, 0, \dots, 0)||| = 1$.

For the remainder of this lecture, $|||\cdot|||$ will denote an arbitrary unitarily invariant norm. Next, let us consider some familiar examples.

Example 6.16 (ℓ_2 operator norm). The ℓ_2 operator norm, also known as the *spectral norm*, is defined by

$$\|\boldsymbol{A}\|_2 := \sigma_1(\boldsymbol{A}) \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n.$$

The singular values of a matrix are numbered in decreasing order; i.e., by convention $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$.

The ℓ_2 operator norm is unitarily invariant. ▪

Example 6.17 (Frobenius norm). The Frobenius norm, also known as the Hilbert–Schmidt norm, is defined by

$$\|\boldsymbol{A}\|_{\mathrm{F}} := \left(\sum_{i=1}^n \sigma_i(\boldsymbol{A})^2\right)^{1/2} \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n.$$

The Frobenius norm is unitarily invariant. ▪

Example 6.18 (Schatten p-norms). For $1 \le p < \infty$, define

$$\|\boldsymbol{A}\|_{S_p} := \left(\sum_{i=1}^n \sigma_i(\boldsymbol{A})^p\right)^{1/p} \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n.$$

The functions $\|\cdot\|_{S_p}$, commonly known as Schatten norms, define unitarily invariant norms on the space $\mathbb{M}_n$. The Schatten 1-norm, also known as the *trace norm* or the *nuclear norm*, corresponds to the sum of the singular values of the matrix. The norm $\|\cdot\|_{S_\infty}$ coincides with the spectral norm (Example 6.16). ▪

Example 6.19 (Ky Fan matrix norm). For each $k = 1, \dots, n$, define

$$\|\boldsymbol{A}\|_{(k)} := \sum_{i=1}^k \sigma_i(\boldsymbol{A}) \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n.$$

The functions $\|\cdot\|_{(k)}$ define unitarily invariant norms on $\mathbb{M}^n$, known as the Ky Fan matrix norms. ▪

Exercise 6.20 (Why unitarily invariant?). Provide an explanation as to why each of the examples above is unitarily invariant.

As in the case for symmetric gauge functions, combinations of Schatten and Ky Fan matrix norms are also unitarily invariant.

6.4 Characterization of unitarily invariant norms

Observe that all the matrix norms in the examples above are completely determined by the singular values of the matrix. This is a joint consequence of the singular value decomposition (SVD) and unitary invariance. Indeed, if $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ is an SVD of the matrix $\boldsymbol{A} \in \mathbb{M}_n$, then $|||\boldsymbol{A}||| = |||\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*||| = |||\boldsymbol{\Sigma}|||$. This observation sets the ground for us to establish the strong connection that exists between symmetric gauge functions and unitarily invariant norms.

Theorem 6.21 (Unitarily invariant norms: Characterization). There is a one-to-one correspondence between symmetric gauge functions and unitarily invariant norms. This correspondence is specified by the following statements:

1. Given a symmetric gauge function $\Phi : \mathbb{R}^n \to \mathbb{R}_+$, the function $|||\cdot|||_\Phi : \mathbb{M}^n \to \mathbb{R}_+$ defined by $|||\boldsymbol{A}|||_\Phi := \Phi(\boldsymbol{\sigma}(\boldsymbol{A}))$ is a unitarily invariant norm on $\mathbb{M}^n$.
2. Given a unitarily invariant norm $|||\cdot|||$ on $\mathbb{M}^n$, the map $\Phi : \mathbb{R}^n \to \mathbb{R}$ defined by $\Phi(\boldsymbol{x}) := |||\mathrm{diag}(\boldsymbol{x})|||$ is a symmetric gauge function on $\mathbb{R}^n$.

In fact, these operations are mutual inverses.

Proof. **(1).** Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a symmetric gauge function, and let $|||\cdot|||_\Phi$ be defined as in (1). Positive definiteness of $|||\cdot|||_\Phi$ follows from the facts that Φ is positive definite and the zero matrix is the only matrix whose singular values are all zero. Being a norm, Φ is positive homogeneous, and since multiplying a matrix by a real number α scales its singular values by $|\alpha|$, the function $|||\cdot|||_\Phi$ is also positive homogeneous.

It remains to show that $|||\cdot|||_\Phi$ satisfies the triangle inequality and is unitarily invariant. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$ be $n \times n$ real matrices. By Exercise 6.22, we know that $\boldsymbol{\sigma}(\boldsymbol{A}) + \boldsymbol{\sigma}(\boldsymbol{B}) \prec_\omega \boldsymbol{\sigma}(\boldsymbol{A}) + \boldsymbol{\sigma}(\boldsymbol{B})$. Since the symmetric gauge function Φ is isotone,

$$\Phi(\boldsymbol{\sigma}(\boldsymbol{A}+\boldsymbol{B})) \le \Phi(\boldsymbol{\sigma}(\boldsymbol{A}) + \boldsymbol{\sigma}(\boldsymbol{B})) \le \Phi(\boldsymbol{\sigma}(\boldsymbol{A})) + \Phi(\boldsymbol{\sigma}(\boldsymbol{B})).$$

We conclude that $|||\cdot|||_\Phi$ is a norm. Note that the first and second inequalities above follow from monotonicity and convexity of Φ, respectively.

Finally, observe that multiplying any given matrix by a unitary matrix does not alter its singular values. Indeed, for any $\boldsymbol{A} \in \mathbb{M}_n$, its norm $|||\boldsymbol{A}|||_\Phi$ is completely determined by its singular values. It follows that $|||\cdot|||_\Phi$ is unitarily invariant.

(2). Let $|||\cdot|||$ be a unitarily invariant norm, and let Φ be defined as in (2). It is immediate that Φ inherits all the norm properties from $|||\cdot|||$. To show unitary invariance of $|||\cdot|||_\Phi$, we first let $\boldsymbol{D} = \mathrm{diag}(\pm 1, \ldots, \pm 1)$. The matrix $\boldsymbol{D}$ is unitary, so we have that

$$\Phi(\boldsymbol{D}\boldsymbol{x}) = |||\mathrm{diag}(\boldsymbol{D}\boldsymbol{x})||| = |||\boldsymbol{D}\mathrm{diag}(\boldsymbol{x})||| = |||\mathrm{diag}(\boldsymbol{x})|||.$$

Similarly, since permutation matrices are unitary, we have that

$$\Phi(\boldsymbol{P}\boldsymbol{x}) = |||\boldsymbol{P}\mathrm{diag}(\boldsymbol{x})||| = |||\mathrm{diag}(\boldsymbol{x})|||$$

for any permutation matrix $\boldsymbol{P} \in \mathbb{M}_n$. Finally, note that $|||\cdot|||_\Phi$ is normalized, as $\sigma(\mathrm{diag}(1, 0, \ldots, 0)) = (1, 0, \ldots, 0)$ and Φ is normalized. ∎

Problem 6.22 (Singular values of the sum). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{C})$. Show that the vectors of singular values $\boldsymbol{\sigma}(\boldsymbol{A})$ and $\boldsymbol{\sigma}(\boldsymbol{B})$ satisfy the submajorization inequality $\boldsymbol{\sigma}(\boldsymbol{A})+\boldsymbol{\sigma}(\boldsymbol{B}) \prec_\omega \boldsymbol{\sigma}(\boldsymbol{A}) + \boldsymbol{\sigma}(\boldsymbol{B})$.

Next, we state the Fan dominance theorem for unitarily invariant matrices, analogous to Theorem 6.6. As in the vector case, this theorem emphasizes the importance of Ky Fan norms in the theory of unitarily invariant norms.

Theorem 6.23 (Fan dominance: Matrix case). Fix $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$. The following statements are equivalent:

1. $\|\boldsymbol{A}\|_{(k)} \le \|\boldsymbol{B}\|_{(k)}$ for each $k = 1, \dots, n$.
2. $|||\boldsymbol{A}||| \le |||\boldsymbol{B}|||$ for every unitarily invariant norm.

Proof. The theorem follows from Theorem 6.6 and from the characterization of unitarily invariant norms in terms of symmetric gauge functions. ∎

6.5 Duality for unitarily invariant norms

In the last section, we defined unitarily invariant norms and established a characterization in terms of symmetric gauge functions. Now, we build on this connection to develop duality theory for unitarily invariant norms and establish a generalization of the Hölder inequality.

Definition 6.24 The dual $|||\cdot|||^*$ of a unitarily invariant norm $|||\cdot|||$ on $\mathbb{M}_n$ is given by

$$|||\boldsymbol{B}|||^* := \max\{\langle \boldsymbol{B}, \boldsymbol{A}\rangle : |||\boldsymbol{A}||| \le 1\} \quad \text{for each } \boldsymbol{B} \in \mathbb{M}_n.$$

Recall that $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \operatorname{Tr}(\boldsymbol{B}^*\boldsymbol{A})$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$.

Exercise 6.25 (Involution). Let $|||\cdot||| : \mathbb{R}^n \to \mathbb{R}$ be a unitarily invariant norm. Prove that $(\Phi^*)^* = \Phi$.

Exercise 6.26 (Dual unitarily invariant norms). Prove that $|||\cdot|||^*$ is unitarily invariant if and only if $|||\cdot|||$ is unitarily invariant.

Exercise 6.27 (Dual norm inequality). For each unitarily invariant norm $|||\cdot|||$,

$$|\langle \boldsymbol{B}, \boldsymbol{A}\rangle| \le |||\boldsymbol{B}|||^* \cdot |||\boldsymbol{A}||| \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n.$$

Verify this statement.

In Section 6.3 we saw that, given a symmetric gauge function Φ on $\mathbb{R}^n$, we can define a unitarily invariant norm on $\mathbb{M}_n$ via $|||\boldsymbol{A}|||_\Phi := \Phi(\boldsymbol{\sigma}(\boldsymbol{A}))$. The following theorem describes how the dual of the unitarily invariant norm $|||\cdot|||_\Phi$ relates to the dual Φ^* of the symmetric gauge function used to define it.

Theorem 6.28 (Von Neumann duality). The dual $|||\cdot|||_\Phi^*$ of the unitarily invariant norm $|||\cdot|||_\Phi$ associated to a symmetric gauge function Φ is the unitarily invariant norm $|||\cdot|||_{\Phi^*}$ associated to the dual Φ^* of the symmetric gauge function. That is,

$$|||\boldsymbol{B}|||_\Phi^* = |||\boldsymbol{B}|||_{\Phi^*} \quad \text{for all } \boldsymbol{B} \in \mathbb{M}_n.$$

Proof. To prove this result, we use von Neumann's trace theorem. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$. Then

$$\max\{\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}) : \boldsymbol{U}, \boldsymbol{V} \in \mathbb{M}_n \text{ are unitary}\} = \sum_{i=1}^n \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B}).$$

Exercise 6.29 invites you to prove this result.

As a consequence, we may calculate that

$$\begin{aligned}
|||\boldsymbol{B}|||_{\Phi}^{*} &= \max\{\operatorname{tr}(\boldsymbol{B}^{*}\boldsymbol{A}) : |||\boldsymbol{A}|||_{\Phi} \leq 1\} \\
&= \max\{\operatorname{tr}(\boldsymbol{B}^{*}\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}) : \boldsymbol{U},\boldsymbol{V} \in \mathbb{M}_n \text{ are unitary}, |||\boldsymbol{A}|||_{\Phi} \leq 1\} \\
&= \max\left\{\sum_{i=1}^{n} \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B}) : \Phi(\sigma(\boldsymbol{A})) \leq 1\right\} \\
&= \max\{\langle \sigma(\boldsymbol{A}),\ \sigma(\boldsymbol{B})\rangle : \Phi(\sigma(\boldsymbol{A})) \leq 1\} \\
&\leq \Phi^{*}(\sigma(\boldsymbol{B})) \\
&= |||\boldsymbol{B}|||_{\Phi^{*}}.
\end{aligned}$$

The second equality above is justified by unitary invariance of $|||\cdot|||_{\Phi}^{*}$.

The inequality $|||\boldsymbol{B}|||_{\Phi^{*}} \leq |||\boldsymbol{B}|||_{\Phi}^{*}$ follows from a similar line of reasoning and is left as an exercise for the reader. ∎

Exercise 6.29 (von Neumann trace theorem: Singular values). Provide a proof for the von Neumann trace theorem for singular values used in the proof of Theorem 6.28.

In light of the previous theorem, we can easily find expressions for the dual norms of our examples from section 6.3.

Example 6.30 (Schatten norms). Let $p, q \in [1, \infty]$ be Hölder conjugates. Then

$$\|\boldsymbol{B}\|_{S_p}^{*} = \|\boldsymbol{B}\|_{S_q}.$$

Recall that the dual to the ℓ_p norm is the ℓ_q norm, where q is the Hölder conjugate of p. Therefore, the dual norm to the Schatten p-norm is the Shatten q-norm. Indeed, the Schatten p- and q-norms coincide with the ℓ_p and ℓ_q norms of the vector of singular values. In particular, the Schatten 2-norm is self dual, while the Schatten norms $\|\cdot\|_{S_1}$ and $\|\cdot\|_{S_\infty}$ form a dual pair. ▪

Example 6.31 (Ky Fan norms). For each $k = 1, \ldots, n$, the dual norm of the Ky Fan k-norm is given by

$$\|\boldsymbol{B}\|_{(k)}^{*} = \max\left\{\|\boldsymbol{B}\|_{(1)}, \frac{1}{k}\|\boldsymbol{B}\|_{(n)}\right\}.$$

Note that the Ky Fan 1-norm $\|\cdot\|_{(1)}$ corresponds to the spectral norm, while $\|\cdot\|_{(n)}$ corresponds to trace norm. ▪

We end this lecture with a generalization of Hölder's inequality for unitarily invariant norms.

Theorem 6.32 (Generalized Hölder inequality). Let $|||\cdot|||$ be unitarily invariant, and let $p, q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$|||\,|\boldsymbol{A}\boldsymbol{B}|\,||| \leq |||\boldsymbol{A}|^{p}|||^{1/p}\,|||\boldsymbol{B}|^{q}|||^{1/q} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n,$$

where $|\boldsymbol{M}| := (\boldsymbol{M}^{*}\boldsymbol{M})^{1/2}$.

Proof sketch. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$. By Exercise 6.33 and isotonicity of symmetric gauge functions, we have that $\Phi(\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{B})) \leq \Phi(\boldsymbol{\sigma}(\boldsymbol{A}) \odot \boldsymbol{\sigma}(\boldsymbol{B}))$ for all symmetric gauge functions Φ. The result then follows from the characterization of unitarily invariant norms (Theorem 6.21) and Theorem 6.14. ∎

Refer to [Bha97, Cor. IV.2.6] for a more detailed proof.

Exercise 6.33 (Singular values of the product). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$. Show that $\boldsymbol{\sigma}(\boldsymbol{AB}) \prec_\omega \boldsymbol{\sigma}(\boldsymbol{A}) \odot \boldsymbol{\sigma}(\boldsymbol{B})$.

In this lecture, we discussed two important and closely related families of norms, symmetric gauge functions on $\mathbb{R}^n$ and unitarily invariant norms on the space of $n \times n$ real matrices $\mathbb{M}_n$. We established generalizations of Hölder-type inequalities for these two families of norms by exploiting their defining properties and their connection to each other. These results exemplify how considering families of norms with invariance properties allows us to prove results for a broad class of norms which can then specialize into results of interest when applied to particular examples.

Notes

This material is adapted from Bhatia's book [Bha97, Chap. IV].

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

7. Matrix Inequalities via Complex Analysis

Date: 25 January 2022 **Scribe:** Yixuan (Roy) Wang

In this lecture, we develop a new approach for proving matrix inequalities based on the theory of complex interpolation. As motivation, we begin with the duality theorem for Schatten p-norms, and we recall some of the difficulties that arose in the proof. We recast the duality theorem as a theorem about interpolation, which suggests the possibility of deriving this result using general tools for studying interpolation problems. We develop a powerful approach based on the Hadamard three-lines theorem, a quantitative version of the maximum principle for analytic functions. As an illustration of these ideas, we provide an alternative derivation of the duality theorem for Schatten p-norms.

Agenda:

1. Motivation
2. Interpolation inequalities
3. Maximum modulus principle
4. The three-lines theorem
5. Example: Duality of Schatten norms

7.1 Motivation: Real analysis is not always enough

To begin, let us present some basic inequalities for Schatten norms, including the main duality theorem. We will see that this result can be framed as a kind of interpolation inequality. This observation opens the door to using complex interpolation theory.

7.1.1 Schatten norm inequalities

We first recall the definition of the Schatten p-norms of a complex matrix.

Definition 7.1 (Schatten norms). The Schatten p-norms $\|\cdot\|_p$ for $p \in [1, \infty]$ are defined as

$$\|\boldsymbol{A}\|_p := \left(\sum\nolimits_{i=1}^{n} \sigma_i(\boldsymbol{A})^p\right)^{1/p} \qquad \text{for } 1 \le p < \infty;$$
$$\|\boldsymbol{A}\|_\infty := \sigma_1(\boldsymbol{A}).$$

Here, $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$ is an $n \times n$ complex matrix. The function σ_i returns the ith largest singular value of a matrix.

For powers $p \in [1, \infty)$, we can also write the Schatten p-norm as a trace:

$$\|\boldsymbol{A}\|_p = (\operatorname{tr} |\boldsymbol{A}|^p)^{1/p} \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n(\mathbb{C}).$$

Recall that the power binds before the trace.

As usual, $|\boldsymbol{A}| := (\boldsymbol{A}^*\boldsymbol{A})^{1/2}$ is the matrix absolute value. The pth power of a positive-semidefinite matrix is defined via the usual functional calculus (i.e., by raising the eigenvalues to the pth power).

In Lecture 6, we characterized the unitarily invariant norms on the matrix space $\mathbb{M}_n(\mathbb{C})$, and we established a one-to-one correspondence with the symmetric gauge functions on the vector space $\mathbb{R}^n$. In particular, the Schatten p-norms are analogous with the vector l_p norms. By the duality for ℓ_p vector norms and the von Neumann duality theorem, we derived a duality relation for Schatten norms: $\|\cdot\|_p^* = \|\cdot\|_q$ where the indices satisfy the conjugacy relation $1/p + 1/q = 1$.

Theorem 7.2 (Duality relation for Schatten p-norms). For all indices $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, the dual norm of the Schatten p-norm is the Schatten q-norm. The duality is equivalent to a Hölder-type inequality.

$$|\operatorname{tr}(\boldsymbol{A}^*\boldsymbol{B})| \le \|\boldsymbol{A}\|_p \cdot \|\boldsymbol{B}\|_q \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{C}). \tag{7.1}$$

As unspooled in Lecture 6, the proof of the duality of Schatten p-norms involves Birkhoff's theorem on doubly stochastic matrices, the von Neumann trace theorem, the von Neumann duality theorem, and finally duality for ℓ_p vector norms. It is fairly complicated, and we spent four lectures building up to the theorem.

In addition to its complexity, this type of reasoning does not offer a very flexible approach to proving matrix inequalities. We may encounter related problems that we cannot solve directly from these ideas. For example, consider the following Schwarz-type inequality.

Theorem 7.3 (A Schwarz-type inequality). Fix $p \ge 1$. For all $\boldsymbol{B} \in \mathbb{M}_n(\mathbb{C})$ and all $\boldsymbol{A}_j \in \mathbb{M}_n(\mathbb{C})$ for $j = 1, 2, \cdots, m$, we have

$$\left\| \sum\nolimits_{j=1}^m \boldsymbol{A}_j^* \boldsymbol{B} \boldsymbol{A}_j \right\|_p \le \left\| \sum\nolimits_{j=1}^m \boldsymbol{A}_j^* \boldsymbol{A}_j \right\|_{2p} \|\boldsymbol{B}\|_{2p}.$$

Proof. See Problem Set 2. ∎

To prove results of this type, it is valuable to have additional tools. This lecture will show how to use basic methods from complex analysis to derive matrix inequalities, including results like Theorem 7.3.

7.1.2 Interpolation inequalities

As a first step, let us rephrase the duality statement (7.1) as an interpolation inequality.

Proposition 7.4 (Schatten duality: Interpolation form). Theorem 7.2 follows from the statement below. For all positive-semidefinite matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+(\mathbb{C})$, we have

$$\|\boldsymbol{A}^\theta \boldsymbol{B}^{1-\theta}\|_1 \le \|\boldsymbol{A}\|_1^\theta \cdot \|\boldsymbol{B}\|_1^{1-\theta} \quad \text{for } 0 < \theta < 1. \tag{7.2}$$

You can think about $\|\boldsymbol{A}^\theta \boldsymbol{B}^{1-\theta}\|_1$ as the trace norm of a weighted geometric mean of the two positive-semidefinite matrices. The right-hand side is a weighted geometric mean of their trace norms.

Proof. We will establish (7.1) for positive-semidefinite matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+(\mathbb{C})$; Exercise 7.6 asks you to derive the extension for general matrices.

First, fix $p \in (1, \infty)$. We show that (7.2) implies (7.1) by a change of variables. Consider the bijections $\boldsymbol{A} \mapsto \boldsymbol{A}^p$ and $\boldsymbol{B} \mapsto \boldsymbol{B}^q$ where $\theta = 1/p$ and $1 - \theta = 1/q$. We obtain the inequality

$$\|\boldsymbol{A}\boldsymbol{B}\|_1 \le \|\boldsymbol{A}\|_p \cdot \|\boldsymbol{B}\|_q \quad \text{for all psd } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+(\mathbb{C}).$$

Exercise 7.5 asks you to check that $|\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})| \le \|\boldsymbol{A}\boldsymbol{B}\|_1$, which yields the inequality (7.1). The boundary cases $p = 1$ and $p = \infty$ follow from the last display when we take limits. ∎

Exercise 7.5 (The trace norm). Prove that the trace is bounded by the trace norm:

$$|\operatorname{tr} \boldsymbol{M}| \le \|\boldsymbol{M}\|_1 \quad \text{for all } \boldsymbol{M} \in \mathbb{M}_n(\mathbb{C}).$$

Hint: Among many proofs, the easiest one introduces an SVD of the matrix.

Exercise 7.6 (Schatten duality: General case). For the general case when $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{C})$, derive inequality (7.1) from inequality (7.2). **Hint:** Introduce polar factorizations.

In view of this discussion, we may focus on proving the interpolation inequality (7.2). A surprising and powerful approach to this problem is to allow the interpolation parameter θ to take on *complex* values. This insight leads us to consider a complex interpolation inequality, which we can establish using the miracle of complex analysis. The key tool in this argument is a quantitative extension of the maximum modulus principle, called the Hadamard three-lines theorem. The rest of this lecture develops the required background and establishes the interpolation inequality (7.2).

7.2 The maximum modulus principle

In this section, we present the definition of a complex analytic function and prove an essential fact called the mean value formula. Then we establish the simplest version of the maximum modulus principle.

7.2.1 Domains and analytic functions

We begin with some definitions on domains and analytic functions.

Definition 7.7 (Domain). A *domain* $\Omega \subseteq \mathbb{C}$ is an open and connected set.

Definition 7.8 (Analytic function). A function $f : \Omega \to \mathbb{C}$ on a domain $\Omega \subseteq \mathbb{C}$ is *(complex) analytic* if it has a locally convergent power series expansion at each point of the domain. More precisely, consider a closed disc inside the domain:

$$\overline{\mathsf{D}}_r(a) := \{z \in \mathbb{C} : |z - a| \le r\} \subset \Omega.$$

Then there exists a convergent Taylor expansion

$$f(z) = \sum\nolimits_{k=0}^{\infty} c_k (z - a)^k \qquad \text{for all } z \in \overline{\mathsf{D}}_r(a).$$

The coefficients $c_k \in \mathbb{C}$. In particular, $c_0 = f(a)$.

We can show that the Taylor expansion about each point $a \in \Omega$ is uniquely determined, and it converges absolutely on the largest (open) disc $\mathsf{D}_r(a)$ that is contained in Ω. Moreover, when $\overline{\mathsf{D}}_r(a) \subseteq \overline{\mathsf{D}}_R(a) \subset \Omega$, the coefficients in the expansion coincide.

From this definition, we can easily confirm that analytic functions are continuous inside the domain Ω. In fact, a complex function is analytic on a domain if and only if it is differentiable within the domain (also called *holomorphic*).

Example 7.9 (Exponentials). For any complex number $c \in \mathbb{C}$, the function $z \mapsto \mathrm{e}^{cz}$ is analytic on the complex plane $\mathbb{C}$. Moreover, the familiar Taylor series expansion for the exponential converges in the whole complex plane $\mathbb{C}$. ∎

7.2.2 Mean value formula and maximum modulus principle

In this section, we shall establish the maximum modulus principle for analytic functions. Our key tool is the mean value formula.

Proposition 7.10 (Mean value formula). Let $f : \Omega \to \mathbb{C}$ be analytic, and let $\overline{\mathsf{D}}_r(a) \subset \Omega$.

Then we have the formula

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + re^{i\theta})\, d\theta\,. \tag{7.3}$$

Proof. The proof of this proposition is straightforward. We can just integrate the power series expansion of f at $a \in \mathbb{C}$ term by term. The constant order ($k = 0$) term contributes $f(a)$; the higher monomial terms vanish. ∎

We can apply the mean value formula to prove a special case of the maximum modulus principle: the version for a disc.

Proposition 7.11 (Maximum modulus principle: Disc). Let $f : \Omega \to \mathbb{C}$ be analytic, and consider a closed disc contained in the domain: $\overline{\mathsf{D}}_r(a) \subset \Omega$. If the function $z \mapsto |f(z)|$ achieves its maximum over $\overline{\mathsf{D}}_r(a)$ at the point a, then the function f must be constant on the disc $\overline{\mathsf{D}}_r(a)$. That is, $f(z) = f(a)$ for each $z \in \overline{\mathsf{D}}_r(a)$.

Proof. Fix the center $a \in \Omega$ and the radius $r > 0$ of the disc. By the triangle inequality applied to the mean value formula (7.3), we compute

$$|f(a)| \le \frac{1}{2\pi} \int_0^{2\pi} |f(a + re^{i\theta})|\, d\theta \le \frac{1}{2\pi} \int_0^{2\pi} |f(a)|\, d\theta = |f(a)|\,.$$

Therefore, both inequalities hold with equality. Since $|f|$ is continuous, the value $f(a + re^{i\theta})$ has constant phase for all $\theta \in [0, 2\pi)$. Furthermore, the magnitude $|f(a + re^{i\theta})| = |f(a)|$ for all θ. In other words, $\theta \mapsto f(a + re^{i\theta})$ is a constant function. By the mean value formula, the constant value of $f(a + re^{i\theta})$ equals $f(a)$.

Finally, we apply the same argument for each $r_0 < r$ and its associated disc $\overline{\mathsf{D}}_{r_0}(a)$. We conclude that $f(z) = f(a)$ for each $z \in \overline{\mathsf{D}}_r(a)$, because it is on the boundary of some disc $\overline{\mathsf{D}}_{r_0}(a)$ for $r_0 = |z - a| \le r$. ∎

With the maximum modulus principle for the disc at hand, we are in position to prove the general maximum modulus principle on bounded domains.

Theorem 7.12 (Maximum modulus principle: Bounded domain). Let $\Omega \subseteq \mathbb{C}$ be a bounded domain. Assume that $f : \Omega \to \mathbb{C}$ is analytic on Ω and continuous on $\overline{\Omega}$. Then $z \mapsto |f(z)|$ achieves its maximum on the boundary $\partial\Omega$.

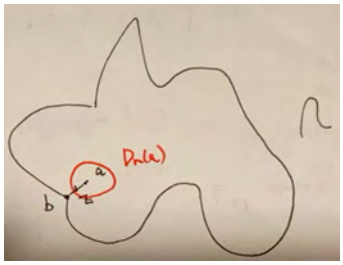

Figure 7.1 Identifying the disc to apply maximum modulus principle on a disc.

Proof. Since the closure $\overline{\Omega}$ is compact and $|f|$ is continuous on $\overline{\Omega}$, we know that $|f|$ attains its maximum on $\overline{\Omega}$ by the extreme value theorem.

We argue by contradiction. Suppose $|f|$ *does not* achieve a maximum on $\partial\Omega$. Let $a \in \Omega$ be a point where $|f(a)| \ge |f(z)|$ for all $z \in \overline{\Omega}$. Let $b \in \partial\Omega$ be the closest point in $\partial\Omega$ to a; such a point exists since Ω is compact.

By assumption, $|f(b)| < |f(a)|$. Since $|f|$ is continuous on the line segment $[a, b]$, there must exist a point $z \in (a, b) \subset \Omega$ in the interior of the domain such that $|f(z)| < |f(a)|$; see Figure 7.1 for an illustration.

Consider a disc $\overline{\mathsf{D}}_r(a)$ that contains z and sits inside Ω. Since $|f|$ achieves its maximum at a, we have that $|f(a)| \ge |f(w)|$ for all $w \in \overline{\mathsf{D}}_r(a)$. By the maximum modulus principle for a disc, f is a constant on $\overline{\mathsf{D}}_r(a)$. But this is a contradiction to the fact that $|f(z)| < |f(a)|$. ∎

Aside: (Maximum modulus principle). In fact, a stronger result is valid. Assume that $f : \Omega \to \mathbb{C}$ is analytic on a *bounded* domain Ω and continuous on $\overline{\Omega}$. If the function $|f|$ attains maximum at any point inside the domain Ω, then the function f is *constant* on the domain.

7.3 Interpolation: The three-lines theorem

The maximum modulus principle (Theorem 7.12) applies to general bounded domains. However, the result can fail for unbounded domains without additional assumptions. For one thing, we do not even know if $|f|$ attains its maximal value; it could be unbounded. Nevertheless, for sufficiently regular functions, we can establish a maximum modulus principle for particular unbounded domains (e.g., strips and wedges).

The next result provides a quantitative version of the maximum modulus principle on the strip. This theorem is attributed to Hadamard.

Theorem 7.13 (Hadamard three-lines). Let $\Omega = \{z : 0 < \operatorname{Re} z < 1\}$ be a vertical strip in the complex plane. Assume that the function $f : \Omega \to \mathbb{C}$ is analytic on Ω, and assume that f is *bounded* and continuous on $\overline{\Omega}$. Define the quantity

$$M(\theta) := \sup_{t \in \mathbb{R}} |f(\theta + \mathrm{i}t)| \quad \text{for } \theta \in [0, 1].$$

Then the function M is log-convex. In particular, we have

$$M(\theta) \le M(0)^{1-\theta} \cdot M(1)^{\theta} \quad \text{for } \theta \in [0, 1]. \tag{7.4}$$

Proof. We will only prove the interpolation inequality (7.4); the general log-convexity statement follows by a scaling argument. Without loss of generality, we can also assume that $M(0)$ and $M(1)$ are strictly positive; otherwise we can add a small positive constant δ to f and take the limit $\delta \downarrow 0$.

Consider the auxiliary function

$$F(z) := f(z) \cdot M(0)^{z-1} M(1)^{-z} \quad \text{for } z \in \overline{\Omega}.$$

Then F is analytic on Ω and continuous on $\overline{\Omega}$. Moreover, for $z = \theta + \mathrm{i}t$ with $\theta, t \in \mathbb{R}$, we can compute

$$|F(\theta + \mathrm{i}t)| = |f(\theta + \mathrm{i}t)| \cdot M(0)^{\theta-1} M(1)^{-\theta}. \tag{7.5}$$

Therefore F is bounded on $\overline{\Omega}$, and $|F(z)| \le 1$ on $\partial\Omega$ by construction.

Claim 7.14 (Auxiliary function is bounded). We claim that $|F(z)| \le 1$ for all $z \in \overline{\Omega}$.

Granted that Claim 7.14 holds, we may quickly complete the argument:

$$\begin{aligned} 1 \ge \sup_{t \in \mathbb{R}} |F(\theta + \mathrm{i}t)| &= [\sup_{t \in \mathbb{R}} |f(\theta + \mathrm{i}t)|] \cdot M(0)^{\theta-1} M(1)^{-\theta} \\ &= M(\theta) \cdot M(0)^{\theta-1} M(1)^{-\theta}. \end{aligned}$$

The first equality is (7.5). Rearrange to reach the bound (7.4).

To prove Claim 7.14, we will use a regularization argument. For $\varepsilon > 0$, consider the function

$$F_\varepsilon(z) = F(z) \cdot \mathrm{e}^{\varepsilon(z^2-1)} \quad \text{for } z \in \overline{\Omega}.$$

Once again, we recognize that F_ε is analytic on Ω and continuous on $\overline{\Omega}$. Moreover, for $z = \theta + \mathrm{i}t$ with $\theta \in [0, 1]$ and $t \in \mathbb{R}$, we can compute

$$|F_\varepsilon(z)| = |F(z)| \cdot \mathrm{e}^{\varepsilon(\theta^2-1-t^2)} \le |F(z)| \cdot \mathrm{e}^{-\varepsilon t^2} \le |F(z)|. \tag{7.6}$$

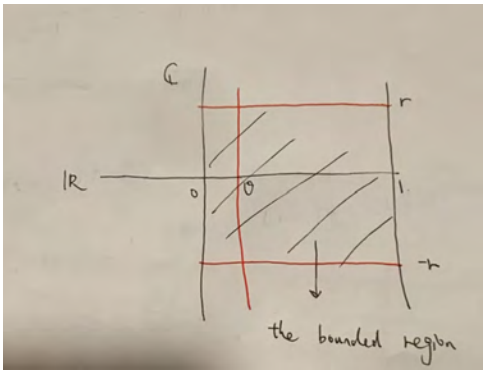

Figure 7.2 Identifying the bounded domain to apply maximum modulus principle.

Therefore, F_ε is bounded on $\overline{\Omega}$, and $|F_\varepsilon(z)| \le 1$ on $\partial\Omega$.

Identify a positive $r > 0$ such that $\mathrm{e}^{\varepsilon r^2} \ge \sup_{z\in\overline{\Omega}} |F(z)|$; see Figure 7.2. Consider the bounded rectangle domain $\mathsf{R} \subset \overline{\Omega}$ with sides defined by the lines $\operatorname{Re} z = 1$ and $\operatorname{Re} z = 0$ and $\operatorname{Im} z = r$ and $\operatorname{Im} z = -r$. By the computation in (7.6), we deduce that

$$|F_\varepsilon(z)| \le |F(z)| \cdot \mathrm{e}^{-\varepsilon r^2} \le 1 \quad \text{for } z \in \overline{\Omega} \text{ with } |\operatorname{Im} z| \ge r.$$

In other words, $|F_\varepsilon| \le 1$ on the part of the strip outside the rectangle R.

What about the rectangle R itself? By the maximum modulus principle, the maximum of $|F_\varepsilon|$ on R must be achieved on the boundary $\partial\mathsf{R}$. According to the last display, $|F_\varepsilon| \le 1$ on the top and bottom of the rectangle. By construction of the auxiliary function and the regularizer, we have $|F_\varepsilon| \le |F| \le 1$ on the left and right sides of the rectangle.

From the last two paragraphs, we determine that $|F_\varepsilon(z)| \le 1$ on $\overline{\Omega}$. Finally, note that

$$1 \ge |F_\varepsilon(z)| = |F(z)| \cdot \mathrm{e}^{\varepsilon(\theta^2 - 1 - t^2)} \to |F(z)| \quad \text{pointwise as } \varepsilon \downarrow 0.$$

This observation completes the proof of Claim 7.14. ∎

Exercise 7.15 (Three-lines: Weaker regularity condition). From the proof of Theorem 7.13, we may infer that the regularity of f plays an important role. The boundedness condition can be relaxed, however, as long as we have the property that $f(z) = o(\mathrm{e}^{cz^2})$ for any positive constant c. That is, f should increase more slowly than any quadratic exponential function. Adapt the proof to address this scenario.

> **Aside: (Analytic functions on the strip: Integral representations).** There is an alternative proof of the theorem via solving the Poisson equation for the strip [Ste56]. Indeed, analytic functions are harmonic, satisfying $\Delta f = 0$. Therefore, we can pass to the Green's function kernel of the strip and use an integral of the boundary values to represent the function f on the whole domain. We can establish the interpolation inequality by bounding this integral.

7.4 Example: Duality for Schatten norms

Finally, with the Hadamard three-lines theorem at hand, we may give an alternative proof of the duality for Schatten norms by establishing Proposition 7.4. Observe that

the interpolation inequality in (7.2) resembles the conclusion (7.4) of the three-lines theorem. This parallel suggests an argument based on complex analysis. As a first step, we explain what it means to apply a complex power to a positive matrix.

Definition 7.16 (Complex power). For a positive-semidefinite matrix $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$, consider the spectral resolution $\boldsymbol{A} = \sum_j \lambda_j \boldsymbol{P}_j$, where $\lambda_j > 0$ and the $\boldsymbol{P}_j$ are orthoprojectors. For a complex number $z \in \mathbb{C}$, we define the complex power

$$\boldsymbol{A}^z := \sum\nolimits_j \lambda_j^z \boldsymbol{P}_j = \sum\nolimits_j \mathrm{e}^{z \log \lambda_j} \boldsymbol{P}_j.$$

This is a standard example of the functional calculus (for normal matrices).

The first step in the argument is to establish the duality between the trace norm (Schatten-1) and the operator norm (Schatten-∞).

Exercise 7.17 (Trace norm and operator norm). By an independent argument, prove that

$$|\operatorname{tr}(\boldsymbol{A}^* \boldsymbol{B})| \le \|\boldsymbol{A}\|_1 \cdot \|\boldsymbol{B}\|_\infty \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{C}).$$

Hint: There are many possible arguments. The easiest approach is to introduce an SVD of $\boldsymbol{A}$, cycle the trace, and use the variational definition of the largest singular value.

Proof of Proposition 7.4. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+(\mathbb{C})$ be positive-semidefinite matrices, and let $\theta \in (0, 1)$. We intend to bound the quantity

$$\operatorname{tr}(\boldsymbol{A}^\theta \boldsymbol{B}^{1-\theta} \boldsymbol{C}) \quad \text{where } \|\boldsymbol{C}\|_\infty = 1.$$

By allowing the parameter θ to take on complex values, we can obtain an analytic function for input to the Hadamard theorem.

Consider the function

$$f(z) := \operatorname{tr}(\boldsymbol{A}^z \boldsymbol{B}^{1-z} \boldsymbol{C}) \quad \text{for } 0 \le \operatorname{Re} z \le 1.$$

By the definition of complex power, we can compute

$$f(z) = \sum\nolimits_{j,k} \lambda_j^z \mu_k^{1-z} \operatorname{tr}(\boldsymbol{P}_j \boldsymbol{Q}_k \boldsymbol{C}),$$

where we have introduced the spectral resolutions $\boldsymbol{A} = \sum_j \lambda_j \boldsymbol{P}_j$ and $\boldsymbol{B} = \sum_k \mu_k \boldsymbol{Q}_k$. We see that f is a linear combination of analytic functions on $\Omega = \mathbb{C}$, Therefore f is analytic and continuous the closure of the strip $\Omega = \{z : 0 < \operatorname{Re} z < 1\}$.

As in the statement of Hadamard's three-lines theorem, define

$$M(\theta) := \sup\nolimits_{t \in \mathbb{R}} |f(\theta + \mathrm{i}t)| \quad \text{for } 0 \le \theta \le 1.$$

We need to check that f is bounded before we invoke the theorem. In fact, each term of the linear combination in the expression for $f(z)$ is bounded since

$$|\lambda_k^z \mu_j^{1-z}| = \lambda_k^{\operatorname{Re} z} \mu_j^{1-\operatorname{Re} z} \le \max\nolimits_j \max\{\lambda_j, \mu_j\}.$$

Therefore, f is indeed bounded. We can apply Theorem 7.13 to obtain the inequality

$$|\operatorname{tr}(\boldsymbol{A}^\theta \boldsymbol{B}^{1-\theta} \boldsymbol{C})| \le M(\theta) \le M(0)^{1-\theta} \cdot M(1)^\theta. \tag{7.7}$$

Let us see why this inequality gives us what we need.

We can easily bound the terms $M(0)$ and $M(1)$. Using Exercise 7.17 and Exercise 7.5, we see that

$$|f(it)| \le \|\boldsymbol{A}^{it}\boldsymbol{B}^{1-it}\|_1 \cdot \|\boldsymbol{C}\|_\infty = \|\boldsymbol{A}^{it}\boldsymbol{B}\boldsymbol{B}^{-it}\|_1 = \|\boldsymbol{B}\|_1.$$

Indeed, $\boldsymbol{A}^{it}$ and $\boldsymbol{B}^{-it}$ are unitary matrices and the Schatten norms are unitarily invariant. We recognize that $M(0) = \sup_{t\in\mathbb{R}} |f(it)| \le \|\boldsymbol{B}\|_1$. Similarly, we have $M(1) \le \|\boldsymbol{A}\|_1$.

Finally, we can introduce these bounds into (7.7):

$$\|\boldsymbol{A}^{\theta}\boldsymbol{B}^{1-\theta}\|_1 = \max\{\operatorname{tr}(\boldsymbol{A}^{\theta}\boldsymbol{B}^{1-\theta}\boldsymbol{C}) : \|\boldsymbol{C}\|_\infty = 1\} \le \|\boldsymbol{A}\|_1^{\theta} \cdot \|\boldsymbol{B}\|_1^{1-\theta}.$$

The first relation follows from Exercise 7.17. This is the required result. ∎

Notes

The idea of using complex interpolation to prove convexity inequalities is due to Thorin. Littlewood described this approach as "the most impudent in mathematics, and brilliantly successful" [Gar07, qtd. p. 135]. The application to proving matrix inequalities is familiar to operator theorists, but it does not appear in the standard books on matrix analysis. The presentation here is due to the instructor.

We have given an independent proof of the maximum modulus principle, adapted to our purpose. The proof of the Hadamard three-lines theorem is also standard; for example, see [Gar07, Prop. 9.1.1]. You will find more information about maximum modulus principles on bounded and unbounded domains in any complex analysis text, such as Ahlfors [Ahl66].

The proof of duality for the Schatten norms is intended as an illustration of these ideas. Theorem 7.3 is due to Lust-Piquard, and the proof via complex interpolation is reported by Pisier & Xu [PX97, Lem. 1.1]. There is a beautiful application of complex interpolation to proving multivariate Golden–Thompson inequalities [SBT17]. The instructor has also used complex interpolation to develop new types of matrix concentration inequalities [Tro18].

Lecture bibliography

[Ahl66] L. V. Ahlfors. *Complex analysis: An introduction of the theory of analytic functions of one complex variable*. Second. McGraw-Hill Book Co., New York-Toronto-London, 1966.

[Gar07] D. J. H. Garling. *Inequalities: a journey into linear analysis*. Cambridge University Press, Cambridge, 2007. DOI: 10.1017/CBO9780511755217.

[PX97] G. Pisier and Q. Xu. "Non-commutative martingale inequalities". In: *Comm. Math. Phys.* 189.3 (1997), pages 667–698. DOI: 10.1007/s002200050224.

[Ste56] E. M. Stein. "Interpolation of linear operators". In: *Transactions of the American Mathematical Society* 83.2 (1956), pages 482–492.

[SBT17] D. Sutter, M. Berta, and M. Tomamichel. "Multivariate trace inequalities". In: *Comm. Math. Phys.* 352.1 (2017), pages 37–58. DOI: 10.1007/s00220-016-2778-5.

[Tro18] J. A. Tropp. "Second-order matrix concentration inequalities". In: *Appl. Comput. Harmon. Anal.* 44.3 (2018), pages 700–736. DOI: 10.1016/j.acha.2016.07.005.

8. Uniform Smoothness and Convexity

Date: 27 January 2022 **Scribe:** Ethan Epperly

Agenda:
1. Convexity and smoothness
2. Uniform smoothness for Schatten norms
3. Proof: Scalar case
4. Proof: Matrix case

In the previous two lectures, we studied a class of unitarily invariant norms known as the Schatten p-norms. Recall that the Schatten p-norm ($1 \le p < \infty$) is

$$\|\boldsymbol{A}\|_p := \left(\sum_{i=1}^n \sigma_i(\boldsymbol{A})^p\right)^{1/p} \quad \text{for } \boldsymbol{A} \in \mathbb{M}_n(\mathbb{F}).$$

Today, we will study the *geometry* of the space of matrices equipped with the Schatten p-norms. In particular, we shall answer the question "How *smooth* is the unit ball of the Schatten p-norm?" In answering this question, we will develop a powerful *uniform smoothness inequality* for Schatten p-norms, which has important applications in random matrix theory and other areas.

8.1 Convexity and smoothness

Let $(\mathsf{X}, \|\cdot\|)$ be a normed linear space. Our first task is to define a concept of smoothness and a dual concept of convexity for the space X. To gain intuition for these concepts, consider the pictoral depiction of a unit ball in Figure 8.1. At the point in orange labeled "very convex", the ball is more pointed and convex. The ball is more smooth and flat at the blue point labeled "very smooth".

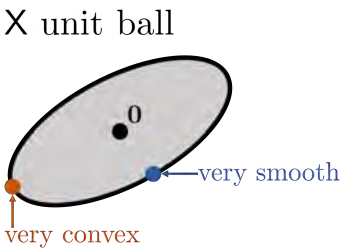

Figure 8.1 The unit ball of a normed linear space $(\mathsf{X}, \|\cdot\|)$ and two points on its boundary where the ball is smooth and convex.

As we suggested, the notions of convexity and smoothness are dual to each other. If a unit ball is very convex, then the dual unit ball will be very smooth. This duality is particularity apparent for polytopes, where the pointed vertices of polytope correspond to the flat facets of its dual. See Figure 8.2 for an illustration with the ℓ_1 and ℓ_∞ balls.

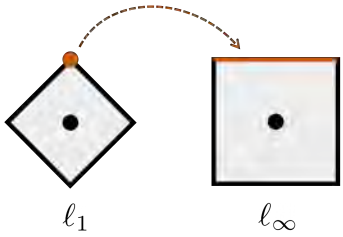

Figure 8.2 Duality of vertices and facets for the ℓ_1 and ℓ_∞ unit balls.

Let us develop a quantitive notion of convexity for the space X. Consider unit-norm vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{X}$. Then, by convexity of $\|\cdot\|$,

$$\left\|\frac{1}{2}(\boldsymbol{x}+\boldsymbol{y})\right\| \le \frac{1}{2}\|\boldsymbol{x}\| + \frac{1}{2}\|\boldsymbol{y}\| = 1.$$

We expect that if X is "very convex", then this inequality will be far from saturated. This motivates us to introduce the *modulus of continuity*:

$$\delta_{\mathsf{X}}(s) := \inf\left\{1 - \left\|\tfrac{1}{2}(\boldsymbol{x}+\boldsymbol{y})\right\| : \|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1,\ \|\boldsymbol{x}-\boldsymbol{y}\| = 2s\right\} \tag{8.1}$$

The modulus of convexity is defined for all $s \in [0,1]$. We illustrate the modulus of convexity in Figure 8.3.

Now, we turn our attention to defining a modulus of smoothness. For unit-norm vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{X}$, convexity implies

$$\frac{1}{2}\|\boldsymbol{x}+\tau\boldsymbol{y}\| + \frac{1}{2}\|\boldsymbol{x}-\tau\boldsymbol{y}\| \le 1 + \tau.$$

Once again, we expect this inequality to be far from saturated if the unit ball is "very smooth" at $\boldsymbol{x}$. We quantify the discrepancy by the modulus of smoothness:

$$\rho_{\mathsf{X}}(\tau) := \sup\left\{\frac{1}{2}\|\boldsymbol{x}+\tau\boldsymbol{y}\| + \frac{1}{2}\|\boldsymbol{x}-\tau\boldsymbol{y}\| - 1 : \|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1\right\}. \tag{8.2}$$

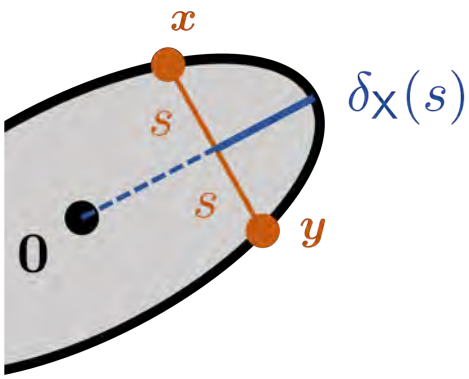

Figure 8.3 (Modulus of convexity). Graphical illustration of $\delta_{\mathsf{X}}(s)$, the modulus of convexity (8.1).

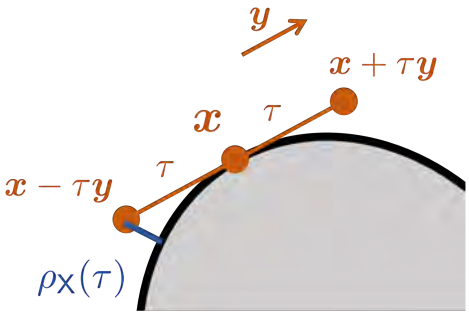

Figure 8.4 (Modulus of smoothness). Graphical illustration of $\rho_{\mathsf{X}}(\tau)$, the modulus of smoothness (8.2).

The modulus of smoothness is defined for all $\tau \in [0, +\infty)$.

The anticipated duality between convexity and smoothness is encapsulated in the following theorem. The basic qualitative result was obtained by M. M. Day, while the quantitative form here due to J. Lindenstrauss [Lin63].

Theorem 8.1 (Lindenstrauss). Let X be a normed linear space and X^* its dual. Then

$$\rho_{\mathsf{X}^*}(\tau) = \sup\{\tau s - \delta_{\mathsf{X}}(s) : s \in [0, 1]\}.$$

In other words, the modulus of smoothness of the dual of X is the Legendre–Fenchel conjugate of the modulus of convexity of X.

Informally, this theorem states that if X is very convex, then its dual X^* is very smooth. The converse is valid as well.

8.2 Uniform smoothness for Schatten norms

The main result of this lecture is an inequality that fully captures the uniform smoothness properties of matrices equipped with the Schatten p-norm ($p \geq 2$). We begin with the statement and its consequences, and then we outline some applications before turning to the proof.

Theorem 8.2 (Tomczak-Jaegermann 1974; Ball–Carlen–Lieb 1994). For $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{M}_n(\mathbb{F})$ and $p \geq 2$, we have

$$\left(\frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}\|_p^p + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_p^p\right)^{2/p} \leq \|\boldsymbol{X}\|_p^2 + (p-1)\|\boldsymbol{Y}\|_p^2. \tag{8.3}$$

The constant $p-1$ is optimal, even for $n = 1$.

Theorem 8.2, and indeed all of the results in this lecture, also hold for rectangular matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{M}^{m\times n}(\mathbb{F})$.

This result was first established by Tomczak-Jaegermann in the more general setting of trace-class operators on Hilbert spaces [Tom74]; she obtained the optimal constant $p-1$ for all even integers p. Ball, Carlen, and Lieb [BCL94] obtained the optimal constant of $p-1$ for all $p \geq 2$. We will present an alternative approach to the theorem, developed in 2021 by the instructor (unpublished).

Several remarks are in order:

- By considering diagonal matrices, we see that the same inequality holds for vectors $\boldsymbol{x}, \boldsymbol{y} \in \ell_p^n$ with $p \geq 2$.

  $$\left(\frac{1}{2}\|\boldsymbol{x}+\boldsymbol{y}\|_p^p + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_p^p\right)^{2/p} \leq \|\boldsymbol{x}\|_p^2 + (p-1)\|\boldsymbol{y}\|_p^2.$$

  In fact, all of the other results in this lecture have parallels for ℓ_p^n spaces and for L_p spaces.
- By Lyapunov's inequality,

  $$\frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}\|_p^2 + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_p^2 \leq \|\boldsymbol{X}\|_p^2 + (p-1)\|\boldsymbol{Y}\|_p^2. \tag{8.4}$$

  Equivalently, invoke the concavity of $t \mapsto t^{2/p}$ and Jensen's inequality.
- The uniform smoothness bound (8.4) is *reversed* for $p \in [1,2]$, yielding a *uniform convexity bound*:

  $$\frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}\|_p^2 + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_p^2 \geq \|\boldsymbol{X}\|_p^2 + (p-1)\|\boldsymbol{Y}\|_p^2 \quad \text{for } p \in [1,2]. \tag{8.5}$$

- For the Schatten 2-norm (i.e., the Frobenius norm), the bounds (8.3) and (8.4) holds with equality. This is the parallelogram law:

  $$\frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}\|_2^2 + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_2^2 = \|\boldsymbol{X}\|_2^2 + \|\boldsymbol{Y}\|_2^2,$$

  which holds for vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a Hilbert space with norm $\|\cdot\|$. See Figure 8.5 or an illustration.
- For $p \approx \log n$, we have $\|\cdot\|_p \approx \|\cdot\|_\infty$. See Exercise 8.5 for a quantitative statement. Plugging this relation into (8.4),

  $$\frac{1}{2}\|\boldsymbol{X}+\boldsymbol{Y}\|_\infty^2 + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_\infty^2 \lessapprox \|\boldsymbol{X}\|_\infty^2 + \log n \cdot \|\boldsymbol{Y}\|_\infty^2.$$

The next set of exercises supports these observations.

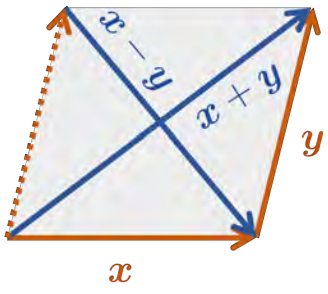

Figure 8.5 Geometry of the parallelogram law.

Exercise 8.3 (Optimality). Prove that (8.3) cannot be improved by replacing $p-1$ by a smaller constant. (**Hint:** Consider $n = 1$. Use a second order expansion for small $\boldsymbol{Y}$.)

Problem 8.4 (Duality). Deduce that the uniform convexity result (8.5) follows as a formal consequence of (8.4).

Exercise 8.5 (Approximating S_∞). For $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{F})$, show that

$$\|\boldsymbol{X}\|_\infty \le \|\boldsymbol{X}\|_p \le n^{1/p}\|\boldsymbol{X}\|_\infty \quad \text{for all } p \ge 1.$$

For all $n \ge 2$, conclude that

$$\|\boldsymbol{X}\|_\infty \le \|\boldsymbol{X}\|_p \le \sqrt{\mathrm{e}}\,\|\boldsymbol{X}\|_\infty \quad \text{for } p = 2\log n. \tag{8.6}$$

As a corollary of the uniform smoothness bound Theorem 8.2, we obtain the modulus of smoothness for the Schatten p-norm:

Corollary 8.6 (Modulus of smoothness). For the normed space $\mathsf{X} = (\mathbb{M}_n(\mathbb{F}), \|\cdot\|_p)$ with $p \ge 2$, the modulus of smoothness satisfies the bound

$$\varrho(\tau) \le \frac{1}{2}(p-1)\tau^2.$$

The result is sharp as $\tau \downarrow 0$ in the sense $\lim_{\tau\downarrow 0} \frac{1}{2}(p-1)\tau^2/\varrho(\tau) = 1$.

Proof. Fix $p \ge 2$ and consider matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\|\boldsymbol{X}\|_p = \|\boldsymbol{Y}\|_p = 1$. By convexity of the function $a \mapsto |a|^2 - 1$, we obtain the numeric inequality

$$a - 1 \le \frac{1}{2}(|a|^2 - 1) \quad \text{for } a \in \mathbb{R}.$$

Applying this inequality for $a = \|\boldsymbol{X} + \boldsymbol{Y}\|_p$ and $a = \|\boldsymbol{X} - \boldsymbol{Y}\|_p$ gives

$$\begin{aligned}\varrho_{\mathsf{X}}(\tau) &\le \frac{1}{2}(\|\boldsymbol{X} + \tau\boldsymbol{Y}\|_p - 1) + \frac{1}{2}(\|\boldsymbol{X} - \tau\boldsymbol{Y}\|_p - 1) \\ &\le \frac{1}{2}\left(\frac{1}{2}\|\boldsymbol{X} + \tau\boldsymbol{Y}\|_p^2 + \frac{1}{2}\|\boldsymbol{X} - \tau\boldsymbol{Y}\|_p^2 - 1\right).\end{aligned}$$

Now apply the uniform smoothness bound (8.4):

$$\varrho_{\mathsf{X}}(\tau) \le \frac{1}{2}\left(\|\boldsymbol{X}\|_p^2 + (p-1)\tau^2\|\boldsymbol{Y}\|_p^2 - 1\right) = \frac{1}{2}(p-1)\tau^2.$$

This is the promised conclusion. The optimality for small τ can be proven using a second-order expansion for small τ, as in Exercise 8.3. ∎

Qualitatively, this result shows that the unit ball of the Schatten p-norm becomes less smooth as p increases. By analogy to the ℓ_p spaces, this result is unsurprising as the ℓ_2 unit ball is round and smooth, whereas the ℓ_∞ unit ball has rough corners. Corollary 8.6 confirms that this intuition carries over to Schatten p-norms. See Figure 8.6 for a schematic drawing.

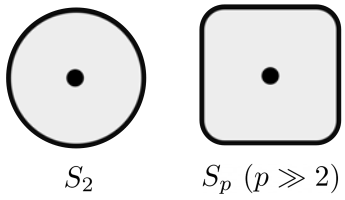

Figure 8.6 The unit ball of the Schatten p-norm becomes rougher as p increases.

Exercise 8.7 (Modulus of convexity). Use Theorem 8.1 and Corollary 8.6 to bound the modulus of convexity for the Schatten p-norm ($1 \le p \le 2$).

8.3 Application: Sum of independent random matrices

The uniform smoothness bound Theorem 8.2 has important implications for random matrix theory and related areas. We shall focus on one application: bounding the expected spectral norm of a sum of independent random matrices.

First, consider an independent family $(X_1, \dots, X_N)$ of centered real random variables. Then, by the additivity of the variance,

$$\mathbb{E}\left|\sum_{i=1}^N X_i\right|^2 = \sum_{i=1}^N \mathbb{E}\,|X_i|^2.$$

Recall that a random variable X is *centered* if it has expectation zero: $\mathbb{E}\,X = 0$.

This calculation carries over immediately to Hilbert spaces. For instance, for an independent family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)$ of centered random variables in the space of matrices $\mathbb{M}_n(\mathbb{F})$ equipped with the Schatten 2-norm,

$$\mathbb{E}\left\|\sum_{i=1}^N \boldsymbol{X}_i\right\|_2^2 = \sum_{i=1}^N \mathbb{E}\,\|\boldsymbol{X}_i\|_2^2. \tag{8.7}$$

This statement can be verified using the bilinearity of the trace inner product $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle := \operatorname{tr}(\boldsymbol{X}^*\boldsymbol{Y})$.

Unfortunately, Schatten 2-norm bounds are often uninformative for computational applications of random matrices [Tro15, pp. 122–123]. With the help of the uniform smoothness bound (8.4), we can obtain a Schatten p-norm inequality version of (8.7). By taking $p \approx \log n$, the result of Exercise 8.5 yields a Schatten ∞-norm bound (i.e., a bound in the spectral norm).

To begin, observe that the uniform smoothness bound (8.4) can be interpreted in probabilistic language. Introduce a random variable $\varepsilon \sim \textsc{uniform}\{\pm 1\}$. The conclusion of (8.4) can be reformulated as an expectation:

A random variable $\varepsilon \sim \textsc{uniform}\{\pm 1\}$ is referred to as a *Rademacher random variable*.

$$\mathbb{E}\,\|\boldsymbol{X} + \varepsilon\boldsymbol{Y}\|_p^2 \le \|\boldsymbol{X}\|_p^2 + (p-1)\cdot \mathbb{E}\,\|\varepsilon\boldsymbol{Y}\|_p^2.$$

We can improve this inequality in the same way that we can upgrade midpoint convexity to the full statement of Jensen's inequality. Indeed, the same bound holds if we replace $\varepsilon\boldsymbol{Y}$ by a general centered random matrix $\boldsymbol{Z}$.

Proposition 8.8 (Ricard–Xu 2016). Let $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{F})$ be a fixed matrix and $\boldsymbol{Z} \in \mathbb{M}_n(\mathbb{F})$ a centered random matrix. For $p \ge 2$,

$$\mathbb{E}\,\|\boldsymbol{X} + \boldsymbol{Z}\|_p^2 \le \|\boldsymbol{X}\|_p^2 + (p-1)\cdot \mathbb{E}\,\|\boldsymbol{Z}\|_p^2.$$

This result was orginally proven by Ricard and Xu [RX16] in the setting of von Neumann algebras. An elementary proof appears in the paper [Hua+21, Lem. A.1]. We present an even simpler proof adapted from [Nao12], which yields a slightly suboptimal constant of $2(p-1)$ in place of $p-1$.

Proof. We compute

$$\begin{aligned}\frac{1}{2}\big(\mathbb{E}\,\|\boldsymbol{X}+\boldsymbol{Z}\|_p^2 + \|\boldsymbol{X}\|_p^2\big) &\le \frac{1}{2}\big(\mathbb{E}\,\|\boldsymbol{X}+\boldsymbol{Z}\|_p^2 + \mathbb{E}\,\|\boldsymbol{X}-\boldsymbol{Z}\|_p^2\big)\\ &\le \|\boldsymbol{X}\|_p^2 + (p-1)\cdot\mathbb{E}\,\|\boldsymbol{Z}\|_p^2.\end{aligned}$$

The first inequality is Jensen's inequality, and the second is uniform smoothness (8.4). Rearranging gives the advertised result with a suboptimal constant $2(p-1)$. ∎

Applying this result iteratively yields inequalities for the expected Schatten norm of a sum of independent random matrices.

Corollary 8.9 (Sums of independent random matrices). Consider an independent family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)$ of independent centered random matrices in $\mathbb{M}_n(\mathbb{F})$. For each $p \ge 2$,

$$\mathbb{E}\left\|\sum_{i=1}^N \boldsymbol{X}_i\right\|_p^2 \le (p-1)\sum_{i=1}^N \mathbb{E}\,\|\boldsymbol{X}_i\|_p^2. \tag{8.8}$$

Consequently, for each $n \geq 3$,

$$\mathbb{E}\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\|_\infty \leq \sqrt{2\mathrm{e}\log n} \cdot \sqrt{\sum_{i=1}^{N} \mathbb{E}\,\|\boldsymbol{X}_i\|_\infty^2} \tag{8.9}$$

Proof sketch. The first conclusion follows from Proposition 8.8 applied iteratively to each $\boldsymbol{X}_k$, conditional on the realization of $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{k-1})$. For the second conclusion, we derive choose $p = 2\log n$ to obtain

$$\begin{aligned}\left(\mathbb{E}\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\|_\infty\right)^2 &\leq \mathbb{E}\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\|_\infty^2 \leq \mathbb{E}\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\|_p^2 \\ &\leq (p-1)\sum_{i=1}^{N} \mathbb{E}\,\|\boldsymbol{X}_i\|_p^2 \leq 2\mathrm{e}\log n \cdot \sum_{i=1}^{N} \mathbb{E}\,\|\boldsymbol{X}_i\|_\infty^2.\end{aligned}$$

The first inequality is Jensen, the second and fourth are the norm comparison (8.6), and the third is (8.8). Taking square roots gives the desired conclusion. ∎

Uniform smoothness can be used to derive matrix concentration inequalities for other random matrix models, such as products of independent random matrices in [Hua+21]. See [Tro15] for a survey of exponential matrix concentration inequalities, based on more sophisticated tools from matrix analysis. The $\sqrt{\log n}$ prefactor in (8.9) is known to be necessary [Tro15, pp. 114–115], but it can sometimes be removed with still more difficult tools [BBv21].

> **Aside: (Uniform smoothness and quantum computation).** As another application, recent work has used uniform smoothness improved error bounds for simulating the time-evolution of quantum systems using Trotter formulas, a workhorse algorithm in quantum computation [CB21].

8.4 Proof: Scalar case

Having discussed applications of the uniform smoothness bound (8.3), we now turn to the task of proving it. To reduce notational clutter, we shall introduce probabilistic notation, as we already saw in (8.3). Let $x, y \in$ be real numbers, and let ε be a Rademacher random variable. Our task is to prove the following inequality:

$$(\mathbb{E}\,|x + \varepsilon y|^p)^{2/p} \leq |x|^2 + (p-1)\cdot|y|^2. \tag{8.10}$$

This is known as the *Gross two-point inequality*. To simplify, we shall assume p is an even natural number. Lifting this restriction is left to the reader (Exercise 8.11 and Problem 8.12).

Our approach will be based on *interpolation*. The idea is straightforward. We define a path between a simple object that we understand and a more complicated object. We obtain inequalities by controlling the derivative along the interpolation path. There are other simpler ways of proving the Gross two-point inequality (8.10), but the proof by interpolation generalizes directly to matrices. In this task, the following lemma will be helpful.

Lemma 8.10 (Mean-value inequality). Let $\varphi : \mathbb{R} \to \mathbb{R}$ be such that φ' is convex. Then, for all $b, a \in \mathbb{R}$,

$$(b-a)(\varphi(b) - \varphi(a)) \leq \frac{1}{2}(b-a)^2(\varphi'(b) + \varphi'(a)).$$

Proof. By the fundamental theorem of calculus and convexity,

$$\begin{aligned}(b-a)(\varphi(b)-\varphi(a)) &= (b-a)^2 \int_0^1 \varphi'((1-\tau)b+\tau a)\,\mathrm{d}\tau \\ &\le (b-a)^2 \int_0^1 \big[(1-\tau)\varphi'(b)+\tau\varphi'(a)\big]\,\mathrm{d}\tau \\ &= \frac{1}{2}(b-a)^2(\varphi'(b)+\varphi'(a)).\end{aligned}$$

This is the desired conclusion. ∎

With this result in hand, we prove the Gross two-point inequality (8.10).

Proof of Theorem 8.2 (scalar case). Assume that p is an even natural number. Define the interpolant

$$u(t) := \mathbb{E}\,|x+\varepsilon\sqrt{t}y|^p \quad \text{for } t \in [0,1]. \tag{8.11}$$

Observe that

$$u(0) = |x|^p \quad \text{and} \quad u(1) = \mathbb{E}\,|x+\varepsilon y|^p.$$

The value $u(1)$ at the endpoint is the left-hand side of Gross's inequality (8.10), raised to the $p/2$ power.

With the help of the mean-value inequality (Lemma 8.10) for the function $\varphi : a \mapsto |a|^{p-1}$, we compute and bound the derivative of the interpolant u:

$$\begin{aligned}\dot{u}(t) &= p \cdot \frac{1}{\sqrt{2}t} \cdot \mathbb{E}\left[(\varepsilon y)(x+\varepsilon y\sqrt{t})^{p-1}\right] \\ &= p \cdot \frac{1}{4t}(2y\sqrt{t}) \cdot \left[\frac{1}{2}(x+y\sqrt{t})^{p-1} - \frac{1}{2}(x-y\sqrt{t})^{p-1}\right] \\ &\le \frac{1}{2}p(p-1)y^2\left[\frac{1}{2}(x+y\sqrt{t})^{p-2} + \frac{1}{2}(x-y\sqrt{t})^{p-2}\right] \\ &= \frac{1}{2}p(p-1)y^2 \cdot \mathbb{E}(x-y\sqrt{t})^{p-2} \\ &\le \frac{1}{2}p(p-1)y^2 \cdot \left[\mathbb{E}(x-y\sqrt{t})^p\right]^{1-2/p} \\ &= \frac{1}{2}p(p-1)y^2 \cdot u(t)^{1-2/p}.\end{aligned}$$

The first inequality is the mean-value inequality Lemma 8.10, and the second inequality is Lyapunov's inequality. We have obtained a differential inequality for the function u.

To solve this inequality, define the function $v(t) = u(t)^{2/p}$ for $t \in [0,1]$. Then $v(0) = |x|^2$, while $v(1)$ is the left-hand side of Gross' inequality (8.10). Finally, we use the fundamental theorem of calculus to compute

$$\begin{aligned}(\mathbb{E}\,|x+\varepsilon y|^p)^{2/p} - |x|^2 = v(1) - v(0) &= \int_0^1 \dot{v}(t)\,\mathrm{d}t \\ &= \int_0^1 \frac{2}{p}u(t)^{2/p-1} \cdot \dot{u}(t)\,\mathrm{d}t \\ &\le (p-1)\int_0^1 y^2\,\mathrm{d}t = (p-1)y^2.\end{aligned} \tag{8.12}$$

The inequality is transferred from the last display. Rearranging gives the Gross two-point inequality (8.10). ∎

Why define the interpolant u as (8.11) with the *square root* dependence $\sqrt{t}$ instead of the more natural-seeming *linear* interpolant

$$u(t) := \mathbb{E}\,|x+\varepsilon t y|^p ?$$

In fact, the linear choice also works, but it gives a suboptimal constant of $2(p-1)$ in place of $p-1$.

Exercise 8.11 (Gross: $p \geq 3$). Modify the proof of Gross' two-point inequality (8.10) to handle the case $p \geq 3$. **Hint:** This is essentially the same proof; the only change is that signum functions arise from the derivative.

Problem 8.12 (Gross: $p \in (2,3)$). Prove Gross' two-point inequality (8.10) for $p \in (2,3)$. **Hint:** The derivative φ' is now *concave*. Modify Lemma 8.10, bounding the integral by the midpoint rule.

8.5 Proof: Matrix case

We shall now extend the proof of Gross' two-point inequality for scalars (8.10) to prove the uniform smoothness bound (8.3) for matrices.

8.5.1 Setup

The first step is a standard tool in matrix analysis; we transform possibly non-Hermitian matrices $\boldsymbol{X}, \boldsymbol{Y}$ to Hermitian matrices by passing to the the *Hermitian dilation*:

$$\boldsymbol{M} \in \mathbb{M}_n(\mathbb{F}) \quad \text{maps to} \quad \mathscr{H}(\boldsymbol{M}) := \begin{bmatrix} \mathbf{0} & \boldsymbol{M} \\ \boldsymbol{M}^* & \mathbf{0} \end{bmatrix} \in \mathbb{H}_{2n}(F).$$

As the following exercise shows, this allows us to assume without loss of generality that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Hermitian.

Exercise 8.13 (Hermitian dilation). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{F})$. Prove the following properties of the Hermitian dilation:

1. The mapping $\mathscr{H} : \mathbb{M}_n(\mathbb{F}) \to \mathbb{H}_{2n}(\mathbb{F})$ is a *real-linear* function. That is, for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n(\mathbb{F})$ and $\alpha, \beta \in \mathbb{R}$,
$$\mathscr{H}(\alpha \boldsymbol{A} + \beta \boldsymbol{B}) = \alpha \mathscr{H}(\boldsymbol{A}) + \beta \mathscr{H}(\boldsymbol{B}).$$
2. The eigenvalues of $\mathscr{H}(\boldsymbol{A})$ are equal to $\pm \sigma_i(\boldsymbol{A})$ for $i = 1, 2, \ldots, n$.
3. Schatten norms of $\boldsymbol{A}$ and $\mathscr{H}(\boldsymbol{A})$ are proportional: $\|\mathscr{H}(\boldsymbol{A})\|_p = 2^{1/p} \|\boldsymbol{A}\|_p$.

For a Hermitian matrix $\boldsymbol{M}$, the Schatten p-norm is given by a trace:

$$\|\boldsymbol{M}\|_p = (\operatorname{tr} |\boldsymbol{M}|^p)^{1/p} \quad \text{where } |\boldsymbol{M}| = (\boldsymbol{M}^* \boldsymbol{M})^{1/2}.$$

For even p and Hermitian $\boldsymbol{M}$, this expression simplifies further: $\|\boldsymbol{M}\|_p = (\operatorname{tr} \boldsymbol{M}^p)^{1/p}$. This reformulation is useful because it will allow us to leverage the linearity and cyclicity of the trace in our calculations.

8.5.2 Trace functions and inequalities

The linearity and cyclicity of the trace allow for many scalar inequalities to be "upgraded" to trace inequalities by systematic arguments. We introduce one such technique, the generalized Klein inequality, in this section. In order to do this, we begin by defining matrix functions:

Definition 8.14 (Standard matrix function). Let $\varphi : \mathbb{R} \to \mathbb{R}$. We extend this function to Hermitian matrices $\varphi : \mathbb{H}_n \to \mathbb{H}_n$ as follows. Let $\boldsymbol{A} \in \mathbb{H}_n$ be a matrix with spectral resolution

$$\boldsymbol{A} = \sum\nolimits_i \lambda_i \boldsymbol{P}_i.$$

Then
$$\varphi(\boldsymbol{A}) := \sum\nolimits_i \varphi(\lambda_i)\boldsymbol{P}_i \in \mathbb{H}_n(\mathbb{F})$$

We will need the following standard derivative computation.

Proposition 8.15 (Derivative of trace power). For each natural number $p \in \mathbb{N}$ and all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n(\mathbb{F})$,
$$\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{tr}(\boldsymbol{A} + t\boldsymbol{H})^p \Big|_{t=0} = p \operatorname{tr}[\boldsymbol{A}^{p-1}\boldsymbol{H}].$$

Proof sketch. The key observation is the identity
$$\boldsymbol{B}^p - \boldsymbol{A}^p = \sum\nolimits_{k=0}^{p-1} \boldsymbol{B}^k(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{p-1-k} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n.$$
This formula can be verified by direct algebraic manipulation. ∎

Exercise 8.16 (Derivative of trace power). Prove Proposition 8.15.

There are a variety of techniques for deriving trace inequalities from scalar inequalities. We shall make use of the following result.

Proposition 8.17 (Generalized Klein inequality). Suppose that $f_k, g_k : \mathbb{R} \to \mathbb{R}$ are functions such that
$$\sum\nolimits_k f_k(a)\, g_k(b) \geq 0 \quad \text{for all } a, b \in \mathbb{R}.$$
Then
$$\sum\nolimits_k \operatorname{tr}\left[f_k(\boldsymbol{A})\, g_k(\boldsymbol{B})\right] \geq 0 \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathbb{F}).$$

Proof. Introduce spectral resolutions
$$\boldsymbol{A} = \sum\nolimits_i \lambda_i \boldsymbol{P}_i \quad \text{and} \quad \boldsymbol{B} = \sum\nolimits_j \mu_j \boldsymbol{Q}_j.$$
Then calculate
$$\begin{aligned}\sum\nolimits_k \operatorname{tr}\left[f_k(\boldsymbol{A})\, g_k(\boldsymbol{B})\right] &= \sum\nolimits_k \operatorname{tr}\left[\left(\sum\nolimits_i f_k(\lambda_i)\boldsymbol{P}_i\right)\left(\sum\nolimits_j g_k(\mu_j)\boldsymbol{Q}_j\right)\right] \\ &= \sum\nolimits_{i,j}\left[\sum\nolimits_k f_k(\lambda_i)\, g_k(\mu_j)\right] \cdot \operatorname{tr}(\boldsymbol{P}_i\boldsymbol{Q}_j).\end{aligned}$$
The bracketed quantity is positive by assumption and $\operatorname{tr}(\boldsymbol{P}_i\boldsymbol{Q}_j)$ is positive because the trace of the product of two positive-semidefinite matrices is positive. Therefore,
$$\sum\nolimits_k \operatorname{tr}\left[f_k(\boldsymbol{A})\, g_k(\boldsymbol{B})\right] \geq 0,$$
as claimed. ∎

Exercise 8.18 (Trace of psd product). Prove that $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \geq 0$ for positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. **Hint:** Write $\boldsymbol{B} = \boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}$ and cycle the trace.

As a corollary, we obtain a trace version of the mean-value inequality, Lemma 8.10.

Corollary 8.19 (Mean-value trace inequality). Let $\varphi : \mathbb{R} \to \mathbb{R}$ be such that $\psi := \varphi'$ is convex. Then for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathbb{F})$,
$$\operatorname{tr}\left[(\boldsymbol{B} - \boldsymbol{A})(\varphi(\boldsymbol{A}) - \varphi(\boldsymbol{B}))\right] \leq \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{B} - \boldsymbol{A})^2\,(\psi(\boldsymbol{B}) + \psi(\boldsymbol{A}))\right].$$

Exercise 8.20 (Mean-value trace inequality). Prove Corollary 8.19 using the mean-value inequality (Lemma 8.10) and the generalized Klein inequality (Proposition 8.17).

Exercise 8.21 (Derivative of a power). Derive the result in Proposition 8.15 using the generalized Klein inequality (Proposition 8.17). Extend this argument to compute the derivative of $\operatorname{tr}|\cdot|^p$ for all $p > 0$.

8.5.3 Proof of Theorem 8.2

With these ingredients, we may complete the proof of the uniform smoothness bound for Schatten p-norms.

Proof of Theorem 8.2. By passing to the Hermitian dilation if necessary, we can assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Hermitian. We again restrict attention to p an even integer. We leave the proof for general $p > 2$ as an exercise.

Define the interpolant

$$u(t) := \operatorname{tr}\left[\boldsymbol{X} + \varepsilon\boldsymbol{Y}\sqrt{t}\right] \quad \text{for } t \in [0,1].$$

We now compute and bound the derivative of u using our computation for the derivative of the trace power (Proposition 8.15) and the mean-value trace inequality (Corollary 8.19):

$$\begin{aligned}
\dot{u}(t) &= p \cdot \frac{1}{2\sqrt{t}}\,\mathbb{E}\operatorname{tr}\left[(\varepsilon\boldsymbol{Y})(\boldsymbol{X} + \varepsilon\boldsymbol{Y}\sqrt{t})^{p-1}\right] \\
&= \frac{1}{2}p \cdot \frac{1}{4t}\operatorname{tr}\left[(2\sqrt{t}\boldsymbol{Y})\left(\left(\boldsymbol{X} + \boldsymbol{Y}\sqrt{t}\right)^{p-1} - \left(\boldsymbol{X} - \boldsymbol{Y}\sqrt{t}\right)^{p-1}\right)\right] \\
&\le \frac{1}{2}p \cdot \operatorname{tr}\left[\boldsymbol{Y}^2\left(\frac{1}{2}(p-1)\left(\boldsymbol{X} + \boldsymbol{Y}\sqrt{t}\right)^{p-2} + \frac{1}{2}(p-1)\left(\boldsymbol{X} + \boldsymbol{Y}\sqrt{t}\right)^{p-2}\right)\right] \\
&= \frac{1}{2}p(p-1) \cdot \mathbb{E}\operatorname{tr}\left[\boldsymbol{Y}^2(\boldsymbol{X} + \varepsilon\boldsymbol{Y}\sqrt{t})^{p-2}\right] \\
&\le \frac{1}{2}p(p-1)\,(\operatorname{tr}\boldsymbol{Y}^p)^{2/p}\,\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{X} + \varepsilon\boldsymbol{Y}\sqrt{t}\right)^p\right]^{1-2/p} \\
&\le \frac{1}{2}p(p-1)\,(\operatorname{tr}\boldsymbol{Y}^p)^{2/p}\left[\mathbb{E}\operatorname{tr}\left(\boldsymbol{X} + \varepsilon\boldsymbol{Y}\sqrt{t}\right)^p\right]^{1-2/p} \\
&= \frac{1}{2}p(p-1)\|\boldsymbol{Y}\|_p^2 \cdot u(t)^{1-2/p}.
\end{aligned}$$

By convention, the trace binds after nonlinear operations. For example, $\operatorname{tr}\boldsymbol{Y}^p = \operatorname{tr}(\boldsymbol{Y}^p)$.

The first equality is Proposition 8.15. The three inequalities are Corollary 8.19, the Hölder inequality for Schatten norms, and Lyapunov's inequality. The remainder of the proof follows from the same calculation (8.12) as the scalar case. ∎

Problems

Problem 8.22 (Matrix Khintchine inequality). The uniform smoothness inequalities in this lecture can be substantially improved using the same method. Consider a self-adjoint random matrix of the form

$$\boldsymbol{X} = \boldsymbol{B} + \sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{A}_i \quad \text{where } \varepsilon_i \text{ are i.i.d. Rademacher.}$$

The self-adjoint matrices $\boldsymbol{B}, \boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{H}_d$ are fixed.

1. For $p \geq 2$, prove the matrix Khintchine inequality:

$$\mathbb{E}\,\|\boldsymbol{X}\|_p^2 \leq \|\boldsymbol{B}\|_p^2 + (p-1)\cdot\left\|\sum\nolimits_{i=1}^n \boldsymbol{A}_i^2\right\|_p.$$

2. For *even* p, show that the constant can be improved to $[(p-1)!!]^{1/p}$.

Notes

Leonard Gross, for whom the Gross two-point inequality (8.10) is named, is a mathematician and mathematical physicist who did foundational work on quantum field theory and statistical physics. He is among the early developers of logarithmic Sobolev inequalities, which can be used to prove concentration inequalities for nonlinear functions of independent random variables [van14, §3].

Interpolation is a very effective tool. Another relatively accessible application uses interpolation to prove comparison inequalities for Gaussian processes, which can be used to prove sharp bounds on the expected spectral norm of a standard Gaussian random matrix [Ver18, §§7.2–7.3]. An advanced application of these ideas can be used to develop matrix concentration inequalities which sometimes circumvent the logarithmic factor we obtained in (8.9); see [BBv21].

The Hermitian dilation has an old history in matrix analysis and operator theory. Jordan first used the Hermitian dilation in 1874 in his discovery of the singular value decomposition. (The singular value decomposition was also discovered independently the year prior by Beltrami.) The Hermitian dilation was popularized by Wielandt and is often referred to as the Jordan–Wielandt matrix for this reason. See [SS90, pp. 34–35] for a discussion of this history.

Lecture bibliography

[BCL94] K. Ball, E. A. Carlen, and E. H. Lieb. "Sharp Uniform Convexity and Smoothness Inequalities for Trace Norms". In: *Invent Math* 115.1 (Dec. 1994), pages 463–482. DOI: 10.1007/BF01231769.

[BBv21] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel. "Matrix Concentration Inequalities and Free Probability". In: *arXiv:2108.06312 [math]* (Aug. 2021). arXiv: 2108.06312 [math].

[CB21] C.-F. Chen and F. G. S. L. Brandão. "Concentration for Trotter Error". Available at https://arXiv.org/abs/2111.05324. Nov. 2021. arXiv: 2111.05324 [math-ph, physics:quant-ph].

[Hua+21] D. Huang et al. "Matrix Concentration for Products". In: *Foundations of Computational Mathematics* (2021), pages 1–33.

[Lin63] J. Lindenstrauss. "On the Modulus of Smoothness and Divergent Series in Banach Spaces." In: *Michigan Mathematical Journal* 10.3 (1963), pages 241–252.

[Nao12] A. Naor. "On the Banach-Space-Valued Azuma Inequality and Small-Set Isoperimetry of Alon–Roichman Graphs". In: *Combinatorics, Probability and Computing* 21.4 (July 2012), pages 623–634. DOI: 10.1017/S0963548311000757.

[RX16] É. Ricard and Q. Xu. "A Noncommutative Martingale Convexity Inequality". In: *The Annals of Probability* 44.2 (Mar. 2016), pages 867–882. DOI: 10.1214/14-AOP990.

[SS90] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. 1st Edition. Academic Press, 1990.

[Tom74] N. Tomczak-Jaegermann. "The Moduli of Smoothness and Convexity and the Rademacher Averages of the Trace Classes S_p $(1 \leq p < \infty)$". In: *Studia Mathematica* 50.2 (1974), pages 163–182.

[Tro15] J. A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pages 1–230.

[van14] R. van Handel. *Probability in High Dimension*. Technical report. Princeton University, June 2014. DOI: 10.21236/ADA623999.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

9. Additive Perturbation Theory

Date: 1 February 2022 **Scribe:** Salvador Rey Gomez De La Cruz

Matrix perturbation theory studies how a function of a matrix changes as the matrix changes. In particular, we may ask how the eigenvalues of an Hermitian matrix change under additive perturbation. That is, for Hermitian matrices $\boldsymbol{A}, \boldsymbol{E}$ of the same size, how do the eigenvalues of the perturbed matrix $\boldsymbol{A} + \boldsymbol{E}$ compare with the eigenvalues of $\boldsymbol{A}$? Today, we will develop Lidskii's theorem, a result that provides a very detailed answer to this question. This is one of the most important applications of majorization theory in matrix analysis.

Agenda:
1. Poincaré theorem
2. Courant–Fischer–Weyl minimax principle
3. Weyl monotonicity principle
4. Lidskii theorem

9.1 Variational principles

The first step toward Lidskii's theorem is a set of variational principles, which describe the eigenvalues of a matrix as solutions to optimization problems. Our first result is due to Poincaré. This theorem gives upper and lower bounds on eigenvalues in terms of quadratic forms. Afterward, we will see that these bounds lead to exact representations of the eigenvalues.

Theorem 9.1 (Poincaré). Let $\boldsymbol{A} \in \mathbb{H}_n$, and let $\mathsf{L} \subseteq \mathbb{C}^n$ be a subspace with dimension k. There are unit vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{L}$ such that

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle \le \lambda_k^{\downarrow}(\boldsymbol{A}) \quad \text{and}$$
$$\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y} \rangle \ge \lambda_k^{\uparrow}(\boldsymbol{A}).$$

The larger the dimension of the subspace L, the smaller the quadratic can be. Similarly, the larger the dimension, the larger the quadratic form can be.

Proof. The proof relies on dimension counting. Let $(\lambda_i^{\downarrow}, \boldsymbol{u}_i)$ be the eigenpairs of $\boldsymbol{A}$. The set $(\boldsymbol{u}_i : i = 1, \dots n)$ comprises an orthonormal basis of $\mathbb{C}^n$. In other words, $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij}$.

Consider the subspace $\mathsf{S} = \operatorname{span}\{\boldsymbol{u}_k, \dots, \boldsymbol{u}_n\}$. Observe that $\dim \mathsf{L} + \dim \mathsf{S} = n + 1$. Since the total dimension is greater than n, the intersection $\mathsf{L} \cap \mathsf{S}$ is a *nontrivial* subspace. Since the subspace is nontrivial, it contains a unit-norm vector $\boldsymbol{x} \in \mathsf{L} \cap \mathsf{S}$. We may express this vector in the distinguished basis for the subspace:

$$\boldsymbol{x} = \sum\nolimits_{i=k}^{n} \alpha_i \boldsymbol{u}_i \quad \text{where } \alpha_i \in \mathbb{C}.$$

Because $\boldsymbol{x}$ is a unit vector and the basis is orthonormal,

$$1 = \|\boldsymbol{x}\|^2 = \sum\nolimits_{i=k}^{n} |\alpha_i|^2.$$

Because $\boldsymbol{u}_i$ are eigenvectors of $\boldsymbol{A}$, we find that

$$\boldsymbol{A}\boldsymbol{x} = \sum\nolimits_{i=k}^{n} \alpha_i \lambda_i^{\downarrow} \boldsymbol{u}_i.$$

We may now evaluate the quadratic form $\langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle$. Indeed,

$$\begin{aligned}\langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle &= \sum\nolimits_{i=k}^{n} \sum\nolimits_{j=k}^{n} \langle \alpha_j \boldsymbol{u}_j,\ \alpha_i \lambda_i^{\downarrow} \boldsymbol{u}_i\rangle \\ &= \sum\nolimits_{j=k}^{n} |\alpha_j|^2 \lambda_j^{\downarrow} \\ &\le \lambda_k^{\downarrow} \sum\nolimits_{j=k}^{n} |\alpha_j|^2 = \lambda_k^{\downarrow}.\end{aligned}$$

The final inequality comes from the fact that the eigenvalues $\lambda_j^{\downarrow}$ are arranged in descending order. This proves the first inequality in the statement of the theorem.

In order to prove the second inequality, we simply replace $\boldsymbol{A}$ with $-\boldsymbol{A}$. This operation yields a unit vector $\boldsymbol{x} \in \mathsf{L}$ with the property that

$$-\langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle \le \lambda_k^{\downarrow}(-\boldsymbol{A}) = -\lambda_k^{\uparrow}(\boldsymbol{A}).$$

In the first line, we used the fact that the eigenvalues of $-\boldsymbol{A}$ are the negatives eigenvalues of $\boldsymbol{A}$. Because of the change in sign, the order reverses. This is the second statement in the theorem. ∎

Theorem 9.1 yields a corollary that provides a minimax representation for the ordered eigenvalues.

Corollary 9.2 (Courant–Fischer–Weyl minimax principle). Let $\boldsymbol{A} \in \mathbb{H}_n$. Then

$$\lambda_k^{\downarrow}(\boldsymbol{A}) = \max_{\substack{\mathsf{L}\subseteq\mathbb{C}^n \\ \dim(\mathsf{L})=k}} \min_{\substack{\|\boldsymbol{x}\|^2=1 \\ \boldsymbol{x}\in\mathsf{L}}} \langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle = \min_{\substack{\mathsf{S}\subseteq\mathbb{C}^n \\ \dim(\mathsf{S})=n-k+1}} \max_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{S}}} \langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle$$

This result represents eigenvalues as quadratic forms. Let us emphasize that these expressions are *linear* in the matrix $\boldsymbol{A}$.

Proof. As before, let $(\lambda_i^{\downarrow}(\boldsymbol{A}), \boldsymbol{u}_i)$ denote the eigenpairs of $\boldsymbol{A}$. By Theorem 9.1, each subspace $\mathsf{L} \subseteq \mathbb{C}^n$ with dimension k contains a unit vector $\boldsymbol{x} \in \mathsf{L}$ such that

$$\lambda_k^{\downarrow}(\boldsymbol{A}) \ge \langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle \ge \min_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{L}}} \langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle.$$

By choosing $\mathsf{L} = \operatorname{span}\{\boldsymbol{u}_1, \dots \boldsymbol{u}_k\}$, we see that equality can obtain.

For the second statement in Theorem 9.2, apply the first statement to the matrix $-\boldsymbol{A}$ instead of $\boldsymbol{A}$. Recall that $\lambda_k^{\downarrow}(-\boldsymbol{A}) = -\lambda_{n-k+1}^{\downarrow}(\boldsymbol{A})$ and that $\min(a) = -\max(a)$. ∎

A direct consequence of Corollary 9.2 is the Rayleigh–Ritz theorem, a fundamental result with wide impact in computational mathematics. In particular, it offers a way of approaching eigenvalue estimates by means of optimization tools.

Corollary 9.3 (Rayleigh–Ritz). Let $\boldsymbol{A} \in \mathbb{H}_n$. Then

$$\begin{aligned}\lambda_1^{\downarrow}(\boldsymbol{A}) &= \max_{\|\boldsymbol{x}\|^2=1} \langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle, \quad \text{and} \\ \lambda_n^{\downarrow}(\boldsymbol{A}) &= \min_{\|\boldsymbol{y}\|^2=1} \langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle.\end{aligned}$$

Proof. To prove this result, simply let $k = n$ and $k = 1$ in the first and second statement of Corollary 9.2, respectively. ∎

Corollary 9.3 has some striking implications:

- The maximum eigenvalue map $\boldsymbol{A} \mapsto \lambda_{\max}(\boldsymbol{A})$ is a convex function on $\mathbb{H}_n$.
- The minimum eigenvalue map $\boldsymbol{A} \mapsto \lambda_{\min}(\boldsymbol{A})$ is a concave function on $\mathbb{H}_n$.

Indeed, we observe that the maximum eigenvalue is the maximum of linear functions of the matrix, so is a convex function. Likewise, the minimum eigenvalue is a minimum of linear functions, so is concave.

Exercise 9.4 (Spectral condition number convex cone). For $c \geq 1$, consider the set $\mathsf{K}(c) := \{\boldsymbol{A} \in \mathbb{H}_n : \kappa_2(\boldsymbol{A}) \leq c\}$ of matrices whose spectral conditional number is bounded by c. Show that $\mathsf{K}(c)$ is a closed convex cone.

Recall that the spectral condition number $\kappa_2(\boldsymbol{A}) := \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$ for each Hermitian matrix $\boldsymbol{A}$.

Exercise 9.5 (Rayleigh–Ritz via diagonalization). Prove Corollary 9.3 by diagonalizing $\boldsymbol{A}$ and computing the quadratic form explicitly.

Exercise 9.6 (Rayleigh quotient). Recall that the Rayleigh quotient is defined as

$$R(\boldsymbol{A}; \boldsymbol{x}) := \frac{\langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle}{\langle \boldsymbol{x},\ \boldsymbol{x}\rangle} \quad \text{for } \boldsymbol{x} \neq \boldsymbol{0}.$$

Prove Corollary 9.3 by differentiating the Rayleigh quotient with respect to $\boldsymbol{x}$ and finding the second-order stationary points.

9.2 Weyl monotonicity principle

The first step toward analyzing additive perturbations is to study a special case where the perturbation is a positive-semidefinite matrix. This result is called the Weyl monotonicity principle.

9.2.1 Positive-semidefinite matrices

First, we recall the the definition of a positive-semidefinite matrix and the concept of the positive-semidefinite order on Hermitian matrices.

Definition 9.7 (Positive semidefinite matrix). A complex Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathbb{C})$ is positive semidefinite (psd) when $\langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle \geq 0$ for all $\boldsymbol{x} \in \mathbb{C}^n$.

Definition 9.8 (Positive semidefinite order). For Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$, we write $\boldsymbol{B} \succcurlyeq \boldsymbol{A}$ when $\boldsymbol{B} - \boldsymbol{A}$ is a positive semidefinite matrix.

9.2.2 Weyl monotonicity

With these definitions out of the way, we can give a rigorous statement of the Weyl monotonicity principle.

Corollary 9.9 (Weyl monotonicity principle). Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n$, and assume that $\boldsymbol{H}$ is psd. Then

$$\lambda_k^{\downarrow}(\boldsymbol{A} + \boldsymbol{H}) \geq \lambda_k^{\downarrow}(\boldsymbol{A}) \quad \text{for each } k = 1, \dots, n.$$

Equivalently, $\boldsymbol{B} \succcurlyeq \boldsymbol{A}$ implies that $\lambda_k^{\downarrow}(\boldsymbol{B}) \geq \lambda_k^{\downarrow}(\boldsymbol{A})$ for each index k.

Warning 9.10 (Eigenvalue increase). The converse of the Weyl monotonicity principle is not true! The conditions $\lambda_k^{\downarrow}(\boldsymbol{B}) \geq \lambda_k^{\downarrow}(\boldsymbol{A})$ do not imply that $\boldsymbol{B} \succcurlyeq \boldsymbol{A}$. ∎

Proof. Let S_{n-k+1} be the subspace that represents $\lambda_k^{\downarrow}(\boldsymbol{A}+\boldsymbol{H})$ in the second (minimax) relation in Corollary 9.2. Then

$$\begin{aligned}\lambda_k^{\downarrow}(\boldsymbol{A}+\boldsymbol{H}) &= \max_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{S}_{n-k+1}}} \langle \boldsymbol{y},\ (\boldsymbol{A}+\boldsymbol{H})\boldsymbol{y}\rangle = \max_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{S}_{n-k+1}}} \left(\langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle + \langle \boldsymbol{y},\ \boldsymbol{H}\boldsymbol{y}\rangle\right) \\ &\geq \max_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{S}_{n-k+1}}} \langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle \\ &\geq \min_{\substack{\mathsf{S}\subseteq\mathbb{C}^n \\ \dim \mathsf{S}=n-k+1}} \max_{\substack{\|\boldsymbol{y}\|^2=1 \\ \boldsymbol{y}\in\mathsf{S}}} \langle \boldsymbol{y},\ \boldsymbol{A}\boldsymbol{y}\rangle \\ &= \lambda_k^{\downarrow}(\boldsymbol{A}).\end{aligned}$$

The first inequality is valid because $\boldsymbol{H}$ is psd. The second inequality occurs because S_{n-k+1} represents $\lambda_k^{\downarrow}(\boldsymbol{A}+\boldsymbol{H})$ via the minimax principle, so the minimum over all subspaces $\mathsf{S}\subseteq\mathbb{C}^n$ with dimension $n-k+1$ has to be smaller. The final equality follows from Corollary 9.2.

In order to prove the equivalence, simply note that $\boldsymbol{B}\succcurlyeq\boldsymbol{A}$ if and only if $\boldsymbol{H}:=\boldsymbol{B}-\boldsymbol{A}\succcurlyeq\boldsymbol{0}$. Then invoke the first result. ∎

Exercise 9.11 (Eigenvalue increase). Consider the family of relations $\lambda_k(\boldsymbol{B})\geq\lambda_k(\boldsymbol{A})$ for each index k. Show that these relations do not determine a partial order on Hermitian matrices. **Hint:** Equality of eigenvalues does not imply equality of matrices.

9.3 The Lidskii theorem

Corollary 9.9 states that perturbation of an Hermitian matrix by a psd matrix increases each of the eigenvalues. Lidskii's theorem describes how the eigenvalues change under an *arbitrary* Hermitian perturbation.

Theorem 9.12 (Lidskii). Let $\boldsymbol{A},\boldsymbol{E}\in\mathbb{H}_n$. Let $1\leq i_1<i_2<\cdots<i_k\leq n$ be distinct indices. Then

$$\sum\nolimits_{j=1}^{k}\left[\lambda_{i_j}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E})-\lambda_{i_j}^{\downarrow}(\boldsymbol{A})\right]\leq\sum\nolimits_{j=1}^{k}\lambda_j^{\downarrow}(\boldsymbol{E})$$

We give a simple proof of Lidskii's theorem due to Li & Mathias [LM99]. This argument reduces the full result to two applications of the Weyl monotonicity principle. See [Bha97] for three alternative proofs of Lidskii's theorem.

Proof. Fix a number k, and choose indeces $i_1<i_2<\cdots<i_n$. Without loss of generality, we may assume that $\lambda_k^{\downarrow}(\boldsymbol{E})=0$. If not, we simply replace $\boldsymbol{E}$ with $\boldsymbol{E}-\lambda_k^{\downarrow}(\boldsymbol{E})\mathbf{I}$. Indeed, this transformation reduces the kth eigenvalue of $\boldsymbol{E}$ to zero, and it shifts both sides of the inequality by an equal amount, namely $-k\lambda_k^{\downarrow}(\boldsymbol{E})$.

Next, introduce the Jordan decomposition $\boldsymbol{E}=\boldsymbol{E}_+-\boldsymbol{E}_-$, where the summands $\boldsymbol{E}_+$ and $\boldsymbol{E}_-$ are commuting psd matrices. Indeed, $\boldsymbol{E}$ admits a spectral representation

$$\boldsymbol{E}=\sum\nolimits_{j=1}^{n}\lambda_j^{\downarrow}\boldsymbol{P}_j.$$

The elements $\boldsymbol{E}_\pm$ of the Jordan decomposition are the matrices

$$\boldsymbol{E}_+=\sum\nolimits_{j=1}^{n}(\lambda_j^{\downarrow})_+\boldsymbol{P}_j\quad\text{and}\quad\boldsymbol{E}_-=\sum\nolimits_{j=1}^{n}(\lambda_j^{\downarrow})_-\boldsymbol{P}_j.$$

The two functions $(a)_\pm := \max\{\pm a, 0\}$ return the positive (resp., negative) part of a number $a \in \mathbb{R}$. Note that $\boldsymbol{E} \preccurlyeq \boldsymbol{E}_+$ and so $\boldsymbol{A} + \boldsymbol{E} \preccurlyeq \boldsymbol{A} + \boldsymbol{E}_+$.

We make the following calculation, invoking the Weyl monotonicity principle twice.

$$\begin{aligned}
\sum\nolimits_{j=1}^{k} \left[\lambda_{i_j}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \lambda_{i_j}^{\downarrow}(\boldsymbol{A})\right] &\le \sum\nolimits_{j=1}^{k} \left[\lambda_{i_j}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}_+) - \lambda_{i_j}^{\downarrow}(\boldsymbol{A})\right] \\
&\le \sum\nolimits_{j=1}^{n} \left[\lambda_{j}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}_+) - \lambda_{j}^{\downarrow}(\boldsymbol{A})\right] \\
&= \operatorname{tr}(\boldsymbol{A}+\boldsymbol{E}_+) - \operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{E}_+) \\
&= \sum\nolimits_{j=1}^{n} \lambda_{j}^{\downarrow}(\boldsymbol{E}_+) \\
&= \sum\nolimits_{j=1}^{k} \lambda_{j}^{\downarrow}(\boldsymbol{E})
\end{aligned}$$

The first inequality follows from Corollary 9.9 because $\boldsymbol{A} + \boldsymbol{E} \preccurlyeq \boldsymbol{A} + \boldsymbol{E}_+$. The second inequality introduces the missing indices; it relies on Corollary 9.9 and the fact that $\boldsymbol{A} \preccurlyeq \boldsymbol{A} + \boldsymbol{E}_+$. Next, we recall that the trace is the sum of eigenvalues, and we rely the fact that the trace is linear. Finally, note that $\lambda_k^{\downarrow}(\boldsymbol{E}) = 0$. Therefore, $\lambda_j^{\downarrow}(\boldsymbol{E}_+) = 0$ for $j \ge k$. Meanwhile, the largest k eigenvalues of $\boldsymbol{E}_+$ and $\boldsymbol{E}$ coincide. This observation completes the proof. ∎

9.4 Consequences of Lidskii's theorem

Theorem 9.12 has many remarkable applications. In this section, we will explore a few of the immediate consequences.

9.4.1 Majorization relations

First, let us rewrite Lidskii's theorem as a majorization relation. If we choose $i_j = j$ for each $j = 1, \dots, k$, then the following majorization relation arises.

$$\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}) \prec \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{E}). \tag{9.1}$$

In the majorization relation, the trace equality follows from the linearity of the trace and the fact that $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{E}) = \sum_{j=1}^{n} \lambda_j(\boldsymbol{A}+\boldsymbol{E})$. In particular, we may consider the $k = 1$ case of the majorization relation:

$$\lambda_1^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \lambda_1^{\downarrow}(\boldsymbol{A}) \le \lambda_1^{\downarrow}(\boldsymbol{E}).$$

This gives an elegant additive bound for the largest eigenvalue.

From the majorization relation (9.1), we may perform two changes of variables to arrive at a pair of relations

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

The first majorization relation is accomplished by rewriting (9.1) with $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{E}$. The second majorization relation follows upon rearranging (9.1), setting $\boldsymbol{B} = \boldsymbol{E}$, changing the sign of $\boldsymbol{A}$, and noting that $\lambda^{\downarrow}(-\boldsymbol{A}) = -\lambda^{\uparrow}(\boldsymbol{A})$. By another change of variables in the last display, we arrive at the equivalent relation

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

These two pairs of majorization relations give detailed information about how the eigenvalues behave when we add or subtract Hermitian matrices.

9.4.2 Norm comparisons

The appearance of majorization relations allows us to activate tools from our study of majorization to derive further results. In particular, we may use symmetric gauge functions to summarize how the eigenvalues change under additive perturbation.

Corollary 9.13 (Eigenvalue norm comparisons). Let $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ be a symmetric gauge function. Then

$$\Phi(\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A})) \le \Phi(\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{E}))$$

Proof. Each symmetric gauge function is isotone. For any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{x} \prec \boldsymbol{y}$, we have $\Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{y})$ because of the isotonicity. Apply this result to (9.1). ∎

Example 9.14 (Weyl perturbation theorem). Consider the symmetric gauge function $\Phi(\cdot) = \|\cdot\|_\infty$. One obtains the inequality

$$\max_j |\lambda_j^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \lambda_j^{\downarrow}(\boldsymbol{A})| \le \max_j |\lambda_j^{\downarrow}(\boldsymbol{E})| = \|\boldsymbol{E}\|_\infty .$$

In other words, the maximum change in each eigenvalue is controlled by the spectral-norm difference between the two matrices. This last relation implies that the eigenvalue function

$$\boldsymbol{\lambda}^{\downarrow} : (\mathbb{H}_n, S_\infty) \to (\mathbb{R}^n, \ell_\infty)$$

is a 1-Lipschitz map between two metric spaces. In particular, $\boldsymbol{\lambda}^{\downarrow}$ is a continuous function on the space of Hermitian matrices.

It is sometimes more convenient to make another change of variables. One obtains

$$\max_j |\lambda_j^{\downarrow}(\boldsymbol{B}) - \lambda_j^{\downarrow}(\boldsymbol{A})| \le \|\boldsymbol{B}-\boldsymbol{A}\|_\infty .$$

This statement describes how the eigenvalues of two Hermitian matrices reflect the difference between the two matrices. ∎

Aside: The eigenvalues of a general $n \times n$ matrix also change continuously. Nevertheless, this statement requires some care because there is no natural way to order the eigenvalues. Furthermore, the proofs involve tools from algebra or complex analysis.

Example 9.15 (Hoffman–Wielandt theorem). Consider the symmetric gauge function $\Phi(\cdot) = \|\cdot\|_2$. Then

$$\sum\nolimits_{j=1}^{n} |\lambda_j^{\downarrow}(\boldsymbol{A}+\boldsymbol{E}) - \lambda_j^{\downarrow}(\boldsymbol{A})|^2 \le \sum\nolimits_{j=1}^{n} |\lambda_j^{\downarrow}(\boldsymbol{E})|^2 = \|\boldsymbol{E}\|_{\mathrm{F}}^2.$$

This implies that the eigenvalue map

$$\boldsymbol{\lambda}^{\downarrow} : (\mathbb{H}_n, S_2) \to (\mathbb{R}^n, \ell_2)$$

is 1-Lipschitz between two Euclidean spaces. Equivalently, by a change of variables,

$$\sum\nolimits_{j=1}^{n} |\lambda_j^{\downarrow}(\boldsymbol{B}) - \lambda_j^{\downarrow}(\boldsymbol{A})|^2 \le \|\boldsymbol{B}-\boldsymbol{A}\|_{\mathrm{F}}^2.$$

This is the form in which the result is usually presented. ∎

Aside: The Hoffman–Wielandt theorem can be extended to normal matrices, but it requires matching the eigenvalues of the two matrices more carefully. For details, see [Bha97, Sec. VI and Eqn. VI.36].

9.4.3 Beyond norms

By using other isotone functions, we may derive many more results on perturbation of eigenvalues. Here is a picayune example.

Example 9.16 (Entropy of eigenvalues). Consider two psd matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ with the same dimension.

$$\operatorname{ent}(\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{A}+\boldsymbol{H}) - \lambda^{\downarrow}(\boldsymbol{A})) \ge \operatorname{ent}(\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{H})).$$

Recall that the entropy of a positive vector is defined as $\operatorname{ent}(\boldsymbol{x}) := -\sum_j x_j \log x_j$, and the *negation* of the entropy is an isotone function. ∎

Notes

Variational principles for eigenvalues and perturbation theory for eigenvalues are perennial topics in matrix analysis. Good references include the books of Bhatia [Bha97; Bha07a], Parlett [Par98], Saad [Saa11b], and Stewart & Sun [SS90]. The classic reference is Kato's magnum opus [Kat95].

Our approach to variational principles is drawn from [Bha97, Chap. III]. Historically, Lidskii's theorem was regarded as a very difficult result, but Li & Mathias [LM99] have really cut to the heart of the matter.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha07a] R. Bhatia. *Perturbation bounds for matrix eigenvalues*. Reprint of the 1987 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. DOI: 10.1137/1.9780898719079.

[Kat95] T. Kato. *Perturbation theory for linear operators*. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995.

[LM99] C.-K. Li and R. Mathias. "The Lidskii-Mirsky-Wielandt theorem–additive and multiplicative versions". In: *Numerische Mathematik* 81.3 (1999), pages 377–413.

[Par98] B. N. Parlett. *The symmetric eigenvalue problem*. Corrected reprint of the 1980 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. DOI: 10.1137/1.9781611971163.

[Saa11b] Y. Saad. *Numerical methods for large eigenvalue problems*. Revised edition of the 1992 original [1177405]. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. DOI: 10.1137/1.9781611970739.ch1.

[SS90] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. 1st Edition. Academic Press, 1990.

10. Multiplicative Perturbation Theory

Date: 3 February 2022 **Scribe:** Eitan Levin

Agenda:
1. Recap of Lidskii's theorem
2. Li–Mathias theorem
3. Consequences
4. Sylvester inertia theorem
5. Ostrowski monotonicity
6. Proof of Li-oMathias

In the last lecture, we studied how the eigenvalues of a Hermitian matrix $\boldsymbol{A}$ change under an additive perturbation $\boldsymbol{A} + \boldsymbol{E}$. This discussion culminated with Lidskii's theorem. In this lecture, we prove a multiplicative analogue of Lidskii's theorem due to Li and Mathias. They consider multiplicative perturbations of the form $\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}$. The proof of the Li–Mathias theorem proceeds in parallel with their proof of Lidskii's theorem, which we recall below.

Throughout this lecture, eigenvalues and singular values are always sorted in decreasing order, so we write $\boldsymbol{\lambda}, \boldsymbol{\sigma}$ instead of $\boldsymbol{\lambda}^{\downarrow}, \boldsymbol{\sigma}^{\downarrow}$ to lighten our notation.

10.1 Recap of Lidskii's theorem

We begin by recalling Lidskii's theorem and the proof strategy from Lecture 9. For all Hermitian matrices $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{H}_n$ and distinct indices $1 \le i_1 < \ldots < i_k \le n$, Lidskii's theorem states that

$$\sum_{j=1}^{k} [\lambda_{i_j}(\boldsymbol{A}+\boldsymbol{E}) - \lambda_{i_j}(\boldsymbol{A})] \le \sum_{j=1}^{k} \lambda_j(\boldsymbol{E}). \tag{10.1}$$

Equivalently, this can be stated in terms of majorization as $\lambda(\boldsymbol{A}+\boldsymbol{E}) - \lambda(\boldsymbol{A}) \prec \lambda(\boldsymbol{E})$. By applying isotone functions to this majorization inequality, one can obtain many more scalar and vector inequalities.

In Lecture 9, we gave a proof of Lidskii's theorem that proceeds along the following lines:

1. Shift the matrix $\boldsymbol{A}$ so that the kth eigenvalue is zero.
2. Extract the positive part $\boldsymbol{E}_+$ of $\boldsymbol{E}$, and reduce to the case $\boldsymbol{E} = \boldsymbol{E}_+ \succcurlyeq \boldsymbol{0}$.
3. Apply Weyl's monotonicity theorem, which states that $\lambda_i(\boldsymbol{A}+\boldsymbol{E}) - \lambda_i(\boldsymbol{A}) \ge 0$ for all $i = 1, \ldots, n$ if $\boldsymbol{E} \succcurlyeq \boldsymbol{0}$.
4. Bound the left-hand side of (10.1) by $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{E}) - \operatorname{tr}(\boldsymbol{A})$, and invoke the additivity of the trace.

In this lecture, we prove multiplicative analogues of the above results, where differences are replaced by ratios, positivity is replaced by expansivity, and the additivity of the trace is replaced by the multiplicativity of the determinant.

10.2 The theorem of Li & Mathias

We consider multiplicative perturbations of a Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n$ of the form $\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}$ for $\boldsymbol{S} \in \mathbb{M}_n$, in order for the perturbed matrix to also be Hermitian. Our main goal in this lecture is to prove the following multiplicative analogue of Lidskii's theorem (10.1), due to Li and Mathias [LM99, Thm. 2.3].

Theorem 10.1 (Li–Mathias). Choose $\boldsymbol{A} \in \mathbb{H}_n$, and $\boldsymbol{S} \in \mathbb{M}_n$, and $k \in \{1, \dots, n\}$. For any k distinct indices $1 \le i_1 < \dots < i_k \le n$ such that $\lambda_{i_j}(\boldsymbol{A}) \neq 0$ for all j, we have

$$\prod_{j=1}^{k} \frac{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_{i_j}(\boldsymbol{A})} \le \prod_{j=1}^{k} \lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{S}). \tag{10.2}$$

Moreover, if $k = n$ and $\boldsymbol{A}$ is invertible, then (10.2) holds with equality.

If the multiplier $\boldsymbol{S}$ is approximately unitary, i.e., $\boldsymbol{S}^*\boldsymbol{S} \approx \mathbf{I}$, then $\lambda_j(\boldsymbol{S}^*\boldsymbol{S}) \approx 1$ for all j. In this case, Theorem 10.1 shows that $\lambda(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) \approx \lambda(\boldsymbol{A})$.

Before proving Theorem 10.1, we will note some of its consequences. First, it would be desirable to express Theorem 10.1 as a majorization inequality, similarly to Lidskii's theorem. To that end, we would like to take the logarithm of both sides of (10.2) to convert the product into a sum. However, this is only well-defined if each factor in the product on the left-hand side of (10.2) is positive. To show that this is indeed the case, at least when $\boldsymbol{S}$ is invertible, we appeal to Sylvester's law of inertia.

10.2.1 Sylvester's inertia theorem

We proceed to state and prove Sylvester's law of inertia after two preliminary definitions.

Definition 10.2 (Inertia). The *inertia* of a Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n$ is the triplet of integers $\operatorname{inertia}(\boldsymbol{A}) := (n_+(\boldsymbol{A}), n_0(\boldsymbol{A}), n_-(\boldsymbol{A}))$ which equal, respectively, the number of positive, zero, and negative eigenvalues of $\boldsymbol{A}$.

Definition 10.3 (Congruence). Two Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$ are *congruent* if there exists an *invertible* matrix $\boldsymbol{S} \in \mathbb{M}_n$ satisfying $\boldsymbol{A} = \boldsymbol{S}^*\boldsymbol{B}\boldsymbol{S}$.

Congruence is an equivalence relation on $\mathbb{H}_n$. It arises from the effect of a change of basis on the symmetric matrix representing a quadratic form. In particular, note that $\boldsymbol{A}$ is congruent to $\boldsymbol{B}$ if and only if $\boldsymbol{B}$ is congruent to $\boldsymbol{A}$, since $\boldsymbol{A} = \boldsymbol{S}^*\boldsymbol{B}\boldsymbol{S}$ for invertible $\boldsymbol{S} \in \mathbb{M}_n$ holds if and only if $\boldsymbol{B} = (\boldsymbol{S}^{-1})^*\boldsymbol{A}\boldsymbol{S}^{-1}$.

Sylvester's theorem [Syl52] states that the inertia is invariant under congruence transformation.

Theorem 10.4 (Sylvester's law of inertia). Two Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$ are congruent if and only if $\operatorname{inertia}(\boldsymbol{A}) = \operatorname{inertia}(\boldsymbol{B})$.

In the setting of Theorem 10.1, Theorem 10.4 implies that if $\boldsymbol{S}$ is invertible then $\operatorname{sgn}(\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})) = \operatorname{sgn}(\lambda_{i_j}(\boldsymbol{A}))$ for all j, hence each factor on the left-hand side of (10.2) is positive.

If $\boldsymbol{S}$ is not invertible, then each factor on the left-hand side of (10.2) is only guaranteed to be nonnegative, by continuity of each such factor in $\boldsymbol{S}$.

Proof of Theorem 10.4. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are congruent, so $\boldsymbol{A} = \boldsymbol{S}^*\boldsymbol{B}\boldsymbol{S}$ for some invertible $\boldsymbol{S} \in \mathbb{M}_n$. Note that $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ for $\boldsymbol{x} \in \mathbb{R}^n$ if and only if $\boldsymbol{B}(\boldsymbol{S}\boldsymbol{x}) = \mathbf{0}$ because $\boldsymbol{S}^*$ is invertible, hence $\boldsymbol{S}$ maps $\ker(\boldsymbol{A})$ isomorphically onto $\ker(\boldsymbol{B})$. In particular, $n_0(\boldsymbol{A}) = \dim\ker(\boldsymbol{A}) = \dim\ker(\boldsymbol{B}) = n_0(\boldsymbol{B})$.

Next, let $(\lambda_i, \boldsymbol{u}_i)$ be orthogonal eigenpairs of $\boldsymbol{A}$, where $\lambda_i > 0$ for $i = 1, \dots, n_+(\boldsymbol{A})$ by definition of n_+. Define $\mathsf{L}_+ = \operatorname{lin}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{n_+(\boldsymbol{A})}\}$. Any nonzero $\boldsymbol{x} \in \mathsf{L}_+$ can be written as $\boldsymbol{x} = \sum_{i=1}^{n_+(\boldsymbol{A})} \alpha_i \boldsymbol{u}_i$ where $\sum_i |\alpha_i|^2 = \|\boldsymbol{x}\|^2 > 0$. Hence,

$$\langle (\boldsymbol{S}\boldsymbol{x}),\ \boldsymbol{B}(\boldsymbol{S}\boldsymbol{x})\rangle = \langle \boldsymbol{x},\ \boldsymbol{S}^*\boldsymbol{B}\boldsymbol{S}(\boldsymbol{x})\rangle = \langle \boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x}\rangle = \sum_{i=1}^{n_+(\boldsymbol{A})} \lambda_i |\alpha_i|^2 > 0.$$

We conclude that $\langle \boldsymbol{y},\ \boldsymbol{B}\boldsymbol{y}\rangle > 0$ for all nonzero $\boldsymbol{y} \in \boldsymbol{S}\mathsf{L}_+$. Since $\boldsymbol{S}$ is invertible and the eigenvectors $\boldsymbol{u}_i$ are orthogonal, $\dim(\boldsymbol{S}\mathsf{L}_+) = \dim \mathsf{L}_+ = n_+(\boldsymbol{A})$. By the Courant–Fischer–Weyl minimax principle (see Lecture 9),

$$\lambda_{n_+(\boldsymbol{A})}(\boldsymbol{B}) = \max_{\substack{\mathsf{L}\subseteq\mathbb{F}^n \\ \dim \mathsf{L} = n_+(\boldsymbol{A})}} \min_{\substack{\boldsymbol{y}\in\mathsf{L} \\ \|\boldsymbol{y}\|=1}} \langle \boldsymbol{y},\ \boldsymbol{B}\boldsymbol{y}\rangle \geq \min_{\substack{\boldsymbol{y}\in\boldsymbol{S}\mathsf{L}_+ \\ \|\boldsymbol{y}\|=1}} \langle \boldsymbol{y},\ \boldsymbol{B}\boldsymbol{y}\rangle > 0,$$

which implies $n_+(\boldsymbol{B}) \geq n_+(\boldsymbol{A})$. Interchanging the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$, we also get $n_+(\boldsymbol{A}) \geq n_+(\boldsymbol{B})$ and thus $n_+(\boldsymbol{A}) = n_+(\boldsymbol{B})$.

Finally, $n_-(\boldsymbol{A}) = n - n_0(\boldsymbol{A}) - n_+(\boldsymbol{A}) = n_-(\boldsymbol{B})$, so we obtain $\operatorname{inertia}(\boldsymbol{A}) = \operatorname{inertia}(\boldsymbol{B})$.

The converse is not needed for this lecture, and is left as an exercise. ∎

Exercise 10.5 (Sylvester inertia theorem: Converse). Prove the remaining direction in Theorem 10.4. That is, two Hermitian matrices with the same inertia are congruent.

10.2.2 Consequences of the Li–Mathias theorem

We may now derive some consequences of Theorem 10.1. First, when $\boldsymbol{S}$ and $\boldsymbol{A}$ are invertible, then Theorem 10.1 is equivalent to the majorization inequality

$$\left(\log\left(\frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})}\right)\right)_{j=1}^n \prec \left(\log \lambda_j(\boldsymbol{S}^*\boldsymbol{S})\right)_{j=1}^n. \tag{10.3}$$

For a positive-definite matrix $\boldsymbol{A} \succ \boldsymbol{0}$, we can further rewrite this expression in the form $\log \lambda(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) - \log \lambda(\boldsymbol{A}) \prec \log \lambda(\boldsymbol{S}^*\boldsymbol{S})$.

Just as in the additive case, we can obtain many scalar inequalities from (10.3) by applying isotone functions to both sides. For example,

$$\max\nolimits_{j=1,\dots,n}\left|\log\frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})}\right| \leq \max\nolimits_{j=1,\dots,n}|\log\lambda_j(\boldsymbol{S}^*\boldsymbol{S})| = \|\boldsymbol{S}\|^2,$$
$$\sum\nolimits_{j=1}^n \frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})} \leq \sum\nolimits_{j=1}^n \lambda_j(\boldsymbol{S}^*\boldsymbol{S}) = \|\boldsymbol{S}\|_F^2.$$

Theorem 10.1 also implies multiplicative perturbation bounds for singular values.

Corollary 10.6 (Li–Mathias for singular values). Choose $\boldsymbol{S}, \boldsymbol{T} \in \mathbb{M}_n$ and $k \in \{1, \dots, n\}$. For any distinct indices $1 \leq i_1 < \dots < i_k \leq n$ such that $\sigma_{i_j}(\boldsymbol{T}) > 0$ for all j, we have

$$\prod\nolimits_{j=1}^k \frac{\sigma_{i_j}(\boldsymbol{T}\boldsymbol{S})}{\sigma_{i_j}(\boldsymbol{T})} \leq \prod\nolimits_{j=1}^k \sigma_j(\boldsymbol{S}).$$

Moreover, if $k = n$ and $\boldsymbol{T}$ is invertible then the above inequality holds with equality.

Proof. Set $\boldsymbol{A} = \boldsymbol{T}^*\boldsymbol{T}$ in Theorem 10.1. Take the square roots of both sides of (10.2). ∎

Once again, Corollary 10.6 can be restated in terms of majorization. This leads to a classic result from operator theory.

Corollary 10.7 (Gel'fand–Naimark). If $\boldsymbol{S}, \boldsymbol{T} \in \mathbb{M}_n$ are invertible, then

$$\log \sigma(\boldsymbol{T}\boldsymbol{S}) - \log \sigma(\boldsymbol{T}) \prec \log \sigma(\boldsymbol{S}).$$

As before, we can draw many further consequences by applying isotone functions.

10.3 Ostrowski monotonicity

To prove Theorem 10.1, we need a multiplicative analogue of the Weyl monotonicity theorem, which played a core role in the proof of Lidskii's theorem. Weyl monotonicity shows that additive perturbation by a positive semidefinite matrix increases all the eigenvalues. Analogously, multiplicative perturbation by an expansive matrix stretches all the eigenvalues:

Theorem 10.8 (Ostrowski monotonicity). Suppose $\boldsymbol{S} \in \mathbb{M}_n$ satisfies $\boldsymbol{S}^*\boldsymbol{S} \succcurlyeq \mathbf{I}$. For any $\boldsymbol{A} \in \mathbb{H}_n$ and any $j \in \{1, \dots, n\}$ such that $\lambda_j(\boldsymbol{A}) \neq 0$, we have

$$\frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})} \geq 1.$$

Proof. The hypothesis $\boldsymbol{S}^*\boldsymbol{S} \succcurlyeq \mathbf{I}$ implies that $\lambda_n(\boldsymbol{S}^*\boldsymbol{S}) \geq 1$ and in particular, that $\boldsymbol{S}$ is invertible.

To begin, note that $\lambda_j(\boldsymbol{A} - \lambda_j(\boldsymbol{A})\mathbf{I}) = 0$. Sylvester's law of inertia (Theorem 10.4) then implies

$$0 = \lambda_j(\boldsymbol{S}^*(\boldsymbol{A} - \lambda_j(\boldsymbol{A})\mathbf{I})\boldsymbol{S}) = \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S} - \lambda_j(\boldsymbol{A})\boldsymbol{S}^*\boldsymbol{S}).$$

If $\lambda_j(\boldsymbol{A}) > 0$, then the upper bound in Weyl's pertrubation inequality (see Lecture 9) gives

$$\begin{aligned} 0 = \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S} - \lambda_j(\boldsymbol{A})\boldsymbol{S}^*\boldsymbol{S}) &\leq \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) + \lambda_1(-\lambda_j(\boldsymbol{A})\boldsymbol{S}^*\boldsymbol{S}) \\ &= \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) - \lambda_j(\boldsymbol{A})\lambda_n(\boldsymbol{S}^*\boldsymbol{S}). \end{aligned}$$

Similarly, if $\lambda_j(\boldsymbol{A}) < 0$ then the lower bound in Weyl's inequality gives

$$\begin{aligned} 0 = \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S} - \lambda_j(\boldsymbol{A})\boldsymbol{S}^*\boldsymbol{S}) &\geq \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) + \lambda_n(-\lambda_j(\boldsymbol{A})\boldsymbol{S}^*\boldsymbol{S}) \\ &= \lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) - \lambda_j(\boldsymbol{A})\lambda_n(\boldsymbol{S}^*\boldsymbol{S}). \end{aligned}$$

Dividing by $\lambda_j(\boldsymbol{A})$ and using the fact $\lambda_n(\boldsymbol{S}^*\boldsymbol{S}) \geq 1$, we arrive at the desired result. ∎

Exercise 10.9 (Ostrowski: Quantitative version). Prove the following strengthening of Theorem 10.8. For any invertible $\boldsymbol{S} \in \mathbb{M}_n$ and any $\boldsymbol{A} \in \mathbb{H}_n$, we have $\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}) = \theta_j\lambda_j(\boldsymbol{A})$ for some $\theta_j \in [\lambda_n(\boldsymbol{S}^*\boldsymbol{S}), \lambda_1(\boldsymbol{S}^*\boldsymbol{S})]$.

This relation can be viewed as a quantitative version of Sylvester's law of inertia, which is how Ostrowski originally presented this result in [Ost59].

10.4 Proof of the Li–Mathias theorem

We are now ready to prove Theorem 10.1 by substituting multiplicative analogues for their additive counterparts in the proof of Lidskii's theorem (after a few reductions specific to the multiplicative case).

Proof of Theorem 10.1. If $k = n$ and $\boldsymbol{A}$ is invertible, then

$$\prod_{j=1}^{n} \frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})} = \frac{\det(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\det(\boldsymbol{A})} = \frac{\det(\boldsymbol{S}^*)\det(\boldsymbol{A})\det(\boldsymbol{S})}{\det(\boldsymbol{A})} = \det(\boldsymbol{S}^*\boldsymbol{S}),$$

where we used the multiplicativity of the determinant. It remains to prove the inequality (10.2).

We begin with a series of reductions. First, we may assume both $\boldsymbol{A}$ and $\boldsymbol{S}$ are invertible because both sides of the inequality (10.2) are continuous at $(\boldsymbol{A}, \boldsymbol{S})$ whenever $\lambda_{i_j}(\boldsymbol{A}) \neq 0$ for all j, and any square matrix is a limit of invertible matrices.

Here we use the continuity of the eigenvalues, which follows from Weyl's perturbation inequality in Lecture 9.

Second, let $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be an SVD of $\boldsymbol{S}$, where $\boldsymbol{\Sigma} = \operatorname{diag}(s_1, \dots, s_n)$ is the diagonal matrix of singular values of $\boldsymbol{S}$, arranged in descending order. We may then assume $\boldsymbol{S} = \boldsymbol{\Sigma}$ because eigenvalues are invariant under conjugation by unitary matrices, so

$$\frac{\lambda_j(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_j(\boldsymbol{A})} = \frac{\lambda_j\Big(\boldsymbol{V}[\boldsymbol{\Sigma}^*(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})\boldsymbol{\Sigma}]\boldsymbol{V}^*\Big)}{\lambda_j(\boldsymbol{A})} = \frac{\lambda_j\big(\boldsymbol{\Sigma}^*(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})\boldsymbol{\Sigma}\big)}{\lambda_j(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})}.$$

Thus, if we can prove the result for $\boldsymbol{S} = \boldsymbol{\Sigma}$ and arbitrary invertible $\boldsymbol{A}$, then after changing variables $\boldsymbol{A} \mapsto \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$ we obtain the result for the original $\boldsymbol{S}$ and arbitrary invertible $\boldsymbol{A}$.

Third, we may assume $s_k = 1$. Indeed, if we replace $\boldsymbol{S}$ by $s_k^{-1}\boldsymbol{S}$ in the inequality (10.2), then both sides are scaled by s_k^{-2k}. Note that $s_k > 0$ because $\boldsymbol{S}$ is invertible.

To summarize, it suffices to prove (10.2) for invertible $\boldsymbol{A}$ and $\boldsymbol{S} = \operatorname{diag}(s_1, \dots, s_n)$ where $s_1 \geq \dots \geq s_k = 1 \geq \dots \geq s_n > 0$. To that end, define the matrix

$\widehat{\boldsymbol{S}}$ is the "expansive part" of $\boldsymbol{S}$, analogous to the positive part $\boldsymbol{E}_+$ of an additive perturbation $\boldsymbol{E}$.

$$\widehat{\boldsymbol{S}} = \operatorname{diag}(s_1, \dots, s_k, \underbrace{1, \dots, 1}_{n-k \text{ times}}).$$

Observe that

$$\mathbf{I} \preccurlyeq (\boldsymbol{S}^{-1}\widehat{\boldsymbol{S}})^*(\boldsymbol{S}^{-1}\widehat{\boldsymbol{S}}) = \operatorname{diag}(1, \dots, 1, s_{k+1}^{-1}, \dots, s_n^{-1}).$$

Now, Ostrowski monotonicity gives

$$1 \leq \frac{\lambda_{i_j}\Big((\boldsymbol{S}^{-1}\widehat{\boldsymbol{S}})^*\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}(\boldsymbol{S}^{-1}\widehat{\boldsymbol{S}})\Big)}{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})} = \frac{\lambda_{i_j}(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}.$$

Because $\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})/\lambda_{i_j}(\boldsymbol{A}) > 0$ by Sylvester's law of inertia, we deduce that

$$\prod_{j=1}^k \frac{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_{i_j}(\boldsymbol{A})} \leq \prod_{j=1}^k \frac{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_{i_j}(\boldsymbol{A})} \cdot \frac{\lambda_{i_j}(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})} = \prod_{j=1}^k \frac{\lambda_{i_j}(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\lambda_{i_j}(\boldsymbol{A})}.$$

Since $\widehat{\boldsymbol{S}}^*\widehat{\boldsymbol{S}} \succcurlyeq \mathbf{I}$, Ostrowski monotonicity also gives $\lambda_j(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})/\lambda_j(\boldsymbol{A}) \geq 1$ for all j. Therefore,

$$\begin{aligned}\prod_{j=1}^k \frac{\lambda_{i_j}(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\lambda_{i_j}(\boldsymbol{A})} \leq \prod_{j=1}^n \frac{\lambda_j(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\lambda_j(\boldsymbol{A})} = \frac{\det(\widehat{\boldsymbol{S}}^*\boldsymbol{A}\widehat{\boldsymbol{S}})}{\det(\boldsymbol{A})} = \det(\widehat{\boldsymbol{S}}^*\widehat{\boldsymbol{S}}) \\ = \prod_{j=1}^k s_j^2 = \prod_{j=1}^k \lambda_j(\boldsymbol{S}^*\boldsymbol{S}),\end{aligned}$$

where we used the multiplicativity of the determinant similarly to the proof of the equality case above. This establishes the desired inequality (10.2). ∎

Exercise 10.10 (Li–Mathias: Lower bound). In the setting of Theorem 10.1, prove the lower bound

$$\prod_{j=1}^k \frac{\lambda_{i_j}(\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S})}{\lambda_{i_j}(\boldsymbol{A})} \geq \prod_{j=1}^k \lambda_{n-j+1}(\boldsymbol{S}^*\boldsymbol{S}).$$

Note that this implies a corresponding lower bound in Corollary 10.6. **Hint:** First assume $\boldsymbol{S}$ is invertible and apply Theorem 10.1.

In Lectures 9–10, we studied the change in the eigenvalues and singular values under perturbations. In the next lecture, we expand our scope to consider the change in the eigen*spaces* of an Hermitian matrix under perturbation.

Notes

This lecture is adapted from the paper [LM99] of Li & Mathias. The proofs of Sylvester's theorem and Ostrowski's theorem are drawn from Horn & Johnson [HJ13]. See Bhatia's books [Bha97; Bha07a] for some additional discussion of multiplicative perturbation.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha07a] R. Bhatia. *Perturbation bounds for matrix eigenvalues*. Reprint of the 1987 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. DOI: 10.1137/1.9780898719079.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[LM99] C.-K. Li and R. Mathias. "The Lidskii-Mirsky-Wielandt theorem–additive and multiplicative versions". In: *Numerische Mathematik* 81.3 (1999), pages 377–413.

[Ost59] A. M. Ostrowski. "A quantitative formulation of Sylvester's law of inertia". In: *Proceedings of the National Academy of Sciences of the United States of America* 45.5 (1959), page 740.

[Syl52] J. Sylvester. "A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.23 (1852), pages 138–142.

11. Perturbation of Eigenspaces

Date: 8 February 2022 **Scribe:** Taylan Kargin

Agenda:
1. Motivation
2. Principle angles
3. Sylvester equations
4. Eigenspace perturbation

The last two lectures covered the perturbation theory for eigenvalues of Hermitian matrices. We will continue our discussion with perturbation of the eigenspaces of Hermitian matrices, which has a rather different character. We will start out with the motivation demonstrating the major challenging in understand the behavior of eigenspaces under perturbation. This observation leads us to study the principle angles between subspaces and the solutions of Sylvester equations. Finally, we will use these tools to develop a result on the perturbation of eigenspaces associated with well-separated eigenvalues.

11.1 Motivation

Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$ be Hermitian matrices of the same dimension. From Lidskii's theorem (Lecture 9), if $\boldsymbol{A}$ and $\boldsymbol{B}$ are close, then the vectors $\boldsymbol{\lambda}(\boldsymbol{A})$ and $\boldsymbol{\lambda}(\boldsymbol{B})$ of decreasingly ordered eigenvalues are also close. One might ask the following question. Are the eigenspaces of $\boldsymbol{A}$ and $\boldsymbol{B}$ close as well? Let us investigate by posing a simple example.

Example 11.1 (Bad eigenspaces). For small $\varepsilon > 0$, consider the matrices

$$\boldsymbol{A} = \begin{bmatrix} 1+\varepsilon & 0 \\ 0 & 1-\varepsilon \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix}.$$

These matrices represent two different perturbations of the identity matrix. Each of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ has eigenvalues $1 \pm \varepsilon$. Furthermore, the two matrices are close with respect to every UI norm.

One might expect that the eigenspace of $\boldsymbol{A}$ associated with the $1+\varepsilon$ eigenvalue is close to the eigenspace of $\boldsymbol{B}$ associated with the $1+\varepsilon$ eigenvalue and similarly for eigenspaces associated with $1-\varepsilon$ eigenvalue. The eigenpairs of the matrices are easily determined.

$$\begin{aligned} &\boldsymbol{A} \text{ has eigenpairs} && \left(1+\varepsilon, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) && \text{and} && \left(1-\varepsilon, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right); \\ &\boldsymbol{B} \text{ has eigenpairs} && \left(1+\varepsilon, \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) && \text{and} && \left(1-\varepsilon, \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right). \end{aligned}$$

For each eigenvalue, the eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$ are very far apart. This happens in spite of the fact that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are close to each other in every UI norm. ∎

As seen from the preceding example, two matrices can be close without having similar eigenspaces. To appreciate why, observe that 2×2 identity matrix has only one distinct eigenvalue, 1, and the associated eigenspace spans all of $\mathbb{C}^2$. Both $\boldsymbol{A}$ and $\boldsymbol{B}$ are perturbations of the identity matrix, and we are looking at eigenvectors associated to very close eigenvalues. When their eigenvalues coalesce at $\varepsilon = 0$, the 1-dimensional eigenspaces combine to form a 2-dimensional eigenspace with a continuous family of

eigenvectors. By making perturbations of the identity in different directions, we can force the two matrices to have strikingly different eigenvectors.

This observations suggest the idea that eigenspaces associated to *well-separated* eigenvalues may be insensitive to perturbations. To illustrate this effect, let us consider the matrices

$$\boldsymbol{A} = \begin{bmatrix} 1+\varepsilon & 0 & 0 \\ 0 & 1-\varepsilon & 0 \\ 0 & 0 & \lambda \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 1 & \varepsilon & 0 \\ \varepsilon & 1 & 0 \\ 0 & 0 & \lambda \end{bmatrix}.$$

where $\lambda \gg 1$. In that case, the eigenvector associated with the eigenvalue λ is not affected by the instability of the eigenspaces with eigenvalues $1 \pm \varepsilon$. Perturbation theory for eigenspaces builds on this insight.

11.2 Principle angles between subspaces

In this section, we introduce the concept of principle angles between subspaces to quantify what it means for eigenspaces to be similar to each other or different from each other.

11.2.1 Geometric approach

We start by defining the acute angle between a pair of vectors and then we will build up from there to notions of closeness between subspaces.

Definition 11.2 (Acute angle between vectors). Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$ be vectors with unit ℓ_2-norm. The acute angle $\theta \in [0, \pi/2]$ between $\boldsymbol{u}$ and $\boldsymbol{v}$ is determined by the relation $\cos\theta(\boldsymbol{u}, \boldsymbol{v}) = |\langle \boldsymbol{u}, \boldsymbol{v} \rangle|$.

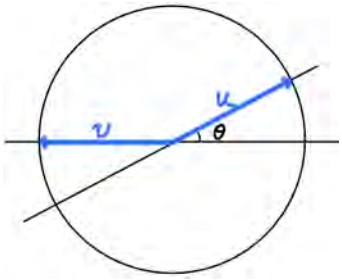

Figure 11.1 Acute angle between vectors

Figure 11.1 demonstrates that the acute angle between vectors is simply the smaller angle between the 1-dimensional subspaces spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$. Now, the question that we have to pose, which goes back to work of [Jor75], is how to extend this idea of finding the smallest angle to subspaces.

We begin with a mechanism for parameterizing the subspaces. Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{C}^{n \times k}$ be tall matrices, each one with orthonormal columns. We emphasize that the columns of the concatenation $[\boldsymbol{X}, \boldsymbol{Y}]$ do not need to be orthonormal. We can parameterize unit vectors in the ranges of these matrices as follows.

$$\{\boldsymbol{X}\boldsymbol{u} \,:\, \|\boldsymbol{u}\| = 1\} \quad \text{and} \quad \{\boldsymbol{Y}\boldsymbol{v} \,:\, \|\boldsymbol{v}\| = 1\}.$$

These sets parametrize the unit spheres in the range of $\boldsymbol{X}$ and in the range of $\boldsymbol{Y}$. Now, the acute angle between two unit vectors from the ranges $\boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{Y}\boldsymbol{v}$ satisfies

$$\cos\theta(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{Y}\boldsymbol{v}) = |\langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{Y}\boldsymbol{v} \rangle| = |\boldsymbol{u}^*(\boldsymbol{X}^*\boldsymbol{Y})\boldsymbol{v}|.$$

The *first principle angle* θ_1 between $\mathrm{range}(\boldsymbol{X})$ and $\mathrm{range}(\boldsymbol{Y})$ is defined as the minimum acute angle between the unit vectors in $\mathrm{range}(\boldsymbol{X})$ and $\mathrm{range}(\boldsymbol{Y})$. In order to minimize the acute angle, we need to *maximize* the cosine of this angle:

$$\cos\theta_1(\mathrm{range}(\boldsymbol{X}), \mathrm{range}(\boldsymbol{Y})) := \max_{\|\boldsymbol{u}\|=1, \|\boldsymbol{v}\|=1} |\boldsymbol{u}^*(\boldsymbol{X}^*\boldsymbol{Y})\boldsymbol{v}| = \sigma_1(\boldsymbol{X}^*\boldsymbol{Y})$$

In short, the closest pair of vectors in these two subspaces, $\mathrm{range}(\boldsymbol{X})$ and $\mathrm{range}(\boldsymbol{Y})$, is obtained by taking the first singular value of $\boldsymbol{X}^*\boldsymbol{Y}$.

To continue, let $(\boldsymbol{u}_1, \boldsymbol{v}_1)$ be a pair of unit vectors where the maximum is achieved in the definition of θ_1. The *second principle angle* θ_2 can be defined recursively by

finding the smallest angle between unit vectors in range($\boldsymbol{X}$) and range($\boldsymbol{Y}$), excluding the directions $\boldsymbol{X}\boldsymbol{u}_1$ and $\boldsymbol{Y}\boldsymbol{v}_1$.

$$\cos\theta_2(\mathrm{range}(\boldsymbol{X}), \mathrm{range}(\boldsymbol{Y})) := \max_{\substack{\|\boldsymbol{u}\|=1,\ \|\boldsymbol{v}\|=1 \\ \boldsymbol{u}\perp\boldsymbol{u}_1,\ \boldsymbol{v}\perp\boldsymbol{v}_1}} |\boldsymbol{u}^*(\boldsymbol{X}^*\boldsymbol{Y})\boldsymbol{v}| = \sigma_2(\boldsymbol{X}^*\boldsymbol{Y}).$$

The second relation is a consequence of the Courant–Fischer–Weyl minimax principle.

11.2.2 Algebraic approach

By extending this approach, we arrive at the general definition of principle angles between a pair of subspaces.

Definition 11.3 (Principle angles). Let $\mathsf{E}, \mathsf{F} \subset \mathbb{C}^n$ be subspaces, possibly with different dimensions. Let $\boldsymbol{X}, \boldsymbol{Y}$ be orthonormal matrices such that range($\boldsymbol{X}$) = E and range($\boldsymbol{Y}$) = F. The *ith principle angle* $\theta_i(\mathsf{E}, \mathsf{F}) \in [0, \pi/2]$ between the subspaces E and F is determined by the relation

$$\cos\theta_i(\mathsf{E}, \mathsf{F}) := \sigma_i(\boldsymbol{X}^*\boldsymbol{Y}) \quad \text{for } i = 1, \dots, \min\{\dim\mathsf{E}, \dim\mathsf{F}\}.$$

The principle angles only depend on the two subspaces, even though the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the definition are not uniquely determined. To see why, it is helpful to rewrite this definition in terms of the orthogonal projectors onto the subspaces, as the orthogonal projectors are unique.

Definition 11.4 (Principle angles II). Let $\mathsf{E}, \mathsf{F} \subset \mathbb{C}^n$ be subspaces. Let $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{H}_n$ be the orthogonal projectors onto E and F, respectively. Then the ith principle angle $\theta_i(\mathsf{E}, \mathsf{F}) \in [0, \pi/2]$ is determined by

$$\cos\theta_i(\mathsf{E}, \mathsf{F}) = \sigma_i(\boldsymbol{P}\boldsymbol{Q}) \quad \text{for } i = 1, \dots, \min\{\dim\mathsf{E}, \dim\mathsf{F}\}.$$

Exercise 11.5 (Principle angles: Equivalence). Check that Definitions 11.3 and 11.4 give the same result.

11.2.3 Similarity and distance between subspaces

The concept of principle angles allows us to define various notions of the *similarity* between a pair of subspaces. Two subspaces are similar when the principle angles are small. Equivalently, subspaces are similar when the cosines of the principle angles are large. We can quantify this effect by applying UI norms to the product of projectors.

For example, let $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{H}_n$ be orthogonal projectors, and define the subspaces E = range($\boldsymbol{P}$) and F = range($\boldsymbol{Q}$). The operator norm of the projector product satisfies

$$\|\boldsymbol{P}\boldsymbol{Q}\|^2 = \cos^2\theta_1(\mathsf{E}, \mathsf{F}).$$

When the cosine of the first principle angle is relatively large (that is, $\cos\theta_1 \approx 1$), then the subspaces are very similar. Conversely, if the cosine of the angle is small (that is, $\cos\theta_1 \approx 0$), then the subspaces are very different. Thus, the spectral norm $\|\boldsymbol{P}\boldsymbol{Q}\|$ of the projector product gives one measure of how close the two subspaces are. Other UI norms lead to other measures of similarity.

In parallel, we can measure the *distance* between subspaces by considering the sines of principle angles. For example, $\sin^2\theta_1(\mathsf{E}, \mathsf{F})$ is a measure of distance between E and F. That is, $\sin\theta_1 \approx 0$ when the subspaces are similar, and $\sin\theta_1 \approx 1$ when the

subspaces are very different. It turns out that the these distances can also be expressed in terms of projector products. For subspaces E, F with the same dimension,

$$\|(\mathbf{I} - \boldsymbol{P})\boldsymbol{Q}\|^2 = \sin^2 \theta_1(\mathsf{E}, \mathsf{F}).$$

Other UI norms lead to other notions of distance. This insight allows us to develop metric geometry for subspaces.

Exercise 11.6 (Distances between subspaces). PS3 has some exercises that describe various ways to measure the distance between subspaces.

11.3 Sylvester equations

Before we study perturbation theory for eigenspaces, we need to take another detour to develop the basic theory of Sylvester equations [Syl84]. For now, this digression might seem perplexing. By the end of the lecture, you will see the wonderful connection between these ideas.

11.3.1 Formulation and solvability

We will start with form and solvability properties of Sylvester equation. Let $\boldsymbol{A} \in \mathbb{H}_m$, and $\boldsymbol{B} \in \mathbb{H}_n$ be Hermitian matrices, perhaps with different dimension. Fix a matrix $\boldsymbol{Y} \in \mathbb{C}^{m \times n}$. The *Sylvester equation* with this data is the linear system

$$\text{Find } \boldsymbol{X} \in \mathbb{C}^{m \times n} : \quad \boldsymbol{A}\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B} = \boldsymbol{Y}. \tag{SYL}$$

It is fruitful to rewrite this linear equation using Kronecker products.

$$(\mathbf{I} \otimes \boldsymbol{A}^{\mathsf{T}} - \boldsymbol{B} \otimes \mathbf{I}) \operatorname{vec}(\boldsymbol{X}) = \operatorname{vec}(\boldsymbol{Y}). \tag{11.1}$$

This expression converts the matrix equation into an ordinary linear system, which may be easier to understand.

Exercise 11.7 (Sylvester equation: Tensor form). Check if the equation above is equivalent to (SYL).

Proposition 11.8 (Solvabililty of (SYL)). Let $\boldsymbol{A} \in \mathbb{H}_m$ and $\boldsymbol{B} \in \mathbb{H}_n$. The Sylvester equation (SYL) has a unique solution for all $\boldsymbol{Y} \in \mathbb{C}^{m \times n}$ if and only if $\lambda_i(\boldsymbol{A}) \neq \lambda_j(\boldsymbol{B})$ for all i, j.

Proof. The Sylvester equation (SYL) has a unique solution for each right-hand side $\boldsymbol{Y}$ if and only if the matrix in the Kronecker product formulation (11.1) is nonsingular. According to PS1, the eigenvalues of the tensor $\mathbf{I} \otimes \boldsymbol{A}^{\mathsf{T}} - \boldsymbol{B} \otimes \mathbf{I}$ are the numbers

$$\lambda_i(\boldsymbol{A}^{\mathsf{T}}) - \lambda_j(\boldsymbol{B}) = \lambda_i(\boldsymbol{A}) - \lambda_j(\boldsymbol{B}) \quad \text{for all } i, j.$$

Therefore, the tensor is nonsingular if and only if $\lambda_i(\boldsymbol{A}) \neq \lambda_j(\boldsymbol{B})$ for all i, j. ∎

Warning 11.9 (Equal coefficients). If $\boldsymbol{A} = \boldsymbol{B}$, then (SYL) never has a unique solution. By placing additional assumptions on the solution matrix (e.g., Hermiticity), we can sometimes recover the unique solvability. ▪

Exercise 11.10 (General Sylvester equations). Extend this discussion to $\boldsymbol{A} \in \mathbb{M}_m$ and $\boldsymbol{B} \in \mathbb{M}_n$ that may not be Hermitian.

11.3.2 Integral representation of the solution

In this section, we will develop a remarkable representation for the solution of the Sylvester equation under a stronger spectral separation condition. This approach gives a direct path to bound the norm of the solution operator.

For motivation, we begin with the trivial scalar case of (SYL). Suppose that $b < 0 \le a$. In this case, the equation (SYL) takes the form $ax - xb = y$, and it has the unique solution $x = y(a-b)^{-1}$. We can rewrite this solution in a bizarre way:

$$x = y \int_0^\infty \mathrm{e}^{-t(a-b)}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-ta} y \mathrm{e}^{tb}\,\mathrm{d}t.$$

The power of this formulation is that we can extend it to the matrix case by capitalizing the letters, which is a surprisingly useful technique in matrix analysis.

Theorem 11.11 (Sylvester with sign conditions). Let $\boldsymbol{A} \in \mathbb{H}_m$ and $\boldsymbol{B} \in \mathbb{H}_n$. Assume $\lambda_i(\boldsymbol{A}) \ge 0$ and $\lambda_j(\boldsymbol{B}) < 0$ for all i, j. Then, for an arbitrary $\boldsymbol{Y} \in \mathbb{C}^{m\times n}$, the solution $\boldsymbol{X}$ to (SYL) can be written as

$$\boldsymbol{X} = \int_0^\infty \mathrm{e}^{-t\boldsymbol{A}} \boldsymbol{Y} \mathrm{e}^{t\boldsymbol{B}}\,\mathrm{d}t.$$

Proof. For a square matrix $\boldsymbol{M} \in \mathbb{M}_n$, we can use the Taylor series expansion of the exponential function to see that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^{t\boldsymbol{M}} = \boldsymbol{M}\mathrm{e}^{t\boldsymbol{M}} = \mathrm{e}^{t\boldsymbol{M}}\boldsymbol{M}.$$

Introduce this ansatz into (SYL):

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B} &= \int_0^\infty \left[\boldsymbol{A}\mathrm{e}^{-t\boldsymbol{A}}\boldsymbol{Y}\mathrm{e}^{t\boldsymbol{B}} - \mathrm{e}^{-t\boldsymbol{A}}\boldsymbol{Y}\mathrm{e}^{t\boldsymbol{B}}\boldsymbol{B}\right]\,\mathrm{d}t \\
&= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\left[-\mathrm{e}^{-t\boldsymbol{A}}\boldsymbol{Y}\mathrm{e}^{t\boldsymbol{B}}\right]\,\mathrm{d}t \\
&= \boldsymbol{Y} - \lim_{t\to\infty} \mathrm{e}^{-t\boldsymbol{A}}\boldsymbol{Y}\mathrm{e}^{t\boldsymbol{B}} = \boldsymbol{Y}.
\end{aligned}$$

The limit in the second to the last equality vanishes. Indeed, $\boldsymbol{A}$ has positive eigenvalues, so $\mathrm{e}^{-t\boldsymbol{A}}$ is bounded. Since $\boldsymbol{B}$ has strictly negative eigenvalues, $\mathrm{e}^{t\boldsymbol{B}} \to \boldsymbol{0}$. ∎

11.3.3 Norm of the solution operator

Using Theorem 11.11, we can quickly bound the norm of the solution to the Sylvester equation (SYL).

Corollary 11.12 (Norm of solution operator). Let $\boldsymbol{A} \in \mathbb{H}_m$ and $\boldsymbol{B} \in \mathbb{H}_n$. For $s \in \mathbb{R}$ and $\varepsilon > 0$, assume $\lambda_i(\boldsymbol{A}) \ge s$ and $\lambda_j(\boldsymbol{B}) \le s - \varepsilon$ for all i, j. For each UI norm $|||\cdot|||$, the solution $\boldsymbol{X}$ to (SYL) satisfies the inequality $|||\boldsymbol{X}||| \le \varepsilon^{-1}|||\boldsymbol{Y}|||$.

Proof. Without loss of generality, we can take $s = 0$. Indeed, the mappings $\boldsymbol{A} \mapsto \boldsymbol{A} - s\mathbf{I}$, and $\boldsymbol{B} \mapsto \boldsymbol{B} - s\mathbf{I}$, leave the left-hand side of the Sylvester equation (SYL) invariant. Take the norm of integral representation in Theorem 11.11, and invoke the triangle

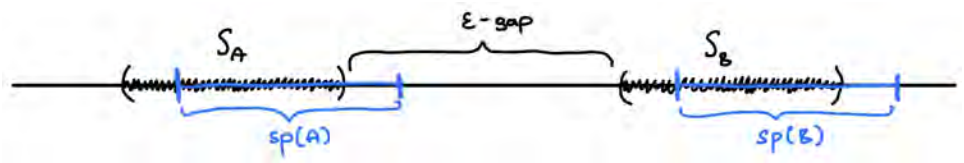

Figure 11.2 (Spectral separation). An ε-spectral gap between $\mathrm{sp}(\boldsymbol{A}) \cap \mathsf{S}_{\boldsymbol{A}}$ and $\mathrm{sp}(\boldsymbol{B}) \cap \mathsf{S}_{\boldsymbol{B}}$

inequality. This step yields the bound

$$\begin{aligned} \|\boldsymbol{X}\| &\le \int_0^\infty \|\mathrm{e}^{-t\boldsymbol{A}} \boldsymbol{Y} \mathrm{e}^{t\boldsymbol{B}}\| \,\mathrm{d}t \\ &\le \int_0^\infty \|\mathrm{e}^{-t\boldsymbol{A}}\| \cdot \|\boldsymbol{Y}\| \cdot \|\mathrm{e}^{t\boldsymbol{B}}\| \,\mathrm{d}t \\ &\le \|\boldsymbol{Y}\| \int_0^\infty 1 \cdot \mathrm{e}^{-t\varepsilon} \mathrm{d}t = \frac{1}{\varepsilon} \|\boldsymbol{Y}\|. \end{aligned}$$

The second inequality holds because each UI norm is an operator ideal norm. The third inequality depends on the assumptions that the spectra of $\boldsymbol{A}$ and $\boldsymbol{B}$ are separated. ∎

Let us take a moment to reinterpret this corollary. Fix the data $\boldsymbol{A} \in \mathbb{H}_m$ and $\boldsymbol{B} \in \mathbb{H}_n$ for the Sylvester equation (SYL), and assume that spectral separation criterion from Corollary 11.12 holds. We can define the linear solution operator $\boldsymbol{\Phi} : \boldsymbol{Y} \mapsto \boldsymbol{X}$ that maps the right-hand side $\boldsymbol{Y}$ of the Sylvester equation to the unique solution $\boldsymbol{X}$. Then Corollary 11.12 show that

$$\|\boldsymbol{\Phi}\| := \max\{\|\boldsymbol{\Phi}(\boldsymbol{Y})\| : \|\boldsymbol{Y}\| \le 1\} \le \varepsilon^{-1}.$$

In other words, the norm of the solution operator is controlled by the spectral separation of the data $\boldsymbol{A}, \boldsymbol{B}$.

11.4 Perturbation theory for eigenspaces

At last, we are prepared to describe how eigenspaces change under perturbation. These results hinge on the theory of Sylvester equations.

Let $\boldsymbol{A} \in \mathbb{H}_n$ be an Hermitian matrix with spectral resolution $\boldsymbol{A} = \sum_{\lambda \in \mathrm{sp}(\boldsymbol{A})} \lambda \boldsymbol{P}_\lambda$. Recall that $\mathrm{sp}(\boldsymbol{A}) = \{\lambda \in \mathbb{R} : \det(\boldsymbol{A} - \lambda \mathbf{I}) = 0\}$ denotes the set of eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{P}_\lambda \in \mathbb{H}_n$ is the unique orthogonal projector onto the eigenspace associated with the eigenvalue λ. For a set $\mathsf{S} \subset \mathbb{R}$, define the restricted spectral projector $\boldsymbol{P}_{\boldsymbol{A}}(\mathsf{S}) = \sum_{\lambda \in \mathrm{sp}(\boldsymbol{A}) \cap \mathsf{S}} \boldsymbol{P}_\lambda$ which acts as the orthogonal projector onto the invariant subspace spanned by all the eigenvalues in S.

Our goal is to argue that if $\mathsf{S}_{\boldsymbol{A}}$ and $\mathsf{S}_{\boldsymbol{B}}$ are well-separated sets and $\boldsymbol{A}$ and $\boldsymbol{B}$ are close, then $\boldsymbol{P}_{\boldsymbol{A}}(\mathsf{S}_{\boldsymbol{A}})$ and $\boldsymbol{P}_{\boldsymbol{B}}(\mathsf{S}_{\boldsymbol{B}})$ are nearly orthogonal. The following theorem formalizes this idea. This result is essentially due to Davis & Kahan [DK70].

Theorem 11.13 (Perturbation of eigenspaces). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$. Let $\mathsf{S}_{\boldsymbol{A}}, \mathsf{S}_{\boldsymbol{B}} \subset \mathbb{R}$ be *intervals* with $\mathrm{dist}(\mathsf{S}_{\boldsymbol{A}}, \mathsf{S}_{\boldsymbol{B}}) > \varepsilon > 0$ as in Figure 11.2. Introduce the restricted spectral projectors $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{P}_{\boldsymbol{A}}(\mathsf{S}_{\boldsymbol{A}})$ and $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{P}_{\boldsymbol{B}}(\mathsf{S}_{\boldsymbol{B}})$. Then

$$\|\boldsymbol{P}_{\boldsymbol{A}} \boldsymbol{P}_{\boldsymbol{B}}\| \le \frac{1}{\varepsilon} \|\boldsymbol{P}_{\boldsymbol{A}} (\boldsymbol{A} - \boldsymbol{B}) \boldsymbol{P}_{\boldsymbol{B}}\| \le \frac{1}{\varepsilon} \|\boldsymbol{A} - \boldsymbol{B}\|$$

for any UI norm $|||\cdot|||$.

Let us sketch the idea behind the argument and then give a more rigorous treatment. Since a matrix commutes with its spectral projectors, observe that

$$\boldsymbol{Y}_0 = \boldsymbol{P}_{\boldsymbol{A}}(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{A}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{B}}) - (\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{B}})\boldsymbol{B}.$$

Therefore, $\boldsymbol{X}_0 = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{B}}$ is the solution to a Sylvester equation. Aside from the zero eigenvalue, the coefficient matrices have spectral separation because the sets $\mathsf{S}_{\boldsymbol{A}}$ and $\mathsf{S}_{\boldsymbol{B}}$ are separated. Heuristically, we should be able to invoke Corollary 11.12 to bound the norm of $\boldsymbol{X}_0$. This gives a bound on the similarity between the two spectral subspaces in terms of the norm of $\boldsymbol{Y}_0$, which reflects the discrepancy $\boldsymbol{A} - \boldsymbol{B}$ between the matrices.

Proof. To make this argument solid, we introduce factorizations $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{Q}_{\boldsymbol{A}}\boldsymbol{Q}_{\boldsymbol{A}}^*$ and $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{Q}_{\boldsymbol{B}}\boldsymbol{Q}_{\boldsymbol{B}}^*$ where $\boldsymbol{Q}_{\boldsymbol{A}}$ has orthonormal columns and $\boldsymbol{Q}_{\boldsymbol{B}}$ has orthonormal columns. Define the matrix

$$\boldsymbol{Y} = \boldsymbol{Q}_{\boldsymbol{A}}^*(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{Q}_{\boldsymbol{B}}.$$

Since $\boldsymbol{Q}_{\boldsymbol{A}}$ and $\boldsymbol{Q}_{\boldsymbol{B}}$ are orthonormal matrices, we can write

$$\begin{aligned}
\boldsymbol{Y} &= (\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{Q}_{\boldsymbol{A}})\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{B}} - \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{B}}(\boldsymbol{Q}_{\boldsymbol{B}}^*\boldsymbol{Q}_{\boldsymbol{B}}) \\
&= \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{B}} - \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{B}}\boldsymbol{Q}_{\boldsymbol{B}} \\
&= \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{Q}_{\boldsymbol{B}} - \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{P}_{\boldsymbol{B}}\boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{B}} \\
&= (\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{A}})(\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{Q}_{\boldsymbol{B}}) - (\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{Q}_{\boldsymbol{B}})(\boldsymbol{Q}_{\boldsymbol{B}}^*\boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{B}}).
\end{aligned}$$

Indeed, the spectral projectors commute with their matrices $\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{A}}$ and $\boldsymbol{P}_{\boldsymbol{B}}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{P}_{\boldsymbol{B}}$. Therefore, the matrix $\boldsymbol{X} = \boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{Q}_{\boldsymbol{B}}$ solves the Sylvester equation with data $\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{A}}$ and $\boldsymbol{Q}_{\boldsymbol{B}}^*\boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{B}}$.

By construction, all eigenvalues of $\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{A}\boldsymbol{Q}_{\boldsymbol{A}}$ belong to $\mathsf{S}_{\boldsymbol{A}}$, and all eigenvalues of $\boldsymbol{Q}_{\boldsymbol{B}}^*\boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{B}}$ belong to $\mathsf{S}_{\boldsymbol{B}}$. Therefore, the eigenvalues are separated by ε. By Corollary 11.12, we have that

$$|||\boldsymbol{Q}_{\boldsymbol{A}}^*\boldsymbol{Q}_{\boldsymbol{B}}||| = |||\boldsymbol{X}||| \le \frac{1}{\varepsilon}|||\boldsymbol{Y}||| = \frac{1}{\varepsilon}|||\boldsymbol{Q}_{\boldsymbol{A}}^*(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{Q}_{\boldsymbol{B}}|||. \tag{11.2}$$

Since $|||\cdot|||$ is UI, the left- and right-hand sides of (11.2) do not change if we introduce the orthonormal matrices $\boldsymbol{Q}_{\boldsymbol{A}}$ and $\boldsymbol{Q}_{\boldsymbol{B}}^*$. By doing so, the spectral projectors appear, and we recognize that

$$\begin{aligned}
|||\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{B}}||| &\le \frac{1}{\varepsilon}|||\boldsymbol{P}_{\boldsymbol{A}}(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{P}_{\boldsymbol{B}}||| \\
&\le \frac{1}{\varepsilon}|||\boldsymbol{P}_{\boldsymbol{A}}||| \cdot |||\boldsymbol{A} - \boldsymbol{B}||| \cdot |||\boldsymbol{P}_{\boldsymbol{B}}||| = \frac{1}{\varepsilon}|||\boldsymbol{A} - \boldsymbol{B}|||.
\end{aligned}$$

The second inequality depends on the operator ideal property of UI norms. ∎

The result of Theorem 11.13 verifies our intuition from the beginning of the lecture. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are close to each other, the spectral projectors for well-separated eigenvalues are almost orthogonal to each other. Quantitatively, the UI norm of the projector product $\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{B}}$ is very small, which reflects *dissimilarity* of the subspaces $\operatorname{range}(\boldsymbol{P}_{\boldsymbol{A}})$ and $\operatorname{range}(\boldsymbol{P}_{\boldsymbol{B}})$. In particular, when $\boldsymbol{A} = \boldsymbol{B}$, the eigenvectors associated with different eigenvalues are orthogonal.

Notes

James Joseph Sylvester (1814–1897) was an English mathematician who made important contributions in matrix theory, number theory and combinatorics. His work includes the proof of Sylvester's law of inertia and the study of Sylvester equations.

The idea of using Sylvester equations to study the perturbation of eigenspaces is usually attributed to Davis & Kahan, who worked out these ideas in an important paper [DK70]. The presentation in this lecture is more modern. It is modeled on Bhatia's book [Bha97, Chap. VII]. We have chosen to emphasize the variational perspective on principal angles because it seems more intuitive.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[DK70] C. Davis and W. M. Kahan. "The Rotation of Eigenvectors by a Perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pages 1–46. eprint: https://doi.org/10.1137/0707001. DOI: 10.1137/0707001.

[Jor75] C. Jordan. "Essai sur la géométrie à n dimensions". In: *Bulletin de la Société mathématique de France* 3 (1875), pages 103–174.

[Syl84] J. J. Sylvester. "Sur l'équation en matrices px= xq". In: *CR Acad. Sci. Paris* 99.2 (1884), pages 67–71.

12. Positive Linear Maps

Date: 10 February 2022 **Scribe:** Ruizhi Cao

In the first half of the course, we discussed majorization and its consequences, and then we turned to the study of perturbation theory for eigenvalues and eigenspaces. In the second half of the course, we are going to talk about positive-semidefinite matrices and operations on positive-semidefinite matrices. This lecture introduces the positive-semidefinite order, and it defines a class of linear maps that respect the order. We will study the properties of these *positive linear maps*.

Agenda:

1. Positive-semidefinite order
2. Positive linear maps
3. Examples of positive maps
4. Basic properties
5. Convexity inequalities
6. Russo–Dye theorem

12.1 Positive-semidefinite order

We will first remind the reader of the definition of a positive-semidefinite (psd) matrix. Then we introduce the psd order, a natural partial order relation on the self-adjoint matrices.

Definition 12.1 (Positive semidefinite; positive definite). For an $n \times n$ complex matrix $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$, we say that $\boldsymbol{A}$ is *positive semidefinite (psd)* if we have

$$\langle \boldsymbol{u},\ \boldsymbol{A}\boldsymbol{u} \rangle \geq 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{C}^n.$$

For $\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C})$, we say that $\boldsymbol{A}$ is *positive definite (pd)* if $\langle \boldsymbol{u},\ \boldsymbol{A}\boldsymbol{u} \rangle > 0$ for all *nonzero* vectors $\boldsymbol{u} \in \mathbb{C}^n$.

Warning: The situation is more delicate for matrices over the real field. In that case, the definition of a psd matrix must also include a symmetry assumption. ∎

The most basic fact about psd matrices is the conjugation rule, which states that conjugation preserves the psd property.

Proposition 12.2 (Conjugation rule). Let $\boldsymbol{A} \in \mathbb{M}_n$, and let $\boldsymbol{X} \in \mathbb{C}^{n \times k}$.

1. If $\boldsymbol{A}$ is psd, then the matrix $\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}$ is also psd.
2. If $\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}$ is a psd matrix and $\boldsymbol{X}$ is surjective (has rank n), then $\boldsymbol{A}$ is psd.

Exercise 12.3 (Conjugation rule). Prove Proposition 12.2.

Exercise 12.4 (The psd cone). Show that the set $\mathbb{M}_n^+$ of psd matrices forms a proper cone. Recall that a proper cone is closed, convex, pointed, and solid.

Definition 12.5 (Semidefinite order). For two self-adjoint matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n^{\mathrm{sa}}$, we write $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ when $\boldsymbol{B} - \boldsymbol{A}$ is psd. The relation "$\preccurlyeq$" is a partial order on $\mathbb{M}_n^{\mathrm{sa}}$ since the cone $\mathbb{M}_n^{\mathrm{sa}}$ is proper. We call this the *psd order*, also known as the *semidefinite order* or *Loewner order*. Similarly, we write $\boldsymbol{A} \prec \boldsymbol{B}$ when $\boldsymbol{B} - \boldsymbol{A}$ is pd.

Recall that a matrix $\boldsymbol{A} \in \mathbb{M}_n$ is *self-adjoint (s.a.)* when $\boldsymbol{A}^* = \boldsymbol{A}$.

It is natural to ask what kind of functions respect the psd order. That is, we are interested in functions $\boldsymbol{F} : \mathbb{M}_n^{\mathrm{sa}} \to \mathbb{M}_k^{\mathrm{sa}}$ for which

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \quad \text{implies} \quad \boldsymbol{F}(\boldsymbol{A}) \preccurlyeq \boldsymbol{F}(\boldsymbol{B}).$$

A function that satisfies this type of relation is said to be *monotone* with respect to the psd order.

We can also ask what kind of functions are "convex" with respect to the psd order. This amounts to the relation

$$\boldsymbol{F}\big(\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\big) \preccurlyeq \tfrac{1}{2}\boldsymbol{F}(\boldsymbol{A}) + \tfrac{1}{2}\boldsymbol{F}(\boldsymbol{B}) \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n^{\mathrm{sa}}. \tag{12.1}$$

In other words, the function of an average is bounded by the average of the function values.

In this lecture, we will consider *linear* functions on matrices that are monotone with respect to the psd order. For a linear function $\boldsymbol{F}$, the convexity relation (12.1) is always valid, so we will focus on montonicity for now. In the next lecture, we will study nonlinear functions that are monotone or convex.

12.2 Positive linear maps

For today, the key concept is a positive linear map, which is a linear function that maps each psd matrix to another psd matrix, perhaps with a different dimension.

Definition 12.6 (Positive linear map). A linear map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ is *positive* when $\boldsymbol{A} \succcurlyeq \mathbf{0}$ implies that $\boldsymbol{\Phi}(\boldsymbol{A}) \succcurlyeq \mathbf{0}$ for all $\boldsymbol{A} \in \mathbb{M}_n$. We say that $\boldsymbol{\Phi}$ is *strictly positive* when $\boldsymbol{A} \succ \mathbf{0}$ implies that $\boldsymbol{\Phi}(\boldsymbol{A}) \succ \mathbf{0}$.

Exercise 12.7 (Positive maps are monotone). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a *linear* map. Show that the following conditions are equivalent.

1. $\boldsymbol{\Phi}$ is positive.
2. $\boldsymbol{A} \succcurlyeq \boldsymbol{B}$ implies $\boldsymbol{\Phi}(\boldsymbol{A}) \succcurlyeq \boldsymbol{\Phi}(\boldsymbol{B})$ for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n^{\mathrm{sa}}$.

Definition 12.8 (Unital, trace preserving). A linear map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ is *unital* if $\boldsymbol{\Phi}(\mathbf{I}_n) = \mathbf{I}_k$. The map is *trace preserving (TP)* if $\operatorname{tr} \boldsymbol{\Phi}(\boldsymbol{A}) = \operatorname{tr} \boldsymbol{A}$ for all $\boldsymbol{A} \in \mathbb{M}_n$.

Exercise 12.9 (Unital, trace preserving). Show that a linear map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ is unital if and only if its adjoint $\boldsymbol{\Phi}^* : \mathbb{M}_k \to \mathbb{M}_n$ is trace preserving.

We equip matrices with the trace inner product: $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^* \boldsymbol{B})$. Adjoints of linear maps are defined with respect to this inner product.

Compare the definition of a unital linear map and a trace-preserving linear map with the analogous definitions for linear functions on vectors. Recall that these concepts arose in our study of majorization.

12.3 Examples of positive linear maps

There are numerous examples for positive linear maps. Some of the common positive linear maps are listed below.

Example 12.10 (Scalar-valued positive linear maps). Here are some examples of positive linear maps that take numerical values.

1. **Trace.** The trace functional $\varphi : \mathbb{M}_n \to \mathbb{C}$ given by $\varphi(\boldsymbol{A}) := \operatorname{tr} \boldsymbol{A}$ is positive and trace preserving.
2. **Normalized trace.** The normalized trace functional $\varphi : \mathbb{M}_n \to \mathbb{C}$ is defined as

$$\varphi(\boldsymbol{A}) := \overline{\operatorname{tr}}\, \boldsymbol{A} = \frac{1}{n} \operatorname{tr} \boldsymbol{A}.$$

The normalized trace is positive and unital.

3. **Sum of entries.** The linear functional $\varphi : \mathbb{M}_n \to \mathbb{C}$ given by

$$\varphi(\boldsymbol{A}) \coloneqq \sum\nolimits_{i,j} a_{ij}$$

is positive. As usual, $a_{ij} \in \mathbb{C}$ are the entries of $\boldsymbol{A}$. Indeed, we can write $\varphi(\boldsymbol{A}) = \langle \mathbf{1},\ \boldsymbol{A}\mathbf{1} \rangle$, where $\mathbf{1} \in \mathbb{C}^n$ is the vector of ones. The scalar $\langle \mathbf{1},\ \boldsymbol{A}\mathbf{1} \rangle$ is positive whenever $\boldsymbol{A}$ is psd.

4. **Linear functionals.** Each linear functional $\varphi : \mathbb{M}_n \to \mathbb{C}$ can be parameterized as

$$\varphi(\boldsymbol{A}) \coloneqq \operatorname{tr}(\boldsymbol{P}\boldsymbol{A}) \quad \text{where } \boldsymbol{P} \in \mathbb{M}_n.$$

The linear functional φ is positive if and only if $\boldsymbol{P}$ is psd. It is unital if and only if $\operatorname{tr} \boldsymbol{P} = 1$.

Positive semidefinite matrices with trace one are called *density matrices*.

Are there other distinguished positive linear functionals? ▪

Example 12.11 (Matrix-valued positive linear maps). Next, we consider positive linear maps from matrices to matrices.

1. **Scalar matrices.** The linear function $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ given by

$$\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq (\overline{\operatorname{tr}}\boldsymbol{A})\, \mathbf{I}_k$$

is positive and unital. It is trace preserving if and only if $k = n$.

2. **Diagonal.** The linear function $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ given by

$$\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq \operatorname{diag}(\boldsymbol{A}) = \sum\nolimits_j a_{jj} \mathbf{E}_{jj}$$

is positive, unital, and trace preserving.

3. **Conjugation.** Let $\boldsymbol{X} \in \mathbb{C}^{n \times k}$. The linear function $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ given by

$$\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq \boldsymbol{X}^* \boldsymbol{A} \boldsymbol{X}$$

is positive. Furthermore, if $\boldsymbol{X}$ is orthonormal, then $\boldsymbol{\Phi}$ is unital. In this case, the conjugation operation extracts a principal submatrix (with respect to some basis).

The conjugation map is the primitive example of a *completely positive linear map*. See Problem Set 3 for definitions and discussion.

4. **Pinching.** Let $(\boldsymbol{P}_i)$ be a family of orthogonal projectors on $\mathbb{M}_n$ that are mutually orthogonal ($\boldsymbol{P}_i \boldsymbol{P}_j = \delta_{ij} \boldsymbol{P}_i$) and that partition the identity ($\sum_j \boldsymbol{P}_j = \mathbf{I}_n$). The function $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_n$ given by

$$\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq \sum\nolimits_j \boldsymbol{P}_j \boldsymbol{A} \boldsymbol{P}_j$$

is called a *pinching*. The pinching operation $\boldsymbol{\Phi}$ is positive, unital, and trace preserving.

Can you think of more examples? ▪

Example 12.12 (Tensors and positive linear maps). There are also several natural examples associated with tensor product operations.

1. **Tensors.** Fix a psd matrix $\boldsymbol{B} \in \mathbb{M}_k$. The function $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_{nk}$ given by

$$\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq \boldsymbol{B} \otimes \boldsymbol{A}$$

is positive. Likewise, $\boldsymbol{\Phi}(\boldsymbol{A}) \coloneqq \boldsymbol{A} \otimes \boldsymbol{B}$ is positive.

2. **Schur products.** Let $\boldsymbol{B} \in \mathbb{M}_n$ be psd. The Schur product map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_n$ given by
$$\boldsymbol{\Phi}(\boldsymbol{A}) := \boldsymbol{B} \odot \boldsymbol{A}$$
is positive, where $\odot$ denotes the Schur product (entrywise product). This result is a consequence of the Schur product theorem, which states that the Schur product of two psd matrices is a psd matrix. To prove this claim, note that $\boldsymbol{B} \cdot \boldsymbol{A}$ is a principal submatrix of $\boldsymbol{B} \otimes \boldsymbol{A}$.
3. **Partial traces.** Consider the linear space $\mathbb{M}_n \otimes \mathbb{M}_k$ of tensor products of matrices. For an elementary tensor $\boldsymbol{B} \otimes \boldsymbol{A}$, we define the partial trace with respect to the first factor:
$$\mathrm{tr}_1(\boldsymbol{B} \otimes \boldsymbol{A}) := (\mathrm{tr}\, \boldsymbol{B})\, \boldsymbol{A}.$$
We extend this map to all tensor products by linearity. The partial trace is a positive linear map. The partial trace tr_2 with respect to the second factor is defined in an analogous fashion.

Can you think of more examples? ∎

Exercise 12.13 (Cone of positive linear maps). Show that the class of positive linear maps $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ is a closed convex cone.

12.4 Properties of positive linear maps

Positive linear maps satisfy many elegant properties. We begin with some results which show that positive linear maps preserve basic algebraic properties of matrices.

Proposition 12.14 (Self-adjointness). Let $\boldsymbol{\Phi}$ be a positive linear map. For each self-adjoint $\boldsymbol{A}$, the image $\boldsymbol{\Phi}(\boldsymbol{A})$ is also self-adjoint.

Proof. The self-adjoint matrix $\boldsymbol{A}$ has a Jordan decomposition:
$$\boldsymbol{A} = \boldsymbol{A}_+ - \boldsymbol{A}_- \quad \text{where each of } \boldsymbol{A}_\pm \text{ is psd.}$$
By linearity of the map $\boldsymbol{\Phi}$, we can decompose $\boldsymbol{\Phi}(\boldsymbol{A})$ as
$$\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{\Phi}(\boldsymbol{A}_+) - \boldsymbol{\Phi}(\boldsymbol{A}_-) \in \mathbb{M}_n^{\mathrm{sa}}.$$
Indeed, the difference of two psd matrices is self-adjoint. ∎

Recall that if the spectral resolution of $\boldsymbol{A}$ is $\boldsymbol{A} = \sum_j \lambda_j \boldsymbol{P}_j$, then
$$\boldsymbol{A}_+ = \sum\nolimits_j (\lambda_j)_+ \boldsymbol{P}_j$$
$$\boldsymbol{A}_- = \sum\nolimits_j (\lambda_j)_- \boldsymbol{P}_j.$$
As usual, $(a)_+ := \max\{a, 0\}$ and $(a)_- := \max\{-a, 0\}$ for $a \in \mathbb{R}$, P_j is the spectral projector.

Proposition 12.15 (Adjoints). Let $\boldsymbol{\Phi}$ be a positive linear map. Then $\boldsymbol{\Phi}(\boldsymbol{A}^*) = \boldsymbol{\Phi}(\boldsymbol{A})^*$ for each matrix $\boldsymbol{A}$.

Proof. Each matrix $\boldsymbol{A} \in \mathbb{M}_n$ has a Cartesian decomposition:
$$\boldsymbol{A} = \boldsymbol{H} + \mathrm{i}\boldsymbol{S} \quad \text{where each } \boldsymbol{H}, \boldsymbol{S} \in \mathbb{M}_n^{\mathrm{sa}}.$$
By linearity of $\boldsymbol{\Phi}$ and Proposition 12.14,
$$\boldsymbol{\Phi}(\boldsymbol{A}^*) = \boldsymbol{\Phi}(\boldsymbol{H} - \mathrm{i}\boldsymbol{S}) = \boldsymbol{\Phi}(\boldsymbol{H}) - \mathrm{i}\boldsymbol{\Phi}(\boldsymbol{S}) = \big(\boldsymbol{\Phi}(\boldsymbol{H}) + \mathrm{i}\boldsymbol{\Phi}(\boldsymbol{S})\big)^* = \big(\boldsymbol{\Phi}(\boldsymbol{A})\big)^*.$$
This is the required result. ∎

The s.a. matrices $\boldsymbol{H}$ and $\boldsymbol{S}$ are given by the expressions
$$\boldsymbol{H} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*);$$
$$\boldsymbol{S} = \frac{1}{2\mathrm{i}}(\boldsymbol{A} - \boldsymbol{A}^*).$$

12.5 Convexity inequalities

It is fruitful to think about a *unital* positive linear map as a generalization of the expectation of a random variable. Here are several parallels.

Properties	Expectation	Unital, positive map
Linearity	$\mathbb{E}[\alpha X + Y] = \alpha\,\mathbb{E}[X] + \mathbb{E}[Y]$	$\boldsymbol{\Phi}(\alpha \boldsymbol{A} + \boldsymbol{B}) = \alpha\boldsymbol{\Phi}(\boldsymbol{A}) + \boldsymbol{\Phi}(\boldsymbol{B})$
Unital	$\mathbb{E}[1] = 1$	$\boldsymbol{\Phi}(\mathbf{I}) = \mathbf{I}$
Positive	$X \geq 0$ implies $\mathbb{E}[X] \geq 0$	$\boldsymbol{A} \succcurlyeq \mathbf{0}$ implies $\boldsymbol{\Phi}(\boldsymbol{A}) \succcurlyeq \mathbf{0}$

Similarly, a unital, trace-preserving, positive linear map is an analogous to a doubly stochastic matrix, which is a particularly nice kind of averaging operation.

The key fact about expectation is Jensen's inequality, which describes how expectation interacts with convex functions. Likewise, unital positive linear maps satisfy some important convexity theorems. This section develops these ideas.

12.5.1 Schur complements

Before we begin, we must remind the reader about the definition of the Schur complement and the core theorem on Schur complements of psd matrices.

Theorem 12.16 (Schur complements). Assume that $\boldsymbol{A} \succ \mathbf{0}$ is pd matrix. Then

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{H} \end{bmatrix} \succcurlyeq \mathbf{0} \quad \text{if and only if} \quad \boldsymbol{H} - \boldsymbol{B}^*\boldsymbol{A}^{-1}\boldsymbol{B} \succcurlyeq \mathbf{0}.$$

The matrix $\boldsymbol{H} - \boldsymbol{B}^*\boldsymbol{A}^{-1}\boldsymbol{B}$ is called the *Schur complement* of the block matrix with respect to the block $\boldsymbol{A}$.

Proof. By block Gaussian elimination, we have

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{B}^*\boldsymbol{A}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{H} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\boldsymbol{A}^{-1}\boldsymbol{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H} - \boldsymbol{B}^*\boldsymbol{A}^{-1}\boldsymbol{B} \end{bmatrix}. \tag{12.2}$$

The right hand side of (12.2) is positive semidefinite. Apply the conjugation rule to complete the proof. ∎

Schur complements arise from partial Gaussian elimination, from Cholesky decomposition, and from partial least-squares. They also describe conditioning of jointly Gaussian random variables.

12.5.2 The Kadison inequality

Our first result describes how the square function interacts with a unital positive linear map. This result is analogous to the application of Jensen's inequality to the square function.

Theorem 12.17 (Kadison inequality). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a *unital* positive linear map. Then

$$\boldsymbol{\Phi}(\boldsymbol{A})^2 \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}^2) \quad \text{for all s.a. } \boldsymbol{A} \in \mathbb{M}_n^{\mathrm{sa}}.$$

Note that this result only applies to *self-adjoint* matrices.

Warning 12.18 (Larger powers?). The analog of Kadison's inequality is false for powers greater than two. On the other hand, there is a version of Lyapunov's inequality that holds for higher powers. ▪

Proof. Each s.a. matrix $\boldsymbol{A}$ admits a spectral resolution:

$$\boldsymbol{A} = \sum\nolimits_j \lambda_j \boldsymbol{P}_j \quad \text{where } \lambda_j \in \mathbb{R}.$$

As usual, the spectral projectors are mutually orthogonal and decompose the identity. With this representation, the square $\boldsymbol{A}^2$ satisfies

$$\boldsymbol{A}^2 = \sum\nolimits_j \lambda_j^2 \boldsymbol{P}_j.$$

By linearity, we have

$$\boldsymbol{\Phi}(\boldsymbol{A}) = \sum\nolimits_j \lambda_j \boldsymbol{\Phi}(\boldsymbol{P}_j) \quad \text{and} \quad \boldsymbol{\Phi}(\boldsymbol{A}^2) = \sum\nolimits_j \lambda_j^2 \boldsymbol{\Phi}(\boldsymbol{P}_j).$$

Note that $\boldsymbol{P}_j \succcurlyeq \mathbf{0}$ implies that $\boldsymbol{\Phi}(\boldsymbol{P}_j) \succcurlyeq \mathbf{0}$. Since $\boldsymbol{\Phi}$ is unital, we also have

$$\sum\nolimits_j \boldsymbol{\Phi}(\boldsymbol{P}_j) = \boldsymbol{\Phi}\left(\sum\nolimits_j \boldsymbol{P}_j\right) = \boldsymbol{\Phi}(\mathbf{I}) = \mathbf{I}.$$

These facts play a key role in the argument.

The key idea is to form a block matrix and to argue that it is psd. By linearity of $\boldsymbol{\Phi}$ and the preceding displays, we may calculate that

$$\begin{bmatrix} \mathbf{I} & \boldsymbol{\Phi}(\boldsymbol{A}) \\ \boldsymbol{\Phi}(\boldsymbol{A}) & \boldsymbol{\Phi}(\boldsymbol{A}^2) \end{bmatrix} = \sum\nolimits_j \begin{bmatrix} \boldsymbol{\Phi}(\boldsymbol{P}_j) & \lambda_j \boldsymbol{\Phi}(\boldsymbol{P}_j) \\ \lambda_j \boldsymbol{\Phi}(\boldsymbol{P}_j) & \lambda_j^2 \boldsymbol{\Phi}(\boldsymbol{P}_j) \end{bmatrix} = \sum\nolimits_j \begin{bmatrix} 1 & \lambda_j \\ \lambda_j & \lambda_j^2 \end{bmatrix} \otimes \boldsymbol{\Phi}(\boldsymbol{P}_j) \succcurlyeq \mathbf{0}.$$

Indeed, the matrix formed from the eigenvalues is psd because

$$\begin{bmatrix} 1 & \lambda \\ \lambda & \lambda^2 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda \end{bmatrix} \begin{bmatrix} 1 \\ \lambda \end{bmatrix}^* \succcurlyeq \mathbf{0} \quad \text{for } \lambda \in \mathbb{R}.$$

The matrices $\boldsymbol{\Phi}(\boldsymbol{P}_j)$ are also psd, and tensor products preserve positivity.

Now, by the Schur complement theorem, the Schur complement of the block $\mathbf{I}$ in the block matrix is also psd. That is,

$$\mathbf{0} \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}^2) - \boldsymbol{\Phi}(\boldsymbol{A})\mathbf{I}^{-1}\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{\Phi}(\boldsymbol{A}^2) - \boldsymbol{\Phi}(\boldsymbol{A})^2.$$

This is equivalent to the statement of Kadison's inequality. ∎

12.5.3 The Choi inequality

Choi's inequality describes how unital positive linear maps interact with the matrix inverse. This result is analogous to the application of Jensen's inequality to the reciprocal.

Theorem 12.19 (Choi inequality). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a unital, *strictly* positive linear map. Then

$$\boldsymbol{\Phi}(\boldsymbol{A})^{-1} \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}^{-1}) \quad \text{for all pd } \boldsymbol{A} \in \mathbb{M}_n.$$

Note that this result only holds for *positive definite* matrices.

Warning 12.20 (Smaller powers?). The analog of Choi's inequality is false for powers that are less than -1. ▪

Proof sketch. We form a block matrix and argue that it is psd:

$$\begin{bmatrix} \mathbf{\Phi}(\boldsymbol{A}) & \mathbf{I} \\ \mathbf{I} & \mathbf{\Phi}(\boldsymbol{A}^{-1}) \end{bmatrix}.$$

The key new observation is that

$$\begin{bmatrix} \lambda & 1 \\ 1 & \lambda^{-1} \end{bmatrix} = \begin{bmatrix} \lambda^{1/2} \\ \lambda^{-1/2} \end{bmatrix} \begin{bmatrix} \lambda^{1/2} \\ \lambda^{-1/2} \end{bmatrix}^* \succcurlyeq \mathbf{0} \quad \text{for } \lambda > 0.$$

The rest of the details are similar to the proof of Kadison's inequality. ■

Example 12.21 (Diagonals). Use the theorems above, we have the following results.

1. For a self-adjoint matrix $\boldsymbol{A} \in \mathbb{M}_n^{\mathrm{sa}}$, the square of the diagonal entries of $\boldsymbol{A}$ is entrywise bounded by the diagonal entries of $\boldsymbol{A}^2$. That is, we have

   $$\operatorname{diag}(\boldsymbol{A})^2 \preccurlyeq \operatorname{diag}(\boldsymbol{A}^2) \quad \text{for each s.a. } \boldsymbol{A}.$$

   Indeed, $\operatorname{diag} : \mathbb{M}_n \to \mathbb{M}_n$ is a positive unital map. Apply Kadison's inequality.
2. For a pd matrix $\boldsymbol{A} \in \mathbb{M}_n$, the inverse of the diagonal entries of $\boldsymbol{A}$ is entrywise bounded by the diagonal entries of $\boldsymbol{A}^{-1}$. That is,

   $$\operatorname{diag}(\boldsymbol{A})^{-1} \preccurlyeq \operatorname{diag}(\boldsymbol{A}^{-1}) \quad \text{for pd } \boldsymbol{A}.$$

   This result follows from the Choi inequality because diag is *strictly* positive and unital.

You may wish to develop analogous results for other unital positive linear maps from our list of examples. ▪

12.6 Russo–Dye theorem

In the previous section, we studied some basic properties of positive linear maps. Apart from these properties, we also want to know how much a positive map can dilate a matrix. This question leads us to equip linear maps with a norm and to find expressions for the norm of a positive linear map.

Theorem 12.22 (Russo–Dye). Let $\mathbf{\Phi}$ be a *unital* positive linear map. Then

$$\|\mathbf{\Phi}\| := \sup\{\|\mathbf{\Phi}(\boldsymbol{A})\| : \|\boldsymbol{A}\| \le 1\} = 1,$$

where $\|\cdot\|$ is the spectral norm (Schatten ∞-norm).

Corollary 12.23 (Russo–Dye). Let $\mathbf{\Phi}$ be a positive linear map. Then $\|\mathbf{\Phi}\| = \|\mathbf{\Phi}(\mathbf{I})\|$.

We will prove these results in the upcoming subsections. There are many interesting applications of these results. See Problem Set 3 for examples.

12.6.1 Contractions

To begin, we need a basic fact from operator theory.

A *contraction* is a matrix $\boldsymbol{K} \in \mathbb{M}_n$ that satisfies $\|\boldsymbol{K}\| \leq 1$.

Proposition 12.24 (Contractions). Each contraction $\boldsymbol{K} \in \mathbb{M}_n$ can be written as an average of two unitary matrices:

$$\boldsymbol{K} = \tfrac{1}{2}(\boldsymbol{U} + \boldsymbol{V}) \quad \text{where } \boldsymbol{U}, \boldsymbol{V} \in \mathbb{M}_n \text{ are unitary.}$$

Exercise 12.25 (Contractions). Prove Proposition 12.24. **Hint:** Each singular value σ_j of $\boldsymbol{K}$ satisfies $\sigma_j \in [0, 1]$. Observe that $\sigma_j = \cos(\theta_j) = \frac{1}{2}\big(\mathrm{e}^{\mathrm{i}\theta_j} + \mathrm{e}^{-\mathrm{i}\theta_j}\big)$ for a number $\theta_j \in [0, 2\pi)$.

12.6.2 Proof of Theorem 12.22

Let us establish the Russo–Dye theorem. We begin with the case of a unitary matrix. We will show that $\|\boldsymbol{\Phi}(\boldsymbol{U})\| \leq 1$ for each unitary $\boldsymbol{U}$. Adapting the proof of the Kadison inequality, we can show that the following block matrix is psd:

$$\begin{bmatrix} \mathbf{I} & \boldsymbol{\Phi}(\boldsymbol{U}) \\ \boldsymbol{\Phi}(\boldsymbol{U})^* & \mathbf{I} \end{bmatrix} \succcurlyeq \mathbf{0}.$$

For this purpose, recall that each unitary matrix has a spectral resolution where the eigenvalues are complex numbers with modulus one. By the Schur complement theorem, we have

$$\boldsymbol{\Phi}(\boldsymbol{U})^* \boldsymbol{\Phi}(\boldsymbol{U}) \preccurlyeq \mathbf{I}.$$

This is equivalent to $\|\boldsymbol{\Phi}(\boldsymbol{U})\| \leq 1$.

For a general matrix $\boldsymbol{A} \in \mathbb{M}_n$ with $\|\boldsymbol{A}\| \leq 1$, note that $\boldsymbol{A}$ is a contraction. Thus, we can write

$$\boldsymbol{A} = \tfrac{1}{2}\big(\boldsymbol{U} + \boldsymbol{V}\big)$$

for some unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. By linearity of $\boldsymbol{\Phi}$ and the triangle inequality for the norm, we have

$$\|\boldsymbol{\Phi}(\boldsymbol{A})\| = \|\tfrac{1}{2}\boldsymbol{\Phi}(\boldsymbol{U}) + \tfrac{1}{2}\boldsymbol{\Phi}(\boldsymbol{V})\| \leq \tfrac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{U})\| + \tfrac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{V})\| \leq 1.$$

We have shown that

$$\|\boldsymbol{\Phi}\| = \sup\{\|\boldsymbol{\Phi}(\boldsymbol{A})\| : \|\boldsymbol{A}\| \leq 1\} \leq 1.$$

To finish, note that $\boldsymbol{\Phi}(\mathbf{I}) = \mathbf{I}$ because $\boldsymbol{\Phi}$ is unital. Therefore, we may conclude that $\|\boldsymbol{\Phi}\| \geq \|\boldsymbol{\Phi}(\mathbf{I})\| = 1$.

12.6.3 Proof of Corollary 12.23

We now turn to the proof of the corollary. First, assume that $\boldsymbol{\Phi}$ is *strictly* positive. In this case, the matrix $\boldsymbol{P} := \boldsymbol{\Phi}(\mathbf{I})$ is positive definite. We can form another linear map $\boldsymbol{\Psi}$ by the expression

$$\boldsymbol{\Psi}(\boldsymbol{A}) := \boldsymbol{P}^{-1/2} \boldsymbol{\Phi}(\boldsymbol{A}) \boldsymbol{P}^{-1/2}.$$

By the conjugation rule, the map $\boldsymbol{\Psi}$ is positive and linear. It is also unital because

$$\boldsymbol{\Psi}(\mathbf{I}) := \boldsymbol{P}^{-1/2} \boldsymbol{\Phi}(\mathbf{I}) \boldsymbol{P}^{-1/2} = \boldsymbol{P}^{-1/2} \boldsymbol{P} \boldsymbol{P}^{-1/2} = \mathbf{I}.$$

Thus, by the Russo–Dye theorem, we conclude that $\|\boldsymbol{\Psi}\| = 1$. For any contraction $\boldsymbol{A} \in \mathbb{M}_n$ with $\|\boldsymbol{A}\| \leq 1$, we have

$$\|\boldsymbol{\Phi}(\boldsymbol{A})\| = \|\boldsymbol{P}^{1/2} \boldsymbol{\Psi}(\boldsymbol{A}) \boldsymbol{P}^{1/2}\| \leq \|\boldsymbol{P}\| \|\boldsymbol{\Psi}(\boldsymbol{A})\| \leq \|\boldsymbol{P}\| = \|\boldsymbol{\Phi}(\mathbf{I})\|.$$

Considering $\boldsymbol{A} = \mathbf{I}$, we realize that $\|\boldsymbol{\Phi}\| = \|\boldsymbol{\Phi}(\mathbf{I})\|$.

To complete the argument, we we assume that $\boldsymbol{\Phi}$ is positive, but maybe not strictly positive. Consider the following family of strictly positive linear maps:

$$\boldsymbol{\Phi}_\varepsilon(\boldsymbol{A}) := \boldsymbol{\Phi}(\boldsymbol{A}) + \varepsilon(\overline{\mathrm{tr}}\, \boldsymbol{A})\mathbf{I} \quad \text{for } \varepsilon > 0.$$

We can apply the result of the last paragraph to see that

$$\|\boldsymbol{\Phi}_\varepsilon\| = \|\boldsymbol{\Phi}_\varepsilon(\mathbf{I})\| = \|\Phi(\mathbf{I}) + \varepsilon\mathbf{I}\|.$$

By continuity of the norm, we can let $\varepsilon \downarrow 0$ to conclude that $\|\boldsymbol{\Phi}\| = \|\boldsymbol{\Phi}(\mathbf{I})\|$ for any positive linear map $\boldsymbol{\Phi}$.

Notes

This lecture is adapted from Bhatia's book on positive-definite matrices [Bha07b, Chap. 2].

Lecture bibliography

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

13. Matrix Monotonicity and Convexity

Date: 15 February 2022 **Scribe:** Nico Christianson

Agenda:
1. Monotonicity and Convexity
2. Examples
3. Matrix Jensen

In the last lecture, we examined positive linear maps, a class of matrix-valued linear maps that respect the positive semidefinite order. Now, we go beyond linear maps, establishing notions of *monotonicity* and *convexity* for nonlinear matrix-valued functions with respect to the semidefinite order. It is found that these properties are more restrictive than their scalar counterparts. Yet, this restriction brings with it considerable added structure. In particular, we demonstrate that matrix convex functions satisfy a remarkable generalization of Jensen's inequality to non-commutative "matrix convex combinations."

13.1 Basic definitions and properties

We begin by recalling some notions from previous lectures. A linear map $\mathbf{\Phi} : \mathbb{M}_n \to \mathbb{M}_m$ is said to be *positive* if and only if it enjoys the monotonicity property

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \quad \text{implies} \quad \mathbf{\Phi}(\boldsymbol{A}) \preccurlyeq \mathbf{\Phi}(\boldsymbol{B})$$

for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$. It follows trivially by linearity that a positive linear map is convex; i.e.,

$$\mathbf{\Phi}(\tau \boldsymbol{A} + \bar{\tau}\boldsymbol{B}) \preccurlyeq \tau \mathbf{\Phi}(\boldsymbol{A}) + \bar{\tau}\mathbf{\Phi}(\boldsymbol{B}) \quad \text{for all } \tau \in [0,1],$$

where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$ and $\bar{\tau} := 1-\tau$. In this lecture, we examine the class of *nonlinear* functions that exhibit monotonicity and convexity with respect to the positive semidefinite order.

13.1.1 Standard matrix functions

We begin with several definitions.

Definition 13.1 Let $\mathsf{I} \subseteq \mathbb{R}$ be an interval of the real line. We define $\mathbb{H}_n(\mathsf{I})$ as the set of Hermitian matrices with all eigenvalues lying in the interval I. That is,

$$\mathbb{H}_n(\mathsf{I}) = \{\boldsymbol{A} \in \mathbb{M}_n : \boldsymbol{A} = \boldsymbol{A}^* \text{ and } \lambda_i(\boldsymbol{A}) \in \mathsf{I} \text{ for each } i = 1, \dots, n\}.$$

It follows via Rayleigh-Ritz that $\mathbb{H}_n(\mathsf{I})$ is convex.

With $\mathbb{H}_n(\mathsf{I})$ defined, we can provide a definition of standard matrix functions that extends to functions whose domain is a proper subset of the real line, such as the inverse, the square root, the logarithm, and entropy.

Definition 13.2 (Standard matrix function). Let $f : \mathsf{I} \to \mathbb{R}$ be a function on an interval $\mathsf{I} \subseteq \mathbb{R}$. For each $n \in \mathbb{N}$, define a matrix function $f : \mathbb{H}_n(\mathsf{I}) \to \mathbb{H}_n$ via the spectral resolution. That is, for any $n \in \mathbb{N}$ and each matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathsf{I})$, we define

$$f(\boldsymbol{A}) = \sum_{i=1}^n f(\lambda_i)\boldsymbol{P}_i \quad \text{where} \quad \boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{P}_i.$$

Note that standard matrix functions are unitarily equivariant. That is, if $\boldsymbol{U} \in \mathbb{M}_n$ is unitary, then $f(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}) = \boldsymbol{U}^* f(\boldsymbol{A})\boldsymbol{U}$.

13.1.2 Monotonicity and Convexity

This broadened definition of standard matrix functions enables a natural generalization of the properties of monotonicity [Löw34] and convexity [Kra36] to matrices.

These properties are sometimes called *operator monotonicity and convexity*.

Definition 13.3 (Matrix monotonicity; Loewner 1934). A function $f : \mathsf{I} \to \mathbb{R}$ is *matrix monotone* on I when $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ implies $f(\boldsymbol{A}) \preccurlyeq f(\boldsymbol{B})$ for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathsf{I})$ and all $n \in \mathbb{N}$.

Definition 13.4 (Matrix convexity; Kraus 1936). A function $f : \mathsf{I} \to \mathbb{R}$ is *matrix convex* on I when

$$f(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{B}) \preccurlyeq \tau f(\boldsymbol{A}) + \bar{\tau} f(\boldsymbol{B}) \quad \text{for all } \tau \in [0, 1],$$

for all matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathsf{I})$ and all $n \in \mathbb{N}$, where $\bar{\tau} = 1 - \tau$. We say that a function $g : \mathsf{I} \to \mathbb{R}$ is *matrix concave* when $-g$ is matrix convex.

13.1.3 Basic properties

We briefly report some of the properties that are immediate from the definitions of matrix monotonicity and convexity.

Proposition 13.5 (Matrix monotonicity). The following statements hold true.

The key takeaway of Proposition 13.5 is that the class of matrix monotone functions forms a closed, convex cone. As we will see later, matrix monotone functions are in fact a proper subset of the cone of scalar monotone functions.

1. **Scalar case.** If f is matrix monotone on I, then f is increasing on I.
2. **Convex cone.** If f, g are matrix monotone on I, then for all $\alpha, \beta \geq 0$, the weighted combination $\alpha f + \beta g$ is matrix monotone on I.
3. **Closure.** If $f_k \to f$ pointwise on I and each f_k is matrix monotone on I, then the limit f is also matrix monotone on I.

Proof sketch. The first two properties can be verified easily. The third property requires use of the fact that the cone of psd matrices is closed. ∎

Exercise 13.6 (Matrix convexity). Verify analogous properties for the class of matrix convex functions. That is, characterize the scalar case, and check that the class of matrix convex functions forms a closed, convex cone.

13.2 Examples

In this section, we detail a variety of functions that are (or are not) matrix monotone and convex.

13.2.1 Affine functions

Consider the affine function $f(t) = \alpha t + \beta$ with $\alpha, \beta \in \mathbb{R}$. The function f is matrix convex on $\mathbb{R}$ regardless of the choice of α, β. Moreover, f is matrix monotone on $\mathbb{R}$ so long as it is increasing, i.e., if $\alpha \geq 0$.

Increasing affine functions are the *unique* examples of matrix monotone functions on the entire real line!

13.2.2 Square function

Consider the square function $f(t) = t^2$. This function is *not* matrix monotone on $\mathbb{R}_+$. To see this, consider the matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_2(\mathbb{R}_+)$ defined as

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

Clearly $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$. Yet

$$\boldsymbol{A}^2 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \not\preccurlyeq \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix} = \boldsymbol{B}^2.$$

On the other hand, the square function f is matrix convex on the entire real line. More generally, positive quadratics $f(t) = \alpha t^2 + \beta t + \gamma$ with $\alpha \geq 0$ and $\beta, \gamma \in \mathbb{R}$ compose the full class of matrix convex functions on $\mathbb{R}$.

We prove the matrix convexity of the square function in the next proposition.

Proposition 13.7 (Square is matrix convex). The square function $t \mapsto t^2$ is matrix convex on $\mathbb{R}$.

Proof. We establish midpoint matrix convexity, which can be extended in a straightforward manner to general matrix convexity. Consider arbitrary $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$, and observe that

$$\mathbf{0} \preccurlyeq (\boldsymbol{A} - \boldsymbol{B})^2 = \boldsymbol{A}^2 + \boldsymbol{B}^2 - \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A} \tag{13.1}$$

since the eigenvalues of a squared matrix are positive. Rearranging (13.1), we obtain

$$\boldsymbol{A}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{A} \preccurlyeq \boldsymbol{A}^2 + \boldsymbol{B}^2. \tag{13.2}$$

Therefore, it holds that

$$\left(\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\right)^2 = \tfrac{1}{4}\left(\boldsymbol{A}^2 + \boldsymbol{B}^2 + \boldsymbol{A}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{A}\right) \preccurlyeq \tfrac{1}{2}\boldsymbol{A}^2 + \tfrac{1}{2}\boldsymbol{B}^2$$

where the semidefinite relation follows via (13.2). This calculation establishes midpoint convexity. ∎

Exercise 13.8 (Square: Matrix convexity). Extend the preceding proof to obtain general matrix convexity of the square function. That is, establish that

$$(\tau\boldsymbol{A} + \bar{\tau}\boldsymbol{B})^2 \preccurlyeq \tau\boldsymbol{A}^2 + \bar{\tau}\boldsymbol{B}^2 \quad \text{for all } \tau \in [0, 1],$$

where $\bar{\tau} = 1 - \tau$.

13.2.3 Inverse

The inverse function $f(t) = t^{-1}$ is matrix convex on $\mathbb{R}_{++} := (0, \infty)$, and its negative $g(t) = -t^{-1}$ is matrix monotone on $\mathbb{R}_{++}$. We prove these properties separately in the next two propositions.

Proposition 13.9 (Inverse: Monotonicity). The negative inverse function $t \mapsto -t^{-1}$ is matrix monotone on $\mathbb{R}_{++}$.

Proof. By the conjugation rule, $\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{B}$ implies that $\mathbf{I} \preccurlyeq \boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}$. Since all eigenvalues of the identity are one, the inverse reverses this relation, so that $\mathbf{I} \succcurlyeq (\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})^{-1} = \boldsymbol{A}^{1/2}\boldsymbol{B}^{-1}\boldsymbol{A}^{1/2}$. Applying the conjugation rule once more gives $\boldsymbol{A}^{-1} \succcurlyeq \boldsymbol{B}^{-1}$. Last, negation reverses the positive-semidefinite order. ∎

Proposition 13.10 (Inverse: Convexity). The inverse function $t \mapsto t^{-1}$ is matrix convex on $\mathbb{R}_{++}$.

Proof. For positive definite matrices $\boldsymbol{A}, \boldsymbol{B} \succ \mathbf{0}$, it follows from the Schur complement theorem that

$$\begin{bmatrix} \boldsymbol{A} & \mathbf{I} \\ \mathbf{I} & \boldsymbol{A}^{-1} \end{bmatrix} \succcurlyeq \mathbf{0} \quad \text{and} \quad \begin{bmatrix} \boldsymbol{B} & \mathbf{I} \\ \mathbf{I} & \boldsymbol{B}^{-1} \end{bmatrix} \succcurlyeq \mathbf{0}.$$

Convexity of the positive-semidefinite cone implies that, for any $\tau \in [0,1]$ and $\bar{\tau} = 1 - \tau$,

$$\begin{bmatrix} \tau \boldsymbol{A} + \bar{\tau} \boldsymbol{B} & \mathbf{I} \\ \mathbf{I} & \tau \boldsymbol{A}^{-1} + \bar{\tau} \boldsymbol{B}^{-1} \end{bmatrix} \succcurlyeq \mathbf{0}. \tag{13.3}$$

Applying the Schur complement theorem to (13.3), we obtain that

$$(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{B})^{-1} \preccurlyeq \tau \boldsymbol{A}^{-1} + \bar{\tau} \boldsymbol{B}^{-1},$$

which establishes matrix convexity. ∎

13.2.4 Power functions

We report (without proof) the matrix monotonicity and convexity statuses of the family of power functions $f(t) = t^p$. Proof sketches for these facts will be presented in a Lecture 15, on integral representations of matrix monotone and convex functions.

Fact 13.11 (Powers: Monotonicity). The following power functions (and only these) are matrix monotone.

- The power function $f(t) = t^p$ is matrix monotone on $\mathbb{R}_+$ for $p \in [0,1]$.
- The power function $f(t) = -t^p$ is matrix monotone on $\mathbb{R}_{++}$ for $p \in [-1,0]$.

The former result is due to Löwner [Löw34], though it is sometimes referred to as the Löwner–Heinz theorem. ∎

Fact 13.12 (Powers: Convexity). The following power functions (and only these) are matrix convex.

- The power function $f(t) = t^p$ is matrix convex on $\mathbb{R}_+$ for $p \in [1,2]$.
- The power function $f(t) = -t^p$ is matrix convex on $\mathbb{R}_+$ for $p \in [0,1]$.
- The power function $f(t) = t^p$ is matrix convex on $\mathbb{R}_{++}$ for $p \in [-1,0]$.

These results follow as a consequence of Fact 13.11 by general considerations that will be discussed in Lecture 15. ∎

13.2.5 Logarithm

The function $f(t) = \log t$ is matrix monotone and concave on $\mathbb{R}_{++}$. We leave the proof of this fact as a series of exercises.

Exercise 13.13 (Logarithm: Integral representation). For each $a > 0$, verify that

$$\log a = \int_0^\infty \left[(1+\lambda)^{-1} - (a+\lambda)^{-1} \right] \mathrm{d}\lambda.$$

Show that this integral representation extends to positive definite matrices via capitalization. For any positive definite $\boldsymbol{A} \succ \mathbf{0}$,

$$\log \boldsymbol{A} = \int_0^\infty \left[(1+\lambda)^{-1}\mathbf{I} - (\boldsymbol{A} + \lambda \mathbf{I})^{-1} \right] \mathrm{d}\lambda.$$

Exercise 13.14 (Inverse: Monotonicity and Convexity). For $\lambda \geq 0$, show that the map $f(t) = (\lambda + t)^{-1}$ is matrix convex on $\mathbb{R}_{++}$, and that its negative $-f$ is matrix monotone on $\mathbb{R}_{++}$. **Hint:** The arguments follow those for the inverse.

Exercise 13.15 (Logarithm: Monotonicity and Concavity). Show that the logarithm $f(t) = \log t$ is matrix monotone and matrix *concave* on $\mathbb{R}_{++}$. **Hint:** apply the previous exercises and the fact that matrix monotone and convex functions compose convex cones that are closed under pointwise limits.

13.2.6 Entropy

The negative entropy $f(t) = t \log t$ is matrix convex on $\mathbb{R}_+$. We leave the proof of this fact as an exercise.

Exercise 13.16 (Entropy: Convexity). Prove that the negative entropy is matrix convex on $\mathbb{R}_+$. **Hint:** Use the identity

$$t \log t = \int_0^\infty \left[t(1+\lambda)^{-1} - t(t+\lambda)^{-1} \right] \mathrm{d}\lambda$$

and follow the route charted by Exercises 13.13, 13.14, and 13.15.

13.2.7 Exponential

The exponential $f(t) = \mathrm{e}^t$ is *neither* matrix monotone *nor* matrix convex on any interval. The exponential's failure to be matrix monotone or convex implies that the classes of matrix monotone and convex functions are *strictly* smaller than their scalar counterparts.

13.3 The matrix Jensen inequality

A striking fact about convexity in the scalar setting is that it is a self-improving property. That is, by simply assuming that

$$f(\tau a + \bar{\tau} b) \le \tau f(a) + \bar{\tau} f(b) \quad \text{for all } \tau \in [0,1]$$

and all $a, b \in \mathsf{I}$, where $\bar{\tau} = 1 - \tau$, we can obtain Jensen's inequality. For any collection of $a_i \in \mathsf{I}$ and $(p_i)_{i=1}^m$ with $\sum_{i=1}^m p_i = 1$ and $p_i \ge 0$, we have

$$f\left(\sum\nolimits_{i=1}^m p_i a_i \right) \le \sum\nolimits_{i=1}^m p_i f(a_i).$$

Even more, it holds that

$$f(\mathbb{E}\, X) \le \mathbb{E}\, f(X)$$

for all integrable random variables X taking values in I.

As it turns out, and as we shall prove in this section, convexity is self-improving in the matrix setting. Moreover, this self-improvement is even more dramatic, extending beyond simple scalar convex combinations to noncommutative "matrix convex combinations," which we define as follows.

Definition 13.17 (Matrix convex combination). Let $(\boldsymbol{A}_i)_{i=1}^m$ be a collection of self-adjoint matrices in $\mathbb{H}_n$. Let $(\boldsymbol{K}_i)_{i=1}^m$ consist of general matrices in $\mathbb{M}_n$ that satisfy $\sum_{i=1}^m \boldsymbol{K}_i^* \boldsymbol{K}_i = \mathbf{I}$. In analogy to the scalar case, the sum

$$\sum\nolimits_{i=1}^m \boldsymbol{K}_i^* \boldsymbol{A}_i \boldsymbol{K}_i$$

is called a *matrix convex combination* of the matrices $(\boldsymbol{A}_i)_{i=1}^m$.

The condition $\sum_{i=1}^m \boldsymbol{K}_i^* \boldsymbol{K}_i = \mathbf{I}$ on the "coefficients" of a matrix convex combination is analogous to the normalization condition $\sum_{i=1}^m p_i = 1$ in the scalar case. Moreover, scalar convex combinations can be recovered by selecting $\boldsymbol{K}_i = p_i^{1/2} \mathbf{I}$ for each $i = 1, \dots, m$. However, matrix convex combinations significantly generalize the scalar case, as illustrated in the following example.

Example 13.18 (A matrix convex combination). Consider the matrices

$$\boldsymbol{K}_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{K}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Clearly it holds that $\boldsymbol{K}_1^*\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{K}_2 = \mathbf{I}$. For arbitrary matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_2$, it can be verified that

$$\boldsymbol{K}_1^*\boldsymbol{A}\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{B}\boldsymbol{K}_2 = \begin{bmatrix} b_{22} & 0 \\ 0 & a_{11} \end{bmatrix},$$

which is obviously not a scalar convex combination of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. ∎

The remarkable main result of this section is that matrix convexity is self-improving to an extent far beyond the scalar case. That is, matrix convexity of a function $f : \mathsf{I} \to \mathbb{R}$ implies a more general form of Jensen's inequality that holds for arbitrary matrix convex combinations [HP82; HP03].

Theorem 13.19 (Matrix Jensen inequality; Hansen–Pedersen 1982, 2003). Fix a matrix convex function $f : \mathsf{I} \to \mathbb{R}$. Let $(\boldsymbol{A}_i)_{i=1}^m$ be a collection of matrices, each residing in $\mathbb{H}_n(\mathsf{I})$, and let $(\boldsymbol{K}_i)_{i=1}^m$ consist of matrices in $\mathbb{M}_n$ that satisfy the normalization condition $\sum_{i=1}^m \boldsymbol{K}_i^*\boldsymbol{K}_i = \mathbf{I}$. Then

$$f\left(\sum\nolimits_{i=1}^m \boldsymbol{K}_i^*\boldsymbol{A}_i\boldsymbol{K}_i\right) \preccurlyeq \sum\nolimits_{i=1}^m \boldsymbol{K}_i^* f(\boldsymbol{A}_i)\boldsymbol{K}_i.$$

Proof of Theorem 13.19. We present the proof in the case that $m = 2$. That is, we will prove that

$$f\left(\boldsymbol{K}_1^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{A}_2\boldsymbol{K}_2\right) \preccurlyeq \boldsymbol{K}_1^* f(\boldsymbol{A}_1)\boldsymbol{K}_1 + \boldsymbol{K}_2^* f(\boldsymbol{A}_2)\boldsymbol{K}_2$$

when $\boldsymbol{K}_1^*\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{K}_2 = \mathbf{I}$. The case of general m follows a similar argument, and in fact follows as a corollary of this case.

To begin, we introduce that block matrix

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_2 \end{bmatrix} \in \mathbb{H}_{2n}(\mathsf{I}).$$

The key idea in the proof is to lift our attention to this block matrix. We will reinterpret the matrix convex combinations in the matrix Jensen inequality in terms of operations on the block matrix that involve simple averages, unitary conjugation, and positive linear maps. By this mechanism, we can access the definition of matrix convexity and exploit the unitary equivariance of standard matrix functions.

We preface the proof with four tricks that will be employed in the execution of the proof. First, note that it is straightforward to apply a standard matrix function to a block diagonal matrix. If $\boldsymbol{T}, \boldsymbol{M} \in \mathbb{H}_n(\mathsf{I})$ and $f : \mathsf{I} \to \mathbb{R}$, we simply have

$$f\left(\begin{bmatrix} \boldsymbol{T} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \end{bmatrix}\right) = \begin{bmatrix} f(\boldsymbol{T}) & \mathbf{0} \\ \mathbf{0} & f(\boldsymbol{M}) \end{bmatrix}. \tag{13.4}$$

In view of this fact, it is helpful to develop methods for extracting the block-diagonal part of a block matrix.

Indeed, we can represent the block diagonal pinching of a matrix as a simple average of unitary conjugates. That is, defining the unitary block diagonal matrix

$$\boldsymbol{U} := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix},$$

we have the identity

$$\frac{1}{2}\begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix} + \frac{1}{2}\boldsymbol{U}^*\begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} \end{bmatrix} \tag{13.5}$$

where $\boldsymbol{B} \in \mathbb{M}_n$ is an arbitrary matrix.

Third, we shall represent the matrix convex combination via unitary conjugation. Observe that $\boldsymbol{K}_1{}^*\boldsymbol{K}_1 + \boldsymbol{K}_2{}^*\boldsymbol{K}_2 = \mathbf{I}$ implies that the block matrix $\begin{bmatrix} \boldsymbol{K}_1{}^* & \boldsymbol{K}_2{}^* \end{bmatrix}^*$ has orthonormal columns; thus by extending the columns into an orthonormal basis, we can extend it into a unitary matrix $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{L}_1 \\ \boldsymbol{K}_2 & \boldsymbol{L}_2 \end{bmatrix} \quad \text{where} \quad \boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^* = \mathbf{I}.$$

Conjugating $\boldsymbol{A}$ with $\boldsymbol{Q}$, we have

$$\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{K}_1{}^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2{}^*\boldsymbol{A}_2\boldsymbol{K}_2 & * \\ * & * \end{bmatrix} \tag{13.6}$$

where the asterices indicate blocks of the matrix that are irrelevant to the argument.

Last, recall that the map $[\cdot]_{11}$ that extracts the top left $(1,1)$ block of a block matrix is a positive linear map, and hence it preserves the positive-semidefinite order. For our purposes, this map extracts the top left block of $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$ specified in (13.6). For example, it holds that

$$\begin{aligned} [\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}]_{11} &= \boldsymbol{K}_1{}^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2{}^*\boldsymbol{A}_2\boldsymbol{K}_2; \\ [\boldsymbol{Q}^*f(\boldsymbol{A})\boldsymbol{Q}]_{11} &= \boldsymbol{K}_1{}^*f(\boldsymbol{A}_1)\boldsymbol{K}_1 + \boldsymbol{K}_2{}^*f(\boldsymbol{A}_2)\boldsymbol{K}_2. \end{aligned}$$

The second relation exploits the fact that $\boldsymbol{A}$ is a block diagonal matrix so we can identify $f(\boldsymbol{A})$.

Armed with these tricks, we may now proceed with the proof. We will maintain the quantities of interest in the $(1,1)$ block of the block matrix, while the remaining blocks remain at our discretion. To begin, observe that

$$f(\boldsymbol{K}_1{}^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2{}^*\boldsymbol{A}_2\boldsymbol{K}_2) = f([\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}]_{11})$$

by the identity (13.6). Since each block matrix in the pinching identity (13.5) has the same first diagonal block, it then follows that

$$f([\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}]_{11}) = f\left(\left[\tfrac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \tfrac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right]_{11}\right).$$

The identity (13.5) ensures that $\frac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \frac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}$ is block diagonal. Then using the first identity (13.4), the application of a standard matrix function to a block diagonal matrix, we can pull the map $[\cdot]_{11}$ outside f:

$$f\left(\left[\tfrac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \tfrac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right]_{11}\right) = \left[f\left(\tfrac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \tfrac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right)\right]_{11}.$$

By matrix convexity of f, it holds that

$$f\left(\tfrac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \tfrac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right) \preccurlyeq \tfrac{1}{2}f(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}) + \tfrac{1}{2}f(\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}).$$

This inequality transmits through the positive linear map $[\cdot]_{11}$. Thus,

$$\left[f\left(\tfrac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \tfrac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right)\right]_{11} \preccurlyeq \left[\tfrac{1}{2}f(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}) + \tfrac{1}{2}f(\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U})\right]_{11}.$$

Since standard matrix functions commute with unitary conjugation, we have

$$\left[\tfrac{1}{2} f(\boldsymbol{Q}^* \boldsymbol{A} \boldsymbol{Q}) + \tfrac{1}{2} f(\boldsymbol{U}^* (\boldsymbol{Q}^* \boldsymbol{A} \boldsymbol{Q}) \boldsymbol{U})\right]_{11} = \left[\tfrac{1}{2} \boldsymbol{Q}^* f(\boldsymbol{A}) \boldsymbol{Q} + \tfrac{1}{2} \boldsymbol{U}^* \boldsymbol{Q}^* f(\boldsymbol{A}) \boldsymbol{Q} \boldsymbol{U}\right]_{11}.$$

To complete the argument, we reverse our course to undo each of the steps. Once more, using the fact that each block matrix in the pinching identity (13.5) has the same first diagonal block, we conclude that

$$\begin{aligned}\left[\tfrac{1}{2} \boldsymbol{Q}^* f(\boldsymbol{A}) \boldsymbol{Q} + \tfrac{1}{2} \boldsymbol{U}^* \boldsymbol{Q}^* f(\boldsymbol{A}) \boldsymbol{Q} \boldsymbol{U}\right]_{11} &= [\boldsymbol{Q}^* f(\boldsymbol{A}) \boldsymbol{Q}]_{11} \\ &= \boldsymbol{K}_1{}^* f(\boldsymbol{A}_1) \boldsymbol{K}_1 + \boldsymbol{K}_2{}^* f(\boldsymbol{A}_2) \boldsymbol{K}_2.\end{aligned}$$

Sequencing the preceding displays, we have obtained that

$$f(\boldsymbol{K}_1{}^* \boldsymbol{A}_1 \boldsymbol{K}_1 + \boldsymbol{K}_2{}^* \boldsymbol{A}_2 \boldsymbol{K}_2) \preccurlyeq \boldsymbol{K}_1{}^* f(\boldsymbol{A}_1) \boldsymbol{K}_1 + \boldsymbol{K}_2{}^* f(\boldsymbol{A}_2) \boldsymbol{K}_2,$$

which is the desired result. ∎

Notes

This lecture is adapted from Bhatia's book [Bha97, Chap. V] and from the instructor's monograph [Tro15] on matrix concentration.

The proof of the matrix Jensen inequality is drawn from Hansen & Pedersen's second paper [HP03]. Using a similar type of argument, Davis [Dav57] had long since proved a weaker version of the matrix Jensen inequality (Theorem 13.19) under the extra assumptions that $0 \in \mathsf{I}$ and $f(0) = 0$. By bringing in deeper tools, Choi [Cho74] strengthened the result further by removing the conditions, and he extended it to a wider class of averaging operations.

The significance of the Hansen–Pedersen result is that it admits a direct (although clever) proof. They developed this argument as the first step in their proof of Loewner's integral theorem, a deep result on matrix monotone and matrix convex convex functions. In fact, Loewner's theorem is the key ingredient in the proof of Choi's generalization. We will return to these matters in Lecture 15.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Cho74] M.-D. Choi. "A Schwarz inequality for positive linear maps on C^*-algebras". In: *Illinois Journal of Mathematics* 18.4 (1974), pages 565 –574. DOI: 10.1215/ijm/1256051007.

[Dav57] C. Davis. "A Schwarz inequality for convex operator functions". In: *Proc. Amer. Math. Soc.* 8 (1957), pages 42–44. DOI: 10.2307/2032808.

[HP82] F. Hansen and G. K. Pedersen. "Jensen's Inequality for Operators and Löwner's Theorem." In: *Mathematische Annalen* 258 (1982), pages 229–241.

[HP03] F. Hansen and G. K. Pedersen. "Jensen's Operator Inequality". In: *Bulletin of the London Mathematical Society* 35.4 (2003), pages 553–564.

[Kra36] F. Kraus. "Über konvexe Matrixfunktionen." ger. In: *Mathematische Zeitschrift* 41 (1936), pages 18–42. URL: http://eudml.org/doc/168648.

[Löw34] K. Löwner. "Über monotone Matrixfunktionen". In: *Mathematische Zeitschrift* 38 (1934), pages 177–216. URL: http://eudml.org/doc/168495.

[Tro15] J. A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pages 1–230.

14. Monotonicity: Differential Characterization

Date: 17 February 2022 **Scribe:** Nicholas H. Nelsen

In this lecture, we continue to analyze the monotonicity and convexity of nonlinear standard matrix functions. Motivated by scalar tests for monotonicity and convexity, we seek analogous results in the matrix function setting. Loewner's theorem, the main result of the lecture, gives a precise characterization of matrix monotonicity; it is necessary and sufficient to check that a particular kernel matrix—associated with the nonlinear function in question and its derivatives—is positive semidefinite. The remainder of the lecture is devoting to proving this result. To that end, we take a detour to first discuss differentiation in normed vector spaces, explicitly characterize these derivatives for standard matrix functions, and finally give some concrete examples where Loewner's result may be applied.

Agenda:

1. Recap
2. Derivatives and differences
3. Loewner's matrix
4. Fréchet derivatives
5. Daleckii–Krein theorem
6. Proof of Loewner theorem
7. Examples

14.1 Recap

The previous two lectures introduced matrix monotonicity and matrix convexity, first in the context of positive linear maps (i.e., linear functions on matrices) and then in the context of standard matrix functions (which have a more rigid structure but are nonlinear). We obtained important convexity inequalities, such as Choi's inequality for positive linear maps and the matrix Jensen inequality for matrix convex functions.

To set the stage for this lecture, we next recall the following definitions. For an interval $\mathsf{I} \subseteq \mathbb{R}$, we define $\mathbb{H}_n(\mathsf{I}) := \{\boldsymbol{A} \in \mathbb{H}_n : \ \lambda_i(\boldsymbol{A}) \in \mathsf{I} \text{ for all } i\}$.

This set I usually serves as the domain of the standard matrix functions we will consider.

Definition 14.1 (Matrix monotonicity; Loewner 1934). A function $f : \mathsf{I} \to \mathbb{R}$ is *matrix monotone on* I if

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \quad \text{implies} \quad f(\boldsymbol{A}) \preccurlyeq f(\boldsymbol{B})$$

for every $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathsf{I})$ and every $n \in \mathbb{N}$.

The action of f on a self-adjoint matrix is interpreted in the sense of standard matrix functions.

Definition 14.2 (Matrix convexity; Kraus 1936). A function $f : \mathsf{I} \to \mathbb{R}$ is *matrix convex on* I if

$$f(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{B}) \preccurlyeq \tau f(\boldsymbol{A}) + \bar{\tau} f(\boldsymbol{B}) \text{ for all } \tau \in [0, 1]\,,$$

where $\bar{\tau} := 1 - \tau$, for every $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathsf{I})$ and every $n \in \mathbb{N}$.

Today, we will primarily focus on analyzing matrix monotonicity.

14.2 Differential characterizations

Many properties of a differentiable, scalar function may be deduced directly from its derivative. Motivated by this observation, we first recap scalar derivative tests for monotonicity and convexity. These tests are generalized to the matrix function setting, through the machinery of divided differences, in the main results of the lecture: Loewner's theorem and the theorem of Aujla & Vasudeva.

14.2.1 Scalar case

We begin our study of monotonicity and convexity by drawing intuition from the scalar setting. Recall that, if $f : \mathsf{I} \to \mathbb{R}$ is differentiable, then

1. f is *increasing on* I if and only if $f'(t) \geq 0$ for all $t \in \mathsf{I}$,
2. f is *convex on* I if and only if $f(t) - f(s) \geq f'(s)(t - s)$ for all $t, s \in \mathsf{I}$.

These conditions are useful in practice because they can be easily verified, provided the first derivative exists. However, it turns out that the first (scalar) derivative does not carry enough information to characterize monotonicity and convexity when univariate functions are lifted to the space of self-adjoint matrices. It is then natural to ask:

Are there analogous differential characterizations in the matrix setting?

The answer is both remarkable and beautiful.

14.2.2 Divided differences

To answer this question, we turn to an object initiated by Newton that arises frequently in numerical analysis and approximation theory.

Definition 14.3 (Divided difference). Let $f : \mathsf{I} \to \mathbb{R}$ be continuously differentiable on the open interval $\mathsf{I} \subseteq \mathbb{R}$. Then the *first divided difference* $f^{[1]} : \mathsf{I} \times \mathsf{I} \to \mathbb{R}$ is the bivariate function

$$f^{[1]}(s, t) := \begin{cases} \dfrac{f(s) - f(t)}{s - t}, & s \neq t, \\ f'(t), & s = t. \end{cases}$$

Intuitively, the first divided difference records the slopes of secants and tangents to f. Tabulating the bivariate function $f^{[1]}$ at a set of points produces *Loewner's matrix*.

Definition 14.4 (Loewner Matrix). Let $f : \mathsf{I} \to \mathbb{R}$ be continuously differentiable on the open interval $\mathsf{I} \subseteq \mathbb{R}$. For $n \in \mathbb{N}$, let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{H}_n(\mathsf{I})$ be any diagonal matrix with entries in I. Then the *Loewner matrix* of f at $\mathbf{\Lambda}$ is

$$\boldsymbol{L}_f(\mathbf{\Lambda}) := f^{[1]}(\mathbf{\Lambda}) := \left[f^{[1]}(\lambda_i, \lambda_j) \right]_{i,j=1\dots,n} \in \mathbb{H}_n.$$

This is essentially the "kernel matrix" of the function $f^{[1]}$ on the dataset of points $(\lambda_k : k = 1, \dots, n)$. This observation leads to connections with the theory of *positive-definite functions*, discussed in Lecture 18.

The first divided difference and Loewner's matrix play a key role in generalizing scalar differential characterizations of monotonicity and convexity to the matrix setting.

14.2.3 Matrix setting

We are now ready to present the main results of this lecture.

Theorem 14.5 (Loewner 1934). Assume that $f : \mathsf{I} \to \mathbb{R}$ is continuously differentiable on an open interval $\mathsf{I} \subseteq \mathbb{R}$. Then

$$f \text{ is matrix monotone on } \mathsf{I} \quad \text{if and only if} \quad f^{[1]}(\mathbf{\Lambda}) \succcurlyeq \mathbf{0}$$

for every diagonal $\mathbf{\Lambda} \in \mathbb{H}_n(\mathsf{I})$ and for every $n \in \mathbb{N}$.

We see that $\boldsymbol{L}_f$ in the matrix setting plays the same role as does f' in the scalar setting. We will prove Loewner's theorem later in Section 14.4.

There is an analogous result [AV95] for matrix convexity, albeit established much later than Loewner's result.

Theorem 14.6 (Aujla–Vasudeva 1995). Assume that $f : \mathsf{I} \to \mathbb{R}$ is continuously differentiable on an open interval $\mathsf{I} \subseteq \mathbb{R}$. Then

$$f \text{ is matrix convex on } \mathsf{I} \quad \text{if and only if} \quad f(\boldsymbol{B}) - f(\boldsymbol{A}) \succcurlyeq f^{[1]}(\boldsymbol{A}) \odot (\boldsymbol{B} - \boldsymbol{A})$$

for every $\boldsymbol{B} \in \mathbb{H}_n(\mathsf{I})$, every diagonal $\boldsymbol{A} \in \mathbb{H}_n(\mathsf{I})$, and every $n \in \mathbb{N}$.

Recall that "$\odot$" is the *Schur* or *Hadamard* entrywise product with respect to the standard basis.

This theorem strongly parallels the differential characterization of scalar convexity.

Exercise 14.7 (Convexity: Differential characterization). Prove Theorem 14.6. **Hint:** Imitate the proof of Theorem 14.5.

We conclude by stating another result that relates matrix convexity to matrix monotonicity. This theorem plays an important role in more thorough treatments of the Loewner theory.

Theorem 14.8 (Bendat–Sherman). Assume that $f : \mathsf{I} \to \mathbb{R}$ is twice continuously differentiable on the open interval $\mathsf{I} \subseteq \mathbb{R}$ and matrix convex. Then for every $\mu \in \mathsf{I}$, the function $g_\mu : \mathsf{I} \to \mathbb{R}$ defined by

$$\lambda \mapsto g_\mu(\lambda) := f^{[1]}(\mu, \lambda)$$

is matrix monotone.

The proof may be found in [Bha97, Theorem V.3.10].

14.3 Derivatives of standard matrix functions

Standard matrix functions are obtained by lifting a scalar function to matrices. It is important to understand the differentiability properties of the standard matrix function. To do this, we first define the Fréchet and Gâteaux derivative of maps between spaces of matrices. The Daleckii–Krein formula (Theorem 14.13) instantiates these derivatives explicitly for continuously differentiable standard matrix functions, and it will be used in the proof of Loewner's theorem in the next section.

14.3.1 Fréchet derivatives

To prove Theorem 14.5, we need to precisely describe what it means to differentiate functions on spaces of matrices, and more generally, vector-valued maps. The key idea is that the derivative is a *linear approximation*.

Definition 14.9 (Fréchet derivative). Let I be an open interval of $\mathbb{R}$, and let $\boldsymbol{F} : \mathbb{H}_n(\mathsf{I}) \to \mathbb{H}_n$ be a matrix-valued function on self-adjoint matrices. The *Fréchet derivative* of $\boldsymbol{F}$ at a point $\boldsymbol{A} \in \mathbb{H}_n(\mathsf{I})$ is the linear map

$$\mathrm{D}\boldsymbol{F}(\boldsymbol{A}) : \mathbb{H}_n \to \mathbb{H}_n$$

defined by the property

$$\frac{\|\boldsymbol{F}(\boldsymbol{A} + \boldsymbol{H}) - \boldsymbol{F}(\boldsymbol{A}) - \mathrm{D}\boldsymbol{F}(\boldsymbol{A})\boldsymbol{H}\|}{\|\boldsymbol{H}\|} \to 0$$

as $\boldsymbol{H} \to \boldsymbol{0}$ in any norm on $\mathbb{H}_n$, provided that such an object $\mathrm{D}\boldsymbol{F}(\boldsymbol{A})$ exists.

Aside: Although specialized to our our self-adjoint matrix setting, the definition of Fréchet derivative easily extends to general infinite-dimensional Banach spaces [CP77].

Note that although the Fréchet derivative of $\boldsymbol{F}$ at a point $\boldsymbol{A}$ is linear, the mapping

$$\mathrm{D}\boldsymbol{F}\colon \boldsymbol{A} \mapsto \mathrm{D}\boldsymbol{F}(\boldsymbol{A})$$

is nonlinear in general. This parallels the usual notion of scalar derivative from calculus. Indeed, for $f : \mathsf{I} \to \mathbb{R}$ continuously differentiable, f' is a nonlinear function but the Fréchet derivative

$$\mathrm{D}f(t)\colon h \mapsto f'(t)h$$

is just the linear operator of scalar multiplication by $f'(t)$, for $t \in \mathsf{I}$, which can be identified with $f'(t)$ itself.

Exercise 14.10 (Derivative). Show that, if it exists, the Fréchet derivative is unique.

Exercise 14.11 (Linear map). Show that if $\boldsymbol{F}\colon \boldsymbol{A} \mapsto \boldsymbol{\Phi}\boldsymbol{A}$ is linear, that is, $\boldsymbol{\Phi}\colon \mathbb{H}_n \to \mathbb{H}_n$ is a linear map, then $\mathrm{D}\boldsymbol{F}(\boldsymbol{A}) = \boldsymbol{\Phi}$ for every $\boldsymbol{A} \in \mathbb{H}_n$.

It is useful to relate the Fréchet derivative to another kind of derivative, the *Gâteaux derivative*, that generalizes directional derivatives to Banach space. Indeed, *if it exists*, the Fréchet derivative of $\boldsymbol{F}$ at $\boldsymbol{A}$ parametrizes all the Gâteaux derivatives of $\boldsymbol{F}$ at $\boldsymbol{A}$ in the sense that

$$\mathrm{D}\boldsymbol{F}(\boldsymbol{A})\boldsymbol{H} = \frac{\mathrm{d}}{\mathrm{d}t}\big[\boldsymbol{F}(\boldsymbol{A} + t\boldsymbol{H})\big]\Big|_{t=0} \tag{14.1}$$

for every $\boldsymbol{H}$. The right hand side is called the *Gâteaux derivative* of $\boldsymbol{F}$ at $\boldsymbol{A}$ in the direction $\boldsymbol{H}$.

Aside: If $\boldsymbol{F}$ is real-valued, this is usually called the *first variation* or *variational derivative* in the calculus of variations.

Warning 14.12 (Directional derivatives). The converse is false. A map can be Gâteaux differentiable in every direction without being Fréchet differentiable. This phenomenon already arises for functions on $\mathbb{R}^2$. ▪

14.3.2 Daleckii–Krein formula

The next result gives a concrete characterization of the Fréchet derivative of a standard matrix function.

Theorem 14.13 (Daleckii–Krein). Assume that $f : \mathsf{I} \to \mathbb{R}$ is continuously differentiable on the open interval $\mathsf{I} \subseteq \mathbb{R}$. For diagonal $\boldsymbol{\Lambda} \in \mathbb{H}_n(\mathsf{I})$,

$$\mathrm{D}f(\boldsymbol{\Lambda})\boldsymbol{H} = f^{[1]}(\boldsymbol{\Lambda}) \odot \boldsymbol{H} \quad \text{for every} \quad \boldsymbol{H} \in \mathbb{H}_n\,. \tag{14.2}$$

Moreover, for $\boldsymbol{A} \in \mathbb{H}_n(\mathsf{I})$ with spectral decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$,

$$\mathrm{D}f(\boldsymbol{A})\boldsymbol{H} = f^{[1]}(\boldsymbol{A}) \odot_{\boldsymbol{A}} \boldsymbol{H} =: \boldsymbol{U}\big[f^{[1]}(\boldsymbol{\Lambda}) \odot (\boldsymbol{U}^*\boldsymbol{H}\boldsymbol{U})\big]\boldsymbol{U}^*\,. \tag{14.3}$$

The following lemma gives the derivative of a polynomial standard matrix function and is used in a limiting argument in the proof of the Daleckii–Krein formula.

Lemma 14.14 (Polynomial: Derivative). Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial. Then for $\boldsymbol{H} \in \mathbb{H}_n$ and diagonal $\boldsymbol{\Lambda} \in \mathbb{H}_n$,

$$\mathrm{D}f(\boldsymbol{\Lambda})\boldsymbol{H} = f^{[1]}(\boldsymbol{\Lambda}) \odot \boldsymbol{H}\,. \tag{14.4}$$

Proof. Notice that both sides of (14.4) are linear in f since differentiation and point evaluation are linear operations. Hence, without loss of generality, we may consider the reduction to the monomial $f : x \mapsto x^p$ for each $p \in \mathbb{Z}_+$.

To this end, recall the algebraic identity

$$\boldsymbol{B}^p - \boldsymbol{C}^p = \sum\nolimits_{k=0}^{p-1} \boldsymbol{B}^k(\boldsymbol{B} - \boldsymbol{C})\boldsymbol{C}^{p-1-k}$$

that holds for every $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{M}_n$. Using this, for $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n$ we compute

$$\begin{aligned}\frac{(\boldsymbol{A}+t\boldsymbol{H})^p - \boldsymbol{A}^p}{t} &= \frac{1}{t}\sum\nolimits_{k=0}^{p-1}(\boldsymbol{A}+t\boldsymbol{H})^k(t\boldsymbol{H})\boldsymbol{A}^{p-1-k} \\ &\to \sum\nolimits_{k=0}^{p-1}\boldsymbol{A}^k\boldsymbol{H}\boldsymbol{A}^{p-1-k}\end{aligned}$$

as $t \downarrow 0$. The right hand side of the display is the Gâteux derivative of f at $\boldsymbol{A}$ in the direction $\boldsymbol{H}$, and it is possible to verify by Taylor's theorem in Banach space [CP77] that the Fréchet derivative of f exists and hence must agree with the Gâteux derivative as in (14.1). In particular, this limit is zero when $p = 0$ (the constant function) since the left-hand side in the above display is identically equal to the zero matrix in this case.

Let $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1 \dots, \lambda_n)$. As $f^{[1]}(\boldsymbol{\Lambda}) = \mathbf{0}$ for $p = 0$, we have already verified (14.4) for $p = 0$. Now let $p \in \mathbb{N}$. Since the diagonal entries of $\boldsymbol{\Lambda}$ may repeat, we must consider two cases, depending on whether the entries are the same or different.

For the first case, consider pairs $i, j \in \{1, \dots, n\}$ such that $\lambda_i \neq \lambda_j$. We compute

$$\begin{aligned}\big[\mathrm{D}f(\boldsymbol{\Lambda})\boldsymbol{H}\big]_{i,j} &= \Big[\sum\nolimits_{k=0}^{p-1}\boldsymbol{\Lambda}^k\boldsymbol{H}\boldsymbol{\Lambda}^{p-1-k}\Big]_{i,j} \\ &= \langle \boldsymbol{\delta}_i,\ \sum\nolimits_{k=0}^{p-1}\boldsymbol{\Lambda}^k\boldsymbol{H}\boldsymbol{\Lambda}^{p-1-k}\boldsymbol{\delta}_j\rangle \\ &= \sum\nolimits_{k=0}^{p-1}\langle \boldsymbol{\Lambda}^k\boldsymbol{\delta}_i,\ \boldsymbol{H}(\boldsymbol{\Lambda}^{p-1-k}\boldsymbol{\delta}_j)\rangle \\ &= \sum\nolimits_{k=0}^{p-1}\lambda_i^k h_{ij}\lambda_j^{p-1-k} \\ &= \Big(\sum\nolimits_{k=0}^{p-1}\lambda_i^k\lambda_j^{p-1-k}\Big)h_{ij} \\ &= \left(\frac{\lambda_i^p - \lambda_j^p}{\lambda_i - \lambda_j}\right)h_{ij}\,.\end{aligned}$$

The last equality follows from an algebraic identity and uses the convention $0^0 = 1$. Notice that since $\lambda_i \neq \lambda_j$,

$$\frac{\lambda_i^p - \lambda_j^p}{\lambda_i - \lambda_j} = \big[f^{[1]}(\boldsymbol{\Lambda})\big]_{ij}.$$

We have attained the desired result.

For the second case, consider pairs $i, j \in \{1, \dots, n\}$ such that $\lambda_i = \lambda_j$. In particular, this includes the diagonal $i = j$. If $\lambda_i = \lambda_j = 0$, then $f'(\lambda_j) = f'(0) = \mathbb{1}_{\{p=1\}}$. Using this observation and continuing from the calculation above,

$$\begin{aligned}\big[\mathrm{D}f(\boldsymbol{\Lambda})\boldsymbol{H}\big]_{i,j} &= \sum\nolimits_{k=0}^{p-1}\lambda_j^k\lambda_j^{p-1-k}h_{ij} \\ &= \sum\nolimits_{k=0}^{p-1}\lambda_j^{p-1}h_{ij} \\ &= p\lambda_j^{p-1}h_{ij} \\ &= f'(\lambda_j)h_{ij} \\ &= \big[f^{[1]}(\boldsymbol{\Lambda})\big]_{ij}h_{ij}\end{aligned}$$

as asserted. ∎

Exercise 14.15 (Power: Divided differences). Verify the algebraic identity

$$\sum\nolimits_{k=0}^{p-1}a^k b^{p-1-k} = \frac{a^p - b^p}{a - b}$$

for $a \neq b \in \mathbb{R}$ and $p \in \mathbb{N}$, with the convention that $0^0 = 1$.

We can now prove Theorem 14.13; a full treatment may be found in [Sim19, Chapter 5] or in [Bha97, Chapter V].

Proof sketch: Daleckii–Krein Formula. The argument proceeds by approximation. We focus on the case where the eigenvalues of diagonal $\mathbf{\Lambda} \in \mathbb{H}_n(\mathsf{I})$ are distinct. By an extension of the Stone–Weierstrass theorem [Sim19, Proposition 5.4], there exists a sequence of polynomials $(f_n : n \in \mathbb{N})$ that converge to f uniformly on compact subintervals of I. Furthermore, the sequence $(f_n' : n \in \mathbb{N})$ of derivatives converges to f' uniformly on compact subintervals of I. It follows that

$$f_n^{[1]}(\lambda_i, \lambda_j) \to f^{[1]}(\lambda_i, \lambda_j) \quad \text{as } n \to \infty$$

for each i, j. By Lemma 14.14, for each $\boldsymbol{H} \in \mathbb{H}_n$,

$$\big[\mathrm{D} f_n(\mathbf{\Lambda})\boldsymbol{H}\big]_{ij} = f_n^{[1]}(\lambda_i, \lambda_j)[\boldsymbol{H}]_{ij} \to f^{[1]}(\lambda_i, \lambda_j)[\boldsymbol{H}]_{ij} \text{ as } n \to \infty$$

for each i, j. Since the convergence holds entrywise, it also holds in any UI matrix norm by a characteristic polynomial argument. By [Sim19, Lemma 5.5(a)], the Gâteaux derivative of f exists and equals the limit of the left-hand side above, which yields (14.2) if it can be shown that the Fréchet derivative exists. This step is found in the proof of [Bha97, Theorem V.3.3].

Equation (14.3) follows from *unitary equivariance* of standard matrix functions. More explicitly, it is easy to verify using the definition of polynomial standard matrix functions and the calculation in the proof of Lemma 14.14 that

$$\mathrm{D} f_n(\boldsymbol{A})\boldsymbol{H} = \boldsymbol{U}\big[f_n^{[1]}(\mathbf{\Lambda}) \odot (\boldsymbol{U}^*\boldsymbol{H}\boldsymbol{U})\big]\boldsymbol{U}^* .$$

A similar limiting argument as above concludes the proof sketch. ∎

14.4 Proof of Loewner's theorem

We may now prove Theorem 14.5. Under the hypotheses, let us first assume that f is matrix monotone on I. We need to show that $f^{[1]}(\mathbf{\Lambda}) \succcurlyeq \mathbf{0}$ for every diagonal $\mathbf{\Lambda} \in \mathbb{H}_n(\mathsf{I})$, where $n \in \mathbb{N}$. To this end, consider a specific perturbation matrix $\boldsymbol{H} := \mathbf{1}\mathbf{1}^* \succcurlyeq \mathbf{0}$. Then

$$\mathbf{\Lambda} + t\boldsymbol{H} \succcurlyeq \mathbf{\Lambda} \quad \text{for } t \geq 0$$

by definition of the psd order. Since I is open and the eigenvalue map $\boldsymbol{\lambda}(\cdot)$ is continuous, there exists $\delta > 0$ such that $\mathbf{\Lambda} + t\boldsymbol{H} \in \mathbb{H}_n(\mathsf{I})$ for all $0 \leq t < \delta$. By matrix monotonicity of f on I,

$$f(\mathbf{\Lambda} + t\boldsymbol{H}) \succcurlyeq f(\mathbf{\Lambda}) \quad \text{when} \quad 0 \leq t < \delta .$$

Since the psd cone is closed, we further deduce

$$\mathrm{D} f(\mathbf{\Lambda})\boldsymbol{H} = \lim_{t \downarrow 0} \frac{f(\mathbf{\Lambda} + t\boldsymbol{H}) - f(\mathbf{\Lambda})}{t} \succcurlyeq \mathbf{0} ,$$

where we are allowed to equate the Fréchet and Gâteaux derivatives of f above because existence of the Fréchet derivative is guaranteed by Theorem 14.13. Again by the Daleckii–Krein formula,

$$\mathbf{0} \preccurlyeq \mathrm{D} f(\mathbf{\Lambda})\boldsymbol{H} = f^{[1]}(\mathbf{\Lambda}) \odot \boldsymbol{H} = f^{[1]}(\mathbf{\Lambda}) .$$

The forward implication is valid.

For the reverse implication, we assume $f^{[1]}(\mathbf{\Lambda}) \succcurlyeq \mathbf{0}$ for every diagonal $\mathbf{\Lambda} \in \mathbb{H}_n(\mathsf{I})$. Let $\boldsymbol{A}_0, \boldsymbol{A}_1 \in \mathbb{H}_n(\mathsf{I})$ be arbitrary and satisfy $\boldsymbol{A}_1 \succcurlyeq \boldsymbol{A}_0$. We need to show that $f(\boldsymbol{A}_1) \succcurlyeq f(\boldsymbol{A}_0)$. The argument proceeds by interpolation. Introduce

$$\boldsymbol{A}_\tau := (1-\tau)\boldsymbol{A}_0 + \tau \boldsymbol{A}_1 \in \mathbb{H}_n(\mathsf{I}) \quad \text{for } \tau \in [0,1].$$

The set $\mathbb{H}_n(\mathsf{I})$ is convex by the Rayleigh–Ritz representation of eigenvalues.

By assumption, the τ-derivative satisfies

$$\dot{\boldsymbol{A}}_\tau := \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{A}_\tau = \boldsymbol{A}_1 - \boldsymbol{A}_0 \succcurlyeq \mathbf{0}\,.$$

Write the eigenvalue decompositions of the parameterized matrix as $\boldsymbol{A}_\tau = \boldsymbol{U}_\tau \mathbf{\Lambda}_\tau \boldsymbol{U}_\tau^*$. We use the fundamental theorem of calculus to obtain

$$\begin{aligned} f(\boldsymbol{A}_1) - f(\boldsymbol{A}_0) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau}\big[f(\boldsymbol{A}_\tau)\big]\,\mathrm{d}\tau \\ &= \int_0^1 \mathrm{D}f(\boldsymbol{A}_\tau)\dot{\boldsymbol{A}}_\tau\,\mathrm{d}\tau \\ &= \int_0^1 \boldsymbol{U}_\tau\big[f^{[1]}(\mathbf{\Lambda}_\tau) \odot (\boldsymbol{U}_\tau^*\dot{\boldsymbol{A}}_\tau\boldsymbol{U}_\tau)\big]\boldsymbol{U}_\tau^*\,\mathrm{d}\tau \succcurlyeq \mathbf{0}. \end{aligned}$$

The second equality is the chain rule; the third equality is from the Daleckii–Krein formula (14.3); and the last inequality follows from the Schur product theorem. Indeed, by hypothesis, $f^{[1]}(\mathbf{\Lambda}_\tau) \succcurlyeq \mathbf{0}$, while $\boldsymbol{U}_\tau^*\dot{\boldsymbol{A}}_\tau\boldsymbol{U}_\tau \succcurlyeq \mathbf{0}$ by the conjugation rule. Therefore, $f^{[1]}(\mathbf{\Lambda}_\tau) \odot (\boldsymbol{U}_\tau^*\dot{\boldsymbol{A}}_\tau\boldsymbol{U}_\tau) \succcurlyeq \mathbf{0}$. This is what we needed to show.

14.5 Examples

We conclude with two illustrative examples where Loewner's theorem may be applied.

Example 14.16 (Rational function). For $\mathsf{I} \subseteq \mathbb{R}$ and real coefficients $a, b, c, d \in \mathbb{R}$ such that $ad - bc > 0$ and $-d/c \notin \mathsf{I}$, define the rational function $f : \mathsf{I} \to \mathbb{R}$ by

$$t \mapsto f(t) := \frac{at+b}{ct+d}\,.$$

Aside: Such rational functions arise, for example, in the aptly named Loewner framework for interpolatory model reduction of dynamical systems [Ben+17, Chapter 8].

Notice that f is continuous on I, and

$$t \mapsto f'(t) = \frac{ad-bc}{(ct+d)^2}$$

is also continuous on I. Hence Loewner's result (Theorem 14.5) applies. For $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and $n \in \mathbb{N}$, the Loewner matrix of f is

$$f^{[1]}(\mathbf{\Lambda}) = \left[\frac{ad-bc}{(c\lambda_i + d)(c\lambda_j + d)}\right]_{i,j=1,\dots,n}.$$

This formula agrees with $f'(\lambda_j)$ when $\lambda_i = \lambda_j$. Introduce the matrix

$$\boldsymbol{S} := \sqrt{ad-bc}\,\operatorname{diag}\Big((c\lambda_k + d)^{-1}\Big)_{k=1,\dots,n}.$$

Observe that

$$f^{[1]}(\mathbf{\Lambda}) = \boldsymbol{S}^*(\mathbf{1}\mathbf{1}^*)\boldsymbol{S} = (\mathbf{1}^*\boldsymbol{S})^*(\mathbf{1}^*\boldsymbol{S}) \succcurlyeq \mathbf{0}\,.$$

We may conclude that the rational function f is matrix monotone on I. ∎

Using Loewner's theorem, by verifying that the Loewner matrix is psd we were able to conclude matrix monotonicity (in this case, of rational functions). The reverse direction is also of interest, where monotonicity can be informative about positive semi-definiteness.

This is useful in the study of kernel methods and positive definite functions [Bha07b, Chapter 5].

Example 14.17 (Logarithm). Since $t \mapsto \log t$ is continuously differentiable and matrix monotone on $(0, \infty)$, it follows by Theorem 14.5 that

$$\left[\frac{\mathbb{1}_{\{\lambda_i=\lambda_j\}} + \log\lambda_i - \log\lambda_j}{\lambda_i \mathbb{1}_{\{\lambda_i=\lambda_j\}} + \lambda_i - \lambda_j}\right]_{i,j=1,\dots,n} \succcurlyeq \mathbf{0}$$

for any $\{\lambda_1, \dots, \lambda_n\} \subset (0, \infty)$ and $n \in \mathbb{N}$. These matrices arises in the study of logarithmic means, for example. ▪

Notes

This lecture is adapted from Bhatia's book [Bha97, Chap. V] with some elements from [Bha07b, Chap. 5].

Lecture bibliography

[AV95] J. S. Aujla and H. Vasudeva. "Convex and monotone operator functions". In: *Annales Polonici Mathematici*. Volume 62. 1. 1995, pages 1–11.

[Ben+17] P. Benner et al. *Model reduction and approximation: theory and algorithms*. SIAM, 2017.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[CP77] "CHAPTER 6. Calculus in Banach Spaces". In: *Functional Analysis in Modern Applied Mathematics*. Volume 132. Mathematics in Science and Engineering. Elsevier, 1977, pages 87–105. DOI: https://doi.org/10.1016/S0076-5392(08)61248-5.

[Sim19] B. Simon. *Loewner's theorem on monotone matrix functions*. Springer, Cham, 2019. DOI: 10.1007/978-3-030-22422-6.

15. Monotonicity: Integral Characterization

Date: 22 February 2022 **Scribe:** Joel A. Tropp

In the last lecture, we showed that smooth matrix monotone functions are characterized by a differential property. The Loewner matrix, which packages divided differences of the function, must be a psd matrix. In this lecture, we will discuss a complementary characterization of matrix monotone functions as those that can be expressed using a certain integral representation. This approach allows us to identify several new examples of matrix monotone functions. Moreover, the description yields some general theorems on matrix convexity, including some matrix Lyapunov inequalities that go well beyond the Kadison and Choi inequalities.

Agenda:

1. Integral representation of matrix monotone functions
2. Uniqueness
3. Geometric approach
4. Monotonicity on the positive reals
5. Filtering with a positive linear map
6. Integral representation of matrix convex functions

15.1 Recap

Last time, we developed characterizations of continuously differentiable matrix monotone and matrix convex functions in terms of derivative properties. Let $\mathsf{I} \subset \mathbb{R}$ be an open interval. Each continuously differentiable function $f : \mathsf{I} \to \mathbb{R}$ induces a family of Loewner matrices. For a diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{H}_n(\mathsf{I})$ and each $n \in \mathbb{N}$, we define

$$\boldsymbol{L}_f(\boldsymbol{\Lambda}) := f^{[1]}(\boldsymbol{\Lambda}) := \left[f^{[1]}(\lambda_j, \lambda_k)\right]_{j,k=1,\dots,n}.$$

As usual, $f^{[1]}(a, b)$ denotes the divided difference of f at the points $a, b \in \mathsf{I}$.

In the theory of matrix monotonicity and convexity, the Loewner matrix plays a role similar to the derivative in the scalar setting. Indeed, under the smoothness assumption on f, we have presented two theorems.

1. **Monotonicity.** The function f is matrix monotone on I if and only if
$$\boldsymbol{L}_f(\boldsymbol{A}) \succcurlyeq \mathbf{0} \quad \text{for each diagonal } \boldsymbol{A} \in \mathbb{H}_n(\mathsf{I}).$$
We require this condition to hold for each $n \in \mathbb{N}$. The $n = 1$ case simply states that f' is positive on I. This result is due Loewner.
2. **Convexity.** The function f is matrix convex on I if and only if
$$\boldsymbol{L}_f(\boldsymbol{A}) \odot (\boldsymbol{B} - \boldsymbol{A}) \succcurlyeq \boldsymbol{B} - \boldsymbol{A} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n(\mathbf{I}) \text{ with } \boldsymbol{A} \text{ diagonal.}$$
As before, this condition is required for each $n \in \mathbb{N}$. The $n = 1$ case corresponds with the characterization of convexity for a smooth, scalar function. This result is due to Aujla & Vasudeva.

Later in this lecture, we will also discuss some other relationships between matrix monotone and matrix convex functions that do not have scalar counterparts.

15.2 Integral representations of matrix monotone functions

In Lecture 13, we established that the logarithm is a matrix monotone function on $\mathbb{R}_{++} := (0, \infty)$ by the following argument. First, we observed that

$$\log a = \int_0^\infty \left[(1 + \lambda)^{-1} - (a + \lambda)^{-1}\right] \mathrm{d}\lambda \quad \text{for } a > 0.$$

We have a similar integral representation for the logarithm of a pd matrix:

$$\log \boldsymbol{A} = \int_0^\infty \left[(1+\lambda)^{-1}\mathbf{I} - (\boldsymbol{A} + \lambda\mathbf{I})^{-1} \right] \mathrm{d}\lambda \quad \text{for } \boldsymbol{A} > \mathbf{0}.$$

Since the function $t \mapsto -(t+\lambda)^{-1}$ is matrix monotone on $\mathbb{R}_{++}$, the integral representation shows that the logarithm is the limit of positive sums of matrix monotone functions. Since the cone of matrix monotone functions is closed under pointwise limits, the logarithm also is matrix monotone.

This argument may seem concocted. In fact, this approach is based on a penetrating insight that *every* matrix monotone function admits an integral representation. This is a famous result of Loewner, which is regarded as one of the deepest theorems in matrix analysis. We present the result and some immediate consequences.

15.2.1 Loewner's integral theorem

The form of the integral representation of a matrix monotone function depends superficially on the interval where it is defined. We begin with the standard open interval $(-1, +1)$. Later, we will see that results for this interval can be transferred directly to other intervals.

Warning: In this theory, we must take care about whether the endpoints of the interval are included because matrix monotone functions need not be continuous at the endpoints. This result is framed for an open interval to avoid this problem. ∎

Theorem 15.1 (Matrix monotonicity: Standard interval; Loewner 1934). Let $f : (-1, +1) \to \mathbb{R}$ be a *nonconstant* function.

1. If f is matrix monotone on $(-1, +1)$, then f admits a *unique* representation of the form
$$f(t) = \alpha + \beta \int_{[-1,+1]} \frac{t}{1 - \lambda t} \, \mathrm{d}\mu(\lambda), \tag{15.1}$$
where $\alpha \in \mathbb{R}$ and $\beta > 0$ and μ is a Borel probability measure on $[-1, +1]$.
2. Conversely, suppose that f satisfies (15.1). Then f is a matrix monotone function on $(-1, +1)$.

The claim (2) is called the "easy" direction of Loewner's theorem. We verify this result in Section 15.2.2. As we will see, the easy direction already offers a powerful tool. Section 15.2.3 contains several examples.

Meanwhile, the claim (1) is called the "hard" direction of Loewner's theorem. In particular, this result implies that every matrix monotone function f on $(-1, +1)$ is *analytic* in the interval $(-1, +1)$. This fact is striking because scalar monotone functions do not even need to be continuous. Introducing deeper ideas from complex analysis, one may prove that the function f continues analytically into the upper half-plane, where it defines a Herglotz function; see [Bha97, Chap. V] or [Sim19] for more discussion about this point.

Simon [Sim19] outlines 11 different proofs (!) of the hard part of Loewner's theorem. Each one approaches the summit from a different direction, and each one requires a long and arduous climb. In Section 15.2.4, we give the easy proof of the uniqueness statement. The existence claim is the difficult step. In Section 15.3, we will summarize a geometric existence proof, but there is no space here for all of the details.

15.2.2 Proof of Theorem 15.1: "Easy" direction

Let us supply the short proof of claim (2) in Loewner's theorem.

Exercise 15.2 (Matrix monotone functions: Elementary examples). For each $\lambda \in [-1, +1]$, confirm that $f : t \mapsto t/(1 - \lambda t)$ is a matrix monotone function on $(-1, +1)$. **Hint:**

Simplify the function algebraically so that you can invoke the fact that the negative inverse is operator monotone.

The constant function $t \mapsto \alpha$ is matrix monotone on any interval. For each $\lambda \in [-1, +1]$, the function $t \mapsto t/(1 - \lambda t)$ is matrix monotone on the interval $(-1, +1)$ by Exercise 15.2. By the usual limiting arguments (simple functions, dominated convergence), we see that the integral in (15.1) represents a matrix monotone function. Indeed, matrix monotone functions are closed under positive linear combinations and pointwise limits. Altogether, the expression (15.1) represents a matrix monotone function on the standard open interval $(-1, +1)$.

15.2.3 First examples

Once we recognize that matrix monotone functions admit integral representations of the type (15.1), we can seek out representations for particular functions. The "easy" direction of Loewner's theorem confirms that these integrals describe matrix monotone functions.

Example 15.3 (Logarithms and friends). Using basic tools from integral calculus, we quickly confirm that

$$\log(1 + t) = \int_{-1}^{0} \frac{t}{1 - \lambda t} \, \mathrm{d}\lambda \quad \text{for } t \in (-1, +1).$$

Theorem 15.1(2) now implies that $t \mapsto \log(1 + t)$ is matrix monotone on $(-1, +1)$. A similar calculation reveals that $t \mapsto -\log(1 - t)$ is matrix monotone on $(-1, +1)$. From here, we can derive further results of interest:

$$\operatorname{arctanh}(t) = \frac{1}{2} \log\left(\frac{1 + t}{1 - t}\right) \quad \text{is matrix monotone on } (-1, +1).$$

Indeed, matrix monotone functions compose a convex cone. ∎

Problem 15.4 (Tangent). Show that the function $t \mapsto \tan(\pi t/2)$ has an integral representation of the form (15.1), hence is matrix monotone on $(-1, +1)$. **Hint:** You can derive this fact from the definite integral

$$\frac{\pi}{2} \tan\left(\frac{\pi t}{2}\right) = \int_0^{\infty} \frac{\lambda^t - 1}{\lambda - \lambda^{-1}} \cdot \frac{\mathrm{d}\lambda}{\lambda} \quad \text{for } t \in (-1, +1).$$

This non-obvious statement appears as [GR07, Formula 3.274(3)]. There are more elementary ways to prove that the tangent is matrix monotone; cf. Lecture 18.

15.2.4 Proof of Theorem 15.1: Uniqueness

To prove that integral representations are unique, we must argue that the integrands are rich enough to approximate all continuous functions.

Exercise 15.5 (Elementary monotone functions: Totality). Establish that the collection

$$\{\lambda \mapsto 1\} \cup \{\lambda \mapsto t(1 - \lambda t)^{-1} : t \in (-1, +1)\}$$

is total in the space $\mathsf{C}[-1, +1]$ of continuous functions, equipped with the supremum norm. **Hint:** This is a direct application of the Stone–Weierstrass theorem because the linear span is an algebra.

In a normed space, a subset is *total* if its linear span is dense.

To prove the uniqueness claim, let f be a matrix monotone function on $(-1, +1)$ with a representation of the form (15.1). Since $\alpha = f(0)$ and $\beta = f'(0) > 0$, we can shift and scale f to make $\alpha = 0$ and $\beta = 1$.

Suppose that there are two Borel probability measures μ, ν on $[-1, +1]$ for which

$$f(t) = \int_{-1}^{+1} \frac{t}{1-\lambda t} \, \mathrm{d}\mu(t) = \int_{-1}^{+1} \frac{t}{1-\lambda t} \, \mathrm{d}\nu(t) \quad \text{for all } t \in (-1, +1).$$

If these two measures are different, they can be separated by a continuous function $h : [-1, +1] \to \mathbb{R}$. That is,

$$\int_{-1}^{+1} h(\lambda) \, \mathrm{d}\mu(\lambda) < \int_{-1}^{+1} h(\lambda) \, \mathrm{d}\nu(\lambda).$$

Since μ and ν are both probability measures, we may shift h to ensure that it has zero integral (with respect to Lebesgue measure): $\int_{-1}^{+1} h(\lambda) \, \mathrm{d}\lambda = 0$.

By Exercise 15.5, the functions $\lambda \mapsto t(1-\lambda t)^{-1}$ are total in the space of continuous functions on $[-1, +1]$ with a zero integral. Therefore, we can approximate h in the supremum norm by a linear combination. For example, with some real coefficients $c_1, \dots, c_k$, we have

$$\left| h(\lambda) - \sum\nolimits_{k=1}^{n} \frac{c_k t_k}{1-\lambda t_k} \right| \leq \varepsilon \quad \text{for all } \lambda \in [-1, +1].$$

By choosing ε sufficiently small, we find that

$$\begin{aligned} \sum\nolimits_{k=1}^{n} c_k f(t_k) &= \sum\nolimits_{k=1}^{n} \int_{-1}^{+1} \frac{c_k t_k}{1-\lambda t_k} \, \mathrm{d}\mu(\lambda) \\ &< \sum\nolimits_{k=1}^{n} \int_{-1}^{+1} \frac{c_k t_k}{1-\lambda t_k} \, \mathrm{d}\nu(\lambda) = \sum\nolimits_{k=1}^{n} c_k f(t_k). \end{aligned}$$

This contradiction forces us to conclude that the measures are the same.

15.3 The geometric approach to Loewner's theorem

Many theorems on integral representation have an elegant geometric interpretation as averages of extreme points of a compact, convex set. Hansen & Petersen [HP82] developed a proof of Theorem 15.1 based on this strategy. The argument relies on a complicated interplay between the smoothness and convexity properties of matrix monotone functions, so we cannot give all of the details here. See [Bha97, Chap. V.4] or [Sim19, Chap. 28] for complete arguments.

15.3.1 Slicing the cone

As we have discussed, the matrix monotone functions on $(-1, +1)$ compose a convex cone, closed under pointwise limits. It is convenient to normalize these functions, which amounts to taking a slice through the cone; see Figure 15.1. It can be shown (not easy!) that every matrix monotone function is differentiable. Therefore, we may introduce the set

$$\mathsf{B} := \{f : (-1, +1) \to \mathbb{R} : f \text{ is matrix monotone, } f(0) = 0, \text{ and } f'(0) = 1\}.$$

Furthermore, one may confirm that each *nonconstant* matrix monotone function g satisfies $g'(0) > 0$. Therefore, the function $(g(t) - g(0))/g'(0) \in \mathsf{B}$.

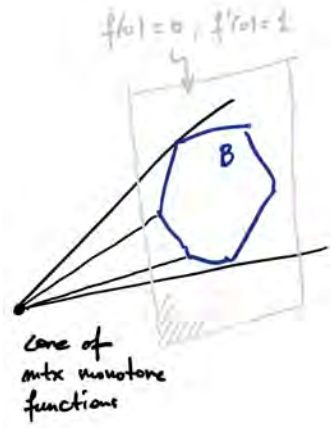

Figure 15.1 (Base of cone). A base B of the cone of matrix monotone functions on $(-1, +1)$.

As it happens, the set B is compact with respect to the topology of pointwise convergence. Roughly speaking, this point follows from the fact that matrix monotone functions in B are bounded above and below:

$$\frac{t}{1+t} \le f(t) \le 0 \qquad \text{when } -1 < t \le 0;$$
$$0 \le f(t) \le \frac{t}{1-t} \qquad \text{when } 0 \le t < 1.$$

This claim also requires a fair amount of argument.

15.3.2 The Krein–Milman theorem

Since B is a compact and convex set, we can activate a famous representation theorem that generalizes Minkowski's theorem on extreme points (Lecture 5).

Theorem 15.6 (Krein–Milman). Let B be a compact, convex subset of a locally convex topological linear space. Then the set coincides with the *closed* convex hull of its extreme points:

$$\mathsf{B} = \overline{\operatorname{conv}}(\operatorname{ext}(\mathsf{B})).$$

In general, the closure is required, and there is a possibility that the extreme points are uninformative (e.g., the extreme points could be dense in B!). Nevertheless, there are many settings where we can explicitly identify the extreme points of the set B. The limit of convex combinations can often be realized as an integral over the extreme points, which leads to an integral representation of the elements of B.

15.3.3 Elementary monotone functions are extreme

As it happens, the elementary matrix monotone functions we have been studying are all extreme points of the set B. We will prove this claim directly using an argument attributed to Boutet de Monvel [Sim19, Thm. 28.12].

Theorem 15.7 (Elementary monotone functions: Extremality). For each $\lambda \in [-1, +1]$, introduce the function

$$\varphi_\lambda(t) = \frac{t}{1-\lambda t} \quad \text{for } t \in (-1, +1).$$

The function φ_λ is an extreme point of B.

In fact, the family $(\varphi_\lambda : \lambda \in [-1, +1])$ exhausts the set of extreme points of B. This claim is significantly harder to prove, so we must take it for granted.

Proof. Fix $\lambda \in [-1, +1]$, and note that $\varphi_\lambda \in \mathsf{B}$. The key to the proof is to observe that the 2×2 Loewner matrix associated with φ_λ always has rank one:

$$\varphi_\lambda^{[1]}(s,t) = \begin{bmatrix} (1-\lambda s)^{-1} \\ (1-\lambda t)^{-1} \end{bmatrix} \begin{bmatrix} (1-\lambda s)^{-1} \\ (1-\lambda t)^{-1} \end{bmatrix}^* \quad \text{for all } s, t \in (-1, +1).$$

Therefore, the Loewner matrix lies in an extreme ray of the psd cone.

Suppose now that we can write $\varphi_\lambda = \frac{1}{2} f + \frac{1}{2} g$ where $f, g \in \mathsf{B}$. The Loewner matrix is linear in the function, so

$$\varphi_\lambda^{[1]}(s,t) = \tfrac{1}{2} f^{[1]}(s,t) + \tfrac{1}{2} g^{[1]}(s,t) \quad \text{for all } s, t \in (-1, +1).$$

But the Loewner matrix $\varphi_\lambda^{[1]}(s,t)$ lies in an extreme ray of the psd cone, so the two matrices on the right are both positive scalar multiples of it. In particular,

$$f^{[1]}(s,t) = \alpha(s,t) \cdot \varphi_\lambda^{[1]}(s,t) \quad \text{for some } \alpha(s,t) \geq 0.$$

To complete the argument, we select $s = 0$ and write out the entries of the matrices. From the normalization $f(0) = 0$ and $f'(0) = 1$, it emerges that

$$\begin{bmatrix} 1 & f(t)/t \\ f(t)/t & f'(t) \end{bmatrix} = \alpha(0,t) \cdot \begin{bmatrix} 1 & (1-\lambda t)^{-1} \\ (1-\lambda t)^{-1} & (1-\lambda t)^{-2} \end{bmatrix} \quad \text{for all } t \in (-1,+1).$$

Inspecting the top-left entry, we realize that $\alpha(0,t) = 1$, regardless of the choice of t. As a consequence, the top-right entry yields $f(t) = t(1-\lambda t)^{-1}$ for all $t \in (-1,+1)$. In other terms, $f = \varphi_\lambda$, and we conclude that φ_λ is an extreme point of B. ∎

15.3.4 From extreme points to integrals

Granted the claim that $(\varphi_\lambda : \lambda \in [-1,+1])$ is the complete family of extreme points of B, we can establish the integral representation. Let $f \in \mathsf{B}$ be a function. The Krein–Milman theorem yields a sequence of approximations

$$\int_{-1}^{+1} \varphi_\lambda(t) \, \mathrm{d}\mu_n(\lambda) \to f(t) \quad \text{as } n \to \infty \text{ for each } t \in (-1,+1).$$

In this expression, μ_n is a probability measure supported on n points in $[-1,+1]$, so it describes a convex combination of n functions φ_{λ_k} with $\lambda_k \in [-1,+1]$. By weak-$*$ compactness of the simplex of probability measures on the compact space $[-1,+1]$, we can extract a subsequence that converges to a probability measure μ on $[-1,+1]$. For this limiting measure,

$$\int_{-1}^{+1} \varphi_\lambda(t) \, \mathrm{d}\mu(\lambda) = f(t) \quad \text{for each } t \in (-1,+1).$$

For a nonconstant matrix monotone function g on $(-1,+1)$, we apply this statement to the function $f(t) = (g(t) - g(0))/g'(0)$ to complete our sketch of the proof of the "hard" direction of Theorem 15.1.

15.4 Matrix monotone functions on the positive real line

Loewner's result gives an integral representation of the matrix monotone functions on the interval $(-1,+1)$. As we will see, integral representations for other classes of matrix monotone functions follow via simple transformations. We will focus on the most useful case: matrix monotone functions on the positive real line.

15.4.1 Fractional linear transformations

We can move among open intervals in the real line using fractional linear transformations. In this section, we will show that fractional linear transformations preserve matrix monotonicity.

Exercise 15.8 (Fractional linear transformations). A fractional linear transformation is a function of the form

$$\psi(t) := \frac{\alpha t + \beta}{\gamma t + \delta} \quad \text{for } t \neq -\delta/\gamma \text{ and where } \alpha, \beta, \gamma, \delta \in \mathbb{R}.$$

When the determinant $\alpha\delta - \beta\gamma > 0$, check that ψ is increasing on the intervals $(-\infty, -\delta/\gamma)$ and $(-\delta/\gamma, +\infty)$.

Show that there is an increasing fractional linear transformation that maps a finite or semi-infinite interval (c, d) onto a finite or semi-infinite interval (a, b) for $a, b, c, d \in \overline{\mathbb{R}}$. We do not allow both endpoints of an interval to be infinite.

Show that an increasing, surjective fractional linear transformation $\psi : (c, d) \to (a, b)$ is matrix monotone.

Exercise 15.9 (Matrix monotonicity: Composition). Suppose that f, g are matrix monotone. Check that $h(t) = f(g(t))$ is matrix monotone, provided that the domains and codomains are compatible.

Exercise 15.10 (Matrix monotonicity: Change of domain). Let $f : (a, b) \to \mathbb{R}$ be a matrix monotone function on a finite or semi-infinite interval. Let $\psi : (c, d) \to (a, b)$ be an increasing fractional linear transformation that is surjective. Confirm that $f \circ \psi$ is matrix monotone on (c, d).

15.4.2 Loewner's theorem for the strictly positive reals

Using these results, we can transfer Theorem 15.1 to obtain a parallel result for matrix monotone functions on $\mathbb{R}_{++}$.

Corollary 15.11 (Loewner: Strictly positive reals). Consider a function $g : \mathbb{R}_{++} \to \mathbb{R}$. The function g is matrix monotone on $\mathbb{R}_{++}$ if and only if it admits a (unique) representation of the form

$$g(t) = \alpha + \int_{[0,1]} \frac{t-1}{\lambda + (1-\lambda)t} \, \mathrm{d}\nu(\lambda) \quad \text{for all } t > 0.$$

If the measure ν has an atom at 0, it contributes a pole at zero. If the measure ν has an atom at 1, it contributes a linear component.

In this expression, $\alpha \in \mathbb{R}$ and ν is a finite, positive Borel measure on $[0, 1]$.

Proof sketch. Let us introduce the fractional linear transformation ψ that maps the strictly positive real line $\mathbb{R}_{++}$ onto the standard open interval $(-1, +1)$. The map ψ and its inverse ψ^{-1} are increasing functions of the form

$$\psi(t) = \frac{t-1}{t+1} \quad \text{for } t \in \mathbb{R}_{++} \quad \text{where} \quad \psi^{-1}(s) = \frac{1+s}{1-s} \quad \text{for } s \in (-1, +1).$$

We can construct the matrix monotone function $f(s) = g(\psi^{-1}(s))$ on $(-1, +1)$. Apply Theorem 15.1 to f to obtain an integral representation. Then make the change of variables $s = \psi(t)$ to transfer this representation back to the function g. Make the affine change of variables $\lambda \mapsto 2\lambda - 1$ to shift the domain of integration from $[-1, +1]$ to $[0, 1]$. The result of this process is quoted above. ∎

Example 15.12 (Logarithm, again). We can use the "easy" direction of Corollary 15.11 to verify that particular functions are matrix monotone. As an example, observe that

$$\log t = \int_0^1 \frac{t-1}{\lambda + (1-\lambda)t} \, \mathrm{d}\lambda \quad \text{for all } t > 0.$$

Therefore, the logarithm is matrix monotone. The representation of the logarithm that we saw before is related to this one by a further change of variables in the integral. ▪

Exercise 15.13 (Loewner: Strictly positive reals). It is sometimes convenient to change variables in Corollary 15.11. For a matrix monotone function $g : \mathbb{R}_{++} \to \mathbb{R}$, show that

$$g(t) = \alpha + \gamma t + \int_{\mathbb{R}_{++}} \frac{\lambda(t-1)}{\lambda + t} \, \mathrm{d}\mu(\lambda) \quad \text{for all } t > 0.$$

where $\alpha \in \mathbb{R}$ and $\gamma \geq 0$ and μ is a finite, positive Borel measure on $\mathbb{R}_{++}$. This is closer to our original presentation of the logarithm in terms of an integral.

15.4.3 Loewner's theorem for the positive reals

So far, we have studied matrix monotone functions on an open interval. In this setting, the function need not have a limiting value at the endpoints. This is the case for the logarithm and the tangent function. When the function does take a limiting value, we can obtain integral representations that are valid up to and including the endpoint.

Corollary 15.14 (Loewner: Positive reals). Consider a function $g : \mathbb{R}_+ \to \mathbb{R}$ that is *continuous* at $t = 0$. The function g is matrix monotone on $\mathbb{R}_+$ if and only if it admits a (unique) representation of the form

$$g(t) = \beta + \gamma t + \int_{(0,1)} \frac{t}{\lambda + (1-\lambda)t} \cdot \frac{\mathrm{d}\nu(\lambda)}{\lambda} \quad \text{for all } t \geq 0.$$

In this expression, $\beta \in \mathbb{R}$ and $\gamma \geq 0$, and ν is a finite, positive Borel measure on $(0, 1)$.

Proof sketch. By continuity, $g(0) = \lim_{t \downarrow 0} g(t)$. Taking a limit of the integral representation in Corollary 15.11, we find that

$$g(0) = \alpha - \int_0^1 \frac{\mathrm{d}\nu(\lambda)}{\lambda} =: \beta > -\infty.$$

Adding and subtracting this quantity in the representation in Corollary 15.11, we obtain the statement after some simplification. ∎

Exercise 15.15 (Loewner: Positive reals). It is often convenient to make another change of variables in Corollary 15.14. For a *continuous* matrix monotone function $g : \mathbb{R}_+ \to \mathbb{R}$, show that

$$g(t) = \beta + \gamma t + \int_{\mathbb{R}_{++}} \frac{\lambda t}{\lambda + t} \, \mathrm{d}\mu(\lambda) \quad \text{for all } t \geq 0.$$

Here, $\beta \in \mathbb{R}$ and $\gamma \geq 0$ and μ is a finite, positive Borel measure on $\mathbb{R}_{++}$.

Exercise 15.15 suggests that we look for functions that have integral representations with this form. Here are two examples.

Example 15.16 (Shifted logarithm: Monotonicity). Beginning with Example 15.12, we can make the change of variables $t \mapsto 1 + t$ and $1 - \lambda \mapsto \lambda^{-1}$ to obtain the representation

$$\log(1 + t) = \int_1^\infty \frac{\lambda t}{\lambda + t} \cdot \lambda^{-2} \, \mathrm{d}\lambda.$$

Therefore, the shifted logarithm $t \mapsto \log(1 + t)$ is matrix monotone on $\mathbb{R}_+$. ▪

Example 15.17 (Powers: Monotonicity). For $r \in (0, 1]$, we can use contour integration to show that

$$t^r = \frac{\sin(\pi r)}{\pi} \int_0^\infty \frac{\lambda t}{\lambda + t} \cdot \lambda^{r-2} \, \mathrm{d}\lambda.$$

Therefore, the power functions $t \mapsto t^r$ for $r \in (0, 1]$ are matrix monotone on $\mathbb{R}_+$.

We can deduce that negative powers are matrix monotone using the composition rule (Exercise 15.9). Indeed, since $t \mapsto -t^{-1}$ is matrix monotone on $\mathbb{R}_{++}$, the power function $t \mapsto -t^{-r}$ is matrix monotone on $\mathbb{R}_{++}$ for each $r \in (0, 1]$. ▪

15.5 Integral representations of matrix convex functions

Just as we developed integral representations for matrix monotone functions, we can develop integral representations for matrix convex functions. These results follow from the representations of matrix monotone functions by means of some general principles. For brevity, we will only consider matrix convex functions on the positive real line.

15.5.1 Relations between matrix monotonicity and convexity

In the scalar setting, monotone functions and convex functions are independent concepts. The main connection is that a differentiable convex function has a monotone derivative. In the matrix setting, however, there is an intricate web of relations between matrix monotone and matrix convex functions. This section outlines some of the basic facts, along with direct proofs.

Warning: Some proofs of Loewner's theorem (including the one we have outlined) depend on the convexity and concavity results in this section, so it may be circular to use integral representations to establish these statements. ∎

Theorem 15.18 (Matrix monotonicity and concavity). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a (continuous) function on the *positive* real line.

1. If f is matrix monotone, then f is matrix concave.
2. If f is matrix concave and takes *positive* values, then f is matrix monotone.

Proof. For the second part, assume that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is matrix concave and *positive*. First, suppose that $\boldsymbol{A} \prec \boldsymbol{B}$. For scalars $\tau \in (0,1)$ and $\bar{\tau} = 1 - \tau$, since f is concave and positive,

$$\begin{aligned} f(\tau \boldsymbol{B}) &= f\big(\tau \boldsymbol{A} + \bar{\tau} \cdot (\tau/\bar{\tau})(\boldsymbol{B} - \boldsymbol{A})\big) \\ &\succcurlyeq \tau \cdot f(\boldsymbol{A}) + \bar{\tau} \cdot f\big((\tau/\bar{\tau})(\boldsymbol{B} - \boldsymbol{A})\big) \succcurlyeq \tau \cdot f(\boldsymbol{A}). \end{aligned}$$

Since f is continuous, we may take $\tau \uparrow 1$. We conclude that $\boldsymbol{A} \prec \boldsymbol{B}$ implies $f(\boldsymbol{A}) \preccurlyeq f(\boldsymbol{B})$. To handle the case where $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$, observe that $f(\boldsymbol{A}) \preccurlyeq f(\boldsymbol{B} + \varepsilon \mathbf{I})$ for $\varepsilon > 0$, and take limits as $\varepsilon \downarrow 0$ to resolve that f is matrix monotone.

For the first part, the key to this argument is the following claim, which we will establish in a moment.

Claim 15.19 (Monotonicity: Submatrix). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous matrix monotone function. For a psd block matrix $\boldsymbol{A}$, we have $f([\boldsymbol{A}]_{11}) \succcurlyeq [f(\boldsymbol{A})]_{11}$.

Recall that $[\cdot]_{11}$ extracts the top-left $(1,1)$ block of a block matrix, and it is a positive linear map.

We assume that $f : \mathbb{R}_+ \to \mathbb{R}$ is matrix monotone. Granted the claim, we introduce a unitary block matrix that generates convex combinations:

$$\boldsymbol{U}_\tau := \begin{bmatrix} \tau^{1/2}\mathbf{I} & \bar{\tau}^{1/2}\mathbf{I} \\ -\bar{\tau}^{1/2}\mathbf{I} & \tau^{1/2}\mathbf{I} \end{bmatrix} \quad \text{for } \tau \in [0,1] \text{ with } \bar{\tau} = 1 - \tau.$$

For psd matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$, construct the psd block diagonal matrix $\boldsymbol{A} = \boldsymbol{A}_1 \oplus \boldsymbol{A}_2$. Using the claim, we calculate that

$$\begin{aligned} f(\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2) &= f\big([\boldsymbol{U}_\tau \boldsymbol{A} \boldsymbol{U}_\tau^*]_{11}\big) \succcurlyeq \big[f(\boldsymbol{U}_\tau \boldsymbol{A} \boldsymbol{U}_\tau^*)\big]_{11} \\ &= \big[\boldsymbol{U}_\tau f(\boldsymbol{A}) \boldsymbol{U}_\tau^*\big]_{11} = \tau f(\boldsymbol{A}_1) + \bar{\tau} f(\boldsymbol{A}_2). \end{aligned}$$

In other words, f is matrix concave.

To prove Claim 15.19, we introduce another block matrix that helps diagonalize a block matrix:

$$\boldsymbol{T}_\varepsilon := \begin{bmatrix} \varepsilon^{1/2}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\varepsilon^{-1/2}\mathbf{I} \end{bmatrix} \quad \text{for } \varepsilon > 0.$$

For a psd block matrix $\boldsymbol{A}$, a short calculation reveals that

$$\boldsymbol{A} \preccurlyeq \boldsymbol{A} + \boldsymbol{T}_\varepsilon^* \boldsymbol{A} \boldsymbol{T}_\varepsilon = \begin{bmatrix} (1+\varepsilon) \cdot [\boldsymbol{A}]_{11} & \mathbf{0} \\ \mathbf{0} & (1+\varepsilon^{-1}) \cdot [\boldsymbol{A}]_{22} \end{bmatrix}.$$

Of course, $[\cdot]_{22}$ extracts the $(2,2)$ block of a block matrix.

Apply the monotone function f to this relation. Since $[\cdot]_{11}$ is a positive linear map, it preserves the psd order on self-adjoint matrices. Therefore,

$$[f(\boldsymbol{A})]_{11} \preccurlyeq f\big((1+\varepsilon) \cdot [\boldsymbol{A}]_{11}\big).$$

Take the limit as $\varepsilon \downarrow 0$ using the assumption that f is continuous. ∎

Next, we point out that a matrix convex function on the positive real line induces another matrix monotone function.

Proposition 15.20 (Monotonicity from convexity). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a matrix convex function with $f(0) \le 0$. Then $g(t) = f(t)/t$ is matrix monotone on $\mathbb{R}_{++}$.

Proof. Assume that $\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{B}$. Since $\boldsymbol{B}^{-1/2} \boldsymbol{A} \boldsymbol{B}^{-1/2} \preccurlyeq \mathbf{I}$, we see that the matrix $\boldsymbol{K} = \boldsymbol{B}^{-1/2} \boldsymbol{A}^{1/2}$ is a contraction. Let $\boldsymbol{L} = (\mathbf{I} - \boldsymbol{K}^* \boldsymbol{K})^{1/2}$. The matrix Jensen inequality (Lecture 13) implies that

$$f(\boldsymbol{A}) = f(\boldsymbol{K}^* \boldsymbol{B} \boldsymbol{K}) = f(\boldsymbol{K}^* \boldsymbol{B} \boldsymbol{K} + \boldsymbol{L}^* \mathbf{0} \boldsymbol{L}) \preccurlyeq \boldsymbol{K}^* f(\boldsymbol{B}) \boldsymbol{K} + \boldsymbol{L}^* f(\mathbf{0}) \boldsymbol{L} \preccurlyeq \boldsymbol{K}^* f(\boldsymbol{B}) \boldsymbol{K}.$$

Conjugate both sides by $\boldsymbol{A}^{-1/2}$ to see that

$$g(\boldsymbol{A}) = \boldsymbol{A}^{-1/2} f(\boldsymbol{A}) \boldsymbol{A}^{-1/2} \preccurlyeq \boldsymbol{B}^{-1/2} f(\boldsymbol{B}) \boldsymbol{B}^{-1/2} = g(\boldsymbol{B}).$$

Indeed, standard matrix functions of the same matrix commute. ∎

Problem 15.21 (Convexity from monotonicity). A converse of Proposition 15.20 also holds. Assume that $f : \mathbb{R}_+ \to \mathbb{R}$ is continuous and $f(0) \le 0$. If $g(t) = f(t)/t$ is matrix monotone on $\mathbb{R}_{++}$, then $f : \mathbb{R}_+ \to \mathbb{R}$ is matrix convex. **Hint:** Using a dilation argument, show that f is matrix convex if and only if $f(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}) \preccurlyeq \boldsymbol{P} f(\boldsymbol{A}) \boldsymbol{P}$ for all orthoprojectors $\boldsymbol{P}$ and all psd matrices $\boldsymbol{A}$. Show that monotonicity of g implies this condition.

Exercise 15.22 (Powers: Convexity). We can identify convexity properties of the power functions using Proposition 15.20 and Problem 15.21. Determine whether $t \mapsto t^r$ is matrix convex or matrix concave for each $r \in [-1, 2]$.

15.5.2 Integral representation

We are now prepared to give an integral representation of matrix convex functions on the positive real line. Results of this type are usually attributed to Bendat & Sherman.

Corollary 15.23 (Matrix convexity: Integral representation). Consider a (continuous) matrix convex function $f : \mathbb{R}_+ \to \mathbb{R}$. Then f admits the representation

$$f(t) = \alpha + \beta t + \gamma t^2 + \int_{\mathbb{R}_{++}} \frac{\lambda t(t-1)}{\lambda + t} \, \mathrm{d}\mu(\lambda) \quad \text{for all } t \ge 0.$$

The coefficient $\gamma \ge 0$, and the Borel measure μ is finite and positive.

Proof. We may shift f so that $f(0) = 0$. Proposition 15.20 implies that $g(t) = f(t)/t$ is matrix monotone on $\mathbb{R}_{++}$. Exercise 15.13 yields an integral representation for g on $\mathbb{R}_{++}$. Multiply through by t, and remove the shift from f. Since f is continuous, the representation is also valid at $t = 0$. ∎

Exercise 15.24 (Entropy). Using an integral representation for the logarithm, find an integral representation for the negative entropy function $\operatorname{negent}(t) := t \log t$ for $t \geq 0$.

Exercise 15.25 (Matrix convexity: Other intervals). Consider a matrix convex function $f : \mathsf{I} \to \mathbb{R}$ on an open interval I. Bendat & Sherman proved that the divided difference $f^{[1]}(s, \cdot) : \mathsf{I} \to \mathbb{R}$ is matrix monotone for each $s \in \mathsf{I}$. Use this fact to derive an integral representation for the matrix convex function f.

15.6 Application: Matrix Jensen and Lyapunov inequalities

The integral representations from this lecture have remarkable consequences for matrix analysis. In this section, we will use them to derive a stronger matrix Jensen inequality. In turn, this inequality allows us to develop a satisfactory extension of Lyapunov's inequality.

15.6.1 Matrix Jensen for a unital, positive linear maps

In Lecture 13, we established the matrix Jensen inequality, which describes how matrix convex functions interact with matrix convex combinations. Earlier, we argued that all unital, positive linear maps can be regarded as averaging operators. The integral representations for matrix convex functions now permit us to derive a Jensen inequality for any unital, positive map.

Theorem 15.26 (Matrix Jensen: Positive linear maps). Consider a matrix convex function $f : \mathbb{R}_+ \to \mathbb{R}$, and let $\mathbf{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a *unital* positive linear map. Then

$$f(\mathbf{\Phi}(\boldsymbol{A})) \preccurlyeq \mathbf{\Phi}(f(\boldsymbol{A})) \quad \text{for all psd } \boldsymbol{A} \in \mathbb{H}_n^+.$$

In particular, this inequality is valid when $-f$ is matrix monotone.

We can trace this kind of result to work of Davis [Dav57] and Choi [Cho74]. See also Ando's paper [And79].

Proof. Corollary 15.23 implies that

$$f(t) = \alpha + \beta t + \gamma t^2 + \int_{\mathbb{R}_{++}} \left[t - (\lambda + 1) + \frac{\lambda(\lambda+1)}{\lambda + t} \right] \cdot \lambda \, \mathrm{d}\mu(\lambda).$$

The coefficient $\gamma \geq 0$. In particular, for a psd matrix $\boldsymbol{A}$,

$$f(\boldsymbol{A}) = \alpha \mathbf{I} + \beta \boldsymbol{A} + \gamma \boldsymbol{A}^2 + \int_{\mathbb{R}_{++}} \left[\boldsymbol{A} - (\lambda+1)\mathbf{I} + \lambda(\lambda+1)(\lambda \mathbf{I} + \boldsymbol{A})^{-1} \right] \cdot \lambda \, \mathrm{d}\mu(\lambda).$$

Kadison's inequality and Choi's inequality yield

$$\mathbf{\Phi}(\boldsymbol{A}^2) \preccurlyeq \big(\mathbf{\Phi}(\boldsymbol{A})\big)^2 \quad \text{and} \quad \mathbf{\Phi}\big((\lambda \mathbf{I} + \boldsymbol{A})^{-1}\big) \preccurlyeq \big(\lambda \mathbf{I} + \mathbf{\Phi}(\boldsymbol{A})\big)^{-1}.$$

Apply $\mathbf{\Phi}$ to the integral representation of $f(\boldsymbol{A})$ to obtain

$$\begin{aligned} \mathbf{\Phi}(f(\boldsymbol{A})) &\preccurlyeq \alpha \mathbf{I} + \beta \mathbf{\Phi}(\boldsymbol{A}) + \gamma \big(\mathbf{\Phi}(\boldsymbol{A})\big)^2 \\ &\quad + \int_{\mathbb{R}_{++}} \left[\mathbf{\Phi}(\boldsymbol{A}) - (\lambda+1)\mathbf{I} + \lambda(\lambda+1)\big(\lambda \mathbf{I} + \mathbf{\Phi}(\boldsymbol{A})\big)^{-1} \right] \cdot \lambda \, \mathrm{d}\mu(\lambda) \\ &= f(\mathbf{\Phi}(\boldsymbol{A})). \end{aligned}$$

We have repeatedly used the linearity, the unital property, and the continuity of $\mathbf{\Phi}$. The semidefinite inequality follows from Kadison's and Choi's inequalities. ∎

Exercise 15.27 (Choi's convexity theorem). Theorem 15.26 holds for every matrix convex function f, regardless of its domain. Prove it. **Hint:** Use Exercise 15.25.

Exercise 15.28 (Matrix Jensen). Deduce that the matrix Jensen inequality from Lecture 13 is a special case of Choi's convexity theorem (Exercise 15.27). **Hint:** Consider the block diagonal matrix $\boldsymbol{A} = \boldsymbol{A}_1 \oplus \boldsymbol{A}_2 \oplus \cdots \oplus \boldsymbol{A}_n$. Show that the matrix convex combination is a unital, positive linear map on this matrix.

15.6.2 Matrix Lyapunov inequalities

Theorem 15.26 has remarkable consequences when applied to the most common matrix convex functions [And79, Cor. 4.2]. These results give additional credence to the idea that unital, positive maps are averaging operators.

Corollary 15.29 (Matrix Lyapunov). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a *unital*, positive linear map. For each $p \geq 1$,

$$\boldsymbol{\Phi}(\boldsymbol{A}) \preccurlyeq \left[\boldsymbol{\Phi}(\boldsymbol{A}^p)\right]^{1/p} \quad \text{for all psd } \boldsymbol{A} \in \mathbb{H}_n^+.$$

Warning: It is tempting to raise both sides of this inequality to the pth power, but the resulting statement is false for $p > 2$. ■

Proof. For $r \in (0, 1)$, the function $t \mapsto t^r$ is matrix concave on $\mathbb{R}_{++}$. This statement follows when we combine Example 15.17 and Theorem 15.18; alternatively, you can make a direct inspection of the integral representation for the power. Theorem 15.26 now implies that

$$\boldsymbol{\Phi}(\boldsymbol{A}^r) \succcurlyeq \boldsymbol{\Phi}(\boldsymbol{A})^r.$$

Make the change of variables $\boldsymbol{A} \mapsto \boldsymbol{A}^{1/r}$, and select $r = 1/p$ where $p \geq 1$. ■

Exercise 15.30 (Matrix Lyapunov: Exponential). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_n$ be a unital, strictly positive linear map. Prove that

$$\boldsymbol{\Phi}(\boldsymbol{A}) \preccurlyeq \log(\boldsymbol{\Phi}(\exp(\boldsymbol{A}))) \quad \text{for all } \boldsymbol{A} \in \mathbb{H}_n.$$

Hint: Note that $t \mapsto \log(\varepsilon + t)$ is matrix concave for $t \geq 0$ and $\varepsilon > 0$. Choose an appropriate value of ε; there is no need to take a limit.

Exercise 15.31 (Matrix Lyapunov: More powers). Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a unital, positive linear map. *For* $p \in [1/2, 1]$, derive the matrix Lyapunov inequality

$$\left[\boldsymbol{\Phi}(\boldsymbol{A}^p)\right]^{1/p} \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}) \quad \text{for all psd } \boldsymbol{A} \in \mathbb{H}_n^+.$$

Exercise 15.32 (Matrix entropy: Filtering). Recall that the entropy $\operatorname{ent}(t) := -t \log t$ for $t \geq 0$ is matrix concave. Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a unital positive linear map. Show that

$$\operatorname{ent}(\boldsymbol{\Phi}(\boldsymbol{A})) \succcurlyeq \boldsymbol{\Phi}(\operatorname{ent}(\boldsymbol{A})). \quad \text{for all psd } \boldsymbol{A} \in \mathbb{H}_n^+.$$

In other words, entropy decreases when we filter by a unital, positive linear map.

Notes

Loewner's theorem is a classic result that has attracted a significant amount of attention, and it has been the subject of several books, notably [Sim19]. Nevertheless, there does not seem to be a short, accessible treatment of these ideas. Nor does there appear to be a self-contained proof of Loewner's theorem that can be presented in a single lecture. This lecture is the instructor's attempt to set out the main facts about integral representation of matrix monotone and matrix convex functions in a way that might

be independently useful or support further study. It draws heavily on Bhatia [Bha97, Chap. V] and Simon [Sim19], but the organization is new.

The proof of uniqueness in Loewner's theorem is adapted from Simon's book [Sim19, Chap. 1]. The geometric proof, via Krein–Milman, is due to Hansen & Pedersen [HP82]. We have sketched some of the ideas from the argument, following Bhatia [Bha97, Chap. V] and Simon [Sim19, Chap. 28]. The proof that elementary monotone functions are monotone is adapted from an argument of Boutet de Monvel, which appears in Simon's book.

Simon [Sim19, Chap. 1] makes it clear that Loewner's theorem can be transferred from one open interval to another. We have expanded on this point to show how the formulation for the strictly positive real line relates to the formulation on the standard open interval. The method of condensing the result for the positive real line appears in Bhatia's book [Bha97, pp. 144–145].

Some of the examples in this lecture appear in Bhatia's books [Bha97, Chap. V] and [Bha07b, Chaps. 4, 5], while some of them were identified by the instructor to support the presentation.

The results on filtering with a positive linear map date back to work of Davis [Dav57], Choi [Cho74], and Ando [And79].

Lecture bibliography

[And79] T. Ando. "Concavity of certain maps on positive definite matrices and applications to Hadamard products". In: *Linear Algebra and its Applications* 26 (1979), pages 203–241.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[Cho74] M.-D. Choi. "A Schwarz inequality for positive linear maps on C^*-algebras". In: *Illinois Journal of Mathematics* 18.4 (1974), pages 565 –574. DOI: 10.1215/ijm/1256051007.

[Dav57] C. Davis. "A Schwarz inequality for convex operator functions". In: *Proc. Amer. Math. Soc.* 8 (1957), pages 42–44. DOI: 10.2307/2032808.

[GR07] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Seventh. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX). Elsevier/Academic Press, Amsterdam, 2007.

[HP82] F. Hansen and G. K. Pedersen. "Jensen's Inequality for Operators and Löwner's Theorem." In: *Mathematische Annalen* 258 (1982), pages 229–241.

[Sim19] B. Simon. *Loewner's theorem on monotone matrix functions*. Springer, Cham, 2019. DOI: 10.1007/978-3-030-22422-6.

16. Matrix Means

Date: 24 February 2022 **Scribe:** Jing Yu

In this lecture, we consider the question about how to "average" a pair of psd matrices. We introduce a class of matrix means, and we give a complete characterization of these functions. This topic may seem like a departure from the recent lectures. In fact, the means that we study are in one-to-one correspondence with a class of matrix monotone functions.

We begin by formalizing the notion of a scalar mean, which describes an average of two positive numbers. Following the ideas of Kubo & Ando [KA79], we explain how to extend this construction to matrices. Then we show that (bivariate) means can be described by (univariate) representer functions. This observation leads to a satisfying theory in both the scalar and matrix settings.

Agenda:

1. Scalar means
2. Axioms for matrix means
3. Representers for scalar means
4. Representers for matrix means
5. Matrix perspective transformations
6. Matrix means from representers
7. Integral representations

16.1 Scalar means

What does it mean to "average" two positive numbers? Although the arithmetic mean is the most widely used notion, you may have encountered several other ways to compute an average of numbers. To draw a clear distinction with the arithmetic mean, it is common to call these functions "means" rather than averages.

Example 16.1 (Scalar means). Here are several important means defined for strictly positive numbers $a, b > 0$.

- **Arithmetic mean:** The arithmetic mean is a familiar example that arises in statistics and probability:
$$(a, b) \mapsto \tfrac{1}{2}(a + b).$$
The arithmetic mean is the best constant approximation to a random variable that takes values a, b with equal probability.
- **Geometric mean:** The formula for the area of a rectangle leads to the notion of the geometric mean:
$$(a, b) \mapsto \sqrt{ab}.$$
This mean also appears in the study of inequalities and in functional analysis because it is connected to the convexity of the exponential function.
- **Harmonic mean:** The harmonic mean arises in the analysis of electrical circuits because of Kirchhoff's law. It computes an average via the rule
$$(a, b) \mapsto \left(\tfrac{1}{2}a^{-1} + \tfrac{1}{2}b^{-1}\right)^{-1}.$$
We can extend from strictly positive numbers to positive numbers by taking limits.
- **Logarithmic mean:** The logarithmic mean is a less common example, although it sometimes arises in the study of heat transfer. It takes the form
$$(a, b) \mapsto \frac{a - b}{\log a - \log b} = \int_0^1 a^r b^{1-r} \,\mathrm{d}r.$$

The second formula is valid for numbers that are positive, but not necessarily strictly positive.

Other examples include the family of binomial means and the family of power means. Have you seen other types of means? ▪

What are the properties common to these examples that can justify their interpretation as the mean of two numbers? By extracting the key features, we may define a class of scalar means.

Definition 16.2 (Scalar mean). A function $M : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ on pairs of *positive* numbers is called a *scalar mean* if it has the following properties.

1. **Strict positivity.** The mean $M(a, b) > 0$ for all $a, b > 0$.
2. **Ordering.** The mean lies between the values of its arguments:

$$\begin{aligned} 0 \le a \le b \quad &\text{implies} \quad a \le M(a, b) \le b; \\ 0 \le b \le a \quad &\text{implies} \quad b \le M(a, b) \le a. \end{aligned}$$

3. **Monotonicity.** The sections of the mean are increasing:

$$\begin{aligned} a &\mapsto M(a, b) \text{ is increasing for each } b \ge 0; \\ b &\mapsto M(a, b) \text{ is increasing for each } a \ge 0. \end{aligned}$$

4. **Positive homogeneity.** A positive scalar can pass through the mean:

$$M(\lambda a, \lambda b) = \lambda M(a, b) \quad \text{for each } \lambda \ge 0 \text{ and all } a, b \ge 0.$$

5. **Continuity.** The mean $(a, b) \mapsto M(a, b)$ is continuous on $\mathbb{R}_+ \times \mathbb{R}_+$.

We say that a mean is *symmetric* if it also satisfies $M(a, b) = M(b, a)$ for all $a, b \ge 0$. This property is typical, but we will not insist on it.

We remark that these properties are interrelated, so they are not fully independent from each other. For instance, the ordering property already implies strict positivity. For a symmetric mean, we do not need to make separate hypotheses about the behavior of the first and second argument. It is also common to define the mean for strictly positive numbers only, since we can use continuity to obtain the value of the mean when one of the arguments is zero.

Exercise 16.3 (Scalar mean: Examples). Confirm that each item in Example 16.1 is a symmetric scalar mean in the sense of Definition 16.2.

16.2 Matrix means

We may now define the concept of a matrix mean by generalizing the axioms for a scalar mean. This approach to matrix means is a hybrid between the classic paper of Kubo & Ando [KA79] and the presentation in Bhatia [Bha07b, Chap. 4.1]. After giving the definition, we look at some of the most basic examples.

Aside: This approach is not the only way to construct a sensible notion of a matrix mean. In particular, Hiai & Kosaki [HK99] have developed another elegant theory.

16.2.1 Axioms

We begin with an axiomatization of matrix means. For the most part, this task is straightforward. Let us emphasize that we only consider means of psd matrices in the same way that we only consider means of positive numbers. The ordering of real

numbers is replaced by the psd order. The only difficulty arises from the generalization of the positive homogeneity property, which we will discuss after the definition.

Definition 16.4 (Matrix mean). Fix $n \in \mathbb{N}$. A function $\boldsymbol{M} : \mathbb{H}_n^+ \times \mathbb{H}_n^+ \to \mathbb{H}_n^+$ on pairs of *psd* matrices is called a *matrix mean* on $\mathbb{H}_n^+$ if it has the following properties.

1. **Strict positivity.** The mean $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) \succ \mathbf{0}$ for all $\boldsymbol{A}, \boldsymbol{B} \succ \mathbf{0}$.
2. **Ordering.** The mean lies between the values of its arguments:

$$\begin{aligned} \mathbf{0} \preccurlyeq \boldsymbol{A} \preccurlyeq \boldsymbol{B} \quad &\text{implies} \quad \boldsymbol{A} \preccurlyeq \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) \preccurlyeq \boldsymbol{B}; \\ \mathbf{0} \preccurlyeq \boldsymbol{B} \preccurlyeq \boldsymbol{A} \quad &\text{implies} \quad \boldsymbol{B} \preccurlyeq \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) \preccurlyeq \boldsymbol{A}. \end{aligned}$$

3. **Monotonicity.** The sections of the mean are increasing:

$$\begin{aligned} \boldsymbol{A}_1 \preccurlyeq \boldsymbol{A}_2 \quad &\text{implies} \quad \boldsymbol{M}(\boldsymbol{A}_1, \boldsymbol{B}) \preccurlyeq \boldsymbol{M}(\boldsymbol{A}_2, \boldsymbol{B}) \text{ for each } \boldsymbol{B} \succcurlyeq \mathbf{0}; \\ \boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2 \quad &\text{implies} \quad \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}_1) \preccurlyeq \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}_2) \text{ for each } \boldsymbol{A} \succcurlyeq \mathbf{0}. \end{aligned}$$

4. **Conjugation.** For all $\boldsymbol{A}, \boldsymbol{B} \succcurlyeq \mathbf{0}$, conjugation passes through the mean:

$$\boldsymbol{M}(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}, \boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) = \boldsymbol{X}^*\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{X} \quad \text{for each } \boldsymbol{X} \in \mathbb{M}_n.$$

5. **Continuity.** The mean $(\boldsymbol{A}, \boldsymbol{B}) \mapsto \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$ is continuous on $\mathbb{H}_n^+ \times \mathbb{H}_n^+$.

We say that a matrix mean is *symmetric* if it also satisfies the identity $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{M}(\boldsymbol{B}, \boldsymbol{A})$ for all $\boldsymbol{A}, \boldsymbol{B} \succcurlyeq \mathbf{0}$.

As in the case of scalar means, we often define a matrix mean for strictly positive matrices. The continuity requirement allows us to extend the matrix mean to all psd matrices. For brevity, we will not give any details on these continuity arguments.

We can think about the conjugation axiom for a matrix mean as a counterpart to the positive homogeneity property of a scalar mean. For each $\lambda > 0$, the function $a \mapsto \lambda a$ is a bijection on the positive numbers. Likewise, for each *nonsingular* $\boldsymbol{X} \in \mathbb{M}_n$, the congruence $\boldsymbol{A} \mapsto \boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}$ is a bijection on the psd matrices. Therefore, it is natural to ask that the mean preserve simultaneous congruence. The extension to all $\boldsymbol{X} \in \mathbb{M}_n$ follows from a short continuity argument.

The conjugation axiom has some striking implications. For example, it ensures that the mean of two scalar matrices must also be a scalar matrix (i.e., a multiple of the identity matrix).

Exercise 16.5 (Matrix mean: Scalar matrices). Let $\boldsymbol{M}$ be a matrix mean on $\mathbb{H}_n^+$, as in Definition 16.4. For all scalars $\alpha, \beta \geq 0$, prove that

$$\boldsymbol{M}(\alpha \mathbf{I}_n, \beta \mathbf{I}_n) = \gamma \mathbf{I}_n \quad \text{for some } \gamma = \gamma(\alpha, \beta) \geq 0.$$

Hint: Assume $\boldsymbol{M}(\alpha \mathbf{I}, \beta \mathbf{I}) = \boldsymbol{A}$. Apply the congruence property by writing $\mathbf{I} = \boldsymbol{Q}^*\boldsymbol{Q}$ for a unitary matrix $\boldsymbol{Q} \in \mathbb{M}_n$. Deduce that $\boldsymbol{A}$ must be a scalar matrix by averaging over all unitary $\boldsymbol{Q}$.

Definition 16.4 provokes several questions. For example, do matrix means exist? What are some interesting examples? Are matrix means interpretable as "averages" of matrices? Can we characterize matrix means? To answer these questions, we first present some examples of matrix means.

16.2.2 Basic examples

The most transparent example of a matrix mean is the matrix extension of the arithmetic mean.

Definition 16.6 (Matrix arithmetic mean). The *matrix arithmetic mean* is the function

$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) := \tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{B}) \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+.$$

Exercise 16.7 (Matrix arithmetic mean). Confirm that the matrix arithmetic mean is symmetric, and it satisfies all the axioms in Definition 16.4.

There are two very basic, but not so obvious, examples of a matrix mean that merit explicit mention.

Definition 16.8 (Left and right matrix means). The *left matrix mean* and *right matrix mean* are respectively given by the expressions

$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{A} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+;$$
$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{B} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+.$$

Exercise 16.9 (Matrix left and right means). Confirm that the left and right matrix means satisfy the axioms in Definition 16.4, but they are not symmetric.

16.2.3 Matrix harmonic mean

Next, we turn to another example that also turns out to be truly fundamental.

Definition 16.10 (Matrix harmonic mean). The *matrix harmonic mean* is defined as

$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) := \left(\tfrac{1}{2}\boldsymbol{A}^{-1} + \tfrac{1}{2}\boldsymbol{B}^{-1}\right)^{-1} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

We extend this definition to all psd matrices by taking limits:

$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) := \lim_{\varepsilon \downarrow 0} \boldsymbol{M}(\boldsymbol{A} + \varepsilon \mathbf{I}, \boldsymbol{B} + \varepsilon \mathbf{I}) \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+.$$

Exercise 16.11 (Harmonic mean: Projectors). Compute the harmonic mean $\boldsymbol{M}(\boldsymbol{P}, \boldsymbol{B})$ for an orthogonal projector $\boldsymbol{P} \in \mathbb{H}_n^+$ and a positive-definite matrix $\boldsymbol{B} \in \mathbb{H}_n^{++}$.

Exercise 16.12 (Matrix harmonic mean). Note that the matrix harmonic mean is symmetric. Confirm that the matrix harmonic mean satisfies all the axioms in Definition 16.4. **Hint:** Assume that the arguments are positive definite, and recall that the matrix inverse is order reversing. Use continuity to extend the axioms to all psd matrices and to the case of psd conjugation.

It is valuable to rewrite the harmonic mean using a related function called the parallel sum [AJD69].

Definition 16.13 (Parallel sum). The *parallel sum* of two positive-definite matrices is defined as

$$(\boldsymbol{A} : \boldsymbol{B}) := \left(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1}\right)^{-1} \quad \text{where } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

We extend this definition to psd matrices via continuity. It is evident that $2(\boldsymbol{A} : \boldsymbol{B})$ coincides with the harmonic mean of $\boldsymbol{A}$ and $\boldsymbol{B}$.

The parallel sum is intimately connected to Schur complements. Indeed, for

positive-definite matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}$,

$$\boldsymbol{A} : \boldsymbol{B} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} = \boldsymbol{B}(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{A} = \boldsymbol{A} - \boldsymbol{A}(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{A}.$$

We can recognize the latter matrix as a Schur complement:

$$\boldsymbol{A} : \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A} \\ \boldsymbol{A} & \boldsymbol{A} + \boldsymbol{B} \end{bmatrix} / (\boldsymbol{A} + \boldsymbol{B}).$$

This representation extends to psd matrices. By symmetry, we also have the relation $\boldsymbol{A} : \boldsymbol{B} = \boldsymbol{B} - \boldsymbol{B}(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{B}$, and we can recognize the Schur complement of another block matrix.

The representation of the parallel sum as a Schur complement gives an alternative proof of the fact that the parallel sum is monotone. Indeed, Schur complements of a psd matrix are increasing with respect to the matrix. We also discover that the parallel sum is concave.

Exercise 16.14 (Parallel sum: Concavity). For psd matrices $\boldsymbol{A}_i, \boldsymbol{B}_i \in \mathbb{H}_n^+$ for $i = 1, 2$, use the connection with Schur complements to prove that

$$(\tau \boldsymbol{A}_1 + \bar{\tau}\boldsymbol{A}_2) : (\tau \boldsymbol{B}_1 + \bar{\tau}\boldsymbol{B}_2) \succcurlyeq \tau\,(\boldsymbol{A}_1 : \boldsymbol{B}_1) + \bar{\tau}\,(\boldsymbol{A}_2 : \boldsymbol{B}_2) \quad \text{for } \tau \in [0, 1].$$

As usual, $\bar{\tau} := 1 - \tau$. In other words, the parallel sum is jointly concave on pairs of psd matrices.

16.2.4 Matrix geometric mean

Beyond the most elementary examples, we may attempt to develop a matrix extension of the geometric mean. We begin with the special case where the two matrices commute. In this instance, it is natural to define the geometric mean as

$$\boldsymbol{A} \sharp \boldsymbol{B} = \boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2} \quad \text{for } \textit{commuting } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+.$$

Indeed, commuting matrices are simultaneously diagonalizable, so this expression amounts to computing the geometric mean of two diagonal matrices, entry by entry.

To extend this formula to all matrices, we realize that the matrix geometric mean $\sharp$ must satisfy the conjugation axiom. For positive-definite $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}$, this condition implies that

$$\begin{aligned} \boldsymbol{A} \sharp \boldsymbol{B} &= (\boldsymbol{A}^{1/2}\mathbf{I}\boldsymbol{A}^{1/2}) \sharp (\boldsymbol{A}^{1/2}\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\boldsymbol{A}^{1/2}) \\ &= \boldsymbol{A}^{1/2}\big(\mathbf{I} \sharp (\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})\big)\boldsymbol{A}^{1/2} = \boldsymbol{A}^{1/2}\big(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\big)^{1/2}\boldsymbol{A}^{1/2}. \end{aligned}$$

We have used the fact that the identity matrix commutes with everything. This expression appears complicated, but we have no choice about it once we accept that matrix means satisfy the conjugation axiom.

These considerations lead to the following definition of the matrix geometric mean.

Definition 16.15 (Matrix geometric mean). For psd matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+$, the *matrix geometric mean* is defined as

$$\boldsymbol{A} \sharp \boldsymbol{B} := \boldsymbol{A}^{1/2} \cdot \big(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\big)^{1/2} \cdot \boldsymbol{A}^{1/2} \quad \text{where } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

We extend this definition to all psd matrices by continuity.

It is clear that the matrix geometric mean is positive. Using the fact that the square-root is a matrix monotone function, it is not hard to show that the matrix geometric mean satisfies the order property. On the other hand, it is not clear that the matrix geometric mean has the other required properties (congruence, monotonicity). Although one may prove these results directly, we will instead develop them as a consequence of a more general theory of matrix means.

16.3 Representer functions for scalar means

Our goal is to develop a characterization of matrix means. To that end, we first return to the scalar case, where we argue that scalar means are in one-to-one correspondence with a class of univariate functions.

16.3.1 Representers from scalar means

When we fix one of its arguments, each scalar mean yields a univariate function. This function has attractive properties of its own.

Definition 16.16 (Scalar mean: Representer). Consider a scalar mean $M : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. The *representer* of the mean is the function

$$f : \mathbb{R}_+ \to \mathbb{R}_+ \quad \text{given by} \quad f(t) := M(1, t) \quad \text{for } t \geq 0.$$

It is easy to provide examples.

Example 16.17 (Scalar mean: Representers). Here are the representer functions of some basic scalar means.

- **Arithmetic mean.** The representer of the arithmetic mean is $f(t) = (1 + t)/2$.
- **Geometric mean.** The representer of the geometric mean is $f(t) = t^{1/2}$.
- **Harmonic mean.** The representer of the harmonic mean is $f(t) = 2t/(t + 1)$.
- **Logarithmic mean.** The representer of the logarithmic mean is the function $f(t) = (t - 1)/\log(t)$.

You may wish to compute the representer functions for general power means and for binomial means. ∎

The axioms for a scalar mean induce several structural constraints on its representer function.

Exercise 16.18 (Scalar mean: Representer properties). Consider a scalar mean $M : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. Introduce the representer function $f(t) = M(1, t)$ for $t \geq 0$. Prove that the representer enjoys the following properties.

1. **Strict positivity.** The representer $f(t) > 0$ for $t > 0$.
2. **Normalization.** The representer take the value $f(1) = 1$.
3. **Monotonicity.** The representer $t \mapsto f(t)$ is increasing.
4. **Subadditivity.** The function $t \mapsto f(t)/t$ is decreasing for $t > 0$.
5. **Continuity.** The representer f is continuous.
6. **Symmetry.** If the mean M is *symmetric*, then $f(t) = t \cdot f(1/t)$ for all $t > 0$.

Most of these results are straightforward. The subadditivity and symmetry properties may take a moment of thought.

16.3.2 Scalar means from representers

Of course, each scalar mean generates a unique representer function. The central question is whether we can reverse the process. That is, can we reconstruct a scalar mean from its representer? The answer is positive, as the next result shows.

Proposition 16.19 (Representers yield scalar means). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function that satisfies properties (1)–(5) in Exercise 16.18. Define the *perspective function* of f:

$$M_f(a, b) := a \cdot f(b/a) \quad \text{for all } a, b > 0.$$

We extend to a function $M_f : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ by taking limits.

Then M_f is a scalar mean with representer f. Indeed, $f \leftrightarrow M_f$ is a bijection between representer functions and scalar means.

Moreover, if f satisfies the symmetry property (5), then the mean M_f is symmetric in its arguments.

Proof. We can easily verify that M_f enjoys the properties of a scalar mean directly from the analogous properties of the representer f. The only technical difficulty arises in verifying the existence of the limit of $M_f(a, b)$ as $a \downarrow 0$ and $b > 0$.

To see that representers and scalar means are in one-to-one correspondence, note that $f(t) = M_f(1, t)$ for all $t > 0$, so that f is the representer of M_f. ∎

16.4 Representation of matrix means

We would like to undertake the same project in the matrix setting. In other words, we wish to express matrix means in terms of representer functions. Although the approach has a lot in common with the scalar setting, the conditions on a matrix representer function are more stringent.

16.4.1 Representers from matrix means

One might imagine that a matrix mean would have more complicated behavior than a scalar mean, and so it might not be possible to characterize it so simply. In fact, we can reduce each matrix mean to a scalar function.

Definition 16.20 (Matrix mean: Representer). Let $\boldsymbol{M} : \mathbb{H}_n^+ \times \mathbb{H}_n^+ \to \mathbb{H}_n^+$ be a matrix mean on $\mathbb{H}_n^+$, as in Definition 16.4. The *representer* of the matrix mean is the scalar function

$$f : \mathbb{R}_+ \to \mathbb{R}_+ \quad \text{for which} \quad f(t) \cdot \mathbf{I}_n = \boldsymbol{M}(\mathbf{I}_n, t \cdot \mathbf{I}_n) \quad \text{for } t \geq 0.$$

This definition uses Exercise 16.5 to ensure that the mean of two scalar matrices is itself a scalar matrix.

Exercise 16.21 (Matrix mean: Representers). Compute the representers for the matrix arithmetic mean, left and right means, the matrix harmonic mean, and the matrix geometric mean.

The axioms for a matrix mean ensure that the representer has many of the same properties as in the scalar setting. The next result collects the easy facts.

Exercise 16.22 (Matrix mean: Representer properties). Consider a matrix mean $\boldsymbol{M}$ on $\mathbb{H}_n^+$. Introduce the representer function $f(t)\mathbf{I} = \boldsymbol{M}(\mathbf{I}, t\mathbf{I})$ for $t \geq 0$. Prove that the representer enjoys the following properties.

1. **Strict positivity.** The representer $f(t) > 0$ for $t > 0$.
2. **Normalization.** The representer take the value $f(1) = 1$.
3. **Monotonicity.** The representer $t \mapsto f(t)$ is monotone.
4. **Subadditivity.** The function $t \mapsto f(t)/t$ is decreasing for $t > 0$.
5. **Continuity.** The representer f is continuous.
6. **Symmetry.** If the mean $\boldsymbol{M}$ is *symmetric*, then $f(t) = t \cdot f(1/t)$ for all $t > 0$.

The arguments are essentially the same as in the scalar case.

16.4.2 Monotonicity of the matrix mean representer

Although we have defined the matrix mean representer in terms of scalar matrices, it actually captures the action of the mean between the identity and any psd matrix. As a consequence, we discover that the matrix mean representer is a matrix monotone function on matrices of an appropriate dimension.

Theorem 16.23 (Matrix mean representer). Consider a matrix mean $\boldsymbol{M}$ on $\mathbb{H}_n^+$ with representer $f : \mathbb{R}_+ \to \mathbb{R}_+$. Then the representer function satisfies

$$f(\boldsymbol{B}) = \boldsymbol{M}(\mathbf{I}, \boldsymbol{B}) \quad \text{for all } \boldsymbol{B} \in \mathbb{H}_n^+.$$

In particular, the representer function is *matrix monotone on* $\mathbb{H}_n^+$.

As a consequence, a scalar mean representer need not be a matrix mean representer because scalar monotonicity ($n = 1$) does not imply matrix monotonicity ($n > 1$).

We begin with a lemma that describes how matrix means interact with orthogonal projectors.

Lemma 16.24 (Matrix mean: Commuting projectors). Under the assumptions of Theorem 16.23, let $\boldsymbol{P} \in \mathbb{H}_n^+$ be an orthogonal projector that commutes with $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+$. Then

$$\boldsymbol{P}\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{P} = \boldsymbol{M}(\boldsymbol{A}\boldsymbol{P}, \boldsymbol{B}\boldsymbol{P})\boldsymbol{P}.$$

That is, the projector commutes with the mean and satisfies a conjugation property.

Proof. First, we show that $\boldsymbol{P}$ commutes with the mean $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$. Indeed, since $\boldsymbol{P}$ commutes with $\boldsymbol{A}$, we have the relation $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}^{1/2}\boldsymbol{P}\boldsymbol{A}^{1/2} \preccurlyeq \boldsymbol{A}$. Likewise, $\boldsymbol{B}\boldsymbol{P} \preccurlyeq \boldsymbol{B}$. By monotonicity of the mean, $\boldsymbol{M}(\boldsymbol{A}\boldsymbol{P}, \boldsymbol{B}\boldsymbol{P}) \preccurlyeq \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$. Conjugate by the projector to see that

$$\boldsymbol{P}\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{P} = \boldsymbol{M}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}, \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}) = \boldsymbol{M}(\boldsymbol{A}\boldsymbol{P}, \boldsymbol{B}\boldsymbol{P}) \preccurlyeq \boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}).$$

We have used the conjugation property of the mean at the first step.

Equivalently, we have shown that

$$\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B}) - \boldsymbol{P}\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{P} \succcurlyeq \mathbf{0}.$$

To see what is going on, assume that $\operatorname{range}(\boldsymbol{P})$ is a subspace spanned by coordinates. Then we are removing a "diagonal block" from the psd matrix $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$, and yet we are left with a psd matrix. This can only happen if the "off-diagonal blocks" are zero too.

Since $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$ is psd, this statement implies that the restriction of the matrix $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$ to $\operatorname{range}(\boldsymbol{P})$ is zero. In particular,

$$(\mathbf{I} - \boldsymbol{P})\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{P} = \mathbf{0} = \boldsymbol{P}\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})(\mathbf{I} - \boldsymbol{P}).$$

The last relation is the conjugate transpose of the first. Rearrange this expression to see that $\boldsymbol{P}$ commutes with $\boldsymbol{M}(\boldsymbol{A}, \boldsymbol{B})$.

By a similar argument, $\boldsymbol{P}$ also commutes with $\boldsymbol{M}(\boldsymbol{A}\boldsymbol{P}, \boldsymbol{B}\boldsymbol{P})$. These two facts complete the proof. ∎

With this lemma at hand, we may now establish Theorem 16.23.

Proof of Theorem 16.23. Fix a matrix $\boldsymbol{B} \in \mathbb{H}_n^+$, and introduce the spectral resolution $\boldsymbol{B} = \sum_i \lambda_i \boldsymbol{P}_i$. Since the projectors decompose the identity,

$$\begin{aligned}\boldsymbol{M}(\mathbf{I}, \boldsymbol{B}) &= \sum\nolimits_i \boldsymbol{M}(\mathbf{I}, \boldsymbol{B})\boldsymbol{P}_i = \sum\nolimits_i \boldsymbol{M}(\boldsymbol{P}_i, \boldsymbol{A}\boldsymbol{P}_i)\boldsymbol{P}_i \\ &= \sum\nolimits_i \boldsymbol{M}(\boldsymbol{P}_i, \lambda_i \boldsymbol{P}_i)\boldsymbol{P}_i = \sum\nolimits_i \boldsymbol{M}(\mathbf{I}, \lambda_i \mathbf{I})\boldsymbol{P}_i \\ &= \sum\nolimits_i f(\lambda_i)\boldsymbol{P}_i = f(\boldsymbol{B}).\end{aligned}$$

We have used Lemma 16.24 in the second and third lines. To pass from the second line to the third, we used the fact that the the range of the projector $\boldsymbol{P}_i$ is an eigenspace of $\boldsymbol{B}$ with eigenvalue λ_i. Last, we have recognized the matrix mean representer, and we applied the definition of a standard matrix function.

Finally, consider $n \times n$ matrices with $\mathbf{0} \preccurlyeq \boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2$. Then

$$f(\boldsymbol{B}_1) = \boldsymbol{M}(\mathbf{I}, \boldsymbol{B}_1) \preccurlyeq \boldsymbol{M}(\mathbf{I}, \boldsymbol{B}_2) = f(\boldsymbol{B}_2).$$

We have invoked the axiom that the matrix mean is monotone on $\mathbb{H}_n^+$. ∎

16.5 Matrix means from matrix representers

As in the scalar case, our next task is to reverse the process and try to construct a matrix mean from a matrix representer function.

16.5.1 Matrix perspective transformations

Our goal is to use the univariate matrix mean representer to construct a bivariate matrix mean. As with the matrix geometric mean, we recognize that the conjugation axiom forces our hand. Therefore, the structure of the matrix mean is already determined by its representer.

Definition 16.25 (Matrix perspective transformation). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a (continuous) function. The *matrix perspective* of f is the bivariate matrix function

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{A}^{1/2} \cdot f(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}) \cdot \boldsymbol{A}^{1/2} \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}. \tag{16.1}$$

In particular,

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A} \cdot f(\boldsymbol{B}\boldsymbol{A}^{-1}) \quad \text{for } \textit{commuting } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

If f is subadditive, we may extend the perspective to $\mathbb{H}_n^+ \times \mathbb{H}_n^+$ by taking limits.

Regardless of the choice of the standard matrix function f, the perspective transformation $\boldsymbol{M}_f$ interacts nicely with conjugation. When the function f has additional properties, the perspective $\boldsymbol{M}_f$ may inherit some of these features. This section contains some elaboration, and we will continue this discussion in the next lecture.

Although the form (16.1) of the perspective transformation is motivated by the conjugation axiom, it is not immediate that the perspective satisfies the conjugation property. The first proposition guarantees that it does.

Proposition 16.26 (Matrix perspective: Conjugation). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a (continuous) function. Then the perspective transformation $\boldsymbol{M}_f$ satisfies the conjugation axiom. For

all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}$,

$$\boldsymbol{M}_f(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}, \boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) = \boldsymbol{X}^*\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{X} \quad \text{for all } \boldsymbol{X} \in \mathbb{M}_n.$$

If f is subadditive, this expression extends to all psd matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+$.

Proof. Assume that $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}$ are positive definite, and fix a nonsingular matrix $\boldsymbol{X} \in \mathbb{M}_n$. The remaining cases will follow from continuity.

The quantity of interest takes the unwieldy form

$$\begin{aligned}\boldsymbol{M}_f(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}, \boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) \\ = (\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X})^{1/2} \cdot f\big((\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X})^{-1/2}(\boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X})(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X})^{-1/2}\big) \cdot (\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X})^{1/2}.\end{aligned}$$

To tame this expression, introduce the matrix $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X})^{-1/2}$. It has the polar factorization $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{U}$ where $\boldsymbol{P} = (\boldsymbol{Y}\boldsymbol{Y}^*)^{1/2}$ and $\boldsymbol{U}$ is unitary. After a short calculation, we find that $\boldsymbol{P} = \boldsymbol{A}^{-1/2}$. Therefore,

$$\begin{aligned}\boldsymbol{M}_f(\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}, \boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) &= \boldsymbol{X}^*\boldsymbol{Y}^{-*} \cdot f(\boldsymbol{Y}^*\boldsymbol{B}\boldsymbol{Y}) \cdot \boldsymbol{Y}^{-1}\boldsymbol{X} \\ &= \boldsymbol{X}^*\boldsymbol{P}^{-1}\boldsymbol{U} \cdot f(\boldsymbol{U}^*(\boldsymbol{P}\boldsymbol{B}\boldsymbol{P})\boldsymbol{U}) \cdot \boldsymbol{U}^*\boldsymbol{P}^{-1}\boldsymbol{X} \\ &= \boldsymbol{X}^*\boldsymbol{A}^{1/2} \cdot f(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}) \cdot \boldsymbol{A}^{1/2}\boldsymbol{X} = \boldsymbol{X}^*\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{X}.\end{aligned}$$

We have used the unitary equivariance of the standard matrix function to eliminate the unitary matrices. ∎

An important corollary of the conjugation invariance property is that the perspective transformation of a matrix monotone function is monotone in both arguments.

Corollary 16.27 (Matrix perspective: Monotonicity). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous *matrix monotone* function. Then each variable of the perspective transformation $\boldsymbol{M}_f$ is matrix monotone on $\mathbb{H}_n^+$:

$$\begin{aligned}\mathbf{0} \preccurlyeq \boldsymbol{A}_1 \preccurlyeq \boldsymbol{A}_2 \quad &\text{implies} \quad \boldsymbol{M}_f(\boldsymbol{A}_1, \boldsymbol{B}) \preccurlyeq \boldsymbol{M}_f(\boldsymbol{A}_2, \boldsymbol{B}) \quad \text{for all } \boldsymbol{B} \succcurlyeq \mathbf{0}; \\ \mathbf{0} \preccurlyeq \boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2 \quad &\text{implies} \quad \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}_1) \preccurlyeq \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}_2) \quad \text{for all } \boldsymbol{A} \succcurlyeq \mathbf{0}.\end{aligned}$$

Proof. First, we remark that a positive matrix monotone function is always subadditive. Therefore, we can take limits to extend the perspective $\boldsymbol{M}_f$ from positive-definite matrices to psd matrices.

Indeed, a matrix monotone function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is concave, so it must be subadditive.

The matrix monotonicity of $\boldsymbol{B} \mapsto \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})$ is an easy consequence of the expression (16.1) for the perspective $\boldsymbol{M}_f$, the conjugation rule, and the matrix monotonicity of f.

As for the other variable, we may assume that $\boldsymbol{A}, \boldsymbol{B}$ are positive definite. Then the conjugation property (Proposition 16.26) ensures that

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{B}^{1/2}\boldsymbol{M}_f(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}, \mathbf{I})\boldsymbol{B}^{1/2}.$$

From the definition (16.1) of the perspective, we find that

$$\boldsymbol{M}_f(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}, \mathbf{I}) = (\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}) \cdot f\big((\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})^{-1}\big).$$

Since f is matrix monotone, the function $t \mapsto t \cdot f(1/t)$ is also matrix monotone. (See Problem 16.28.) Therefore, $\boldsymbol{A} \mapsto \boldsymbol{M}_f(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}, \mathbf{I})$ is matrix monotone. By the conjugation rule, we conclude that $\boldsymbol{A} \mapsto \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})$ is also matrix monotone. ∎

Problem 16.28 (Matrix perspective: Monotonicity property). Suppose that $f : \mathbb{R}_{++} \to \mathbb{R}_+$ is matrix monotone. For each contraction $\boldsymbol{K}$, prove that

$$\boldsymbol{K}^* f(\boldsymbol{A})\boldsymbol{K} \preccurlyeq f(\boldsymbol{K}^*\boldsymbol{A}\boldsymbol{K}) \quad \text{for all psd } \boldsymbol{A}.$$

Deduce that $g(t) = t \cdot f(1/t)$ is matrix monotone on $\mathbb{R}_{++}$. **Hint:** If you exploit the fact that f is matrix concave, then the proof is easy. But this argument may be circular, so you should give an independent proof; see Exercise **??**.

In general, the perspective function $\boldsymbol{M}_f$ treats its two arguments rather differently. In the context of matrix means, it is valuable to understand when the perspective is symmetric in its arguments. The next exercise states the result.

Exercise 16.29 (Matrix perspective: Symmetry). Suppose that $f : \mathbb{R}_+ \to \mathbb{R}_+$ has the property that $f(t) = t \cdot f(1/t)$ for $t > 0$. Prove that

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{M}_f(\boldsymbol{B}, \boldsymbol{A}) \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

Hint: This point follows easily from the conjugation property, much like the result of Corollary 16.27.

16.5.2 Families of matrix means

In this lecture, we will be interested in what happens when we take the perspective transformation of a matrix mean representer. The next definition collects the properties that we need.

Definition 16.30 (Matrix mean representer). We say that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a *matrix mean representer* when f is a (continuous) matrix monotone function that satisfies the normalization $f(1) = 1$.

Definition 16.31 (Matrix mean family). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a matrix mean representer, as in Definition 16.30. Using the matrix perspective, we can define a bivariate function on matrices of any dimension $n \in \mathbb{N}$:

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{A}^{1/2} \cdot f(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}) \cdot \boldsymbol{A}^{1/2} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++}.$$

We extend to psd matrices by continuity. We call $\boldsymbol{M}_f$ a *family of matrix means*.

Our main result states that the matrix perspective of a matrix mean representer generates a family of matrix means. We have already laid most of the groundwork, so this result will follow quickly.

Theorem 16.32 (Matrix representers yield matrix means). Let f be a matrix mean representer, as in Definition 16.30, and let $\boldsymbol{M}_f$ be the associated family of matrix means, as in Definition 16.31.

Then the $\boldsymbol{M}_f$ is a matrix mean on $\mathbb{H}_n^+$ for each $n \in \mathbb{N}$ with matrix mean representer f. Indeed, $f \leftrightarrow \boldsymbol{M}_f$ is a bijection between matrix representer functions and families of matrix means.

Moreover, if $f(t) = t \cdot f(1/t)$ for $t > 0$, then the mean $\boldsymbol{M}_f$ is symmetric in its arguments.

Proof. When f is a matrix mean representer, we can argue that $\boldsymbol{M}_f$ is a matrix mean on matrices of any dimension $n \in \mathbb{N}$.

Positivity. Strict positivity of $\boldsymbol{M}_f$ follows from the fact that a matrix monotone function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(1) = 1$ is strictly positive.

Order. The order properties are straightforward. For example, to see that $\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{B}$ implies that $\boldsymbol{A} \preccurlyeq \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})$, we just need to check that $\mathbf{I} \preccurlyeq f(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})$. But this is a consequence of the fact that $\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2} \preccurlyeq \mathbf{I}$ and the normalization $f(1) = 1$.

To check that $\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{B}$ implies that $\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) \preccurlyeq \boldsymbol{B}$, we invoke the conjugation property (Proposition 16.26) to obtain the equivalent relation $f(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}) \preccurlyeq \mathbf{I}$, which we have already verified. The other cases are essentially the same.

Monotonicity. Corollary 16.27 already establishes the monotonicity property.

Conjugation. The conjugation axiom was obtained in Proposition 16.26.

Continuity. The function $\boldsymbol{M}_f$ is continuous on positive-definite matrices since f is continuous. It is continuous for psd matrices by construction.

Symmetry. Exercise 16.29 requests the proof of the symmetry property.

Bijection. Finally, note that $\boldsymbol{M}_f(\mathbf{I}, t\mathbf{I}) = f(t) \cdot \mathbf{I}$ for all $t > 0$. In other words, f is the matrix mean representer of $\boldsymbol{M}_f$. Since $\boldsymbol{M}_f$ is a matrix mean for each dimension $n \in \mathbb{N}$, the function f must be matrix monotone, continuous, and normalized with $f(1) = 1$. We conclude that $f \leftrightarrow \boldsymbol{M}_f$ is a bijection. ∎

16.5.3 Examples

Examples of families of matrix means include the matrix arithmetic mean, the left and right matrix mean, the matrix harmonic mean, and the matrix geometric mean. There are other important examples.

Example 16.33 (Weighted matrix geometric means). For a parameter $r \in [0, 1]$, we may consider the matrix monotone function $f(t) = t^r$. This function induces a family of weighted geometric means:

$$\boldsymbol{A} \sharp_r \boldsymbol{B} := \boldsymbol{A}^{1/2} \cdot \left(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\right)^r \cdot \boldsymbol{A}^{1/2} \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{++} \text{ and } n \in \mathbb{N}.$$

We can extend to all psd matrices by taking limits. The weighted geometric means interpolate between the left and right mean. With respect to an appropriate geometry on psd matrices, the weighted geometric means $r \mapsto \boldsymbol{A} \sharp_r \boldsymbol{B}$ trace out a geodesic between $\boldsymbol{A}$ and $\boldsymbol{B}$. We recognize that ordinary matrix geometric mean $\boldsymbol{A} \sharp \boldsymbol{B}$ as the midpoint of this geodesic. ▪

Example 16.34 (Matrix logarithmic mean). The function $f(t) = \int_0^1 t^r \, \mathrm{d}r = (t - 1)/\log(t)$ represents the scalar logarithmic mean. This function is matrix monotone and satisfies the normalization $f(1) = 1$, so it induces a family of symmetric matrix means, most easily written using the weighted geometric mean:

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \int_0^1 (\boldsymbol{A} \sharp_r \boldsymbol{B}) \, \mathrm{d}r \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^{+} \text{ and } n \in \mathbb{N}.$$

Is there an alternative expression that makes the role of the logarithm clear? ▪

16.5.4 Integral representations

Theorem 16.32 allows us to draw on Loewner's theory of matrix monotone functions. Recall that every (continuous) matrix monotone function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(1) = 1$ has an integral representation:

$$f(\boldsymbol{A}) = \alpha + \beta\boldsymbol{A} + \int_0^\infty \frac{\lambda \boldsymbol{A}}{\boldsymbol{A} + \lambda \mathbf{I}} \cdot \frac{1 + \lambda}{\lambda} \, \mathrm{d}\mu(\lambda) \quad \text{for all psd } \boldsymbol{A}.$$

The coefficients $\alpha, \beta \geq 0$ and μ is a finite, positive Borel measure on $\mathbb{R}_{++}$. The normalization ensures that $\alpha+\beta+\mu(\mathbb{R}_{++}) = 1$. This fact has a spectacular consequence for matrix means.

Exercise 16.35 (Matrix mean: Integral representation). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous matrix monotone function with $f(1) = 1$. Then the associated family $\boldsymbol{M}_f$ of matrix means takes the form

$$\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \alpha\boldsymbol{A} + \beta\boldsymbol{B} + \int_0^\infty 2(\lambda\boldsymbol{A} : \boldsymbol{B}) \cdot \frac{1+\lambda}{2\lambda} \,\mathrm{d}\mu(\lambda)$$

for all psd $\boldsymbol{A}, \boldsymbol{B}$ with the same dimension. Here, $\boldsymbol{A} : \boldsymbol{B}$ denotes the parallel sum. The coefficients $\alpha, \beta \geq 0$ and μ is a finite, positive Borel measure on $\mathbb{R}_{++}$. The normalization ensures that $\alpha + \beta + \mu(\mathbb{R}_{++}) = 1$.

Find a simplification for the symmetric case where $\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{M}_f(\boldsymbol{B}, \boldsymbol{A})$.

Exercise 16.35 has the remarkable interpretation that every family $\boldsymbol{M}_f$ of matrix means consists of a conic combination of the left mean, the right mean, and a family of harmonic means. We can deduce other significant results as well.

Exercise 16.36 (Minimal and maximal means). The harmonic mean is the minimal matrix mean, while the arithmetic mean is the maximal matrix mean. That is, for a family $\boldsymbol{M}_f$ of symmetric matrix means,

$$2(\boldsymbol{A} : \boldsymbol{B}) \preccurlyeq \boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B}) \preccurlyeq \tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{B})$$

for all psd $\boldsymbol{A}, \boldsymbol{B}$ with the same dimension.

Exercise 16.37 (Matrix means: Concavity). Let $\boldsymbol{M}_f$ be a family of matrix means. For all psd $\boldsymbol{A}_i, \boldsymbol{B}_i$ with the same dimension ($i = 1, 2$),

$$\boldsymbol{M}_f\left(\tau\boldsymbol{A}_1 + \bar{\tau}\boldsymbol{A}_2, \tau\boldsymbol{B}_1 + \bar{\tau}\boldsymbol{B}_2\right) \succcurlyeq \tau\boldsymbol{M}_f(\boldsymbol{A}_1, \boldsymbol{B}_1) + \bar{\tau}\boldsymbol{M}_f(\boldsymbol{A}_2, \boldsymbol{B}_2) \quad \text{for all } \tau \in [0, 1].$$

As usual, $\bar{\tau} := 1 - \tau$. **Hint:** The parallel sum is matrix concave.

We can strengthen the last exercise substantially.

Problem 16.38 (Matrix means: Filtering). Consider a family $\boldsymbol{M}_f$ of matrix means. Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_n$ be a positive linear map (not necessarily unital!). Prove that

$$\boldsymbol{M}_f(\boldsymbol{\Phi}(\boldsymbol{A}), \boldsymbol{\Phi}(\boldsymbol{B})) \succcurlyeq \boldsymbol{\Phi}(\boldsymbol{M}_f(\boldsymbol{A}, \boldsymbol{B})).$$

Hint: The parallel sum satisfies the same concavity property. First, show that this claim holds for a unital, strictly positive linear map by invoking Choi's inequality. Remove the extra conditions by emulating the proof of the Russo–Dye theorem (Lecture 12).

Notes

The results in this lecture are drawn from Kubo & Ando [KA79] and from Bhatia's book [Bha07b, Chap. 4]. The arrangement of material is somewhat different from these sources. It may be conceptually simpler to regard the basic object as a family of means induced by a function and then to derive the properties from this representation.

Lecture bibliography

[AJD69] W. N. Anderson Jr and R. J. Duffin. "Series and parallel addition of matrices". In: *Journal of Mathematical Analysis and Applications* 26.3 (1969), pages 576–594.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[HK99] F. Hiai and H. Kosaki. "Means for matrices and comparison of their norms". In: *Indiana Univ. Math. J.* 48.3 (1999), pages 899–936. DOI: 10.1512/iumj.1999.48.1665.

[KA79] F. Kubo and T. Ando. "Means of positive linear operators". In: *Math. Ann.* 246.3 (1979/80), pages 205–224. DOI: 10.1007/BF01371042.

17. Quantum Relative Entropy

Date: 1 March 2022 **Scribe:** Eray Unsal Atay

Agenda:
1. Scalar entropies
2. Matrix entropies
3. Matrix perspectives
4. Tensors and logarithms
5. Convexity of quantum relative entropy

In this lecture, we study matrix entropy functions, also known as *quantum entropies*. Quantum entropies arise in quantum information theory and quantum statistical physics. They also have remarkable applications in random matrix theory and data science. Quantum entries have deep roots in matrix analysis because matrix monotone and convex functions play an important role in the analysis.

We start with some background from information theory in the scalar setting. We introduce the (scalar) entropy function and the relative entropy, and we discuss their basic properties. These definitions extend to the matrix setting, and we will see that matrix entropies have many properties that parallel those of scalar entropies. Nevertheless, the properties of matrix entropies are far more difficult to establish. We show how to develop these results using matrix perspective transformations, combined with the idea of lifting a matrix problem to tensors.

We will present two major results in this lecture. The first is the convexity of the matrix perspective transformation of a matrix convex function. The second result is the convexity of quantum relative entropy, which represents a nontrivial application of matrix perspectives. This theorem is one of the crown jewels of matrix analysis.

17.1 Entropy and relative entropy

This section provides the definitions of entropy and relative entropy in the scalar setting. The entropy measures the disorder in a probability distribution, while the relative entropy reflects the difference between two probability distributions.

17.1.1 Probability distributions and entropy

We begin with the definition of the probability simplex.

Definition 17.1 (Probability simplex). Define the *probability simplex* Δ_n in $\mathbb{R}^n$ to be the set of positive vectors whose entries add up to one:

$$\Delta_n := \{\boldsymbol{p} \in \mathbb{R}^n : \boldsymbol{p} \geq \mathbf{0} \text{ and } \operatorname{tr}(\boldsymbol{p}) = 1\}.$$

In this expression, $\geq$ is the entrywise inequality. Each vector $\boldsymbol{p} \in \Delta_n$ models a probability distribution on $\{1, \dots, n\}$.

The set Δ_n is closed and convex in $\mathbb{R}^n$. Its extreme points are the standard basis vectors $\boldsymbol{\delta}_i$ for $i = 1, \dots, n$.

The entropy is a function on the probability simplex Δ_n that measures the amount of randomness in a probability distribution.

Definition 17.2 (Entropy). The *entropy* of a probability distribution $\boldsymbol{p} \in \Delta_n$ is

$$\operatorname{ent}(\boldsymbol{p}) := -\sum\nolimits_{i=1}^{n} p_i \log p_i.$$

We instate the convention that $0 \log 0 = 0$.

Recall that *negative* entropy is an isotone (i.e., Schur convex) function. The theory of isotone functions implies that

$$0 \le \operatorname{ent}(\boldsymbol{p}) \le \log n \quad \text{for each } \boldsymbol{p} \in \Delta_n. \tag{17.1}$$

The minimum in (17.1) is achieved when $\boldsymbol{p} = \boldsymbol{\delta}_i$ for some $i \in \{1, \dots, n\}$. These probability distributions describe constant (i.e., deterministic) random variables. The maximum is achieved when $\boldsymbol{p} = n^{-1}\mathbf{1}$, which models a uniform random variable. In other words, the entropy *increases* as the disorder of a probability distribution *increases*.

Let us provide some historical background. The concept of entropy emerged in statistical physics and thermodynamics, due to work by Gibbs and Boltzmann. Later, entropy became one of the core tools in information theory after Shannon [Sha48] found operational interpretations.

Fact 17.3 (Shannon 1948). The entropy $\operatorname{ent}(\boldsymbol{p})$ of a probability distribution $\boldsymbol{p} \in \Delta_n$ is (proportional to) the average number of bits per symbol to encode a sequence of iid random variables that are distributed according to $\boldsymbol{p}$. ∎

In other words, assume that we draw an infinite sequence of independent random variables, each distributed according to $\boldsymbol{p}$. If the distribution is deterministic, we can specify the whole sequence by providing its constant value; asymptotically, this amounts to 0 bits per element of the sequence. On the other hand, for the uniform distribution, it takes $\log_2 n$ bits on average to represent each value in the sequence.

17.1.2 Relative entropy

The key object in today's lecture is the relative entropy, which is a measure of the difference between two probability distributions.

Definition 17.4 (Relative entropy). The *relative entropy* or *Kullback–Leibler (KL) divergence* between two probability distributions $\boldsymbol{p}, \boldsymbol{q} \in \Delta_n$ is given by the expression

$$\mathrm{D}(\boldsymbol{p}; \boldsymbol{q}) := \sum\nolimits_{i=1}^{n} p_i (\log p_i - \log q_i).$$

Exercise 17.5 (Relative entropy). Verify that the relative entropy has the following basic properties.

1. **Positivity.** Explain why $\mathrm{D}(\boldsymbol{p}; \boldsymbol{q}) \ge 0$ for all $\boldsymbol{p}, \boldsymbol{q} \in \Delta_n$.
2. **Unboundedness.** Check that $\mathrm{D}(\boldsymbol{p}; \boldsymbol{q})$ can take the value $+\infty$.
3. **Asymmetry.** Show that $\mathrm{D}(\boldsymbol{p}; \boldsymbol{q})$ is *not* symmetric in its arguments, so it is not a metric.

Relative entropy also has important operational interpretations in information theory and statistics.

Fact 17.6 (Stein's lemma). Fix a distribution $\boldsymbol{q} \in \Delta_n$. If the relative entropy $\mathrm{D}(\boldsymbol{p}; \boldsymbol{q}) < +\infty$, then $\mathrm{D}(\boldsymbol{p}; \boldsymbol{q})^{-1}$ is roughly the number of independent samples we need from the distribution $\boldsymbol{p}$ in order to decide with high probability that $\boldsymbol{p} \neq \boldsymbol{q}$. ∎

The key fact about relative entropy is that it is a convex function.

Exercise 17.7 (Relative entropy is convex). The function $(\boldsymbol{p}; \boldsymbol{q}) \mapsto \mathrm{D}(\boldsymbol{p}; \boldsymbol{q})$ is convex on $\Delta_n \times \Delta_n$. **Hint:** Interpret $(a, b) \mapsto a \log(a/b)$, defined on $\mathbb{R}_{++} \times \mathbb{R}_{++}$, as a perspective transformation of $-\log$.

The result in Exercise 17.7 is the primary motivation for this lecture. It has important applications in optimization theory. For example, it plays a key role in the study of geometric programming and relative entropy programming.

Exercise 17.8 (Information projections). For a closed, convex set $\mathsf{C} \subseteq \Delta_n$, characterize the solution of the minimization problem $\min_{\boldsymbol{p} \in \mathsf{C}} \mathrm{D}(\boldsymbol{p}; \boldsymbol{q})$. The result is analogous with the characterization of the Euclidean projection onto the convex set C.

17.2 Quantum entropy and quantum relative entropy

In this section, we generalize the notions we introduced in the scalar setting to matrices. The quantum entropy reflects the variability of the eigenvalues of a (normalized) psd matrix. The quantum relative entropy describes the discrepancy between two (normalized) psd matrices.

17.2.1 Density matrices and matrix entropy

We start by defining density matrices, which generalize probability distributions.

Definition 17.9 (Density matrices). An $n \times n$ *density matrix* is a psd matrix $\boldsymbol{\varrho} \in \mathbb{H}_n$ with trace one. Define the set $\boldsymbol{\Delta}_n$ of $n \times n$ density matrices:

$$\boldsymbol{\Delta}_n := \{\boldsymbol{\varrho} \in \mathbb{H}_n : \boldsymbol{\varrho} \succcurlyeq \mathbf{0} \text{ and } \operatorname{tr}(\boldsymbol{\varrho}) = 1\}.$$

We can think about a density matrix as a "quantum" probability distribution, which describes the state of a quantum system. The set of density matrices is the quantum version of the probability simplex.

Exercise 17.10 (Density matrices). If $\boldsymbol{\varrho} \in \boldsymbol{\Delta}_n$, show that $\boldsymbol{\lambda}^{\downarrow}(\boldsymbol{\varrho}) \in \Delta_n$. Confirm that the extreme points of $\boldsymbol{\Delta}_n$ are the rank-one density matrices, which are called *pure states*. Pure states take the form $\boldsymbol{\varrho} = \boldsymbol{u}\boldsymbol{u}^*$, where $\|\boldsymbol{u}\| = 1$.

Next, we introduce the entropy of a density matrix.

Definition 17.11 (von Neumann). The *quantum entropy* of a density matrix $\boldsymbol{\varrho} \in \boldsymbol{\Delta}_n$ is

$$\operatorname{ent}(\boldsymbol{\varrho}) := -\operatorname{tr}(\boldsymbol{\varrho} \log \boldsymbol{\varrho}) = -\sum\nolimits_{i=1}^{n} \lambda_i \log \lambda_i,$$

where $\lambda_1, \dots, \lambda_n$ are the (decreasingly ordered) eigenvalues of $\boldsymbol{\varrho}$.

The quantum entropy is given by the eigenvalues of the density matrix, which compose a probability distribution. Quantum entropy then addresses the question "How disordered are the eigenvalues of a density matrix?" In other words, quantum entropy offers a way to measure the disorder in a quantum system with state $\boldsymbol{\varrho}$.

From our discussion, we immediately obtain the following bounds on the quantum entropy.

$$0 \le \operatorname{ent}(\boldsymbol{\varrho}) \le \log n \quad \text{for each } \boldsymbol{\varrho} \in \boldsymbol{\Delta}_n. \tag{17.2}$$

Notice that the bounds on entropy are the same in the scalar case in (17.1) and in the matrix case in (17.2).

The minimum occurs if and only if $\boldsymbol{\varrho}$ is a rank-1 matrix (that is, a pure state). The maximum occurs if and only if $\boldsymbol{\varrho} = n^{-1}\mathbf{I}_n$.

There are many operational interpretations of the quantum entropy in quantum information theory, but they are outside the scope of this lecture.

17.2.2 Matrix relative entropy

Next, we extend the notion of relative entropy to the matrix setting. The next definition is due to Umegaki. It describes one way of comparing the discrepancy between two quantum states.

Definition 17.12 (Umegaki). The (Umegaki) *quantum relative entropy* of two density matrices $\boldsymbol{\varrho}, \boldsymbol{\nu} \in \boldsymbol{\Delta}_n$ is

$$\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu}) := \operatorname{tr}[\boldsymbol{\varrho}(\log \boldsymbol{\varrho} - \log \boldsymbol{\nu})].$$

Aside: There are several other notions of relative entropy in the quantum setting.

Warning 17.13 (Relative entropy and eigenvectors). Unless $\boldsymbol{\varrho}$ and $\boldsymbol{\nu}$ commute, their eigenvalues alone do *not* determine the value of the quantum relative entropy $\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu})$! The interactions between the eigenvectors also play a role. ▪

Exercise 17.14 (Quantum relative entropy is positive). For all $\boldsymbol{\varrho}, \boldsymbol{\nu} \in \boldsymbol{\Delta}_n$, prove that $\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu}) \geq 0$. **Hint:** Use the generalized Klein inequality (Lecture 8).

Quantum relative entropy also has operational interpretations in quantum information theory. For example, we have the quantum extension of Stein's lemma due to Hiai & Petz [HP91] and to Ogawa & Nagaoka [ONo0].

Fact 17.15 (Quantum Stein's lemma). Let $\boldsymbol{\nu} \in \boldsymbol{\Delta}_n$ be a density matrix. If $\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu}) < +\infty$, then $\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu})^{-1}$ is (roughly) the number of unentangled quantum systems, prepared in state $\boldsymbol{\varrho}$, that we must measure to determine that $\boldsymbol{\varrho} \neq \boldsymbol{\nu}$ with high probability. ▪

Having introduced the matrix generalizations of the relative entropy, we can state a major theorem that extends the convexity property of scalar entropy.

Theorem 17.16 (Convexity of quantum relative entropy). The map $(\boldsymbol{\varrho}; \boldsymbol{\nu}) \mapsto \mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu})$ is convex on $\boldsymbol{\Delta}_n \times \boldsymbol{\Delta}_n$.

Exercise 17.17 (Quantum versus scalar). Show that the convexity of quantum relative entropy (Theorem 17.16) implies the convexity of scalar relative entropy (Exercise 17.7). **Hint:** Consider diagonal density matrices.

Unlike the scalar result in Exercise 17.7, Theorem 17.16 is quite hard to prove. It was first obtained by Lindblad [Lin73] using results of Lieb [Lie73a]. In this lecture, we present a proof due to Effros [Eff09] that is based on matrix perspective transformations. A key step in this argument is to lift the problem to tensor products, an idea that first appeared in Ando's beautiful paper [And79] of the convexity of quantum relative entropy and related functions.

Theorem 17.16 has many applications in quantum information theory and quantum statistical physics. In particular, it plays the starring role in the proof that quantum entropy is strongly subadditive [LR73]. The convexity of quantum relative entropy is also the main ingredient in the theory of exponential matrix concentration developed by the lecturer [Tro11; Tro15].

17.3 The matrix perspective transformation

In the scalar setting, the convexity of the relative entropy can be established using perspective transformations (Exercise 17.7). This motivates us to develop a matrix extension of the perspective transformation and to investigate its properties.

17.3.1 Definition and examples

We begin with the definition and some examples.

Definition 17.18 (Matrix perspective transformation). Let $f : \mathbb{R}_{++} \to \mathbb{R}$. Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n^{++}$ be *strictly* positive-definite matrices. The *matrix perspective* of f is the bivariate function

$$\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) := \boldsymbol{A}^{1/2} f(\boldsymbol{A}^{-1/2} \boldsymbol{H} \boldsymbol{A}^{-1/2}) \boldsymbol{A}^{1/2}.$$

In particular, *if $\boldsymbol{A}$ and $\boldsymbol{H}$ commute*, then $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{A} f(\boldsymbol{H}\boldsymbol{A}^{-1})$.

Definition 17.18 may look familiar from Lecture 16. Indeed, if we assume that f is matrix monotone, positive, and satisfies some other inessential properties, then the construction of $\boldsymbol{\Psi}_f$ agrees with the Kubo–Ando matrix mean [KA79].

In this lecture, we study the matrix perspective transformation of a function f that is matrix convex. Let us take a moment to see what these functions look like.

Example 17.19 (Matrix perspectives). Here are some basic examples of matrix perspective transformations for several matrix convex functions $f : \mathbb{R}_{++} \to \mathbb{R}$.

- **Constant.** The function $f(t) = 1$ yields $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{A}$.
- **Identity.** The function $f(t) = t$ yields $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{H}$.
- **Square.** The function $f(t) = t^2$ yields $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{H}$.
- **Inverse.** The function $f(t) = t^{-1}$ yields $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{A}$.

We remark that $f(t) = 1$ and $f(t) = t$ are both matrix monotone, and they yield asymmetric matrix means that Kubo & Ando call the "left mean" and the "right mean." The other two examples, $f(t) = t^2$ and $f(t) = t^{-1}$, are matrix convex but not matrix monotone. These two functions are the extremal examples of matrix convex functions, and they yield perspectives that are attractively symmetrical with each other. ▪

This lecture involves matrix perspective transformations of the negative logarithm and some matrix convex power functions. We omit explicit expressions for now because we will apply the perspective in a particular setting where matters are simpler.

17.3.2 Convexity properties

If $f : \mathbb{R}_{++} \to \mathbb{R}$ is a scalar convex function, then you may recall that the scalar perspective $(a, h) \mapsto af(h/a)$ is a (jointly) convex function on $\mathbb{R}_{++} \times \mathbb{R}_{++}$. This result extends to matrix perspectives, provided that f is matrix convex. The basic result is due to Effros [Eff09]; we present a generalization due to Ebadian et al. [ENG11].

Theorem 17.20 (Matrix perspectives are matrix convex). Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be a *matrix convex* function. Consider *strictly* positive-definite matrices $\boldsymbol{A}_i, \boldsymbol{H}_i \in \mathbb{H}_n^{++}$ for $i = 1, 2$. Then, for all $\tau \in [0, 1]$ with $\bar{\tau} := 1 - \tau$, we have

$$\boldsymbol{\Psi}_f(\tau\boldsymbol{A}_1 + \bar{\tau}\boldsymbol{A}_2; \tau\boldsymbol{H}_1 + \bar{\tau}\boldsymbol{H}_2) \preccurlyeq \tau\boldsymbol{\Psi}_f(\boldsymbol{A}_1; \boldsymbol{H}_1) + \bar{\tau}\boldsymbol{\Psi}_f(\boldsymbol{A}_2; \boldsymbol{H}_2). \tag{17.3}$$

That is, the perspective is (jointly) operator convex on $\mathbb{H}_n^{++} \times \mathbb{H}_n^{++}$.

This theorem tells that the perspective of the averages is "smaller" than the average of the perspectives. It is the key to our proof of Theorem 17.16.

Proof. We will represent the left-hand side of (17.3) as a matrix convex combination, which allows us to invoke the matrix Jensen inequality. For notational convenience,

define

$$\boldsymbol{A} \coloneqq \tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2 \quad \text{and} \quad \boldsymbol{H} \coloneqq \tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2.$$

Introduce the coefficient matrices

$$\boldsymbol{K}_1 \coloneqq \tau^{1/2} \boldsymbol{A}_1^{1/2} \boldsymbol{A}^{-1/2} \quad \text{and} \quad \boldsymbol{K}_2 \coloneqq \bar{\tau}^{1/2} \boldsymbol{A}_2^{1/2} \boldsymbol{A}^{-1/2}.$$

By a short calculation, we find that

$$\boldsymbol{K}_1^* \boldsymbol{K}_1 = \tau \boldsymbol{A}^{-1/2} \boldsymbol{A}_1 \boldsymbol{A}^{-1/2} \quad \text{and} \quad \boldsymbol{K}_2^* \boldsymbol{K}_2 = \bar{\tau} \boldsymbol{A}^{-1/2} \boldsymbol{A}_2 \boldsymbol{A}^{-1/2}.$$

Adding these expressions,

$$\boldsymbol{K}_1^* \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{K}_2 = \mathbf{I}.$$

Thus, the coefficient matrices model a matrix convex combination.

Invoking the matrix Jensen inequality, we determine that

$$\begin{aligned}
\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) &= \boldsymbol{A}^{1/2} f(\boldsymbol{A}^{-1/2} \boldsymbol{H} \boldsymbol{A}^{-1/2}) \boldsymbol{A}^{1/2} \\
&= \boldsymbol{A}^{1/2} f\Big(\boldsymbol{K}_1^* (\boldsymbol{A}_1^{-1/2} \boldsymbol{H}_1 \boldsymbol{A}_1^{-1/2}) \boldsymbol{K}_1 + \boldsymbol{K}_2^* (\boldsymbol{A}_2^{-1/2} \boldsymbol{H}_2 \boldsymbol{A}_2^{-1/2}) \boldsymbol{K}_2 \Big) \boldsymbol{A}^{1/2} \\
&\preccurlyeq \boldsymbol{A}^{1/2} \Big[\boldsymbol{K}_1^* f(\boldsymbol{A}_1^{-1/2} \boldsymbol{H}_1 \boldsymbol{A}_1^{-1/2}) \boldsymbol{K}_1 + \boldsymbol{K}_2^* f(\boldsymbol{A}_2^{-1/2} \boldsymbol{H}_2 \boldsymbol{A}_2^{-1/2}) \boldsymbol{K}_2 \Big] \boldsymbol{A}^{1/2} \\
&= \tau \boldsymbol{\Psi}_f(\boldsymbol{A}_1; \boldsymbol{H}_1) + \bar{\tau} \boldsymbol{\Psi}_f(\boldsymbol{A}_2; \boldsymbol{H}_2).
\end{aligned}$$

This is just what we wanted to prove. ∎

Example 17.21 (Schwarz inequalities). Applying Theorem 17.20 to the perspective of $f : t \mapsto t^{-1}$ already yields an interesting statement. For $\boldsymbol{A}_i, \boldsymbol{H}_i \succ \mathbf{0}$,

$$\begin{aligned}
(\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2)(\tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2)^{-1} (\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2) \\
\preccurlyeq \tau \boldsymbol{A}_1 \boldsymbol{H}_1^{-1} \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2 \boldsymbol{H}_2^{-1} \boldsymbol{A}_2 \quad \text{for } \tau \in [0, 1].
\end{aligned}$$

This type of expression is called a *matrix Schwarz inequality*. In some sense, this quadratic over linear function is the extremal example of a matrix convex function. ▪

17.4 Tensors and logarithms

Ando [And79] had the magnificent idea to prove matrix convexity theorems by lifting formulas involving noncommuting matrices to formulas involving commuting tensors. This mechanism is easy to implement, and it eliminates most of the difficulty from the argument.

17.4.1 Tensors and multiplication operators

In this lecture, we take an approach to tensor product operators that is similar to the abstract approach in Lecture 1. The details are also slightly different from the concrete approach based on Kronecker products. The tensor product we describe here is isomorphic to the other constructions.

We will define elementary tensor product operators as multiplication operators acting on matrices.

Definition 17.22 (Multiplication operators). Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n$. We define the tensor operator $\boldsymbol{A} \otimes \boldsymbol{H} : \mathbb{M}_n \to \mathbb{M}_n$ via left-multiplication with $\boldsymbol{A}$ and right-multiplication with $\boldsymbol{H}$. That is,

$$(\boldsymbol{A} \otimes \boldsymbol{H})(\boldsymbol{M}) = \boldsymbol{A} \boldsymbol{M} \boldsymbol{H} \quad \text{for } \boldsymbol{M} \in \mathbb{M}_n.$$

The elementary tensor operators of the form $\boldsymbol{A} \otimes \boldsymbol{H}$ span the *real* linear space $\mathscr{L}(\mathbb{H}_n \otimes \mathbb{H}_n)$ of (self-adjoint) linear operators on $\mathbb{M}_n$.

Regardless of the choice of $\boldsymbol{A}, \boldsymbol{H}$, it is evident that the tensor operators $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{H}$ commute. That is,

$$(\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{H}) = (\mathbf{I} \otimes \boldsymbol{H})(\boldsymbol{A} \otimes \mathbf{I}) = (\boldsymbol{A} \otimes \boldsymbol{H}).$$

This commutativity property makes these elementary tensor operators quite pleasant to work with.

17.4.2 The logarithm of a tensor operator

Since $\boldsymbol{A} \otimes \boldsymbol{H}$ is a self-adjoint matrix acting on a Hilbert space, it has a spectral resolution, and we can define standard matrix functions in the usual way. In particular, there is an elegant expression for the logarithm of this tensor.

Exercise 17.23 (Logarithm of an elementary tensor operator). Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n^{++}$ be *strictly* positive-definite matrices. Then

$$\log(\boldsymbol{A} \otimes \boldsymbol{H}) = (\log \boldsymbol{A}) \otimes \mathbf{I} + \mathbf{I} \otimes (\log \boldsymbol{H}).$$

Hint: Introduce the spectral resolutions of $\boldsymbol{A}$ and $\boldsymbol{H}$.

This result provides a satisfactory substitute for the familiar properties of the scalar logarithm. There is an analogous result for exponentials.

Exercise 17.24 (Exponential of a tensor sum). Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_n$ be self-adjoint matrices. Prove that

$$\exp(\boldsymbol{A} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{H}) = \exp(\boldsymbol{A}) \otimes \exp(\boldsymbol{H}).$$

17.5 Convexity of matrix trace functions

At this stage, we can provide the proof of Theorem 17.16, which states that the quantum relative entropy is convex. We will also see that this approach yields other remarkable convexity theorems.

17.5.1 Proof of Theorem 17.16

We will show that the quantum relative entropy $(\boldsymbol{A}, \boldsymbol{H}) \mapsto \mathrm{S}(\boldsymbol{A}; \boldsymbol{H})$ is convex on pairs of *strictly* positive-definite matrices. The general result follows from a continuity argument.

We will condense the convexity of the quantum relative entropy from the matrix convexity of $f(t) := -\log t$. The perspective theorem implies that

$$(\boldsymbol{A}, \boldsymbol{H}) \mapsto \boldsymbol{\Psi}_f(\boldsymbol{A} \otimes \mathbf{I}; \mathbf{I} \otimes \boldsymbol{H}) \quad \text{is jointly matrix convex.}$$

We need to see what the perspective $\boldsymbol{\Psi}_f$ looks like. Since $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{H}$ commute, the perspective can be expressed simply as

$$\begin{aligned}
\boldsymbol{\Psi}_f(\boldsymbol{A} \otimes \mathbf{I}; \mathbf{I} \otimes \boldsymbol{H}) &= (\boldsymbol{A} \otimes \mathbf{I}) \cdot f\left((\mathbf{I} \otimes \boldsymbol{H})(\boldsymbol{A} \otimes \mathbf{I})^{-1}\right) \\
&= -(\boldsymbol{A} \otimes \mathbf{I}) \cdot \log\left(\boldsymbol{A}^{-1} \otimes \boldsymbol{H}\right) \\
&= -(\boldsymbol{A} \otimes \mathbf{I})\left[(-\log \boldsymbol{A}) \otimes \mathbf{I} + \mathbf{I} \otimes (\log \boldsymbol{H})\right] \\
&= (\boldsymbol{A} \log \boldsymbol{A}) \otimes \mathbf{I} - \boldsymbol{A} \otimes (\log \boldsymbol{H}).
\end{aligned}$$

We have used basic algebraic properties of tensor product operators, as well as Exercise 17.23.

Fix an arbitrary matrix $\boldsymbol{X} \in \mathbb{M}_n$. Define a (scalar-valued) quadratic form:

As usual, this is the trace inner product on matrices.

$$\varphi_{\boldsymbol{X}}(\boldsymbol{A}; \boldsymbol{H}) := \left\langle \boldsymbol{X},\ \boldsymbol{\Psi}_f(\boldsymbol{A} \otimes \mathbf{I}; \mathbf{I} \otimes \boldsymbol{H})(\boldsymbol{X}) \right\rangle.$$

You should confirm that $\varphi_{\boldsymbol{X}}$ is a convex function of the pair $(\boldsymbol{A}, \boldsymbol{H})$. Indeed, observe that the psd order on $\mathcal{L}(\mathbb{H}_n \otimes \mathbb{H}_n)$ corresponds to increases in this type of quadratic form.

Write out the quadratic form explicitly using the interpretation of the tensor product operator in terms of left- and right-multiplication. We find that

$$\begin{aligned} \varphi_{\boldsymbol{X}}(\boldsymbol{A}; \boldsymbol{H}) &= \langle \boldsymbol{X},\ (\boldsymbol{A} \log \boldsymbol{A})\boldsymbol{X} - \boldsymbol{A}\boldsymbol{X} \log \boldsymbol{H} \rangle \\ &= \operatorname{tr}\left[\boldsymbol{X}^*(\boldsymbol{A} \log \boldsymbol{A})\boldsymbol{X} - \boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X} \log \boldsymbol{H}\right] \quad \text{is convex.} \end{aligned}$$

Choosing $\boldsymbol{X} = \mathbf{I}$, we may conclude that

$$(\boldsymbol{A}, \boldsymbol{H}) \mapsto \operatorname{tr}\left[\boldsymbol{A} \log \boldsymbol{A} - \boldsymbol{A} \log \boldsymbol{H}\right] = \mathrm{S}(\boldsymbol{A}; \boldsymbol{H}) \quad \text{is convex.}$$

This is the statement of Theorem 17.16.

Exercise 17.25 (Alternative perspectives). Prove Theorem 17.16 by applying the same argument to the matrix convex function $f(t) = t \log t$.

17.5.2 Joint concavity of powers

The strategy that we used to prove Theorem 17.16 pays further dividends. By changing the matrix convex function f, we can establish more convexity and concavity theorems. Here is another important example, obtained by Lieb [Lie73a] using a rather difficult argument.

Theorem 17.26 (Lieb 1973). Fix an arbitrary matrix $\boldsymbol{X} \in \mathbb{M}_n$. For any real $r \in (0, 1)$, the function

$$(\boldsymbol{A}, \boldsymbol{H}) \mapsto \operatorname{tr}\left[\boldsymbol{X}^*\boldsymbol{A}^r\boldsymbol{X}\boldsymbol{H}^{1-r}\right]$$

is *concave* on $\mathbb{H}_n^+ \times \mathbb{H}_n^+$.

Proof sketch. The function $f(t) = -t^r$ is matrix convex on $\mathbb{R}_{++}$. Pursue the same reasoning as in the proof of Theorem 17.16. ∎

Notes

This lecture is based on the instructor's monograph [Tro15] with some contributions by Prof. Richard Kueng that appeared in the lecture notes for a previous version of this course. Ando [And79] developed the technique of proving matrix convexity inequalities by lifting to tensor products. We use an implementation of the argument proposed by Effros [Eff09], and extended by Ebadian et al. [ENG11].

Lecture bibliography

[And79] T. Ando. "Concavity of certain maps on positive definite matrices and applications to Hadamard products". In: *Linear Algebra and its Applications* 26 (1979), pages 203–241.

[ENG11] A. Ebadian, I. Nikoufar, and M. E. Gordji. "Perspectives of matrix convex functions". In: *Proceedings of the National Academy of Sciences* 108.18 (2011), pages 7313–7314. DOI: 10.1073/pnas.1102518108.

[Eff09] E. G. Effros. "A matrix convexity approach to some celebrated quantum inequalities". In: *Proceedings of the National Academy of Sciences* 106.4 (2009), pages 1006–1008. DOI: 10.1073/pnas.0807965106.

[HP91] F. Hiai and D. Petz. "The proper formula for relative entropy and its asymptotics in quantum probability". In: *Communications in Mathematical Physics* 143 (1991), pages 99–114.

[KA79] F. Kubo and T. Ando. "Means of positive linear operators". In: *Math. Ann.* 246.3 (1979/80), pages 205–224. DOI: 10.1007/BF01371042.

[Lie73a] E. H. Lieb. "Convex trace functions and the Wigner-Yanase-Dyson conjecture". In: *Advances in Math.* 11 (1973), pages 267–288. DOI: 10.1016/0001-8708(73)90011-X.

[LR73] E. H. Lieb and M. B. Ruskai. "Proof of the strong subadditivity of quantum-mechanical entropy". In: *J. Mathematical Phys.* 14 (1973). With an appendix by B. Simon, pages 1938–1941. DOI: 10.1063/1.1666274.

[Lin73] G. Lindblad. "Entropy, information and quantum measurements". In: *Communications in Mathematical Physics* 33 (1973), pages 305–322.

[ONo00] T. Ogawa and H. Nagaoka. "Strong converse and Stein's lemma in quantum hypothesis testing". In: *IEEE Transactions on Information Theory* 46.7 (2000), pages 2428–2433. DOI: 10.1109/18.887855.

[Sha48] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pages 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[Tro11] J. A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4 (2011), pages 389–434. DOI: 10.1007/s10208-011-9099-z.

[Tro15] J. A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pages 1–230.

18. Positive-Definite Functions

Date: 3 March 2022 **Scribe:** Joel A. Tropp

In this lecture, we will turn to another topic involving the analysis of positive-semidefinite matrices. We will develop the notion of a positive-definite kernel and make a connection between these kernels and positive-definite matrices. Positive-definite kernels play an important role in approximation theory, statistics, machine learning, and physics. Our focus in this lecture is an important class of examples called translation-invariant kernels, which are associated with convolution operators. Positive-definite, translation-invariant kernels are generated by positive-definite functions. We will develop some examples of positive-definite functions. Then we will state and prove Bochner's theorem, which characterizes the continuous positive-definite functions.

Agenda:

1. Positive-definite kernels
2. Positive-definite functions
3. Examples
4. Bochner's theorem
5. Fourier analysis
6. Extensions

18.1 Positive-definite kernels

As we have seen, positive-semidefinite (psd) matrices are a subject that rewards study. The psd property for matrices is defined in terms of quadratic forms:

$$\boldsymbol{A} \in \mathbb{M}_n(\mathbb{C}) \text{ is psd} \quad \text{if and only if} \quad \langle \boldsymbol{u},\ \boldsymbol{A}\boldsymbol{u} \rangle \geq 0 \text{ for all } \boldsymbol{u} \in \mathbb{C}^n. \tag{18.1}$$

We would like to generalize these notions to functions. This section presents the definitions, basic examples, and some applications.

18.1.1 Kernels

A bivariate function is often called a kernel, and we may think about them as acting on functions via integration. In parallel with the concept of a psd matrix, we can introduce the concept of a positive-definite kernel.

Definition 18.1 (Positive-definite kernel). A measurable function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ is called a *kernel* on $\mathbb{R}^d$. The kernel acts on a (suitable) function $h : \mathbb{R}^d \to \mathbb{C}$ by integration:

$$(Kh)(\boldsymbol{x}) \coloneqq \int_{\mathbb{R}^d} K(\boldsymbol{x}, \boldsymbol{y}) h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^d.$$

We say that K is a *positive-definite (pd) kernel* if it satisfies the condition

$$\langle h,\ Kh \rangle \coloneqq \int_{\mathbb{R}^d \times \mathbb{R}^d} \overline{h(\boldsymbol{x})} K(\boldsymbol{x}, \boldsymbol{y}) h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \geq 0 \tag{18.2}$$

for all (suitable) functions $h : \mathbb{R}^d \to \mathbb{C}$.

We are being informal in stating this definition because it does not play a central role in this lecture. One must take care to pose appropriate regularity assumptions on the kernel function K and the test functions h.

In this lecture only, we use the overline for complex conjugation to avoid confusion with convolution operators.

Exercise 18.2 (Pd kernels are Hermitian). If K is a pd kernel, prove that the kernel is also Hermitian: $K(\boldsymbol{x}, \boldsymbol{y}) = \overline{K(\boldsymbol{y}, \boldsymbol{x})}$ for all $\boldsymbol{x}, \boldsymbol{y}$.

Definition 18.3 (Kernel matrix). Let K be a kernel on $\mathbb{R}^d$. For each finite point set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$, we associate a *kernel matrix*:

$$\boldsymbol{K} := \boldsymbol{K}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) := \big[K(\boldsymbol{x}_i, \boldsymbol{x}_j)\big]_{i,j=1,\dots,n}.$$

Under appropriate regularity assumptions, positive-definite kernels induce positive-semidefinite kernel matrices and vice versa.

Exercise 18.4 (Kernels and matrices). Show that the following claims are equivalent.

1. The kernel K is *bounded and continuous* and positive definite for *continuous* test functions h that are in $\mathsf{L}_1(\mathbb{R}^d)$.
2. For each finite point set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, the associated kernel matrix $\boldsymbol{K}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is positive semidefinite.

Hint: The direction **(1 ⇒ 2)** follows when we choose a sequence of functions that tends toward a sum of point masses, located as the data points. The direction **(2 ⇒ 1)** follows when we truncate the integrals to a compact set and approximate by a Riemann sum.

Warning 18.5 (Positive definite?). As with the definition (18.1) of a positive-semidefinite matrix, the definition (18.2) of a positive-definite kernel involves a weak inequality. Nevertheless, it is customary to call these kernels positive definite, rather than "positive semidefinite". The analog of a strictly positive-definite matrix is called a *strictly* positive-definite kernel. ▪

18.1.2 Examples of positive-definite kernels

It is often helpful to think about the value $K(\boldsymbol{x}, \boldsymbol{y})$ of a positive-definite kernel as a measure of the similarity between the points $\boldsymbol{x}$ and $\boldsymbol{y}$. This heuristic is supported by the fact that $K(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, so a point is always positively similar with itself. The kernel matrix $\boldsymbol{K}$ defined in the Exercise 18.4 tabulates the similarities among a family of data points. The next set of examples helps support this point.

Example 18.6 (Positive-definite kernels). Here are a few basic examples of pd kernels that commonly arise.

1. **Inner-product kernel.** The *inner-product kernel* is, perhaps, the simplest positive-definite kernel on $\mathbb{R}^d$. It is defined as

$$K(\boldsymbol{x}, \boldsymbol{y}) := \langle \boldsymbol{x}, \ \boldsymbol{y} \rangle.$$

The associated kernel matrix is usually called the *Gram matrix* of the data points. To see that the kernel is positive-definite, note that

$$\langle h, \ Kh \rangle = \int_{\mathbb{R}^d \times \mathbb{R}^d} \overline{h(\boldsymbol{x})} \langle \boldsymbol{x}, \ \boldsymbol{y} \rangle h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} = \left\| \int_{\mathbb{R}^d} h(\boldsymbol{y}) \boldsymbol{y} \, \mathrm{d}\boldsymbol{y} \right\|^2 \geq 0.$$

The test function h must have sufficient decay to ensure that the integral is defined.

The Gram matrix $\boldsymbol{G}$ of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ takes the form

$$\boldsymbol{G} = [\langle \boldsymbol{x}_j, \ \boldsymbol{x}_k \rangle]_{j,k=1,\dots,n}.$$

2. **Correlation kernel.** The *correlation kernel* is the positive-definite kernel that tabulates the correlations between pairs of vectors:

$$K(\boldsymbol{x}, \boldsymbol{y}) := \frac{\langle \boldsymbol{x}, \ \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

with the understanding that $K(\boldsymbol{x}, \boldsymbol{0}) = K(\boldsymbol{0}, \boldsymbol{y}) = 0$. The associated kernel matrix is the correlation matrix of the data points. The correlation kernel is a positive-definite kernel on $\mathbb{R}^d$ by the same argument as the inner-product kernel.

3. **Gaussian kernel.** For a bandwidth parameter $\alpha > 0$, the *Gaussian kernel* on $\mathbb{R}^d$ is the positive-definite kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) := \mathrm{e}^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2/\alpha}.$$

Note the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y})$ only depends on the difference $\boldsymbol{x} - \boldsymbol{y}$ between its arguments, so it is translation invariant. It is not obvious that this kernel is positive definite; we will establish this fact a little later (Proposition 18.30). The Gaussian kernel is widely used in applications.

There are many other examples of kernels. Our discussion here is limited because we will be focusing on the translation-invariant case. ∎

18.1.3 Applications of kernels

Kernels arise in a wide variety of settings, including approximation theory, statistics, and machine learning. Here are a couple basic applications.

Example 18.7 (Interpolation). Consider scattered data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \dots, n$ where the ordinates $\boldsymbol{x}_i$ are distinct. The interpolation problem asks us to identify a function $f : \mathbb{R}^d \to \mathbb{R}$ that satisfies $f(\boldsymbol{x}_i) = y_i$ for each index i.

Let K be a (real-valued) kernel on $\mathbb{R}^d$. Kernel interpolation poses a model of the form

$$f(\boldsymbol{x}) = \sum\nolimits_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) \quad \text{where each } \alpha_i \in \mathbb{R}.$$

See Figure 18.1 for an illustration. The interpolation problem has a unique solution if and only if the kernel matrix $\boldsymbol{K} = [K(\boldsymbol{x}_j, \boldsymbol{x}_k)]$ is nonsingular. In this case, we obtain the coefficients (α_i) for the model by solving a set of linear equations.

Strictly positive-definite kernels play a role here because they guarantee that the kernel matrix $\boldsymbol{K}$ is strictly positive, hence nonsingular, for every choice of (distinct) data points. Therefore, the kernel provides a unique interpolant f for any set of scattered data. The Gaussian kernel is often used for this purpose. ∎

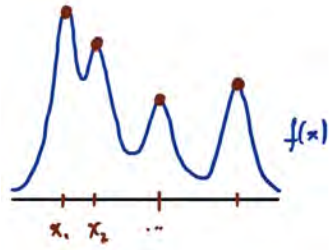

Figure 18.1 (Kernel interpolation). Interpolating real data with a Gaussian kernel. As the bandwidth increases, the kernel interpolant becomes smoother.

Example 18.8 (The kernel trick). Suppose that we have a data analysis method for Euclidean data $(\boldsymbol{x}_i : i = 1, \dots, n)$ in $\mathbb{R}^d$ that only depends on the Gram matrix. For example, principal component analysis (PCA) is an unsupervised learning method that computes features from the data by finding the leading eigenvectors of the Gram matrix and using these eigenvectors to weight the data points. Related examples of Euclidean data analysis methods include canonical correlation analysis (CCA), linear discriminant analysis (LDA), and ridge regression (RR).

> The *kernel trick* posits that we can develop alternative methods for data analysis by replacing the Gram matrix with a pd kernel matrix associated with the data.

For instance, let us summarize how kernel PCA works. Suppose we use a pd kernel to compute a kernel matrix. To find features for the data, we can compute the leading eigenvectors of the kernel matrix. Each eigenvector $\boldsymbol{u} \in \mathbb{R}^d$ leads to a feature of the form $\varphi(\boldsymbol{x}) = \sum_i u_i K(\boldsymbol{x}, \boldsymbol{x}_i)$. This methodology is powerful and widely used.

In summary, we can think about the kernel matrix of a set of data points as a far-reaching generalization of the Gram matrix. The main criticism of kernel methods is that the computational cost of the linear algebra can interfere with the application to large data sets. ∎

18.1.4 Generalizations

A few more remarks are in order.

Remark 18.9 (Bounded + continuous). We often assume that the kernel is bounded and continuous to simplify our exposition. Nevertheless, there are many applications where we must discard these hypotheses.

In physics, kernels often reflect forces of repulsion between particles. For example, the Coulomb (electrical potential) kernel takes the form

$$K(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{4\pi} \|\boldsymbol{x} - \boldsymbol{y}\|^{-1} \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3.$$

This form reflects the fact that two negatively charged particles repel each other, and the force increases as the charges approach each other. The integral operator

$$(K\sigma)(\boldsymbol{x}) = \int_{\mathbb{R}^3} K(\boldsymbol{x}, \boldsymbol{y}) \sigma(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$$

describes the electrical potential induced at a point $\boldsymbol{x} \in \mathbb{R}^3$ by a distribution $\sigma : \mathbb{R}^3 \to \mathbb{R}$ of charge.

Remark 18.10 (Other domains). On an abstract measure space (X, μ), a positive-definite kernel $K : \mathsf{X} \times \mathsf{X} \to \mathbb{C}$ is a (measurable) function that satisfies

$$\int_{\mathsf{X} \times \mathsf{X}} \overline{h(x)} K(x, y) h(y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y) \geq 0 \quad \text{for all (suitable) } h : \mathsf{X} \to \mathbb{C}.$$

This generality can be valuable when working with data that may not take vector values. Using the kernel trick, we can develop methodology for analysis of general data sets by adapting techniques from the Euclidean setting.

18.2 Positive-definite functions

In this lecture, we will make a detailed study of positive-definite, translation-invariant kernels. These kernels are associated with convolution operators. They lead us to introduce and investigate the class of positive-definite functions.

18.2.1 Definitions

A translation-invariant kernel depends only on the difference between its arguments, so it is spatially homogeneous.

Definition 18.11 (Translation-invariant kernel). Consider a measurable function $f : \mathbb{R}^d \to \mathbb{C}$. Introduce a kernel K_f on $\mathbb{R}^d$ via the formula

$$K_f(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x} - \boldsymbol{y}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

A kernel from this class is called *translation-invariant*. The associated integral operator is called a *convolution*:

$$(f * h)(\boldsymbol{x}) := (K_f h)(\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{x} - \boldsymbol{y}) h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^d.$$

This operator is defined for all suitable functions $h : \mathbb{R}^d \to \mathbb{C}$.

Exercise 18.12 (Convolution: Eigenfunctions). Let $f : \mathbb{R}^d \to \mathbb{C}$ be an integrable function. For each $\boldsymbol{t} \in \mathbb{R}^d$, show that the wave $\boldsymbol{x} \mapsto \mathrm{e}^{\mathrm{i}\langle \boldsymbol{t}, \boldsymbol{x} \rangle}$ is an eigenfunction of the convolution operator K_f.

The functions that lead to positive-definite, translation-invariant kernels have a special name. In consonance with the literature, we present this concept in terms of kernel matrices, rather than kernel functions.

Definition 18.13 (Positive-definite function). A function $f : \mathbb{R}^d \to \mathbb{C}$ is called a *positive-definite (pd) function* if it has the following property. For all data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, the kernel matrix $\boldsymbol{K}_f$ associated with the convolution operator K_f is positive semidefinite:

$$\boldsymbol{K}_f \coloneqq \boldsymbol{K}_f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \coloneqq \big[f(\boldsymbol{x}_j - \boldsymbol{x}_k) \big]_{j,k=1,\dots,n} \succcurlyeq \mathbf{0}.$$

We do not require a pd function f to be continuous, but we will typically enforce this property. We will see that boundedness follows as a consequence of the definition.

Exercise 18.14 (Young's inequality). Suppose that $f \in \mathsf{L}_1(\mathbb{R}^d)$. Fix $p \in [1, \infty]$. Prove that

$$\|f * h\|_p \le \|h\|_p \quad \text{for each } h \in \mathsf{L}_p(\mathbb{R}^d).$$

Therefore, the convolution K_f is a bounded operator on L_p. **Hint:** This is an application of Hölder's inequality.

Exercise 18.15 (Positive-definite convolution kernels). Let $f : \mathbb{R}^d \to \mathbb{C}$ be a positive-definite function that is *bounded and continuous*. Show that the associated convolution operator K_f is positive definite for all *continuous* test functions h that are in $\mathsf{L}_1(\mathbb{R})$. **Hint:** Truncate the integral to a compact set, and approximate by a Riemann sum.

Exercise 18.16 (Discontinuous pd functions). Show that the 0–1 indicator function $\mathbb{1}_{\mathbb{Z}}$ of the integers $\mathbb{Z}$ is a pd function on $\mathbb{R}$.

18.2.2 Properties

Individual positive-definite functions have a number of nice features.

Proposition 18.17 (Positive-definite function). Let $f : \mathbb{R}^d \to \mathbb{C}$ be a pd function.

1. **Positivity.** At the origin, $f(\mathbf{0}) \ge 0$.
2. **Symmetry.** We have the relation $f(-\boldsymbol{x}) = \overline{f(\boldsymbol{x})}$ for each $\boldsymbol{x} \in \mathbb{R}^d$.
3. **Boundedness.** The value $|f(\boldsymbol{x})| \le f(\mathbf{0})$ for each $\boldsymbol{x} \in \mathbb{R}^d$.

Proof. For an arbitrary point $\boldsymbol{x} \in \mathbb{R}^d$, we may form the 2×2 kernel matrix $\boldsymbol{K}_f(\mathbf{0}, \boldsymbol{x})$. This matrix is psd:

$$\boldsymbol{K}_f(\mathbf{0}, \boldsymbol{x}) = \begin{bmatrix} f(\mathbf{0}) & f(-\boldsymbol{x}) \\ f(\boldsymbol{x}) & f(\mathbf{0}) \end{bmatrix} \succcurlyeq \mathbf{0}.$$

To prove (1), note that the diagonal entries of a psd matrix are positive. To prove (3), recall that every psd matrix is Hermitian. To obtain (3), we invoke Hadamard's psd criterion, which ensures that $|f(\boldsymbol{x})|^2 \le f(\mathbf{0})^2$. We may take the square-root because $f(\mathbf{0}) \ge 0$. ∎

Hadamard's psd criterion states that $\boldsymbol{A} \succcurlyeq \mathbf{0}$ implies $|a_{ij}|^2 \le a_{ii} a_{jj}$ for all i, j. To prove this, note that each 2×2 principal submatrix of $\boldsymbol{A}$ is psd, so its determinant is positive.

The next exercise shows that continuity of a pd function at the origin is tantamount to continuity everywhere.

Problem 18.18 (Continuity). Prove that a pd function $f : \mathbb{R}^d \to \mathbb{C}$ is uniformly continuous if and only if the real part of f is continuous at the origin. **Hint:** Consider the 3×3 psd kernel matrix $\boldsymbol{K}_f(\mathbf{0}, \boldsymbol{x}, \boldsymbol{y})$. Compute the quadratic form in the vector $\boldsymbol{u} = (z, 1, -1)^*$ for $z \in \mathbb{C}$, and optimize over z.

The class of positive-definite functions has some important stability properties. The last one is distinctive.

Proposition 18.19 (Class of positive-definite functions). The class of pd functions on $\mathbb{R}^d$ satisfies the following properties.

1. **Convex cone.** If f, g are pd functions on $\mathbb{R}^d$, then $\alpha f + \beta g$ is pd for all $\alpha, \beta \geq 0$.
2. **Closedness.** If $(f_n : n \in \mathbb{N})$ is a sequence of pd functions on $\mathbb{R}^d$ that converges pointwise to f, then f is pd.
3. **Multiplication.** If f, g are pd functions on $\mathbb{R}^d$, then fg is pd.

Proof. Point (1) holds because the psd cone is convex, and point (2) holds because the psd cone is closed. To prove (3), we must argue that the entrywise product of two psd matrices remains psd. But this is the statement of Schur's product theorem. ∎

18.3 Examples of positive-definite functions

In this section, we will present several examples of pd functions. To lighten notation, we will restrict our attention to pd functions on the real line, but these examples have straightforward generalizations to higher dimensions.

18.3.1 Complex exponentials

Recall that the eigenfunctions of a convolution operator are the complex exponentials. As a consequence, it should not come as a surprise that complex exponentials provide our first example of a pd function. The theme of our discussion is that the complex exponentials are the building blocks for designing other pd functions.

Proposition 18.20 (Complex exponentials are pd). For each fixed $t \in \mathbb{R}$, the function $x \mapsto \mathrm{e}^{-\mathrm{i}tx}$ for $x \in \mathbb{R}$ is pd.

Proof. To see that the complex exponential is a pd function, we examine the kernel matrix associated with arbitrary points $x_1, \dots, x_n \in \mathbb{R}$. We find that

$$\boldsymbol{K} = \left[\mathrm{e}^{-\mathrm{i}t(x_j - x_k)}\right]_{j,k=1,\dots,n} = \begin{bmatrix} | \\ \mathrm{e}^{-\mathrm{i}tx_j} \\ | \end{bmatrix}_{j=1,\dots,n} \begin{bmatrix} | \\ \mathrm{e}^{-\mathrm{i}tx_k} \\ | \end{bmatrix}^*_{k=1,\dots,n} \succcurlyeq \mathbf{0}. \tag{18.3}$$

In the penultimate expression, the factors are a column vector and its conjugate transpose. This outer product is a psd matrix with rank one. ∎

Exercise 18.21 (Waves are pd). For each fixed $\boldsymbol{t} \in \mathbb{R}^d$, show that $\boldsymbol{x} \mapsto \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{t}, \boldsymbol{x} \rangle}$ for $\boldsymbol{x} \in \mathbb{R}^d$ is a pd function on $\mathbb{R}^d$.

18.3.2 Cosine and sine

Our next example provides a hint about how we can build pd functions from complex exponentials.

Proposition 18.22 (Cosine is pd). For each $t \in \mathbb{R}$, the function $x \mapsto \cos(tx)$ is pd on $\mathbb{R}$.

Proof. The identity $\cos(tx) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}tx} + \mathrm{e}^{-\mathrm{i}tx})$ expresses the cosine as a conic combination of two complex exponentials. The claim follows because each complex exponential is pd, and the class of pd functions is closed under conic combinations. ∎

Exercise 18.23 (Sine is not pd). Show that $x \mapsto \sin(x)$ is not pd on $\mathbb{R}$. **Hint:** Sine is an odd function.

18.3.3 The sine cardinal (sinc) function

We can extend the principle behind the cosine example by allowing for more elaborate combinations, expressed as integrals.

Proposition 18.24 (Sinc is pd). The function $\operatorname{sinc}(x) := \sin(x)/x$ for $x \in \mathbb{R}$ is pd.

We define $\operatorname{sinc}(0) = 1$ to ensure continuity.

Proof. We can write the sinc function as an integral:

$$\operatorname{sinc}(x) = \frac{1}{2} \int_{-1}^{+1} \mathrm{e}^{-\mathrm{i}tx} \, \mathrm{d}t.$$

To conclude that sinc is pd, we write the integral as a limit of Riemann sums. Each Riemann sum is pd because it is a conic combination of complex exponentials. The limit is pd because pd functions are stable under pointwise limits. ∎

As an aside, let us frame an exercise that makes a link between the theory of positive-definite functions and matrix monotone functions. This is just one of many such examples; see [Bha07b, Chap. 5].

Exercise 18.25 (Tangent is matrix monotone). Show that $x \mapsto \tan(x)$ is matrix monotone on $(-\pi/2, +\pi/2)$ by writing the Loewner matrix of the tangent in terms of a kernel matrix associated with the sinc function. **Hint:** $\sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \sin(\beta)\cos(\alpha)$.

18.3.4 Fourier transforms

Now, let us take a massive jump in abstraction before returning to earth. In this section, we consider fully general conic combinations of complex exponentials.

Definition 18.26 (Fourier transform: Positive). Let $f : \mathbb{R} \to \mathbb{C}$ be an integrable function. Its *Fourier transform* $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is the function

$$\widehat{f}(x) := \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} f(t) \, \mathrm{d}t \quad \text{for } x \in \mathbb{R}.$$

More generally, let μ be a finite, complex Borel measure on $\mathbb{R}$. Its Fourier transform $\widehat{\mu} : \mathbb{R} \to \mathbb{C}$ is the function

$$\widehat{\mu}(x) := \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} \, \mathrm{d}\mu(t) \quad \text{for } x \in \mathbb{R}.$$

The first definition involves the measure $\mathrm{d}\mu(t) = f(t) \, \mathrm{d}t$ with density f.

For our purposes, the key insight is that Fourier transforms induce pd functions. This is a natural outcome, but the proof requires a little thought.

Proposition 18.27 (Fourier transforms are pd). Let μ be a finite, *positive* Borel measure on $\mathbb{R}$. Then its Fourier transform $\widehat{\mu}$ is a pd function on $\mathbb{R}$.

Proof. We return to the definition of a pd function. For points $x_1, \dots, x_n \in \mathbb{R}$, form the kernel matrix

$$\boldsymbol{K}_{\widehat{\mu}} = \big[\widehat{\mu}(x_j - x_k)\big]_{j,k} = \int_{\mathbb{R}} \big[\mathrm{e}^{-\mathrm{i}t(x_j - x_k)}\big]_{j,k} \, \mathrm{d}\mu(t).$$

We will show that the kernel matrix $\boldsymbol{K}_{\widehat{\mu}}$ is a well-defined psd matrix, so $\widehat{\mu}$ is a pd function.

For each $t \in \mathbb{R}$, the matrix in the integrand is psd since the complex exponential is a pd function. Therefore, for each *unit* vector $\boldsymbol{v} \in \mathbb{C}^n$,

$$\boldsymbol{v}^* \boldsymbol{K}_{\widehat{\mu}} \boldsymbol{v} = \int_{\mathbb{R}} \boldsymbol{v}^* \left[\mathrm{e}^{-\mathrm{i}t(x_j - x_k)} \right]_{j,k} \boldsymbol{v} \, \mathrm{d}\mu(t) \geq 0.$$

Indeed, the integral of a positive function is positive. The integral is finite because the integrand is bounded by n. ∎

Positive-definite functions that arise from Fourier transforms enjoy some additional regularity properties.

Exercise 18.28 (Fourier transforms: Continuity). Let μ be a finite, positive Borel measure on $\mathbb{R}$. Show that its Fourier transform $\widehat{\mu}$ is a continuous function. **Hint:** Complex exponentials are bounded and continuous; truncate the integral to a compact set.

Problem 18.29 (Riemann–Lebesgue lemma). For an *integrable* function $f \in \mathsf{L}_1(\mathbb{R})$, prove that $\widehat{f}$ is a continuous function that vanishes at infinity. **Hint:** Approximate f by simple functions, and note that the sinc function is a continuous function that vanishes at infinity.

A function $g : \mathbb{R} \to \mathbb{R}$ *vanishes at infinity* if $|g(x)| \to 0$ as $|x| \to \infty$.

18.3.5 Gaussians

The Fourier transform provides us with a powerful tool for detecting other examples of pd functions. Here is a critical example.

Proposition 18.30 (The Gaussian kernel is pd). The function $x \mapsto \mathrm{e}^{-x^2/2}$ for $x \in \mathbb{R}$ is pd.

This result is a consequence of the following fundamental fact, which expresses the Gaussian as a Fourier transform. As we will discuss, Gaussians play a central role in Fourier analysis because they serve as approximate identities.

Fact 18.31 (Gaussian: Fourier transform). The density $\varphi_{b,v}$ of a Gaussian random variable with mean $b \in \mathbb{R}$ and variance $v > 0$ is the function

$$\varphi_{b,v}(t) := \frac{\mathrm{e}^{-(t-b)^2/(2v)}}{\sqrt{2\pi v}} \, \mathrm{d}t \quad \text{for } t \in \mathbb{R}.$$

The density is normalized: $\int_{\mathbb{R}} \varphi_{b,v}(t) \, \mathrm{d}t = 1$. Its Fourier transform is the function

$$\widehat{\varphi}_{b,v}(x) = \mathrm{e}^{\mathrm{i}bx} \cdot \mathrm{e}^{-vx^2/2} \quad \text{for } x \in \mathbb{R}.$$

In other words, the Fourier transform of a Gaussian is a twisted Gaussian. ∎

Proof sketch. By a change of variables, we may assume that $b = 0$ and $v = 1$. To prove that $I := \int_{\mathbb{R}} \varphi_{0,1}(t) \, \mathrm{d}t = 1$, write I^2 as a double integral and change to polar coordinates. To evaluate the Fourier transform $\widehat{\varphi}_{0,1}$, we expand $t \mapsto \mathrm{e}^{-\mathrm{i}tx}$ as a Taylor series. The integrals of the odd terms vanish. The integrals of the even terms are the moments of a standard normal density, which may be evaluated with repeated integration by parts. ∎

18.3.6 Inverse quadratics

As a final example, we consider another Fourier transform that arises in probability theory.

Proposition 18.32 (Inverse quadratic is pd). The function $x \mapsto (1 + x^2)^{-1}$ is pd on $\mathbb{R}$.

Proof. We confirm that the function is pd by writing it as the Fourier transform of (twice) the Laplace density:

$$\frac{1}{1+x^2} = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} \mathrm{e}^{-|t|} \,\mathrm{d}t.$$

For example, use symmetry about $t = 0$ to pass to a cosine integral, and integrate by parts twice. ∎

18.4 Bochner's theorem

All our examples of pd functions derive in one way or another from complex exponentials. This repetition may suggest a lack of creativity, but there is a deeper reason. In fact, every continuous pd function can be represented as a Fourier transform [Boc33].

Theorem 18.33 (Bochner 1933). A *continuous* function f on the real line is positive definite if and only if f is the Fourier transform of a finite, *positive* Borel measure μ on the real line:

$$f(x) = \widehat{\mu}(x) = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} \,\mathrm{d}\mu(t).$$

Proposition 18.27 establishes the "easy" direction: the Fourier transform of a finite, positive measure is a pd function. This result has already paid dividends because it allowed us to identify a number of interesting positive-definite functions.

In this section, we will give a proof the "hard" direction. This result provides a powerful representation for positive-definite functions. It serves as a building block for establishing other difficult theorems in matrix analysis and other fields. In particular, Bochner's theorem has a consequence for probability theory: The characteristic function of a random variable is a continuous pd function and vice versa. This can be used (in a somewhat roundabout way) to prove Lévy's continuity theorem, which is a key step in the standard proof of the central limit theorem.

Beyond that, Bochner's theorem has striking applications in machine learning, where it serves as the foundation of the method of random Fourier features [RR08].

18.4.1 Complex exponentials are extreme points

Before we turn to the proof, let us recall the geometric perspective on integral representations. The pd functions on the real line form a convex cone, closed under pointwise limits. By normalizing the functions, we obtain a compact, convex base of this cone:

$$\mathsf{B} := \{f : f \text{ is pd on } \mathbb{R} \text{ and } f(0) = 1\}.$$

According to the Krein–Milman theorem, we can represent every function $f \in \mathsf{B}$ as a limit of convex combinations of the extreme points. Bochner's theorem states that

$$f(x) = \widehat{\mu}(x) = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} \,\mathrm{d}\mu(t) \quad \text{for a probability measure } \mu.$$

This representation tells us that every extreme point of B must be a complex exponential.

Let us offer an independent proof of the fact that every complex exponential is an extreme point. We argue in the spirit of Boutet de Monvel [Sim19, Thm. 28.12].

Proposition 18.34 (Complex exponentials are extreme). For each $t \in \mathbb{R}$, the complex exponential $e_t : x \mapsto \mathrm{e}^{-\mathrm{i}tx}$ is an extreme point of B.

Proof. Fix the frequency $t \in \mathbb{R}$. We already know that e_t is an element of B. Moreover, the expression (18.3) shows that the 2×2 kernel matrix $\boldsymbol{K}_{e_t}(0, x)$ has *rank one* for each $x \in \mathbb{R}$. This observation is the key to the proof.

Suppose that $e_t = \frac{1}{2} f + \frac{1}{2} g$ where $f, g \in \mathsf{B}$. The kernel matrix is linear in the kernel function, so

$$\boldsymbol{K}_{e_t}(0, x) = \tfrac{1}{2}\boldsymbol{K}_f(0, x) + \tfrac{1}{2}\boldsymbol{K}_g(0, x).$$

But this expression represents the rank-one matrix $\boldsymbol{K}_{e_t}(0, x)$ as an average of two psd matrices (because f, g are pd functions). Recall that each rank-one matrix lies in an extreme ray of the psd cone. Therefore, $\boldsymbol{K}_f = \alpha(x)\boldsymbol{K}_{e_t}$ for some $\alpha(x) \geq 0$. That is,

$$\boldsymbol{K}_f(0, x) = \begin{bmatrix} 1 & f(-x) \\ f(x) & 1 \end{bmatrix} = \alpha(x) \begin{bmatrix} 1 & \mathrm{e}^{\mathrm{i}tx} \\ \mathrm{e}^{-\mathrm{i}tx} & 1 \end{bmatrix} = \alpha(x)\boldsymbol{K}_{e_t}(0, x).$$

We deduce that $\alpha(x) = 1$, and so $f(x) = \mathrm{e}^{-\mathrm{i}tx}$ for all $x \in \mathbb{R}$. Likewise, $g(x) = \mathrm{e}^{-\mathrm{i}tx}$ for all $x \in \mathbb{R}$. We conclude that e_t cannot be written as the average of two distinct functions in B, so it is an extreme point. ∎

18.4.2 Fourier analysis with Gaussians

To prove Bochner's theorem, we introduce some rudiments of Fourier analysis. This task will be easier because we only need to use the main formulas when one of the functions involved is a Gaussian.

For an *integrable* function $f : \mathbb{R} \to \mathbb{C}$, the Fourier transform $\widehat{f}$ and inverse Fourier transform $\check{f}$ are defined explicitly:

$$\widehat{f}(x) := \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} f(t)\,\mathrm{d}t \quad \text{and} \quad \check{f}(t) := \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} f(x)\,\mathrm{d}x \quad \text{for } x \in \mathbb{R}.$$

Among several possible definitions, we have chosen the probabilists' Fourier transform.

The Riemann–Lebesgue lemma states that $\widehat{f}, \check{f} \in \mathsf{C}_0(\mathbb{R})$, the space of continuous functions on $\mathbb{R}$ that vanish at infinity, equipped with the supremum norm.

A few basic properties merit comment. The Fourier transform satisfies an elegant duality property:

$$\int_{\mathbb{R}} f(x)\widehat{h}(x)\,\mathrm{d}x = \int_{\mathbb{R}} \widehat{f}(x)h(x)\,\mathrm{d}x \quad \text{for } f, h \in \mathsf{L}_1(\mathbb{R}). \tag{18.4}$$

The Fourier transform also converts convolution into pointwise multiplication:

$$\widehat{f * h} = \widehat{f} \cdot \widehat{h} \in \mathsf{C}_0(\mathbb{R}) \quad \text{for } f, h \in \mathsf{L}_1(\mathbb{R}). \tag{18.5}$$

Young's inequality ensures that the convolution $f * h$ is an integrable function.

Let us introduce the class of twisted Gaussian functions:

$$\mathsf{G} := \{t \mapsto r\mathrm{e}^{\mathrm{i}ta}\mathrm{e}^{-(t-b)^2/(2v)} \text{ where } r, a, b \in \mathbb{R} \text{ and } v > 0\}.$$

Using Fact 18.31, we can quickly confirm that

$$g \in \mathsf{G} \quad \text{implies} \quad \widehat{g}, \check{g} \in \mathsf{G} \quad \text{and} \quad \widehat{\check{g}} = g.$$

In particular, the Fourier transform is a bijection on G.

The main result that we require is a precursor of Plancherel's identity:

$$\int_{\mathbb{R}} f(t)\overline{g(t)}\,\mathrm{d}t = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(x)\overline{\widehat{g}(x)}\,\mathrm{d}x \quad \text{when } f \in \mathsf{L}_1(\mathbb{R}) \text{ and } g \in \mathsf{G}. \tag{18.6}$$

Indeed, we note that $\overline{g} = \widehat{\check{g}}$, and apply the duality (18.4) to move the Fourier transform to f. Then we pass the complex conjugate through the inverse Fourier transform to obtain the complex conjugate of the Fourier transform and a scale factor.

18.4.3 Approximate identities

The reason that Gaussians suffice for our purposes is that they serve as "approximate identities".

Exercise 18.35 (Gaussian: Approximate identity). Consider centered Gaussian densities:

$$g_\nu(t) := \frac{1}{\sqrt{2\pi\nu}} \cdot \mathrm{e}^{-t^2/(2\nu)} \quad \text{for } \nu > 0.$$

Let $f : \mathbb{R} \to \mathbb{C}$ be a bounded, continuous function. Prove that

$$f(b) = \lim_{\nu \downarrow 0} (f * g_\nu)(b) = \lim_{\nu \downarrow 0} \int_{\mathbb{R}} f(b-t) \cdot g_\nu(t)\,\mathrm{d}t.$$

That is, Gaussians approximate the identity element for the convolution operation. Thus, we can isolate the value of a continuous function by integration against increasingly localized Gaussians. See Figure 18.2. **Hint:** Truncate the integral to the interval $\pm c\sqrt{\nu}$, where c depends on $\sup f$. Apply the mean value theorem to f on this interval.

As $\nu \downarrow 0$, the Fourier transform of g_ν satisfies

$$\widehat{g}_\nu(x) = \mathrm{e}^{-\nu x^2/2} \to 1.$$

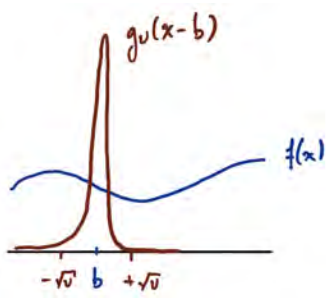

Figure 18.2 (Approximate identity). Using an approximate identity to isolate a function value.

18.4.4 Proof of Bochner's theorem

Let us turn to the proof of Theorem 18.33. This argument is a significant revision of the proofs in [Bha07b, Thm. 5.5.3] and [Sim15, Thm. 6.6.6].

Let $f : \mathbb{R} \to \mathbb{R}$ be the target pd function. Without loss of generality, we may rescale f so that $f(0) = 1$. By hypothesis, the function f is continuous, but it may not be integrable. To repair this defect, we consider a sequence of regularized functions:

$$F_n(t) := f(t) \cdot \mathrm{e}^{-t^2/n} \quad \text{for } t \in \mathbb{R} \text{ and each } n \in \mathbb{N}.$$

We will produce the measure μ that represents f as the limit of the measures μ_n that represent each F_n.

Positive definiteness. To begin, note that F_n is *positive definite and continuous* because it is the product of two pd, continuous functions. The function $F_n \in \mathsf{L}_1(\mathbb{R})$ because f is bounded, and the Gaussian belongs to $\mathsf{L}_1(\mathbb{R})$.

Let us extract the consequence of the pd property that we will need. Choose a test function $g \in \mathsf{G}$. By Exercise 18.15, the convolution kernel induced by F_n is positive definite for this class of test functions. Therefore, we calculate that

$$\begin{aligned} 0 &\le \int_{\mathbb{R}\times\mathbb{R}} \overline{g(s)} F_n(s-t) g(t)\,\mathrm{d}s\,\mathrm{d}t = \int_{\mathbb{R}} \overline{g(s)} \cdot (F_n * g)(s)\,\mathrm{d}s \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \overline{\widehat{g}(x)} \cdot \widehat{(F_n * g)}(x)\,\mathrm{d}x = \frac{1}{2\pi} \int_{\mathbb{R}} |\widehat{g}(x)|^2 \cdot \widehat{F}_n(x)\,\mathrm{d}x. \end{aligned} \tag{18.7}$$

To pass to the second line, we invoked Plancherel's identity (18.6). The last relation is the convolution theorem (18.5).

Positivity. The pd property of F_n implies that its Fourier transform $\widehat{F}_n$ is positive:

$$\widehat{F}_n(x) \ge 0 \quad \text{for all } x \in \mathbb{R}.$$

To prove this, recall that $\widehat{F}_n$ is bounded and continuous because of the Riemann–Lebesgue lemma. Fix a point $b \in \mathbb{R}$, and select a twisted Gaussian function $g_\nu \in \mathsf{G}$ whose Fourier transform satisfies

$$|\widehat{g}_\nu(x)|^2 = \frac{1}{\sqrt{2\pi\nu}} \cdot \mathrm{e}^{-(x-b)^2/2\nu} \quad \text{for } \nu > 0.$$

The function $|\widehat{g}_\nu|^2$ is an approximate identity, centered at b, with width $\sqrt{\nu}$. Introduce $|\widehat{g}_\nu|^2$ into the relation (18.7), and take the limit as $\nu \downarrow 0$ using the fact that $\widehat{F}_n$ is continuous. This demonstrates that $\widehat{F}_n(b) \geq 0$.

Duality. The rest of the proof depends on the complex conjugate of Plancherel's identity (18.6):

$$\int_{\mathbb{R}} g(t) F_n(-t)\,\mathrm{d}t = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{g}(x) \widehat{F}_n(x)\,\mathrm{d}x \quad \text{for } g \in \mathsf{G}. \tag{18.8}$$

What happened to the complex conjugates? Since F_n is pd, we know that $\overline{F_n(t)} = F_n(-t)$ and that $\widehat{F}_n \geq 0$. To exploit this relation, we can choose g to be an approximate identity. This allows us to transfer information between F_n and $\widehat{F}_n$.

Integrability. The next step is to argue that $\widehat{F}_n$ is integrable, which is not obvious *a priori*. More precisely, we show that

$$\frac{1}{2\pi} \int_{\mathbb{R}} \widehat{F}_n(x)\,\mathrm{d}x = 1.$$

To do so, we select the test function

$$g_\nu(t) = \frac{1}{\sqrt{2\pi\nu}} \mathrm{e}^{-t^2/(2\nu)} \quad \text{with} \quad \widehat{g}_\nu(x) = \mathrm{e}^{-\nu x^2/2}.$$

Since F_n is bounded and continuous and g_ν is an approximate identity localized at the origin,

$$\lim_{\nu \downarrow 0} \int_{\mathbb{R}} g_\nu(t) F_n(-t)\,\mathrm{d}t = F_n(0) = 1.$$

Since $\widehat{F}_n \geq 0$ and $\widehat{g}_\nu \uparrow 1$ as $\nu \downarrow 0$, monotone convergence implies that

$$\lim_{\nu \downarrow 0} \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{g}_\nu(x) \widehat{F}_n(x)\,\mathrm{d}x = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{F}_n(x)\,\mathrm{d}x.$$

The last two displays establish the claim because of the identity (18.8).

Measures and limits. Let us introduce a sequence of Borel probability measures on the real line:

$$\mathrm{d}\mu_n(x) \coloneqq \frac{1}{2\pi} \widehat{F}_n(x)\,\mathrm{d}x \quad \text{for each } n \in \mathbb{N}.$$

To verify that μ_n is a probability measure, we rely on the facts that $\widehat{F}_n$ is positive and its integral is normalized. Passing to a subsequence if needed, the sequence $(\mu_n : n \in \mathbb{N})$ of probability measures has a weak-$*$ limit μ. Therefore,

$$\lim_{n \to \infty} \int_{\mathbb{R}} h(x)\,\mathrm{d}\mu_n(x) = \int_R h(x)\,\mathrm{d}\mu(x) \quad \text{for all } h \in \mathsf{C}_0(\mathbb{R}).$$

This claim follows from the Banach–Alaoglu theorem, applied to the dual of the space $\mathsf{C}_0(\mathbb{R})$ of continuous functions that vanish at infinity.

The limit μ is a positive Borel measure with $\mu(\mathbb{R}) \leq 1$.

Fourier transforms. It remains to show that the function f is the Fourier transform of the measure μ. For fixed $b \in \mathbb{R}$, we choose the test functions

$$g_\nu(t) = \frac{1}{\sqrt{2\pi\nu}} \mathrm{e}^{-(x+b)^2/(2\nu)} \quad \text{with} \quad \widehat{g}_\nu(x) = \mathrm{e}^{-\mathrm{i}bx} \cdot \mathrm{e}^{-\nu t^2/2} \quad \text{for } \nu > 0.$$

Using the duality identity (18.8) and taking the limit as $n \to \infty$, we can relate the function f to the measure μ. Indeed,

$$\begin{aligned}\int_{\mathbb{R}} g_\nu(t) f(-t)\,\mathrm{d}t &= \lim_{n\to\infty} \int_{\mathbb{R}} g_\nu(t) F_n(-t)\,\mathrm{d}t \\ &= \lim_{n\to\infty} \int_{\mathbb{R}} \widehat{g}_\nu(x)\,\mathrm{d}\mu_n(x) = \int_{\mathbb{R}} \widehat{g}_\nu(x)\,\mathrm{d}\mu(x).\end{aligned}$$

The first limit is a consequence of monotone convergence because $F_n \uparrow f$ and $g_\nu \geq 0$. The second limit is a consequence of the weak-$*$ convergence of μ_n to μ because $\widehat{g}_\nu \in \mathsf{C}_0(\mathbb{R})$. Last, we take limits as $\nu \downarrow 0$. Indeed,

$$f(b) = \lim_{\nu\downarrow 0} \int_{\mathbb{R}} g_\nu(t) f(-t)\,\mathrm{d}t = \lim_{\nu\downarrow 0} \int_{\mathbb{R}} \widehat{g}_\nu(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}bx}\,\mathrm{d}\mu(x).$$

The first limit is valid because g_ν is an approximate identity, localized at $-b$, and f is bounded and continuous. The second limit follows from dominated convergence because $\widehat{g}_\nu$ is bounded and μ is a finite measure. We have established Bochner's theorem.

18.4.5 Extensions

Bochner's theorem holds in far more general settings. In particular, it is valid for pd functions on $\mathbb{R}^d$. The proof is essentially the same as the proof of Theorem 18.33.

Theorem 18.36 (Bochner). A *continuous* function f on $\mathbb{R}^d$ is positive definite if and only if f is the Fourier transform of a finite, positive Borel measure μ on $\mathbb{R}^d$. That is,

$$f(t) = \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{t},\, \boldsymbol{x}\rangle}\,\mathrm{d}\mu(\boldsymbol{t}).$$

More generally, the theorem holds in abstract settings. We give one such result without proper definitions, and we illustrate it with an example.

Theorem 18.37 (Bochner–Weil). A continuous function f on a locally compact abelian (LCA) group is positive-definite if and only if f is the Fourier transform of a finite, positive Borel measure μ on the dual group.

See [Rud90] for the setting of LCA groups, and see [Rud91] for Raikov's generalization to the setting of commutative Banach algebras.

Example 18.38 (Positive-definite sequences). Consider the LCA group $(\mathbb{Z}, +)$, comprising the integers with addition. A function on $\mathbb{Z}$ is called a sequence, and a sequence $(a_k : k \in \mathbb{Z})$ is *positive definite* when

$$\sum\nolimits_{j,k} \overline{u}_j a_{j-k} u_k \geq 0 \quad \text{for } \boldsymbol{u} : \mathbb{Z} \to \mathbb{C} \text{ with finite support.}$$

Positive-definite sequences arise from positive-definite convolution operators on the integers. These operators correspond with (bi-infinite) Toeplitz matrices.

The dual group $(\mathbb{T}, +)$ is the torus with modular addition. The Fourier transform computes the Fourier series of a measure on the torus. Thus, the Bochner–Weil theorem guarantees that

$$a_k = \int_{[-\pi,+\pi)} \mathrm{e}^{-\mathrm{i}tk}\,\mathrm{d}\mu(t),$$

where μ is a positive, finite Borel measure on $[-\pi, +\pi)$.

This result is called the Carathéodory–Herglotz–Toeplitz theorem. It is actually a precursor to Bochner's work, and it can be established with a rather more elementary argument; for example, see [Bha07b, Thm. 5.5.2]. ∎

Notes

This lecture contains a new presentation of this material. Applications of kernels are extracted from [SSB18]. The procession of examples is drawn from Bhatia's book [Bha07b, Chap. 5]. The proof that complex exponentials are extreme rays of the cone of pd functions seems to be new. The self-contained proof of Bochner's theorem is a significant revision of the proof in Simon's real analysis text [Sim15]. The material on Fourier analysis is adapted from Arbogast & Bona [AB08]. See Rudin's books [Rud90; Rud91] for generalizations of Bochner's theorem.

Lecture bibliography

[AB08] T. Arbogast and J. L. Bona. *Methods of applied mathematics*. ICES Report. UT-Austin, 2008.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[Boc33] S. Bochner. "Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse". In: *Math. Ann.* 108.1 (1933), pages 378–410. DOI: 10.1007/BF01452844.

[RR08] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., 2008, pages 1177–1184. URL: http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf.

[Rud90] W. Rudin. *Fourier analysis on groups*. Reprint of the 1962 original, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1990. DOI: 10.1002/9781118165621.

[Rud91] W. Rudin. *Functional analysis*. Second. McGraw-Hill, Inc., New York, 1991.

[SSB18] B. Schlkopf, A. J. Smola, and F. Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

[Sim15] B. Simon. *Real analysis*. With a 68 page companion booklet. American Mathematical Society, Providence, RI, 2015. DOI: 10.1090/simon/001.

[Sim19] B. Simon. *Loewner's theorem on monotone matrix functions*. Springer, Cham, 2019. DOI: 10.1007/978-3-030-22422-6.

19. Entrywise PSD Preservers

Date: 8 March 2022 **Scribe:** Joel A. Tropp

Agenda:
1. Radial kernels
2. Inner-product kernels
3. Entrywise psd preservers
4. Examples
5. Vasudeva's theorem
6. Extensions

In this lecture, we discuss several classes of kernel functions that are defined on Euclidean spaces of every dimension: radial kernels and inner-product kernels. In the first case, Bochner's theorem leads to a characterization of all radial kernels. To study the second case, we must investigate a new concept. Suppose that we apply a function to each *entry* of a matrix to produce a new matrix. This function is called an entrywise psd preserver if it maps each psd matrix to another psd matrix. These functions also enjoy a beautiful theory that complements the Loewner theory of matrix monotone functions and the Bochner theory of positive-definite functions.

19.1 Families of kernels

Our first object is to extend the work from the last lecture to construct a family of positive-definite convolution kernels that is defined for Euclidean spaces of each dimension. Then we will turn to another family of kernels, based on inner products, that also extend to every dimension.

19.1.1 Positive-definite functions

Recall that a function $f : \mathbb{R}^d \to \mathbb{C}$ is *positive definite (pd)* on $\mathbb{R}^d$ when the matrix

$$\boldsymbol{K}_f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \coloneqq \big[f(\boldsymbol{x}_j - \boldsymbol{x}_k)\big]_{j,k=1,\dots,n} \succcurlyeq \mathbf{0} \quad \text{for all } \boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d.$$

In other words, we consider the convolution kernel K_f induced by the function f. The function f is pd when the kernel matrices $\boldsymbol{K}_f$ associated with the convolution kernel are all psd.

Under mild regularity assumptions, the eigenfunctions of a convolution operator are complex exponentials. This observation suggests that the complex exponentials will play a basic role in characterizing convolution kernels. Indeed, we have the following fundamental result.

Theorem 19.1 (Bochner). A *continuous* function $f : \mathbb{R}^d \to \mathbb{C}$ is positive definite on $\mathbb{R}^d$ if and only if f is the Fourier transform of a finite, positive Borel measure μ. That is,

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\xi}, \boldsymbol{x} \rangle} \,\mathrm{d}\mu(\boldsymbol{\xi}).$$

Bochner's theorem describes an individual pd function, defined on a Euclidean space of a particular dimension. One may also wonder whether there is a way to extend this result to obtain a family of pd convolution kernels in each dimension.

19.1.2 Radial kernels

The key idea is to study a class of convolution kernels that are rotationally invariant. These kernels are spatially and directionally homogeneous.

Definition 19.2 (Radial kernel). A (translation-invariant) kernel $K : \mathbb{R}^d \to \mathbb{C}$ is *radial* if it takes the form

$$K(\boldsymbol{x}, \boldsymbol{y}) := \varphi(\|\boldsymbol{x} - \boldsymbol{y}\|) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where $\varphi : \mathbb{R}_+ \to \mathbb{C}$ is a function. We say that the function φ is *positive-definite radial* on $\mathbb{R}^d$ when the associated kernel matrices

$$\boldsymbol{K}_\varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) := \big[\varphi(\|\boldsymbol{x}_j - \boldsymbol{x}_k\|)\big]_{j,k=1,\dots,n} \succcurlyeq \mathbf{0}$$

for all choices of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ and each $n \in \mathbb{N}$.

In this lecture, $\|\cdot\|$ is the Euclidean norm.

A radial kernel is defined on a Euclidean space of a particular dimension d. Nevertheless, we can also ask when φ is positive-definite radial on $\mathbb{R}^d$ for every $d \in \mathbb{N}$. In this case, we simply say that φ is "positive-definite radial" without additional qualification. Schoenberg provided an elegant answer to this question [Sch38].

Theorem 19.3 (Schoenberg 1938). A *continuous* function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is positive-definite radial if and only if it is given by the Laplace transform of a finite, positive Borel measure μ on $\mathbb{R}_+$. More precisely,

$$\varphi(t) = \int_{\mathbb{R}_+} \mathrm{e}^{-r^2 t^2/2} \,\mathrm{d}\mu(r) \quad \text{for } t \in \mathbb{R}_+.$$

Every pd radial function must take positive values! (Why?)

Proof sketch. The reverse direction is simple. We write the Gaussian as a characteristic function (i.e., a Fourier transform). For $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathrm{e}^{-r^2\|\boldsymbol{x}\|^2/2} = \mathbb{E}[\mathrm{e}^{-\mathrm{i}\langle \boldsymbol{z},\, \boldsymbol{x}\rangle/r}] \quad \text{where} \quad \boldsymbol{z} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_d).$$

We use random variables and expectations to simplify some of the formulas here.

Using dominated convergence,

$$\varphi(\|\boldsymbol{x}\|) = \int_{\mathbb{R}^d} \mathrm{e}^{-r^2\|\boldsymbol{x}\|^2/2} \,\mathrm{d}\mu(r) = \mathbb{E}\left[\int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{z},\, \boldsymbol{x}\rangle/r} \,\mathrm{d}\mu(r)\right].$$

The integral is a pd function, and the average of pd functions is pd.

For the forward direction, we assume that the function $\varphi(\|\cdot\|)$ is pd on $\mathbb{R}^d$ for each $d \in \mathbb{N}$. Use rotational invariance to average over the unit sphere, and apply Bochner's theorem to deduce that

$$\varphi(t) = \int_{\mathbb{R}_+} \Omega_d(rt) \,\mathrm{d}\mu_d(r).$$

The measure μ_d on $\mathbb{R}_+$ is positive, with total mass equal to $\varphi(0)$. The functions

$$\Omega_d(s) := \mathbb{E}[\mathrm{e}^{-\mathrm{i}s\theta_1}] \quad \text{where} \quad \boldsymbol{\theta} \sim \textsc{unif}(\mathbb{S}^{d-1}).$$

In this expression, θ_1 is the first coordinate of a random vector $\boldsymbol{\theta}$ drawn uniformly from the Euclidean unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$.

Since the first coordinate of a spherical vector is almost a centered Gaussian with variance d^{-1}, it is not too hard to argue that

$$\lim_{d\to\infty} \Omega_d(r\sqrt{d}) \to \mathrm{e}^{-r^2/2}.$$

It takes some additional work to show that $\mathrm{d}\mu_d(r\sqrt{d})$ has a weak limit $\mathrm{d}\mu(r)$ in the sense of tempered distributions. Schoenberg's paper [Sch38, Thm. 2] approaches the problem using hard analysis. See Chafaï [Cha13] for a probabilistic argument. ∎

19.1.3 Inner-product kernels

Are there other families of kernels that operate in any dimension? Here is another class, based on inner products instead of Euclidean distances.

Definition 19.4 (Inner-product kernel). For a field $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, an *inner-product kernel* K on $\mathbb{F}^d$ is a function of the form

$$K_\varphi(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \varphi(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^d,$$

where $\varphi : \mathbb{F} \to \mathbb{F}$ is a function. The inner-product kernel is *positive definite* on $\mathbb{F}^d$ when the associated kernel matrices are psd:

$$\boldsymbol{K}_\varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \coloneqq \big[\varphi(\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle)\big]_{j,k=1,\dots,n} \succcurlyeq \mathbf{0}$$

for every choice of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{F}^d$ and all $n \in \mathbb{N}$.

We can characterize pd inner-product kernels more simply.

Exercise 19.5 (Psd preservers). Show that an inner-product kernel K_φ on $\mathbb{F}^d$ is positive definite if and only if

$$\big[a_{jk}\big] \succcurlyeq \mathbf{0} \quad \text{implies} \quad \big[\varphi(a_{jk})\big] \succcurlyeq \mathbf{0} \quad \text{for } \boldsymbol{A} = [a_{jk}] \in \mathbb{M}_d(\mathbb{F}).$$

That is, the function φ is an entrywise psd preserver. **Hint:** A matrix in $\mathbb{M}_d(\mathbb{F})$ is psd if and only if it is the Gram matrix of d points in $\mathbb{F}^d$.

This seems a bit perverse: we are always told that entrywise operations on matrices are unorthodox. Nevertheless, this exercise suggests that entrywise functions that preserve the psd property may merit study. This is the primary object of this lecture.

19.1.4 Applications

Before turning to our work, let us mention some motivating applications of psd inner-product kernels and entrywise psd preservers.

Example 19.6 (Kernel methods). As we briefly discussed in Lecture 18, we can use the kernel trick to design new methods for data analysis. In a method that only uses the Gram matrix of Euclidean data, we can substitute a psd kernel matrix to try to find alternative (non-Euclidean) structure in the data. Some commonly used inner-product kernels include polynomial kernels of the form

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle^p \quad \text{or} \quad K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle)^p \quad \text{for } p \in \mathbb{N}.$$

As we will see, these inner-product kernels are indeed pd. ∎

Example 19.7 (Covariance regularization). Suppose that we have acquired a psd covariance matrix $\boldsymbol{A} \succcurlyeq \mathbf{0}$ whose entries tabulate (estimated) covariances among a family of random variables. In practice, it is common that covariance estimates are inaccurate, so we may wish to process the matrix to mitigate the effects of noise. In some settings, it is important to ensure that the procedure respects the psd property so that the processed matrix is still a covariance matrix.

One class of inexpensive methods applies a scalar function f to each entry of the covariance matrix to obtain $[f(a_{jk})]$. In this context, it is natural to insist that the function f is an entrywise psd preserver. ∎

19.2 Entrywise functions that preserve the psd property

In this section, we initiate our study of scalar functions that act entrywise on a psd matrix to produce another psd matrix. We begin with basic definitions, and then we outline some of the main properties.

19.2.1 Entrywise functions

It is valuable to restrict our attention to matrices whose entries lie within specified sets.

Definition 19.8 (Entrywise matrix function). Let $\mathsf{E} \subseteq \mathbb{F}$. Define the set of $n \times n$ matrices with entries in E:

$$\mathbb{M}_n[\mathsf{E}] := \{\boldsymbol{A} \in \mathbb{M}_n(\mathbb{F}) : a_{jk} \in \mathsf{E} \text{ for all } j, k\}.$$

We can extend a scalar function $f : \mathsf{E} \to \mathbb{F}$ entrywise to matrices:

$$f[\boldsymbol{A}] := [f(a_{jk})] \in \mathbb{M}_n(\mathbb{F}) \quad \text{for each } \boldsymbol{A} = [a_{jk}] \in \mathbb{M}_n[\mathsf{E}].$$

In this lecture, we always use brackets to denote entrywise behavior.

The definition of an entrywise matrix function should be contrasted with our previous definition of a standard matrix function. Indeed, standard matrix functions were defined only for Hermitian matrices, while entrywise functions can be applied to any square (or even rectangular) matrix. It is, perhaps, surprising that there is anything interesting to say about these objects.

19.2.2 Entrywise psd preservers

Our main definition concerns entrywise functions that map certain psd matrices to psd matrices.

Definition 19.9 (Entrywise psd preserver). Let D be an interval ($\mathbb{F} = \mathbb{R}$) or a disc centered on the real line ($\mathbb{F} = \mathbb{C}$). We say that $f : \mathsf{D} \to \mathbb{F}$ is an *entrywise psd preserver (epp)* on $\mathbb{M}_n[\mathsf{D}]$ if

$$\boldsymbol{A} \succcurlyeq \mathbf{0} \quad \text{implies} \quad f[\boldsymbol{A}] \succcurlyeq \mathbf{0} \quad \text{for all } \boldsymbol{A} \in \mathbb{M}_n[\mathsf{D}].$$

If this claim holds true for all $n \in \mathbb{N}$, we simply say that f is an entrywise psd preserver (epp) on D.

Exercise 19.10 (Epps and positivity). If $\mathsf{D} \cap \mathbb{R}_+ = \emptyset$, show that the definition of an epp on $\mathbb{M}_n[\mathsf{D}]$ is vacuous. Otherwise, if $\mathsf{D} \cap \mathbb{R}_+ \neq \emptyset$, exhibit an example of an epp on $\mathbb{M}_n[\mathsf{D}]$.

As with positive linear maps and standard matrix functions, a differentiable epp is also monotone with respect to the psd order.

Problem 19.11 (Differentiable epps are monotone). For simplicity, assume that $\mathsf{D} \subseteq \mathbb{F}$ is an open set. Suppose that $f : \mathsf{D} \to \mathbb{F}$ is differentiable. Show that f is an epp on $\mathbb{M}_n[\mathsf{D}]$ if and only if

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \quad \text{implies} \quad f[\boldsymbol{A}] \preccurlyeq f[\boldsymbol{B}] \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n[\mathsf{D}].$$

Hint: The proof is essentially the same as the proof that a differentiable function is matrix monotone if and only if the Loewner matrix is psd. In this case, the argument is much easier because we have no need of the Daleckii–Krein formula.

19.2.3 Properties

Entrywise psd preservers enjoy some of the same properties as positive-definite functions.

Proposition 19.12 (Entrywise psd preserver). For a disc $\mathsf{D} \subseteq \mathbb{F}$, let $f : \mathsf{D} \to \mathbb{F}$ be an epp.

1. **Positivity.** For $a \in \mathsf{D} \cap \mathbb{R}_+$, the value $f(a) \geq 0$.
2. **Symmetry.** We have $f(z^*) = f(z)^*$ for all $z \in \mathsf{D}$.
3. **Bounds.** For $a, c, z \in \mathsf{D}$ with $a, c \geq 0$,

$$|f(z)|^2 \leq f(a)f(c) \quad \text{when } |z|^2 \leq ac.$$

In particular, $|f(z)| \leq f(a)$ if $|z| \leq a$.

We are using * to denote the complex conjugate.

Proof. Consider a psd matrix $\boldsymbol{A} \in \mathbb{M}_2[\mathsf{D}]$ and its entrywise image $f[\boldsymbol{A}]$:

$$\boldsymbol{A} = \begin{bmatrix} a & z \\ z^* & c \end{bmatrix} \succcurlyeq \mathbf{0} \quad \text{and} \quad f[\boldsymbol{A}] = \begin{bmatrix} f(a) & f(z) \\ f(z^*) & f(c) \end{bmatrix} \succcurlyeq \mathbf{0}.$$

Property (1) holds because the diagonal entries of a psd matrix are positive. Property (2) is a consequence of the symmetry or Hermiticity of a psd matrix. To prove (3), we invoke the Hadamard psd criterion. ∎

Next, we show that epps behave nicely on the strictly positive part of the real line.

Proposition 19.13 (Epp: Restriction to $\mathbb{R}_{++}$). For a disc $\mathsf{D} \subseteq \mathbb{F}$, let $f : \mathsf{D} \to \mathbb{F}$ be an epp. The restriction of $f : \mathsf{D} \cap \mathbb{R}_{++} \to \mathbb{R}_+$ is increasing and continuous.

Proof. Choose numbers $a, c \in \mathsf{D} \cap \mathbb{R}_{++}$ subject to the relation $0 < c < a$. The positivity and boundedness properties ensure that $f(c) = |f(c)| \leq f(a)$. Thus, f is increasing on $\mathsf{D} \cap \mathbb{R}_{++}$.

Define the function $g(x) := \log f(\mathrm{e}^x)$ whenever $\mathrm{e}^x \in \mathsf{D} \cap \mathbb{R}_{++}$. The boundedness property guarantees that $f(\sqrt{ac}) \leq \sqrt{f(a)f(c)}$. Write $a = \mathrm{e}^x$ and $c = \mathrm{e}^y$. Take the logarithm, and recognize the function g:

$$g(\tfrac{1}{2}x + \tfrac{1}{2}y) \leq \tfrac{1}{2}g(x) + \tfrac{1}{2}g(y).$$

In other words, g is midpoint convex, so it is continuous on the interior of its domain. Thus, f is continuous on $\mathsf{D} \cap \mathbb{R}_{++}$. ∎

Proposition 19.14 (Epp: Stability properties). Let $\mathsf{D} \subset \mathbb{F}$ be a disc.

1. **Convex cone.** If f, g are epps on D, then $\alpha f + \beta g$ is an epp on D for all $\alpha, \beta \geq 0$.
2. **Closedness.** If $(f_n : n \in \mathbb{N})$ is a sequence of epps on D that converges pointwise to f, then f is an epp on D.
3. **Multiplication.** If f, g are epps on D, then the pointwise product fg is also an epp on D.

Proof. The claims (1) and (2) are valid because the psd matrices form a convex cone that is closed. Meanwhile, claim (3) is a consequence of the Schur product theorem. ∎

There is only one setting where epps have been completely characterized: the case of 2×2 matrices with strictly positive entries [Vas79, Thm. 2]. We mention this result here because we will use this result to prove a weaker characterization theorem.

Exercise 19.15 (Epps: A special characterization). Show that f is an epp on $\mathbb{M}_2[\mathbb{R}_{++}]$ if and only if $x \mapsto \log f(\mathrm{e}^x)$ is increasing and midpoint convex.

19.3 Examples of entrywise psd preservers

In this section, we will introduce several examples of epps on various domains. For simplicity, we restrict our attention to the real setting from now on. Thus, D is either the whole real line or an interval of the real line.

19.3.1 Monomials

The simplest example of an epp is the function that always returns one.

Proposition 19.16 (Constants are epps). The function $f(t) = 1$ is an epp on $\mathbb{R}$.

Proof. For a psd matrix $\boldsymbol{A} \in \mathbb{M}_n[\mathbb{R}]$, we see that $f[\boldsymbol{A}] = \mathbf{1}\mathbf{1}^{\mathsf{T}}$, which is psd. ∎

Next, we turn our attention to the monomials.

Proposition 19.17 (Monomials are epps). For each $p \in \mathbb{N}$, the function $f(t) = t^p$ is an epp on $\mathbb{R}$.

Proof. For a psd matrix $\boldsymbol{A} \in \mathbb{M}_n[\mathbb{R}]$, we can express

$$f[\boldsymbol{A}] = \underbrace{\boldsymbol{A} \odot \cdots \odot \boldsymbol{A}}_{p \text{ times}} \succcurlyeq \mathbf{0}.$$

As usual, $\odot$ denotes the Schur product. Since $\boldsymbol{A} \succcurlyeq \mathbf{0}$, an iterative application of the Schur product theorem guarantees that the product is psd. ∎

Problem 19.18 (Other powers). Choose a positive number r that is not an integer. Prove that the function $t \mapsto t^r$ is not an epp on $\mathbb{R}_+$. **Hint:** This is hard. One approach is to argue that every epp on $\mathbb{M}_n[\mathbb{R}_+]$ must have $n - 1$ continuous derivatives, and choose $n > r$. See the proofs of Vasudeva's theorem and Bernstein's theorems (below) for some relevant techniques.

19.3.2 Power series

Let us take a step up in abstraction, which will allow us to produce a wealth of additional examples.

Proposition 19.19 (Some power series that are epps). Consider a power series

$$f(t) := \sum\nolimits_{p=0}^{\infty} c_p t^p \quad \text{where } c_p \geq 0 \text{ for } p \in \mathbb{N}. \tag{19.1}$$

If $\mathsf{D} \subseteq \mathbb{R}$ is the domain of convergence, then $f : \mathsf{D} \to \mathbb{R}$ is an epp on D.

Proof. The monomials are epps on $\mathbb{R}$, and the class of epps on $\mathbb{R}$ is a convex cone. Therefore, each polynomial with positive coefficients is an epp on $\mathbb{R}$. Within the domain D of convergence, the partial sums of f are polynomials with positive coefficients that converge pointwise to f. Since the class of epps on D is closed under pointwise limits, we see that f is an epp on D. ∎

Later, we will discuss the class of power series with positive coefficients in greater detail. For the moment, we note a few basic properties.

Exercise 19.20 (Power series with positive coefficients). Suppose that $f : \mathsf{D} \to \mathbb{R}$ has the form (19.1). Then $f \in \mathsf{C}^\infty(\mathsf{D})$, the set of infinitely differentiable functions on D. Furthermore, for each $p \in \mathbb{N}$, the derivative $f^{(p)}$ is also an epp on D.

19.3.3 The exponential and friends

To exploit the ideas from the last section, we first consider some power series related to the exponential function.

Proposition 19.21 (Exponentials are epps). For $a > 0$, the functions $t \mapsto \exp(at)$ and $t \mapsto \sinh(at)$ and $t \mapsto \cosh(at)$ are all epps on $\mathbb{R}$.

Proof. Scaling a matrix by $a > 0$ preserves positivity, so we may assume that $a = 1$. Recall that the exponential, hyperbolic sine, and hyperbolic cosine are entire functions, so their power series converge on the real line:

$$\exp(t) = \sum\nolimits_{p=0}^{\infty} \frac{1}{p!} t^p,$$

$$\cosh(t) = \sum\nolimits_{p=0}^{\infty} \frac{1}{2p!} t^{2p}, \quad \text{and} \quad \sinh(t) = \sum\nolimits_{p=0}^{\infty} \frac{1}{(2p+1)!} t^{2p+1}.$$

Clearly, each of these series has positive coefficients. Invoke Proposition 19.19. ∎

This proposition has an elegant and unexpected outcome. If the matrix $[a_{jk}]$ is psd, then the matrix $[\mathrm{e}^{a_{jk}}]$ is also psd. We can actually strengthen the result further.

Problem 19.22 (Exponentials and cpsd matrices). Suppose that $\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \geq 0$ for all vectors $\boldsymbol{u}$ with $\operatorname{tr}[\boldsymbol{u}] = 0$. Matrices of this type are called *conditionally psd (cpsd)*. Prove that $\exp[\boldsymbol{A}]$ is psd. This result is valid with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.

19.3.4 The logarithm and the inverse

Next, we consider some important power series that converge only on a subinterval of the line.

Proposition 19.23 (Logarithms and inverses that are epps). The inverse $t \mapsto (1-t)^{-1}$ and the logarithm $t \mapsto -\log(1-t)$ are epps on $(-1, +1)$.

Proof. We have the power series expansions

$$(1-t)^{-1} = \sum\nolimits_{p=0}^{\infty} t^p \quad \text{and} \quad -\log(1-t) = \sum\nolimits_{p=1}^{\infty} \frac{1}{p} t^p.$$

These series have positive coefficients, but they converge only in the open interval $\mathsf{D} = (-1, +1)$. Proposition 19.19 implies that these two functions are epps on the interval $(-1, +1)$. ∎

19.3.5 More examples

Several other elementary functions are also epps. We leave these facts as an exercise.

Exercise 19.24 (Elementary functions that are epps). Show that the following functions are epps by inspection of their power series. Make plots to compare these functions.

1. **Secant.** The function $t \mapsto \sec(2t/\pi)$ is an epp on $(-1, +1)$.
2. **Tangent.** The function $t \mapsto \tan(2t/\pi)$ is an epp on $(-1, +1)$.
3. **Arctanh.** The function $t \mapsto \operatorname{arctanh}(t)$ is an epp on $(-1, +1)$.
4. **Arcsine.** The function $t \mapsto \arcsin(t)$ is an epp on $[-1, +1]$.
5. **Log gamma.** The function $t \mapsto \log \Gamma(1-t)$ is an epp on $(-1, +1)$.

Hint: Use the magisterial *Handbook of Mathematical Functions*, prepared by Milton Abramowitz & Irene Stegun [AS64] and updated by Frank Olver et al. [Olv+10]. Or just turn to Wikipedia if you consider it trustworthy.

19.4 Absolutely monotone and completely monotone functions

All our examples of entrywise psd preservers are based on power series with positive coefficients. Although this may appear to reflect a lack of inspiration, there is a deeper reason. In the real setting, it can be shown that every epp is given by such a power series, at least under some regularity conditions. We will discuss results of this type in the next section.

As a preliminary, we need to take some time to develop the theory of power series with positive coefficients. These objects also arise from the study of absolutely monotone functions, and they have a very long history in analysis. In this subject, one must pay close attention to differentiability assumptions and the behavior of a function at the endpoints of its domain. Our exposition is modeled after Widder's classic book [Wid41, Chap. IV]. We elaborate on this discussion below in Section 19.6.

19.4.1 Absolute monotonicity

Bernstein introduced the concept of absolutely monotonicity for an infinitely differentiable function.

Definition 19.25 (Absolutely monotone function). An *infinitely differentiable* function $f : (a, b) \to \mathbb{R}$ on an open interval of the real line is called *absolutely monotone* if its derivatives are positive:

$$f^{(p)}(t) \geq 0 \quad \text{for all } t \in (a, b) \text{ and all } p \in \mathbb{Z}_+.$$

For a *left-closed, right-open* interval, we say that $f : [a, b) \to \mathbb{R}$ is absolutely monotone if f is continuous on $[a, b)$ and absolutely monotone on (a, b).

We allow $a, b \in \{\pm\infty\}$ in this definition.

Exercise 19.26 (Absolute monotonicity: Right-hand derivatives). Suppose that $f : [a, b) \to \mathbb{R}$ is absolutely monotone. This includes the assumption that $f(a) := \lim_{h \downarrow 0} f(a + h)$ exists and is positive. Show that the right-derivative $f'(a) := \lim_{h \downarrow 0} f'(a + h)$ exists and is positive. With this choice of $f'(a)$, confirm that f' is continuous on $[a, b)$. By induction, deduce that $f^{(p)}$ is a positive, continuous function on $[a, b)$.

Exercise 19.27 (Absolute monotonicity: Properties). An absolutely monotone function is positive, increasing, and convex.

Exercise 19.28 (Absolute monotonicity: Stability). On an open or half-open interval, the absolutely monotone functions compose a convex cone. The product of two absolutely monotone functions is absolutely monotone. The derivatives of an absolutely monotone are also absolutely monotone.

19.4.2 Absolutely monotone functions are analytic

In general, the existence of derivatives does not imply that a function has a power series representation. Absolutely monotone functions, however, do enjoy this property. This is called the "little" Bernstein theorem.

Theorem 19.29 (Bernstein). Suppose that $f : [a, b) \to \mathbb{R}$ is absolutely monotone on a half-open interval. Write $r := b - a$. Then

$$f(t) = \sum\nolimits_{p=0}^{\infty} \frac{f^{(p)}(a)}{p!} (t - a)^p \quad \text{for all } t \in (a - r, a + r). \tag{19.2}$$

Moreover, f extends to an analytic function on the open disc $\mathsf{D}_a(r)$ centered at a with radius r.

In particular, if $f : \mathbb{R}_+ \to \mathbb{R}$ is absolutely monotone, then it extends to an entire function whose power series expansion at zero has positive coefficients.

Proof. Exercise 19.26 shows that f has right-derivatives of all orders at $t = a$. For each $n \in \mathbb{N}$, we may expand f as a Taylor series with an integral remainder:

$$f(t) = \sum\nolimits_{p=0}^{n-1} \frac{f^{(p)}(a)}{p!}(t-a)^p + R_n(t) \quad \text{when } a \le t \le c < b.$$

We may express the remainder in the form

$$R_n(t) = \frac{1}{(n-1)!}\int_a^t f^{(n)}(s)(c-s)^{n-1}\left(\frac{t-s}{c-s}\right)^{n-1} \mathrm{d}s.$$

Since the fraction in the integrand decreases as a function of t and the derivatives $f^{(n)}$ are positive on the interval $[a, b)$,

$$\begin{aligned} R_n(t) &\le \left(\frac{t-a}{c-a}\right)^{n-1} \cdot \frac{1}{(n-1)!}\int_a^t f^{(n)}(t)(c-s)^{n-1}\,\mathrm{d}s \\ &\le \left(\frac{t-a}{c-a}\right)^{n-1} \cdot \frac{1}{(n-1)!}\int_a^c f^{(n)}(t)(c-s)^{n-1}\,\mathrm{d}s \\ &= \left(\frac{t-a}{c-a}\right)^{n-1} \cdot R_n(c) \le \left(\frac{t-a}{c-a}\right)^{n-1} \cdot f(c). \end{aligned}$$

Indeed, $R_n(c) \le f(c)$ because absolutely monotonicity ensures that all terms in the partial Taylor series are all positive. We deduce that $R_n(t) \to 0$ for all $t \in [a, c]$. Therefore, the series expansion (19.2) converges for $t \in [a, c]$. Since $c < b$ is arbitrary, we may extend the interval of definition to $[a, b)$.

Finally, note that each summand in the series (19.2) is positive when $t \in [a, a + r)$. In other words, the series converges absolutely when $t \in [a, a + r)$. As a consequence, the series also converges for every $t \in (a - r, a + r)$. In fact, the same conclusion extends to $t \in \mathsf{D}_a(r)$. ∎

19.5 Vasudeva's theorem

We are now prepared to state and prove a partial converse to Proposition 19.19. Every epp on the strictly positive real line can be written as a power series with positive coefficients [Vas79].

Theorem 19.30 (Entrywise psd preservers on $\mathbb{R}_{++}$; Vasudeva 1979). A function $f : \mathbb{R}_{++} \to \mathbb{R}$ is an epp if and only if it is absolutely monotone. That is,

$$f(t) = \sum\nolimits_{p=0}^{\infty} c_p t^p \quad \text{where } c_p \ge 0 \text{ for all } p \in \mathbb{N}.$$

This statement includes the claim that the series converges for all $t > 0$.

In Proposition 19.15, we have already seen that an epp on $\mathbb{R}_{++}$ is necessarily continuous. Theorem 19.30 asserts that an epp on $\mathbb{R}_{++}$ is actually an *analytic* function. In this section, we give a complete proof of Vasudeva's theorem.

If we add a much stronger differentiability assumption, we can reach a similar conclusion for the whole real line.

Corollary 19.31 (Entrywise psd preservers on $\mathbb{R}$). An *analytic* function $f : \mathbb{R} \to \mathbb{R}$ is an epp if and only if it its power series expansion at zero has positive coefficients.

Proof. The reverse implication is Proposition 19.19.

For the forward direction, we assume that f is analytic, so it has a power series expansion about zero that converges on $\mathbb{R}$. Since f is also an epp on the strictly positive line $\mathbb{R}_{++}$, Vasudeva's theorem guarantees that the coefficients in this power series are positive. ∎

Corollary 19.31 is actually true without any regularity assumption, but this claim is somewhat harder to prove. See Section 19.5.4 for related results of this type.

19.5.1 Smooth epps have positive derivatives

The key to the proof of Vasudeva's theorem is a calculation for a smooth epp f on $\mathbb{R}_{++}$. The idea is to exploit the epp property to deduce that derivatives of f are positive. This idea dates back to work of Loewner and Horn [Hor69].

Lemma 19.32 (Smooth epp: Positive derivatives). Consider an *infinitely differentiable* function $f : \mathbb{R}_{++} \to \mathbb{R}$ that is an epp on $\mathbb{R}_{++}$. Then $f^{(p)} \geq 0$ for all $p \in \mathbb{Z}_+$.

Proof. Fix a point $a > 0$. Choose a positive integer $p \in \mathbb{Z}_+$, and write $d = p + 1$. From the assumption that f is an epp, we may extract a family of inequalities:

$$\boldsymbol{u}^* f[a\mathbf{1}\mathbf{1}^* + t\boldsymbol{c}\boldsymbol{c}^*]\boldsymbol{u} \geq 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^d \text{ and } \boldsymbol{c} \in \mathbb{R}^d_{++} \text{ and small } t > 0.$$

Indeed, when t is sufficiently small, $a\mathbf{1}\mathbf{1}^{\mathsf{T}} + t\boldsymbol{c}\boldsymbol{c}^*$ is a psd matrix with strictly positive entries. By the pth order Taylor expansion of f about a with a derivative remainder, we find that

$$\sum\nolimits_{j,k=1}^{d} u_j u_k \left[\sum\nolimits_{r=0}^{p-1} \frac{(tc_jc_k)^r}{r!} f^{(r)}(a) + \frac{(tc_jc_k)^p}{p!} f^{(p)}(a + t\theta_{jk}c_jc_k) \right] \geq 0.$$

The scaling coefficients $\theta_{jk} \in (0, 1)$, but they may depend on everything else.

To continue, we require the entries of $\boldsymbol{c} \in \mathbb{R}^d_{++}$ to be *distinct*. Consider the $d \times d$ Vandermonde matrix

$$\boldsymbol{V} = \begin{bmatrix} 1 & c_1 & c_1^2 & \dots & c_1^p \\ 1 & c_2 & c_2^2 & \dots & c_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & c_d & c_d^2 & \dots & c_d^p \end{bmatrix}. \tag{19.3}$$

Since the entries of $\boldsymbol{c}$ are distinct, the Vandermonde matrix $\boldsymbol{V}$ is nonsingular (Problem 19.33). Therefore, we can find a vector $\boldsymbol{u} \in \mathbb{R}^d$ that is orthogonal to the first $p - 1$ columns but not orthogonal to the pth column. Equivalently, $\sum_j u_j c_j^r = 0$ for $0 \leq r < p$, while $\sum_j u_j c_j^p \neq 0$.

With these choices, the sum in the penultimate display collapses to the form

$$\frac{t^p}{p!} \sum\nolimits_{j,k=1}^{d} u_j u_k \cdot c_j^p c_k^p \cdot f^{(p)}(a + t\theta_{jk}c_jc_k) \geq 0.$$

Clear the leading fraction. Take $t \downarrow 0$ to conclude that $f^{(p)}(a) \geq 0$. ∎

Problem 19.33 (Vandermonde determinants). Prove that the Vandermonde matrix (19.3) satisfies $\det(\boldsymbol{V}) = \prod_{i \neq j}(c_i - c_j)$. **Hint:** Compute the LU factorization of $\boldsymbol{V}$.

19.5.2 Smoothing

Next, we show that it is possible to take an arbitrary epp on $\mathbb{R}_{++}$ and smooth it to obtain an infinitely differentiable epp. This result requires the use of a mollifier, a nice function that can confer its nice properties onto another function.

Fact 19.34 (Mollifier). Fix $\delta \in (0, 1)$. There is a probability density function $h_\delta : \mathbb{R}_{++} \to \mathbb{R}_+$ that has compact support, is infinitely differentiable, has expectation 1, and has variance δ. ∎

Exercise 19.35 (Epp: Smoothing). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a *continuous* function. For $\delta > 0$, introduce the mollifier h_δ, and construct the smoothed function

This integral can be interpreted as a convolution of two functions on the abelian group $(\mathbb{R}_{++}, \times)$.

$$f_\delta(t) := \int_0^\infty f(t/x) \cdot h_\delta(x) \cdot \frac{\mathrm{d}x}{x} \quad \text{for } t \geq 0.$$

Establish the following properties.

1. **Continuity.** The function f_δ is continuous on $\mathbb{R}_+$.
2. **Smoothness.** The function f_δ is infinitely differentiable on $\mathbb{R}_{++}$.
3. **Limits.** As $\delta \downarrow 0$, the sequence $f_\delta \to f$ pointwise.
4. **Epp.** If f is an epp on $\mathbb{R}_{++}$, then f_δ is also an epp on $\mathbb{R}_{++}$.

Hint: The first three parts follow from dominated convergence, perhaps after the change of variables $u = t/x$. For the last part, approximate the integral by a Riemann sum.

19.5.3 Proof of Theorem 19.30

We may now establish Vasudeva's result (Theorem 19.30). The reverse implication is Proposition 19.19. For the forward implication, the basic idea is to smooth the epp. The smoothed epp has positive derivatives, so it must be absolutely monotone. Under some regularity conditions, the limit of absolutely monotone functions remains absolutely monotone, so we can transfer the power series representation of the smoothed epp back to the original epp.

Assume that $f : \mathbb{R}_{++} \to \mathbb{R}_+$ is an epp on $\mathbb{R}_{++}$. By Proposition 19.15, the epp f is continuous and increasing on $\mathbb{R}_{++}$. Thus, we can extend f to a continuous function on $\mathbb{R}_+$ with the limiting value $f(0) = \inf_{t>0} f(t)$.

For each $\delta > 0$, introduce the smoothed function $f_\delta : \mathbb{R}_+ \to \mathbb{R}_+$, as in Exercise 19.35. The function f_δ is an infinitely differentiable epp on $\mathbb{R}_{++}$. Therefore, we may invoke Lemma 19.32 to determine that $f_\delta^{(p)} \geq 0$ on $\mathbb{R}_{++}$ for all $p \in \mathbb{Z}_+$. Since f_δ is continuous on $\mathbb{R}_+$, the "little" Bernstein theorem (Theorem 19.29) implies that f_δ is an entire function whose power series expansion at zero has positive coefficients:

$$f_\delta(t) = \sum\nolimits_{p=0}^\infty c_p(\delta) t^p \quad \text{with } c_p(\delta) \geq 0 \text{ for all } t \in \mathbb{R}.$$

Uniformly in δ, the coefficients are absolutely summable because $\sum_{p=0}^\infty c_p(\delta) = f_\delta(1) \to f(1)$. We will apply some standard results from complex analysis to complete the argument.

Exercise 19.35 implies that $f_\delta \to f$ pointwise as $\delta \downarrow 0$. We need to upgrade the convergence. The family $(f_\delta : 0 < \delta < 1)$ is uniformly bounded on each compact set in $\mathbb{C}$ because the coefficients in the power series for f_δ are uniformly absolutely summable. Passing to a subsequence if necessary, $f_\delta \to f$ uniformly on each compact set in $\mathbb{C}$ [Ahl66, Thm. 12, p. 217].

This result follows from the Arzelà–Ascoli theorem.

Finally, if a sequence of analytic functions converges uniformly on each compact set in $\mathbb{C}$, then the limit is an analytic function [Ahl66, Thm. 1, p. 174]. We deduce that f is analytic. Therefore,

This result is an easy consequence of Morera's theorem.

$$f(t) = \sum\nolimits_{p=0}^{\infty} c_p t^p \quad \text{with } c_p \geq 0 \text{ for all } t \in \mathbb{R}.$$

Indeed, the coefficients must be positive because we have taken the limit of series with positive coefficients. This is the required result.

19.5.4 Variations

Vasudeva's theorem is, perhaps, the simplest of several results that characterize entrywise psd preservers. A key feature of these results is that they do not make strong prior assumptions on the smoothness properties of the epp, but rather extract smoothness as a consequence of the structural assumption.

The first theorem of this type was derived by Schoenberg [Sch38].

Theorem 19.36 (Schoenberg 1938). A *continuous* function $f : [-1, +1] \to \mathbb{R}$ is an epp if and only if it has a representation as a power series about zero with positive coefficients.

Somewhat later, Rudin [Rud59] removed the continuity condition from Schoenberg's theorem.

Theorem 19.37 (Rudin 1959). A function $f : (-1, +1) \to \mathbb{R}$ is an epp if and only if it has a representation as a power series about zero with positive coefficients.

Rudin conjectured that there was an extension of his result to the complex setting, and this result was obtained soon after by Herz [Her63].

Theorem 19.38 (Herz 1963). A function $f : \mathsf{D}_0(1) \to \mathbb{C}$ is an epp if and only if it has a representation of the form

$$f(z) = \sum\nolimits_{p,q \geq 0} c_{pq} z^p (z^*)^q \quad \text{where } c_{pq} \geq 0.$$

See the survey [Bel+18] for more discussion of these results, as well as recent advances.

19.6 *Completely monotone functions

The rest of this lecture is for general education. Absolutely monotone functions have a counterpart where the derivatives alternate sign.

Definition 19.39 (Complete monotonicity). An *infinitely differentiable* function $f : (a, b) \to \mathbb{R}$ on an open interval of the real line is called *completely monotone* if its derivatives alternate sign:

$$(-1)^p f^{(p)} \geq 0 \quad \text{for all } t \in (a, b) \text{ and all } p \in \mathbb{Z}_+.$$

Exercise 19.40 (Absolute versus complete). Show that f is completely monotone on (a, b) if and only if its reversal $t \mapsto f(-t)$ is absolutely monotone on $(-b, -a)$.

In many ways, it is more natural to study completely monotone functions, and we will see that they enjoy a remarkable integral representation.

19.6.1 Differences and derivatives

Completely monotone functions also can be defined without continuity assumptions by using right-difference operators. In this case, it is natural to insist that the domain of the function is the positive (or strictly positive) real line.

Definition 19.41 (Right difference). For $h > 0$, the *right-difference* operator of width h is defined on functions $f : \mathbb{R}_+ \to \mathbb{R}$ via the rule

$$\Delta_h f : a \mapsto f(a+h) - f(a) \quad \text{for } a \in \mathbb{R}_+.$$

Higher-order differences are defined iteratively:

$$\Delta_h^{p+1} f : a \mapsto \Delta_h^p(a+h) - \Delta_h^p(a) \quad \text{for } a \in \mathbb{R}_+.$$

Exercise 19.42 (Higher-order differences). Prove that the iterated differences satisfy

$$\Delta_h^p f(a) = \sum\nolimits_{k=0}^{p} (-1)^{p-k} \binom{p}{k} f(a+kh).$$

With this concept, we can give another definition of complete monotonicity that does not rely on differentiability properties.

Definition 19.43 (Completely monotone function II). A function $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotone if

$$(-1)^p \Delta_h^p f \geq 0 \quad \text{for all } h > 0 \text{ and all } p \in \mathbb{Z}_+.$$

The definition in terms of differences essentially implies the condition on derivatives in the original definition.

Exercise 19.44 (Differences and derivatives). Assume that f is infinitely differentiable on $\mathbb{R}_{++}$ and completely monotone in the sense of Definition 19.43. Prove that the derivatives of f alternate sign: $(-1)^p f^{(p)} \geq 0$ on $\mathbb{R}_{++}$ for each $p \in \mathbb{Z}_+$.

Exercise 19.45 (Completely monotone function: Properties). Suppose that $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotone in the sense of Definition 19.43. Prove that f is positive, decreasing, and convex. Therefore, f is continuous on $\mathbb{R}_{++}$.

Exercise 19.46 (Completely monotone function: Transforms). Suppose that $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotone in the sense of Definition 19.43. Establish the following properties.

1. **Scaling.** For $a > 0$, the scaling af is completely monotone.
2. **Differences.** For $h > 0$, the negative difference $-\Delta_h f$ is completely monotone.
3. **Translation.** For $c > 0$, the shift $t \mapsto f(t+c)$ is completely monotone.
4. **Dilation.** For $a > 0$, the dilation $t \mapsto f(at)$ is completely monotone.

Exercise 19.47 (Completely monotone functions: Convex cone). Prove that the completely monotone functions on $\mathbb{R}_+$ compose a convex cone that is closed under pointwise limits.

19.6.2 Bernstein's theorem on completely monotone functions

As we have mentioned, completely monotone functions have a beautiful integral representation. This result is called the "big" Bernstein theorem. Our proof is drawn from Lax [Lax02, Sec. 14.3].

Theorem 19.48 (Bernstein). A function $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotone as in Definition 19.43 if and only if it is the Laplace transform of a finite, positive measure μ on $\overline{\mathbb{R}}_+$. That is, for some $c \geq 0$,

$$f(t) = c\delta_0(t) + \int_{\mathbb{R}_+} \mathrm{e}^{-tx} \,\mathrm{d}\mu(x) \quad \text{for all } t \in \mathbb{R}_+.$$

This expression can be written more compactly as an integral over $\overline{\mathbb{R}}_+$ with the understanding that $t \mapsto \mathrm{e}^{-\infty \cdot t} = \delta_0(t)$.

A remarkable consequence of Theorem 19.48 is that a completely monotone function is not just continuous on $\mathbb{R}_{++}$ but also *analytic* on $\mathbb{R}_{++}$. Therefore, Definition 19.39 is actually equivalent with Definition 19.43 up to the behavior at $t = 0$.

It is also rather interesting to compare Bernstein's theorem with Bochner's theorem. The former tells us what happens if we take the Laplace transform of a finite, positive measure, while the latter tells us what happens if we take the Fourier transform.

Proof sketch. The reverse direction simply asks us to verify that the integral represents a completely monotone function. This is an easy calculation.

Consider the convex set of normalized completely monotone functions:

$$\mathsf{B} \coloneqq \{f : \mathbb{R}_+ \to \mathbb{R} \text{ with } f(0) = 1 \text{ and } f \text{ completely monotone}\}.$$

Since completely monotone functions are positive and decreasing, every function $f \in \mathsf{B}$ satisfies $0 \leq f(t) \leq 1$ for $t \in \mathbb{R}_+$. After some argument, it follows that the set B is compact in the topology of pointwise convergence.

Let e be an extreme point of B. If $e(t) = 1$ for some $t > 0$, then $e(t) = 1$ because e is convex and decreasing. If $e(t) = 0$ for all $t > 0$, then $e(t) = \delta_0(t)$.

Let us exclude these two cases. Then e must be continuous on $\mathbb{R}_+$. Indeed, a completely monotone function is continuous on $\mathbb{R}_{++}$. But e cannot have a discontinuity at zero, or else it is a convex combination of δ_0 and another function in B.

To continue, observe that there is a point $a_0 > 0$ where $0 < e(a_0) < 1$. Since e is continuous, we have $0 < e(a) < 1$ on the interval $0 < a \leq a_0$. Fix a point $0 < a \leq a_0$, and define two functions

$$f_a(t) \coloneqq \frac{e(t+a)}{e(a)} \quad \text{and} \quad g_a(t) \coloneqq \frac{e(t) - e(t+a)}{1 - e(a)} \quad \text{for } t \in \mathbb{R}_+.$$

Both functions $f_a, g_a \in \mathsf{B}$, and clearly

$$e(t) = e(a) \cdot f_a(t) + (1 - e(a)) \cdot g_a(t) \quad \text{for all } t \in \mathbb{R}_+.$$

Since e is an extreme point, $e = f_a = g_a$. From the definition of f_a, we see that

$$e(t+a) = e(t)e(a) \quad \text{for all } t \in \mathbb{R}_+ \text{ when } 0 < a \leq a_0.$$

The only continuous solution to these equations with $e(0) = 1$ and $0 < e(t) \leq 1$ takes the form $e(t) = \mathrm{e}^{-tx}$ for some $x > 0$.

The integral representation follows from a routine application of the Krein–Milman theorem. ∎

Notes

The presentation in this lecture is new. We have used the survey article [Bel+18] as a guide to the literature, and we have adapted some of the organization from them. Widder's book [Wid41] remains an excellent source for material on Laplace transforms,

absolutely monotone functions, and completely monotone functions. We have revised Vasudeva's proof to make parts of it more transparent. The simple proof of the "big" Bernstein theorem via Krein–Milman is adapted from Lax [Lax02] with some mistakes corrected.

Lecture bibliography

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. For sale by the Superintendent of Documents. U. S. Government Printing Office, Washington, D.C., 1964.

[Ahl66] L. V. Ahlfors. *Complex analysis: An introduction of the theory of analytic functions of one complex variable*. Second. McGraw-Hill Book Co., New York-Toronto-London, 1966.

[Bel+18] A. Belton et al. "A panorama of positivity". Available at https://arXiv.org/abs/1812.05482. 2018.

[Cha13] D. Chafaï. *A probabilistic proof of the Schoenberg theorem*. 2013. URL: https://djalil.chafai.net/blog/2013/02/09/a-probabilistic-proof-of-the-schoenberg-theorem/.

[Her63] C. S. Herz. "Fonctions opérant sur les fonctions définies-positives". In: *Ann. Inst. Fourier (Grenoble)* 13 (1963), pages 161–180. URL: http://aif.cedram.org/item?id=AIF_1963__13__161_0.

[Hor69] R. A. Horn. "The theory of infinitely divisible matrices and kernels". In: *Trans. Amer. Math. Soc.* 136 (1969), pages 269–286. DOI: 10.2307/1994714.

[Lax02] P. D. Lax. *Functional analysis*. Wiley-Interscience, 2002.

[Olv+10] F. W. J. Olver et al., editors. *NIST handbook of mathematical functions*. With 1 CD-ROM (Windows, Macintosh and UNIX). National Institute of Standards and Technology, 2010.

[Rud59] W. Rudin. "Positive definite sequences and absolutely monotonic functions". In: *Duke Math. J.* 26 (1959), pages 617–622. URL: http://projecteuclid.org/euclid.dmj/1077468771.

[Sch38] I. J. Schoenberg. "Metric spaces and positive definite functions". In: *Trans. Amer. Math. Soc.* 44.3 (1938), pages 522–536. DOI: 10.2307/1989894.

[Vas79] H. Vasudeva. "Positive definite matrices and absolutely monotonic functions". In: *Indian J. Pure Appl. Math.* 10.7 (1979), pages 854–858.

[Wid41] D. V. Widder. *The Laplace Transform*. Princeton University Press, Princeton, N. J., 1941.

II. problem sets

1. Multilinear Algebra & Majorization

Date: 4 January 2022

This assignment covers multilinear algebra and majorization. The problems are optional, but keep in mind that you have to do mathematics to learn mathematics!

1. **Kron Job.** Let H be an n-dimensional inner-product space over the field $\mathbb{F}$ with orthonormal basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$. We write the matrix of a linear operator on H with respect to this distinguished basis. We arrange the family $(\mathbf{e}_i \otimes \mathbf{e}_j : 1 \leq i, j \leq n)$ of elementary tensors in $\mathsf{H} \otimes \mathsf{H}$ in lexicographic order. That is, $\mathbf{e}_i \otimes \mathbf{e}_j$ precedes $\mathbf{e}_k \otimes \mathbf{e}_\ell$ when $i < k$ or else $i = k$ and $j < \ell$.

    1. Show that $(\mathbf{e}_i \otimes \mathbf{e}_j : i, j)$ is a basis for $\mathsf{H} \otimes \mathsf{H}$. The easiest way to do this is via the identification of elementary tensors as linear functionals on the space of bilinear forms. When are two of these linear functionals linearly independent?

    2. Suppose that $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are linearly independent vectors. Show that

    $$\boldsymbol{x} \otimes \boldsymbol{y} + \boldsymbol{z} \otimes \boldsymbol{w}$$

    is an elementary tensor if and only if $\boldsymbol{w} = \alpha\, \boldsymbol{y}$ for some scalar $\alpha \in \mathbb{F}$.

    3. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be linear operators on H, so $\boldsymbol{A} \otimes \boldsymbol{B}$ is a linear operator on $\mathsf{H} \otimes \mathsf{H}$. Show that the matrix representation of this operator with respect to the ordered basis for $\mathsf{H} \otimes \mathsf{H}$ takes the form

    $$\mathscr{M}(\boldsymbol{A} \otimes \boldsymbol{B}) = \begin{bmatrix} a_{11}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ \vdots & & \vdots \\ a_{n1}\boldsymbol{B} & \dots & a_{nn}\boldsymbol{B} \end{bmatrix} \in \mathbb{F}^{n^2 \times n^2}$$

    The block matrix is called the *Kronecker product* of $\boldsymbol{A}$ and $\boldsymbol{B}$. It gives a concrete representation of the tensor product $\boldsymbol{A} \otimes \boldsymbol{B}$. We typically omit the $\mathscr{M}$ in this notation.

    4. The operator $\mathrm{vec} : \mathbb{F}^{n \times n} \to \mathbb{F}^{n^2}$ maps a matrix to a vector by concatenating the columns in order from left to right. Show that

    $$(\boldsymbol{A} \otimes \mathbf{I}) \cdot \mathrm{vec}(\boldsymbol{M}) = \mathrm{vec}(\boldsymbol{M}\boldsymbol{A}^{\mathsf{T}}) \quad \text{and} \quad (\mathbf{I} \otimes \boldsymbol{B}) \cdot \mathrm{vec}(\boldsymbol{M}) = \mathrm{vec}(\boldsymbol{B}\boldsymbol{M})$$

    Interpret the tensors $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{B}$ as right- and left-multiplication operators. Without making any further calculations, use this interpretation to explain why $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{B}$ must commute.

    5. Assume for concreteness that $\boldsymbol{A}$ and $\boldsymbol{B}$ are Hermitian. Describe the eigenvalues and eigenvectors of $\boldsymbol{A} \otimes \mathbf{I}$, of $\mathbf{I} \otimes \boldsymbol{B}$, of $\boldsymbol{A} \otimes \boldsymbol{B}$, and of $\boldsymbol{A} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{B}$.

    6. Assume that $\boldsymbol{A}$ is Hermitian. What are the eigenvalues of the operator $\boldsymbol{A} \otimes \mathbf{I} \otimes \cdots \otimes \mathbf{I} + \cdots + \mathbf{I} \otimes \cdots \otimes \mathbf{I} \otimes \boldsymbol{A}$ acting on $\otimes^k \mathsf{H}$? Each of the k terms is a tensor product of $k - 1$ identity operators and one copy of $\boldsymbol{A}$.

7. (*) The *Schur product* of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{A} \cdot \boldsymbol{B} = [a_{ij} b_{ij}]$. Relate the Schur product to the Kronecker product, and use this fact to establish Schur's Product Theorem: If $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite, then $\boldsymbol{A} \cdot \boldsymbol{B}$ is positive semidefinite.

2. **Wedge and vee.** Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathsf{H}$, and let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_k \in \mathsf{H}$.

   1. Verify that the antisymmetric tensor product is related to the determinant:
   $$\langle \boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k,\ \boldsymbol{y}_1 \wedge \cdots \wedge \boldsymbol{y}_k \rangle = \det [\langle \boldsymbol{x}_i,\ \boldsymbol{y}_j \rangle].$$
   2. (*) Verify that the symmetric tensor product is related to the permanent:
   $$\langle \boldsymbol{x}_1 \vee \cdots \vee \boldsymbol{x}_k,\ \boldsymbol{y}_1 \vee \cdots \vee \boldsymbol{y}_k \rangle = \operatorname{per} [\langle \boldsymbol{x}_i,\ \boldsymbol{y}_j \rangle].$$
   3. Show that $\boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{x}_k = \boldsymbol{0}$ if and only if the family $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$ is linearly dependent.
   4. The span of each elementary antisymmetric tensor is a one-dimensional subspace of $\bigwedge^k \mathsf{H}$. Prove that these one-dimensional subspaces of $\bigwedge^k \mathsf{H}$ are in one-to-one correspondence with the k-dimensional subspaces of H.

3. **Majorization mix.** We extend a function $\varphi : \mathbb{R} \to \mathbb{R}$ to a function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ by the definition $[\varphi(\boldsymbol{x})]_i = \varphi(x_i)$ for each index $i = 1, \dots, n$. The function $(a)_+ := \max\{a, 0\}$.

   1. **Xiaoqi's problem.** Is it true that $n^{-1}\mathbf{1} \prec \boldsymbol{x} \prec (1, 0, \dots, 0)$ for every vector $\boldsymbol{x} \in \mathbb{R}^n$ with $\operatorname{tr} \boldsymbol{x} = 1$? Prove it or provide a counterexample.
   2. **DS.** Suppose that the matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive, trace-preserving, and unital. Prove that $\boldsymbol{A}$ is doubly stochastic.
   3. **k-max.** Using duality, show that the function $\boldsymbol{x} \mapsto \sum_{i=1}^k x_i^{\downarrow}$ is convex.
   4. **Submajorization without rearrangement.** Show that $\boldsymbol{x} \prec_w \boldsymbol{y}$ if and only if $\operatorname{tr}(\boldsymbol{x} - t\mathbf{e})_+ \le \operatorname{tr}(\boldsymbol{y} - t\mathbf{e})_+$ for all $t \in \mathbb{R}$.
   5. **Majorization without rearrangement.** Using the previous part, show that $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if $\operatorname{tr} |\boldsymbol{x} - t\mathbf{e}| \le \operatorname{tr} |\boldsymbol{y} - t\mathbf{e}|$ for all $t \in \mathbb{R}$.
   6. **Majorization and convexity.** Using the previous part and results from class, show that $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if $\operatorname{tr} \varphi(\boldsymbol{x}) \le \operatorname{tr} \varphi(\boldsymbol{y})$ for every convex function $\varphi : \mathbb{R} \to \mathbb{R}$.
   7. **Schur "Convexity" theorem.** Suppose that $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a differentiable, permutation invariant function. Consider the property
   $$(x_i - x_j) \left(\frac{\partial \Phi}{\partial x_i}(\boldsymbol{x}) - \frac{\partial \Phi}{\partial x_j}(\boldsymbol{x}) \right) \ge 0 \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \text{ and all } 1 \le i < j \le n.$$
   Prove that Φ is isotone if and only if the condition in the display holds. **Hints:** Reduce to showing that the map preserves majorization for vectors that differ in two coordinates only. Use interpolation, as in class, and the fundamental theorem of calculus.

4. **Ky Fan norms.** In this problem, we explore an important class of matrix norms that will arise later.

1. **Ky Fan maximum principle.** Use Schur's majorization theorem on the diagonal of an Hermitian matrix to establish the following identity for an Hermitian matrix $\boldsymbol{A}$:
$$\sum\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A}) = \max \sum\nolimits_{i=1}^{k} \langle \boldsymbol{u}_i,\ \boldsymbol{A}\boldsymbol{u}_i \rangle$$
where the maximum ranges over all sets $\{\boldsymbol{u}_i : i = 1, \dots, k\}$ of k orthonormal vectors.

2. Use the previous part to prove the following identity for Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n \times n}$, where $1 \le k \le n$:
$$\sum\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A} + \boldsymbol{B}) \le \sum\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A}) + \sum\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{B})$$

3. Let $\boldsymbol{C} \in \mathbb{C}^{m \times n}$ be a general matrix, and consider the Hermitian dilation
$$\mathscr{H}(\boldsymbol{C}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{C} \\ \boldsymbol{C}^* & \mathbf{0} \end{bmatrix}.$$
Let $\sigma_1, \dots, \sigma_r$ be the first r singular values of $\boldsymbol{C}$, where $r = \min\{m, n\}$. Demonstrate that the eigenvalues of $\mathscr{H}(\boldsymbol{C})$ are $\pm\sigma_1, \dots, \pm\sigma_r$, along with an appropriate number of zeros.

4. Use the last two parts to prove the following identity for general matrices $\boldsymbol{C}, \boldsymbol{F} \in \mathbb{C}^{m \times n}$, where $k \le \min\{m, n\}$:
$$\sum\nolimits_{i=1}^{k} \sigma_i^{\downarrow}(\boldsymbol{C} + \boldsymbol{F}) \le \sum\nolimits_{i=1}^{k} \sigma_i^{\downarrow}(\boldsymbol{C}) + \sum\nolimits_{i=1}^{k} \sigma_i^{\downarrow}(\boldsymbol{F})$$

5. Conclude that $\|\boldsymbol{C}\|_{(k)} = \sum_{i=1}^{k} \sigma_i^{\downarrow}(\boldsymbol{C})$ defines a norm, called the Ky Fan k-norm.

5. ***The cone of convexity.** It is often the case that an interesting class of functions forms a convex cone, and we can understand that class well by analyzing the geometry of the cone. Here is an important classical example of this phenomenon. Recall that a *convex cone* is a subset of a (real) linear space that contains all positive multiples and (finite) sums of its elements. The *conic hull* of a set S is given by
$$\operatorname{cone}(\mathsf{S}) := \left\{ \sum\nolimits_{i=1}^{r} \alpha_i \boldsymbol{x}_i \text{ where } \alpha_i \ge 0 \text{ and } \boldsymbol{x}_i \in S \text{ and } r \in \mathbb{N} \right\}.$$

1. Let $\mathsf{I} = [0, 1]$. Consider the class
$$\mathscr{C} := \{ \varphi : \mathsf{I} \to \mathbb{R} \text{ is convex, continuous, and } \varphi(0) = 0. \}$$
Show that $\mathscr{C}$ is a convex cone in $\mathsf{C}(\mathsf{I})$, the Banach space of real-valued, continuous functions on I, equipped with the supremum norm.

2. **Hardy, Littlewood, Pólya.** Consider the subset $\mathsf{S} \subset \mathsf{C}(\mathsf{I})$ consisting of linear functions and convex angle functions:
$$\mathsf{S} = \{ \varphi : t \mapsto t \} \bigcup \{ \varphi : t \mapsto -t \} \bigcup \{ \varphi : t \mapsto (t - \alpha)_+ \text{ for } \alpha \in \mathsf{I} \}.$$
Show that $\operatorname{cone}(\mathsf{S})$ is dense in the cone $\mathscr{C}$ of convex functions.

3. **Karamata.** The dual of $\mathsf{C}(\mathsf{I})$ is the Banach space of signed (Radon) measures on I, equipped with the total variation norm. The dual cone $\mathscr{C}^*$ of $\mathscr{C}$ is defined as the set of all measures μ that satisfy

$$\int_0^1 \varphi \, \mathrm{d}\mu \geq 0 \quad \text{for all } \varphi \in \mathscr{C}.$$

Prove that $\mathscr{C}^*$ contains precisely those measures μ that satisfy

$$\int_0^1 x \, \mathrm{d}\mu = 0 \quad \text{and} \quad \int_0^t \mu([0,x]) \, \mathrm{d}x \geq 0 \quad \text{whenever } 0 \leq t \leq 1.$$

4. What are the extreme rays of the cone $\mathscr{C}$?

2. UI Norms & Variational Principles

Date: 27 January 2022

This assignment covers majorization, rearrangements, symmetric gauge functions, unitarily invariant norms, complex interpolation, variational principles, and spectral functions.

1. **Majorization cone.** Consider the convex cone of ordered positive vectors:

$$K = \{\boldsymbol{x} \in \mathbb{R}^n : x_1 \geq x_2 \geq \cdots \geq x_n \geq 0\}.$$

   1. Determine the dual cone K^*.
   2. Express the condition $\boldsymbol{x} \in K^*$ in terms of a majorization relation.
   3. We say that $\boldsymbol{y}$ dominates $\boldsymbol{x}$ with respect to the K^* order if and only if $\boldsymbol{y} - \boldsymbol{x} \in K^*$. Show that a function $f : K \to \mathbb{R}$ is isotone if and only if $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ whenever $\boldsymbol{y}$ dominates $\boldsymbol{x}$ with respect to the K^* order.

2. **Rearrangements.** Rearrangements of vectors and functions are a classical part of the theory of majorization, and they also play an important role in real analysis. We'll establish some of the basic rearrangement inequalities.

   1. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Prove Chebyshev's rearrangement inequality:

$$\sum_{i=1}^n x_i^{\downarrow} y_i^{\uparrow} \leq \sum_{i=1}^n x_i y_i \leq \sum_{i=1}^n x_i^{\downarrow} y_i^{\downarrow}.$$

   2. (*) Let $f : [0,1] \to \mathbb{R}$ be absolutely integrable. Define the function $m(y) := \mu\{t \in \mathbb{R} : f(t) > y\}$, where μ is the Lebesgue measure. The decreasing rearrangement of f is the function

$$f^{\downarrow}(x) = \sup\{y : m(y) > x\} \quad \text{for } 0 \leq x \leq 1.$$

      Show that f and $f^{\downarrow}$ are equimeasurable:

$$\int_0^1 \varphi(f(x))\,\mathrm{d}x = \int_0^1 \varphi(f^{\downarrow}(x))\,\mathrm{d}x \quad \text{for all } \varphi : \mathbb{R} \to \mathbb{R}$$

      provided that the integrals exist.
   3. (*) For absolutely integrable functions $f, g : [0,1] \to \mathbb{R}$, prove that

$$\int_0^1 f(x)\, g(x)\,\mathrm{d}x \leq \int_0^1 f^{\downarrow}(x)\, g^{\downarrow}(x)\,\mathrm{d}x.$$

3. **Fan minimum principles.** Let $\Phi_{(k)}(\boldsymbol{x}) = \sum_{i=1}^k |x|_i^{\downarrow}$ be the k-max norm on $\mathbb{R}^n$, where $1 \leq k \leq n$. Define the Ky Fan matrix norm $\|\cdot\|_{(k)} = \Phi_{(k)} \circ \boldsymbol{\sigma}$.

   1. For $\boldsymbol{x} \in \mathbb{R}^n$, show that $\Phi_{(1)}(\boldsymbol{x}) \leq \Phi(\boldsymbol{x}) \leq \Phi_{(n)}(\boldsymbol{x})$ for each symmetric gauge function Φ.

2. For each vector $\boldsymbol{x} \in \mathbb{R}^n$, prove that

$$\Phi_{(k)}(\boldsymbol{x}) = \min\{\Phi_{(n)}(\boldsymbol{y}) + k\,\Phi_{(1)}(\boldsymbol{z}) : \boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}\}.$$

3. Determine the dual norm of $\Phi_{(k)}$.
4. For each matrix $\boldsymbol{X} \in \mathbb{C}^{n\times n}$, prove that

$$\|\boldsymbol{X}\|_{(k)} = \min\left\{\|\boldsymbol{Y}\|_{(n)} + k\|\boldsymbol{Z}\|_{(1)} : \boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{Z}\right\}.$$

5. Determine the dual norm of the Ky Fan matrix norm $\|\cdot\|_{(k)}$.

4. **More variational principles.** Variational principles are powerful tools in analysis. One way they are useful is to linearize complicated functions. This problem contains a few additional variational principles.

1. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$. Use the von Neumann trace theorem and the Hermitian dilation to demonstrate that

$$\max\{\operatorname{Re}\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}) : \boldsymbol{U}, \boldsymbol{V} \in \mathbb{C}^{n\times n} \text{ are unitary}\} = \sum\nolimits_{i=1}^{n} \sigma_i(\boldsymbol{A})\,\sigma_i(\boldsymbol{B}).$$

2. For each Hermitian matrix $\boldsymbol{A}$,

$$\sum\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A}) = \max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})$$
$$\sum\nolimits_{i=1}^{k} \lambda_i^{\uparrow}(\boldsymbol{A}) = \min_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})$$

where $\boldsymbol{U}$ ranges over all $n \times k$ matrices with orthonormal columns. What is the analog of this result for singular values?
3. For each positive-semidefinite matrix $\boldsymbol{A}$,

$$\prod\nolimits_{i=1}^{k} \lambda_i^{\downarrow}(\boldsymbol{A}) = \max_{\boldsymbol{U}} \det(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})$$
$$\prod\nolimits_{i=1}^{k} \lambda_i^{\uparrow}(\boldsymbol{A}) = \min_{\boldsymbol{U}} \det(\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U})$$

where $\boldsymbol{U}$ ranges over all $n \times k$ matrices with orthonormal columns. What is the analog of this result for singular values?
4. (*) For each positive-definite matrix $\boldsymbol{A} \in \mathbb{H}_n$,

$$(\det \boldsymbol{A})^{1/n} = \min\{n^{-1}\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) : \det \boldsymbol{B} = 1, \boldsymbol{B} > \boldsymbol{0}\}.$$

This result implies Minkowski's determinant theorem, problem (6)(c)(iv). How?

5. **Interpolation inequalities.** In this problem, we will use complex interpolation to prove some interesting matrix inequalities that are not necessarily easy to obtain otherwise. You will also need to use duality to represent a matrix norm as the supremum of a trace.

1. Let $\boldsymbol{A}, \boldsymbol{B}$ be positive semidefinite. Show that

$$\|\boldsymbol{A}^s\boldsymbol{B}^s\| \le \|\boldsymbol{A}\boldsymbol{B}\|^s \quad \text{for } 0 \le s \le 1.$$

Deduce that $\|\boldsymbol{A}\boldsymbol{B}\|^s \le \|\boldsymbol{A}^s\boldsymbol{B}^s\|$ for $s \ge 1$. Compare with [Bha97, Theorem IX.2.1].

2. Let $\boldsymbol{A}, \boldsymbol{B}$ be positive semidefinite. Use part (a) to conclude that
$$\|\boldsymbol{B}^s \boldsymbol{A}^s \boldsymbol{B}^s\| \leq \|(\boldsymbol{B}\boldsymbol{A}\boldsymbol{B})^s\| \quad \text{for } 0 \leq s \leq 1.$$
A similar result holds for $s \geq 1$.
3. (*) Use anti-symmetric tensor products and majorization techniques to extend (b) to all unitarily invariant norms. The Schatten 1-norm case is the Araki–Lieb–Thirring inequality.
4. Let $\boldsymbol{A}, \boldsymbol{B}$ be positive semidefinite, and let $\boldsymbol{X}$ be arbitrary. For $t \in [0, 1]$, prove that
$$\|\boldsymbol{A}^t \boldsymbol{X} \boldsymbol{B}^{1-t}\| \leq \|\boldsymbol{A}\boldsymbol{X}\|^t \|\boldsymbol{X}\boldsymbol{B}\|^{1-t}.$$
This bound holds for every unitarily invariant norm. Explain how to extend your argument to address this case; this is easier than (c). Compare with [Bha97, Corollary IX.5.3].
5. (*) Let $\boldsymbol{A}_1, \dots, \boldsymbol{A}_m \in \mathbb{M}_n$, and let $\boldsymbol{B} \in \mathbb{M}_n$ be positive semidefinite. For each $p \geq 1$, prove
$$\left(\operatorname{tr}\left|\sum\nolimits_{i=1}^m \boldsymbol{A}_i^* \boldsymbol{B} \boldsymbol{A}_i\right|^p\right)^{1/p} \leq \left(\operatorname{tr}\left|\sum\nolimits_{i=1}^m \boldsymbol{A}_i^* \boldsymbol{A}_i\right|^{2p}\right)^{1/(2p)} \left(\operatorname{tr} \boldsymbol{B}^{2p}\right)^{1/(2p)}.$$
Recall that $|\boldsymbol{M}| \coloneqq (\boldsymbol{M}^*\boldsymbol{M})^{1/2}$, and connect it with the Schatten norm.

6. **Lewis's approach to spectral functions.** Let $\mathbb{H}_n$ be the real linear space of Hermitian matrices of size n, equipped with the trace inner product $\langle \boldsymbol{Y}, \boldsymbol{X} \rangle = \operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{X})$. Write $\boldsymbol{\lambda}(\boldsymbol{A})$ for the decreasingly ordered eigenvalues of an Hermitian matrix $\boldsymbol{A}$. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be permutation-invariant. A *spectral function* on Hermitian matrices is map of the form
$$(f \circ \boldsymbol{\lambda})(\boldsymbol{A}) = f(\lambda_1(\boldsymbol{A}), \lambda_2(\boldsymbol{A}), \dots, \lambda_n(\boldsymbol{A}))$$
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. The *Fenchel conjugate* of f is the (closed, convex) function
$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \left[\langle \boldsymbol{y}, \boldsymbol{x} \rangle - f(\boldsymbol{x})\right] \quad \text{for } \boldsymbol{y} \in \mathbb{R}^n.$$
If f is closed and convex, then $(f^*)^* = f$. Recall Fenchel's inequality:
$$\langle \boldsymbol{y}, \boldsymbol{x} \rangle \leq f^*(\boldsymbol{y}) + f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$
The *subdifferential* of f at the point $\boldsymbol{x}$ is the set
$$\partial f(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{y}, \boldsymbol{x} \rangle = f^*(\boldsymbol{y}) + f(\boldsymbol{x})\}.$$
If f is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. Similarly, for a function $F : \mathbb{H}_n \to \mathbb{R} \cup \{+\infty\}$, the Fenchel conjugate is
$$F^*(\boldsymbol{Y}) = \sup_{\boldsymbol{X} \in \mathbb{H}_n} \left[\langle \boldsymbol{Y}, \boldsymbol{X} \rangle - F(\boldsymbol{X})\right].$$
The definition of the subdifferential of F is analogous with the vector definition.

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be permutation invariant. Explain why the conjugate $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex and permutation invariant.
2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be permutation invariant. Use the von Neumann trace theorem to show
$$(f \circ \boldsymbol{\lambda})^*(\boldsymbol{Y}) = (f^* \circ \boldsymbol{\lambda})(\boldsymbol{Y}).$$
Conclude that if f is closed and convex, then $f \circ \boldsymbol{\lambda}$ also is closed and convex.

3. Consider the map $g : \boldsymbol{x} \mapsto -\sum_{i=1}^n \log x_i$ on $\mathbb{R}^n_{++}$.
   1. Argue that g is closed and convex. What is its conjugate?
   2. What is the spectral function $g \circ \boldsymbol{\lambda}$? What is its domain? What is its conjugate? What does Fenchel's inequality say?
   3. Establish the multiplicative form of the Brunn–Minkowski inequality for positive semidefinite $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n$:

$$\det(\tau \boldsymbol{A} + (1-\tau)\boldsymbol{B}) \geq (\det \boldsymbol{A})^{\tau} (\det \boldsymbol{B})^{1-\tau} \quad \text{for } \tau \in [0,1].$$

   4. (*) Derive the Minkowski determinant theorem as a consequence of (iii):

$$(\det(\boldsymbol{A} + \boldsymbol{B}))^{1/n} \geq (\det \boldsymbol{A})^{1/n} + (\det \boldsymbol{B})^{1/n}.$$

4. Let f be permutation invariant and convex. Show that $\boldsymbol{Y} \in \partial(f \circ \boldsymbol{\lambda})(\boldsymbol{X})$ if and only if $\boldsymbol{\lambda}(\boldsymbol{Y}) \in (\partial f)(\boldsymbol{\lambda}(\boldsymbol{X}))$ and there exists a unitary matrix $\boldsymbol{U}$ with $\boldsymbol{U}^* \boldsymbol{X} \boldsymbol{U} = \operatorname{diag}(\boldsymbol{\lambda}(\boldsymbol{X}))$ and $\boldsymbol{U}^* \boldsymbol{Y} \boldsymbol{U} = \operatorname{diag}(\boldsymbol{\lambda}(\boldsymbol{Y}))$.
5. Let $\boldsymbol{X}$ be an Hermitian matrix with an eigenvalue decomposition $\boldsymbol{X} = \boldsymbol{U} \operatorname{diag}(\boldsymbol{\lambda}(\boldsymbol{X})) \boldsymbol{U}^*$. Suppose that f is differentiable at $\boldsymbol{\lambda}(\boldsymbol{X})$, an interior point of the domain. Show that

$$\nabla(f \circ \boldsymbol{\lambda})(\boldsymbol{X}) = \boldsymbol{U} \operatorname{diag}((\nabla f)(\boldsymbol{\lambda}(\boldsymbol{X}))) \boldsymbol{U}^*.$$

6. Compute the derivative of the map $\frac{1}{2}\|\cdot\|_{\mathrm{F}}^2 = \frac{1}{2}\sum_{i=1}^n \lambda_i^2$ on $\mathbb{H}_n$.
7. Compute the subdifferential of the maximum eigenvalue map λ_1 on $\mathbb{H}_n$.
8. Compute the subdifferential of the Schatten 1-norm $\|\boldsymbol{A}\|_1 = \sum_{i=1}^n |\lambda_i(\boldsymbol{A})|$ on $\mathbb{H}_n$.
9. (*) We can use the same method to study smoothness properties of unitarily invariant norms. Consider the real linear space $\mathbb{C}^{n \times n}$ of complex matrices, equipped with the real inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{Re}} = \operatorname{Re} \operatorname{tr}(\boldsymbol{A}^* \boldsymbol{B})$. Let $\boldsymbol{\sigma}(\boldsymbol{B})$ denote the vector of singular values of $\boldsymbol{B}$. Let Φ be any symmetric gauge function. Find an expression for the subdifferential of $\Phi \circ \boldsymbol{\sigma}$. If Φ is differentiable at $\boldsymbol{\sigma}(\boldsymbol{B})$, what is its derivative $\nabla(\Phi \circ \boldsymbol{\sigma})(\boldsymbol{B})$?

3. Perturbation Theory & Positive Maps

Date: 15 February 2022

This assignment covers principal angles, Sylvester equations, perturbation of eigenspaces, pinching, positive linear maps, and completely positive maps.

1. **Principal angles.** Let E and F be subspaces of $\mathbb{C}^n$. Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be orthogonal projectors on $\mathbb{C}^n$ with ranges E and F. Prove the following statements.

    1. Suppose E and F have the same dimension. The singular values of $\boldsymbol{PQ}$ are the cosines of the principal angles between E and F. Then the nonzero singular values of $(\mathbf{I} - \boldsymbol{P})\boldsymbol{Q}$ are the sines of the nonzero principal angles between E and F.
    2. Suppose E and F have the same dimension. Then the nonzero singular values of $\boldsymbol{P} - \boldsymbol{Q}$ are the nonzero singular values of $(\mathbf{I} - \boldsymbol{P})\boldsymbol{Q}$, each counted twice.
    3. Show that $\|\boldsymbol{P} - \boldsymbol{Q}\| \leq 1$.
    4. If $\|\boldsymbol{P} - \boldsymbol{Q}\| < 1$, then E and F have the same dimension.

2. **Davis–Kahan.** In this problem, we will develop the proof of the most famous perturbation theorem for invariant subspaces. Let $\boldsymbol{A} \in \mathbb{M}_m$ and $\boldsymbol{B} \in \mathbb{M}_n$ be normal matrices. Assume that $|\lambda_i(\boldsymbol{A})| \leq \varrho$ and $|\lambda_j(\boldsymbol{B})| > \varrho$ for all i, j. Consider the Sylvester equation $\boldsymbol{AX} - \boldsymbol{XB} = \boldsymbol{Y}$.

    1. Prove that the Sylvester equation has a unique solution.
    2. Argue that the solution admits the representation

    $$\boldsymbol{X} = \sum\nolimits_{p=0}^{\infty} \boldsymbol{A}^p \boldsymbol{Y} \boldsymbol{B}^{-p-1}.$$

    Hint: The spectral radius formula states that $|\lambda_i(\boldsymbol{A})| = \lim_{p\to\infty} \|\boldsymbol{A}^p\|^{1/p}$.
    3. For each unitarily invariant norm $|||\cdot|||$, deduce that the solution satisfies

    $$|||\boldsymbol{X}||| \leq \frac{1}{\delta} |||\boldsymbol{Y}|||.$$

    4. Now, let $\boldsymbol{A}, \boldsymbol{B}$ be normal matrices of the same dimension. Let $\mathsf{S}_{\boldsymbol{A}}$ and $\mathsf{S}_{\boldsymbol{B}}$ be subsets of the complex plane, separated by an annulus of width δ. Consider the spectral projector $\boldsymbol{P}_{\boldsymbol{A}}$ of $\boldsymbol{A}$ onto the subspace spanned by the eigenvalues listed in $\mathsf{S}_{\boldsymbol{A}}$, as well as the analog $\boldsymbol{P}_{\boldsymbol{B}}$ for $\boldsymbol{B}$. Prove that

    $$|||\boldsymbol{P}_{\boldsymbol{A}} \boldsymbol{P}_{\boldsymbol{B}}||| \leq \frac{1}{\delta} |||\boldsymbol{A} - \boldsymbol{B}|||.$$

    5. Specialize the last result to the case where $\boldsymbol{A}, \boldsymbol{B}$ are Hermitian. Consider $\mathsf{S}_{\boldsymbol{A}} = [a, b]$ and $\mathsf{S}_{\boldsymbol{B}} = (-\infty, a - \delta] \cup [b + \delta, +\infty)$. Interpret the left-hand side as a measure of the size of the sines of the principal angles between $\operatorname{range}(\boldsymbol{P}_{\boldsymbol{A}})$ and $\operatorname{range}(\boldsymbol{P}_{\boldsymbol{B}})^{\perp}$. This is called the sine-theta theorem.

6. What happens when $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{E}$ for a small perturbation $\boldsymbol{E}$? Give bounds on the change to an "isolated" eigenspace of $\boldsymbol{A}$ after perturbation. "Isolated" means that we consider an interval in the spectrum, and assume that the rest of the spectrum is some distance away.

3. **Pinch me.** Noncommutative averaging operations play an important role in matrix analysis. This problem gives an overview of the basic results in this direction. Consider (conformally partitioned) 2×2 block matrices of order $m + n$:

$$\boldsymbol{U} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}.$$

Define the simple *pinching* operation

$$\boldsymbol{\Psi}(\boldsymbol{A}) = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}) = \begin{bmatrix} \boldsymbol{A}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_{22} \end{bmatrix}.$$

1. For each Hermitian $\boldsymbol{A}$, show that $\boldsymbol{\lambda}(\boldsymbol{\Psi}(\boldsymbol{A})) \prec \boldsymbol{\lambda}(\boldsymbol{A})$. Pinching flattens out eigenvalues.
2. For each square matrix $\boldsymbol{A}$, show that $\boldsymbol{\sigma}(\boldsymbol{\Psi}(\boldsymbol{A})) \prec_w \boldsymbol{\sigma}(\boldsymbol{A})$.
3. Suppose that $\{\boldsymbol{P}_j\}$ is a family of orthogonal projectors with the property that $\boldsymbol{P}_i\boldsymbol{P}_j = \delta_{ij}\boldsymbol{P}_i$ and $\sum_j \boldsymbol{P}_j = \mathbf{I}$. Consider the (general) pinching map

$$\boldsymbol{\Psi}(\boldsymbol{A}) = \sum\nolimits_j \boldsymbol{P}_j\boldsymbol{A}\boldsymbol{P}_j.$$

Show that $\boldsymbol{\Psi}$ amounts to a block diagonal restriction in an appropriate basis. Express $\boldsymbol{\Psi}$ as a product of simple pinchings. Prove that $\boldsymbol{\Psi}$ satisfies the relations in (a) and (b).
4. Derive Schur's Theorem for Hermitian matrices, $\operatorname{diag}\boldsymbol{A} \prec \boldsymbol{\lambda}(\boldsymbol{A})$, as a consequence.
5. For positive definite $\boldsymbol{A}$, derive Fischer's inequality $\det \boldsymbol{A} \leq \det \boldsymbol{\Psi}(\boldsymbol{A})$ as a consequence. Conclude that $\det \boldsymbol{A} \leq \prod_i a_{ii}$, a result called Hadamard's determinant inequality.

4. **Positive linear maps.** In this problem, we use ideas from class to explore some more examples and properties of positive maps.

1. Adapt the proof of Kadison's Theorem to prove Choi's Theorem. Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a unital, positive linear map. For a normal matrix $\boldsymbol{A}$,

$$\boldsymbol{\Phi}(\boldsymbol{A})^*\boldsymbol{\Phi}(\boldsymbol{A}) \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}^*\boldsymbol{A})$$
$$\boldsymbol{\Phi}(\boldsymbol{A})\,\boldsymbol{\Phi}(\boldsymbol{A})^* \preccurlyeq \boldsymbol{\Phi}(\boldsymbol{A}^*\boldsymbol{A}).$$

2. (*) It is not always the case that $\boldsymbol{\Phi}(\boldsymbol{A})$ is normal, even if $\boldsymbol{A}$ is normal. Use this observation to construct an example where the two inequalities above in (a) are different.
3. Let $\boldsymbol{B}$ be positive semidefinite. Consider the Schur multiplication operator $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{A} \cdot \boldsymbol{B}$. Use the Russo–Dye Theorem to compute $\|\boldsymbol{\Phi}\|$.
4. Let $\boldsymbol{A}$ be a normal matrix whose eigenvalues are in the (strict) right half-plane. Consider the Lyapunov equation

$$\boldsymbol{A}^*\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A} = \boldsymbol{B}.$$

Extend our discussion about Sylvester equations to find an integral expression for the solution operator $\mathsf{L}_{\boldsymbol{A}} : \boldsymbol{B} \mapsto \boldsymbol{X}$. Use this expression to prove that $\mathsf{L}_{\boldsymbol{A}}$ is a positive linear map. (*) Apply the Russo–Dye Theorem to find an expression for the norm of the operator.

5. (*) Let $\boldsymbol{\Phi}$ be a strictly positive linear map. Assume that $\boldsymbol{A}$ is positive definite and $\boldsymbol{H}$ is Hermitian. Use Kadison's inequality to show that

$$\boldsymbol{\Phi}(\boldsymbol{H}\boldsymbol{A}^{-1}\boldsymbol{H}) \succcurlyeq \boldsymbol{\Phi}(\boldsymbol{H})\boldsymbol{\Phi}(\boldsymbol{A})^{-1}\boldsymbol{\Phi}(\boldsymbol{H})$$

This is a type of Cauchy–Schwarz inequality. **Hint:** Use $\boldsymbol{\Phi}$ and $\boldsymbol{A}$ to construct a unital, positive linear map.

5. **Adjoints.** The adjoint $\boldsymbol{\Phi}^*$ of a linear map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ is defined as the unique linear map $\boldsymbol{\Phi}^* : \mathbb{M}_k \to \mathbb{M}_n$ for which

$$\langle \boldsymbol{\Phi}^*(\boldsymbol{B}),\ \boldsymbol{A} \rangle = \langle \boldsymbol{B},\ \boldsymbol{\Phi}(\boldsymbol{A}) \rangle \quad \text{for all } \boldsymbol{A} \in \mathbb{M}_n \text{ and } \boldsymbol{B} \in \mathbb{M}_k.$$

1. Show that $\boldsymbol{\Phi}$ is trace-preserving if and only if $\boldsymbol{\Phi}^*$ is unital.
2. Compute the adjoint of $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}$.
3. Compute the adjoint of the Schur product map $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{A} \cdot \boldsymbol{B}$.
4. Compute the adjoint of the Lyapunov solution operator $\mathsf{L}_{\boldsymbol{A}}$.

6. **Doubly stochastic maps.** A linear map on matrices is called *doubly stochastic* if it is positive, unital, and trace-preserving.

1. Let p_i be nonnegative numbers that sum to one, and let $\boldsymbol{U}_i$ be unitary. Show that the following map is doubly stochastic:

$$\boldsymbol{\Phi}(\boldsymbol{A}) = \sum\nolimits_{i=1}^{N} p_i \boldsymbol{U}_i \boldsymbol{A} \boldsymbol{U}_i^*.$$

2. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be Hermitian matrices of the same size, and suppose that $\boldsymbol{A} = \boldsymbol{\Phi}(\boldsymbol{B})$ for a doubly stochastic map $\boldsymbol{\Phi}$. Show that

$$\boldsymbol{A} = \sum\nolimits_{i=1}^{N} p_i \boldsymbol{U}_i \boldsymbol{B} \boldsymbol{U}_i^*$$

where p_i are nonnegative numbers that sum to one and $\boldsymbol{U}_i$ are unitary. These objects depend on both $\boldsymbol{\Phi}$ and $\boldsymbol{B}$. **Hints:** Use Birkhoff's Theorem to prove the result in case $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal matrices. Then introduce eigenvalue decompositions of $\boldsymbol{A}$ and $\boldsymbol{B}$.

7. ***Complete Positivity.** Let $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ be a linear map. Consider the extension $\boldsymbol{\Phi}_m : \mathbb{M}_m(\mathbb{M}_n) \to \mathbb{M}_m(\mathbb{M}_k)$ defined by

$$\boldsymbol{\Phi}_m : \begin{bmatrix} \boldsymbol{A}_{11} & \dots & \boldsymbol{A}_{1m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_{m1} & \dots & \boldsymbol{A}_{mm} \end{bmatrix} \longmapsto \begin{bmatrix} \boldsymbol{\Phi}(\boldsymbol{A}_{11}) & \dots & \boldsymbol{\Phi}(\boldsymbol{A}_{1m}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}(\boldsymbol{A}_{m1}) & \dots & \boldsymbol{\Phi}(\boldsymbol{A}_{mm}) \end{bmatrix}$$

That is, we obtain $\boldsymbol{\Phi}_m$ by applying $\boldsymbol{\Phi}$ to each block of an $m \times m$ block matrix whose blocks have size $n \times n$. Since $\mathbb{M}_m(\mathbb{M}_n)$ is isomorphic to $\mathbb{M}_{m \times n}$, we can think about $\boldsymbol{\Phi}_m$ as a linear map on matrices. We say that $\boldsymbol{\Phi}$ is *completely positive* when $\boldsymbol{\Phi}_m$ is a positive map for each $m = 1, 2, 3, \dots$. This statement is equivalent to the condition that $\boldsymbol{\Phi} \otimes \mathbf{I}_m$ is a positive linear map for each $m = 1, 2, 3, \dots$. Completely positive maps play an important role in operator theory.

1. Show that every positive linear functional $\varphi : \mathbb{M}_n \to \mathbb{C}$ takes the form $\varphi(\boldsymbol{A}) = \sum_i \boldsymbol{u}_i^* \boldsymbol{A} \boldsymbol{u}_i$ for some vectors $\boldsymbol{u}_i$.
2. Show that every positive linear functional is a completely positive map.
3. Show that every linear map $\boldsymbol{\Phi} : \mathbb{M}_n \to \mathbb{M}_k$ of the form $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{X}^* \boldsymbol{A} \boldsymbol{X}$ is completely positive.
4. Show that the completely positive maps form a convex cone.
5. Show that every completely positive map has the form $\boldsymbol{\Phi}(\boldsymbol{A}) = \sum_{i=1}^{nk} \boldsymbol{X}_i^* \boldsymbol{A} \boldsymbol{X}_i$. **Hint:** The block matrix $[\boldsymbol{A}_{ij}] \in \mathbb{M}_n(\mathbb{M}_n)$ with blocks $\boldsymbol{A}_{ij} = \mathbf{E}_{ij}$ is positive semidefinite.
6. Consider the conjugate transpose map $\boldsymbol{\Phi}(\boldsymbol{A}) = \boldsymbol{A}^*$ defined on $\mathbb{M}_n$. Is $\boldsymbol{\Phi}$ positive? Compute the norm of $\boldsymbol{\Phi}_n$. **Hint:** Consider the same block matrix from the last part.
7. Let $\boldsymbol{\Phi}$ be a positive linear map. For matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{M}_n$, define the covariance function $\operatorname{Cov}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{\Phi}(\boldsymbol{A}^* \boldsymbol{B}) - \boldsymbol{\Phi}(\boldsymbol{A})^* \boldsymbol{\Phi}(\boldsymbol{B})$. Define the variance $\operatorname{Var}(\boldsymbol{A}) = \operatorname{Cov}(\boldsymbol{A}, \boldsymbol{A}) \succcurlyeq \mathbf{0}$. If $\boldsymbol{\Phi}$ is completely positive, show that

$$\begin{bmatrix} \operatorname{Var}(\boldsymbol{A}) & \operatorname{Cov}(\boldsymbol{A}, \boldsymbol{B}) \\ \operatorname{Cov}(\boldsymbol{A}, \boldsymbol{B})^* & \operatorname{Var}(\boldsymbol{B}) \end{bmatrix} \succcurlyeq \mathbf{0}.$$

III. projects

Projects

Each student will complete a project on foundations or applications of matrix analysis. This project will result in a writeup of *about 8–10 pages*, with content equivalent to a 90-minute lecture on the material, prepared in the lecture note template. For the project, you may select a classic or modern topic in matrix analysis and prepare a new treatment of the material. The list below may help you identify a topic or papers for this study. Alternatively, you can describe an application of matrix analysis in your own field of research or in your own research. There will be some milestones during the term (project selection, initial summary, etc.) to help keep this work on track. It would be preferable for each student to choose a slightly different topic.

Numerical linear algebra

Schur complements arise from Gaussian elimination, from partial solution of least-squares problems, and from conditioning jointly Gaussian random variables. They enjoy a beautiful and detailed theory. To get started, see the book [Zha05], especially the chapter by Ando.

The Lanczos iteration is a remarkable algorithm for reducing a positive-definite matrix to tridiagonal (Jacobi) form. There are deep connections with Gaussian quadrature, moment matching, and continued fractions. See Golub & Meurant [GM10] or Liesen & Strakoš [LS13].

The QR iteration computes the eigenvalues of a real symmetric matrix by repeated QR factorization. This algorithm has a continuous analog, called the *Toda flow*, which can be studied using methods for dynamical systems. A first reference is Lax's algebra book [Lax07, Chap. 18].

Generalized eigenvalue problems involve equations of the form $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{B}\boldsymbol{x}$. These problems lead to the theory of matrix pencils. When the matrices are not normal, generalized Schur factorizations arise. For generalized eigenvalue problems of positive-definite type, there is a very satisfying theory that connects with the generalized singular value decomposition. The generalized SVD provides a mechanism for solving regularized least-squares problems. A first reference is Golub & Van Loan [GVL13].

Eigenvalues and perturbations

Given a matrix, one may ask whether it is possible to narrow down the possible locations of its eigenvalues by examining its entries. The most famous result of this form is the Geršgorin disc theorem, but there are many variants and extensions. See Bhatia [Bha97, Chaps. VI–VIII] or Horn & Johnson [HJ13, Chap. 6] to get started.

Relatedly, one may attempt to use the eigenvalues of the matrix to produce information about the entries of the eigenvectors. The key result in this area may be called the eigenvector–eigenvalue identity. It has been repeatedly rediscovered in various fields of mathematics. For theoretical and historical details, see the recent paper of Denton et al. [Den+22].

We may also ask how much the eigenvalues or eigenvectors of a matrix change under perturbations. This course covers some of the most important results, but this is just the tip of an iceberg. For more information, see Bhatia's books [Bha97; Bha07a] or Saad's book [Saa11b]. The classic reference is Kato [Kat95], which is both wide and deep.

Matrix nearness problems

A recurring theme in matrix analysis is to characterize the "structured" matrix that is closest to a specified matrix. Higham's paper [Hig89] offers a primer on these problems. See also Dhillon & Tropp [DT07].

Among the classic results, the most famous is the Eckart–Young–Mirsky theorem [EY39; Mir60], which states that the truncated singular value decomposition provides the best low-rank approximation with respect to each unitarily invariant norm. Gu proposed a reversed form of the Eckart–Young theorem [Gu15]; his result can be generalized to all unitarily invariant norms using majorization.

The problem of finding a matrix $\boldsymbol{X}$ that minimizes the form $\|\boldsymbol{A} - \boldsymbol{BXC}\|_{\mathrm{F}}$ arises quite often in linear algebra. It is quite easy to solve this problem using differential calculus. The associated problem for the spectral norm, however, is much harder. The basic solution is given by Parrott's theorem [Par78]. This result was generalized and placed in an appropriate context by Davis et al. [DKW82].

Nonnegative and totally nonnegative matrices

Matrices whose entries are nonnegative arise in a huge number of applications, and they are closely associated with the theory of Markov chains. There are several books devoted to the topic of nonnegative matrices [BP94; Sen81; Min88; BR97].

The Perron–Frobenius theory describes the spectral properties of a nonnegative matrix. There are many sources for this material, including [HJ13, Chap. 8].

Combinatorial graphs provide another rich source of nonnegative matrices. The subject of spectral graph theory connects the eigenvalues and eigenvectors of the graph Laplacian with its metric and combinatorial properties. Good references include Spielman's notes [Spi18a] and the books [GR01; Chu97].

A totally positive matrix has the property that every square minor has a positive determinant. These matrices are deeply connected with moment problems in analysis, and they have a long history in matrix analysis. For a recent reference, see Pinkus [Pin10].

Matrix scaling and balancing

A very interesting class of problems arises when we consider scaling of a nonnegative matrix $\boldsymbol{A}$ by diagonal matrices $\boldsymbol{D}_1\boldsymbol{A}\boldsymbol{D}_2$ to ensure that the rows and columns have fixed sums. An iterative algorithm for this problem was invented by Kruithof [Kru37] and rediscovered by Sinkhorn [Sin64]. Recently, this method has achieved new prominence because of a connection with optimal transport problems. For example, see Altschuler et al. [AWR17].

Matrix balancing is another matrix scaling problem that requires us to minimize the ℓ_1 norm of rows by symmetric diagonal scaling. The classic algorithm for this problem is due to Osborne. Schulman & Sinclair [SS17] provided the first proof that Osborne's algorithm succeeds, under appropriate assumptions. Altschuler & Parrilo [AP22] have developed a streamlined and more general approach. Their balancing algorithms can also be used as an ingredient in solving min-mean-cycle problems [AP20].

Recall that positive linear maps offer a noncommutative analog of nonnegative matrices. In particular, positive linear maps that are unital and trace-preserving can be

viewed as generalizations of doubly stochastic matrices. Gurvits [Gur04] has developed an astonishing generalization of the Sinkhorn algorithm to this noncommutative setting. See also his joint paper with Garg et al. [Gar+20] for a complete analysis.

Control and dynamical systems

In the study of dynamical systems, we can use matrix theory to characterize when a discrete or continuous dynamical system is stable. These results are associated with Lyapunov and Schur. For example, see Horn & Johnson [HJ94, Chap. 2].

Several types of matrix equations (Riccati, Sylvester) also arise naturally from problems in control theory. For some discussion of these problems, see Horn & Johnson [HJ94, Chap. 4] or Bhatia [Bha97, Chap. VII].

Matrix functions

What does it mean to apply a (scalar) function to a matrix? For normal matrices, we can simply apply the function to the eigenvalues, leaving the eigenvectors alone. For general matrices, we can invoke the Cauchy integral formula. These two approaches coincide whenever they are both valid. This type of spectral function has a beautiful and rich theory. Higham's book [Hig08] develops the foundations, focusing especially on the numerical aspects. Bhatia's book [Bha97, Chap. X] discusses perturbation of matrix functions. There is also a nice series of papers by Lewis on spectral functions, such as [Lew96], which forms the basis for a homework problem.

Another possible approach is to apply a scalar function to each entry of the matrix. Although this approach seems naïve, it also has a rich theory that dates back to Schoenberg's work [Sch38] on Euclidean distance matrices in the 1930s. See Horn & Johnson [HJ94, Chap. 6] and Bhatia [Bha07b] for a smattering of results, plus additional references. The survey of Belton et al. [Bel+18] also covers this topic.

Integral representations

In analysis, it is often the case that an interesting class of functions can be seen as a convex cone. Within the cone, the extremal functions play a key role because they frame the boundary of the cone. As a consequence, every function in the cone can be expressed as an average of the extremal functions. These averages are typically written as integrals against a probability measure, and they provide a powerful tool for understanding the behavior of the class of functions. The key results in this area include the (strong) Krein–Milman theorem and the Choquet theory.

Examples of this principle include the theorem of Bochner on positive-definite functions, the theorem of Bernstein on completely monotone functions, and the theorem of Loewner on matrix monotone functions. See Simon's book [Sim11] for an introduction to this geometric perspective.

Integral representation theorems are also connected with ideas from complex analysis (e.g., the Herglotz theorem), with interpolation (e.g., Nevanlinna–Pick), and with the theory of moments (e.g., Stieltjes). For an introduction to these connections, see the survey of Berg [Ber08]. For a more comprehensive classical treatment, see the book [BCR84]. Simon [Sim19] has also written an entire book on Loewner's theorem.

A surprising and beautiful application of integral representations arises in the theory of matrix means. Kubo & Ando [KA79] develop a family of matrix means that are induced by matrix monotone functions. Hiai & Kosaki [HK99] develop a different type of matrix mean that is induced by a positive-definite function. See also Bhatia [Bha07b, Chap. 4].

Semidefinite programming

Optimization over the cone of positive-semidefinite matrices plays an central role in various disciplines, including data science, combinatorial optimization, control theory, and quantum information. For an introduction to this field, see Boyd & Vandenberghe [VB96]. For a more in-depth treatment, consider the book of Ben Tal & Nemirovski [BTN01].

The Hausdorff–Toeplitz theorem asserts that the field of values of a matrix composes a convex set in the complex plane. This is the simplest example of a class of "hidden convexity" theorems, which describe convexity that appears in surprising circumstances. Brickman's theorem is a significant generalization, which states that pairs of (positive) quadratic forms always induce a convex set of points in the plane. Barvinok [Bar02, Sec. II.14] describes this result, along with some generalizations. See also the paper [BT10] of Beck & Teboulle.

The results in the last paragraph assert that certain quadratic optimization problems and their Lagrangian dual exhibit strong duality. The S-Lemma is a manifestation of this same idea in control theory. In linear algebra, this result means that we can compute the maximum eigenvalue of an Hermitian matrix, even though the problem is not convex. See [BV04, App. B] for this perspective.

Semidefinite programming also plays a key role in algorithms for solving combinatorial optimization problems by relaxation and rounding. This idea goes back, at least, to the work of Lovász [Lov79], and it became widely known after Goemans & Williamson [GW95] developed their method for solving the MAXCUT problem. For a survey with applications in signal processing, see [Luo+10]. Related results appear in the books [Bar02; BTN01].

In fact, the idea of relaxation and rounding implicitly appears in Grothendieck's work on tensor products, which leads to the famous Grothendieck inequality for matrices. This result was cast in the language of matrices by Lindenstrauss. This is a major topic in functional analysis, but most of the references are technical. There is an elementary treatment of Krivine's proof of Grothendieck's inequality in Vershynin's book [Ver18, Secs. 3.5–3.7]. For some algorithmic work, see [AN06; Tro09].

There is a "noncommutative" extension of the Grothendieck inequality due to Pisier and Haagerup. For an algorithmic proof of this result, see the paper of Naor et al. [NRV14]. Closely related optimization problems arise in the theory of quantum optimal transport [Col+21].

Geometry of positive-definite matrices

The positive-definite matrices, equipped with a metric induced by the geometric mean, compose a Riemannian manifold with remarkable geometric properties. Bhatia [Bha07b, Chap. 6] offers an introduction to this subject. Optimization over this manifold has become a topic of contemporary interest in machine learning and related areas. For example, see Sra's paper [SH15].

Optimal transport distances between Gaussian distributions lead to a different geometric structure on the class of positive-semidefinite matrices, namely the Bures–Wasserstein metric. See Bhatia et al. [BJL19] for a study of this metric space. Optimization with respect to this metric plays an important role in applications of optimal transport; for example, see [Alt+21].

Quantum information theory

In quantum information theory, a positive-semidefinite matrix with trace one is called a *density matrix*. It models the state of a (finite-dimensional) quantum system. As

a consequence, there are rich interactions between matrix analysis and quantum information. See Watrous's book for an introduction [Wat18].

Because of this type of connection, mathematical physics has indeed been a major driver of research on matrix analysis. The joint convexity of quantum relative entropy was derived as part of an effort to understand subadditivity properties of quantum entropy under tensorization. There is also a remarkable duality between subadditivity of entropy and (noncommutative) Brascamp–Lieb inequalities. See Carlen's notes [Car10] for a survey of these ideas.

The joint convexity of quantum relative entropy implies (in fact, is equivalent to) a concavity property of the trace exponential [Tro12]. De Huang developed some striking generalizations of this result using techniques from majorization and complex interpolation; see [Hua19] and the works cited. These results play a key role in the theory of matrix concentration.

Inequalities of Golden–Thompson type compare the matrix exponential of a sum with the product of matrix exponentials. Some results of this type appear in [Bha97, Chap. IX]. For far reaching generalizations with beautiful proofs, see the paper of Sutter et al. [SBT17]. These results also have been used to study matrix concentration.

Hyperbolic polynomials

A hyperbolic polynomial $p : \mathbb{R}^n \to \mathbb{R}$ has the property that its restrictions $t \mapsto p(\boldsymbol{x} - t\mathbf{e})$ to a fixed direction $\mathbf{e}$ have only real roots. This turns out to be a fundamental concept in the geometry of polynomials, which is becoming a topic of increasing interest in computational mathematics. For a digestible introduction, see the paper [Bau+01]. For a classic analysis reference, see the book of Hörmander [Hör94].

The Newton inequalities are, perhaps, the most elementary result that stems from the notion of hyperbolicity. They state that the sequence of elementary symmetric polynomials compose an ultra-log-concave sequence. This result has many striking generalizations. In matrix analysis, the deepest theorem of this type is Alexandrov's inequality for mixed discriminants. See the paper of Shenfeld and Van Handel [SH19] for an easy proof of the latter result. For some applications of mixed discriminants in combinatorics, see Barvinok's book [Bar16].

The geometry of polynomials also stands at the center of some other recent advances in operator theory, combinatorics, and other areas. Marcus, Spielman, & Srivastava used these ideas in their construction of bipartite Ramanujan graphs, their sharp version of the restricted invertibility theorem, and their solution of the Kadison–Singer problem. See [MSS14] for an introduction to this circle of ideas and references to the original work. These results all have implications in matrix analysis.

In a related direction, Anari and collaborators have been using log-concave polynomials for remarkable effect in their work on counting bases of matroids and related problems. This research has implications for Monte Carlo Markov chain methods for sampling from determinantal point processes. The first paper in the sequence is [AGV21].

Projects bibliography

[AN06] N. Alon and A. Naor. "Approximating the cut-norm via Grothendieck's inequality". In: *SIAM J. Comput.* 35.4 (2006), pages 787–803. DOI: 10.1137/S0097539704441629.

[AP20] J. Altschuler and P. Parrilo. "Approximating Min-Mean-Cycle for low-diameter graphs in near-optimal time and memory". Available at https://arxiv.org/abs/2004.03114. 2020.

[AP22] J. Altschuler and P. Parrilo. "Near-linear convergence of the Random Osborne algorithm for Matrix Balancing". In: *Math. Programming* (2022). To appear.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017.

[Alt+21] J. Altschuler et al. "Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 2021.

[AGV21] N. Anari, S. O. Gharan, and C. Vinzant. "Log-concave polynomials, I: entropy and a deterministic approximation algorithm for counting bases of matroids". In: *Duke Math. J.* 170.16 (2021), pages 3459–3504. DOI: 10.1215/00127094-2020-0091.

[BR97] R. B. Bapat and T. E. S. Raghavan. *Nonnegative matrices and applications*. Cambridge University Press, Cambridge, 1997. DOI: 10.1017/CBO9780511529979.

[Bar02] A. Barvinok. *A course in convexity*. American Mathematical Society, Providence, RI, 2002. DOI: 10.1090/gsm/054.

[Bar16] A. Barvinok. *Combinatorics and complexity of partition functions*. Springer, Cham, 2016. DOI: 10.1007/978-3-319-51829-9.

[Bau+01] H. H. Bauschke et al. "Hyperbolic polynomials and convex analysis". In: *Canad. J. Math.* 53.3 (2001), pages 470–488. DOI: 10.4153/CJM-2001-020-6.

[BT10] A. Beck and M. Teboulle. "On minimizing quadratically constrained ratio of two quadratic functions". In: *J. Convex Anal.* 17.3-4 (2010), pages 789–804.

[Bel+18] A. Belton et al. "A panorama of positivity". Available at https://arXiv.org/abs/1812.05482. 2018.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. Analysis, algorithms, and engineering applications. Society for Industrial and Applied Mathematics (SIAM), 2001. DOI: 10.1137/1.9780898718829.

[Ber08] C. Berg. "Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity". In: *Positive Definite Functions: From Schoenberg to Space-Time Challenges*. Castellón de la Plana: University Jaume I, 2008, pages 15–45.

[BCR84] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups*. Theory of positive definite and related functions. Springer-Verlag, New York, 1984. DOI: 10.1007/978-1-4612-1128-0.

[BP94] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Revised reprint of the 1979 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. DOI: 10.1137/1.9781611971262.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha07a] R. Bhatia. *Perturbation bounds for matrix eigenvalues*. Reprint of the 1987 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. DOI: 10.1137/1.9780898719079.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[BJL19] R. Bhatia, T. Jain, and Y. Lim. "On the Bures-Wasserstein distance between positive definite matrices". In: *Expo. Math.* 37.2 (2019), pages 165–191. DOI: 10.1016/j.exmath.2018.01.002.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. DOI: 10.1017/CBO9780511804441.

[Car10] E. Carlen. "Trace inequalities and quantum entropy: an introductory course". In: *Entropy and the quantum*. Volume 529. Contemp. Math. Amer. Math. Soc., Providence, RI, 2010, pages 73–140. DOI: 10.1090/conm/529/10428.

[Chu97] F. R. K. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[Col+21] S. Cole et al. "Quantum optimal transport". Available at https://arXiv.org/abs/2105.06922. 2021.

[DKW82] C. Davis, W. M. Kahan, and H. F. Weinberger. "Norm-preserving dilations and their applications to optimal error bounds". In: *SIAM J. Numer. Anal.* 19.3 (1982), pages 445–469. DOI: 10.1137/0719029.

[Den+22] P. B. Denton et al. "Eigenvectors from eigenvalues: a survey of a basic identity in linear algebra". In: *Bull. Amer. Math. Soc. (N.S.)* 59.1 (2022), pages 31–58. DOI: 10.1090/bull/1722.

[DT07] I. S. Dhillon and J. A. Tropp. "Matrix nearness problems with Bregman divergences". In: *SIAM J. Matrix Anal. Appl.* 29.4 (2007), pages 1120–1146. DOI: 10.1137/060649021.

[EY39] C. Eckart and G. Young. "A principal axis transformation for non-hermitian matrices". In: *Bull. Amer. Math. Soc.* 45.2 (1939), pages 118–121. DOI: 10.1090/S0002-9904-1939-06910-3.

[Gar+20] A. Garg et al. "Operator scaling: theory and applications". In: *Found. Comput. Math.* 20.2 (2020), pages 223–290. DOI: 10.1007/s10208-019-09417-z.

[GR01] C. Godsil and G. Royle. *Algebraic graph theory*. Springer-Verlag, New York, 2001. DOI: 10.1007/978-1-4613-0163-9.

[GW95] M. X. Goemans and D. P. Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *J. Assoc. Comput. Mach.* 42.6 (1995), pages 1115–1145. DOI: 10.1145/227683.227684.

[GM10] G. H. Golub and G. Meurant. *Matrices, moments and quadrature with applications*. Princeton University Press, Princeton, NJ, 2010.

[GVL13] G. H. Golub and C. F. Van Loan. *Matrix computations*. Fourth. Johns Hopkins University Press, Baltimore, MD, 2013.

[Gu15] M. Gu. "Subspace iteration randomization and singular value problems". In: *SIAM J. Sci. Comput.* 37.3 (2015), A1139–A1173. DOI: 10.1137/130938700.

[Gur04] L. Gurvits. "Classical complexity and quantum entanglement". In: *J. Comput. System Sci.* 69.3 (2004), pages 448–484. DOI: 10.1016/j.jcss.2004.06.003.

[HK99] F. Hiai and H. Kosaki. "Means for matrices and comparison of their norms". In: *Indiana Univ. Math. J.* 48.3 (1999), pages 899–936. DOI: 10.1512/iumj.1999.48.1665.

[Hig89] N. J. Higham. "Matrix nearness problems and applications". In: *Applications of matrix theory (Bradford, 1988)*. Volume 22. Inst. Math. Appl. Conf. Ser. New Ser. Oxford Univ. Press, New York, 1989, pages 1–27. DOI: 10.1093/imamat/22.1.1.

[Hig08] N. J. Higham. *Functions of matrices*. Theory and computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. DOI: 10.1137/1.9780898717778.

[Hör94] L. Hörmander. *Notions of convexity*. Birkhäuser Boston, Inc., 1994.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Corrected reprint of the 1991 original. Cambridge University Press, Cambridge, 1994.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[Hua19] D. Huang. "Improvement on a Generalized Lieb's Concavity Theorem". Available at https://arXiv.org/abs/1902.02194. 2019.

[Kat95] T. Kato. *Perturbation theory for linear operators*. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995.

[Kru37] R. Kruithof. "Telefoonverkeersrekening". In: *De Ingenieur* 52 (1937), pp. E15–E25.

[KA79] F. Kubo and T. Ando. "Means of positive linear operators". In: *Math. Ann.* 246.3 (1979/80), pages 205–224. DOI: 10.1007/BF01371042.

[Lax07] P. D. Lax. *Linear algebra and its applications*. second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007.

[Lew96] A. S. Lewis. "Derivatives of spectral functions". In: *Math. Oper. Res.* 21.3 (1996), pages 576–588. DOI: 10.1287/moor.21.3.576.

[LS13] J. Liesen and Z. Strakoš. *Krylov subspace methods*. Principles and analysis. Oxford University Press, Oxford, 2013.

[Lov79] L. Lovász. "On the Shannon capacity of a graph". In: *IEEE Trans. Inform. Theory* 25.1 (1979), pages 1–7.

[Luo+10] Z. Q. Luo et al. "Semidefinite relaxation of quadratic optimization problems". In: *IEEE Signal Process. Mag.* 27.3 (2010), pages 20–34.

[MSS14] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Ramanujan graphs and the solution of the Kadison-Singer problem". In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. III*. Kyung Moon Sa, Seoul, 2014, pages 363–386.

[Min88] H. Minc. *Nonnegative matrices*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1988.

[Mir60] L. Mirsky. "Symmetric gauge functions and unitarily invariant norms". In: *Quart. J. Math. Oxford Ser. (2)* 11 (1960), pages 50–59. DOI: 10.1093/qmath/11.1.50.

[NRV14] A. Naor, O. Regev, and T. Vidick. "Efficient rounding for the noncommutative Grothendieck inequality". In: *Theory Comput.* 10 (2014), pages 257–295. DOI: 10.4086/toc.2014.v010a011.

[Par78] S. Parrott. "On a quotient norm and the Sz.-Nagy - Foiaş lifting theorem". In: *J. Functional Analysis* 30.3 (1978), pages 311–328. DOI: 10.1016/0022-1236(78)90060-5.

[Pin10] A. Pinkus. *Totally positive matrices*. Cambridge University Press, Cambridge, 2010.

[Saa11b] Y. Saad. *Numerical methods for large eigenvalue problems*. Revised edition of the 1992 original [1177405]. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. DOI: 10.1137/1.9781611970739.ch1.

[Sch38] I. J. Schoenberg. "Metric spaces and positive definite functions". In: *Trans. Amer. Math. Soc.* 44.3 (1938), pages 522–536. DOI: 10.2307/1989894.

[SS17] L. J. Schulman and A. Sinclair. "Analysis of a Classical Matrix Preconditioning Algorithm". In: *J. Assoc. Comput. Mach. (JACM)* 64.2 (2017), 9:1–9:23.

[Sen81] E. Seneta. *Nonnegative matrices and Markov chains*. Second. Springer-Verlag, New York, 1981. DOI: 10.1007/0-387-32792-4.

[SH19] Y. Shenfeld and R. van Handel. "Mixed volumes and the Bochner method". In: *Proc. Amer. Math. Soc.* 147.12 (2019), pages 5385–5402. DOI: 10.1090/proc/14651.

[Sim11] B. Simon. *Convexity*. An analytic viewpoint. Cambridge University Press, Cambridge, 2011. DOI: 10.1017/CBO9780511910135.

[Sim19] B. Simon. *Loewner's theorem on monotone matrix functions*. Springer, Cham, 2019. DOI: 10.1007/978-3-030-22422-6.

[Sin64] R. Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *Ann. Math. Statist.* 35 (1964), pages 876–879. DOI: 10.1214/aoms/1177703591.

[Spi18a] D. Spielman. "Yale CPSC 662/AMTH 561 : Spectral Graph Theory". Available at http://www.cs.yale.edu/homes/spielman/561/561schedule.html. 2018.

[SH15] S. Sra and R. Hosseini. “Conic geometric optimization on the manifold of positive definite matrices”. In: *SIAM J. Optim.* 25.1 (2015), pages 713–739. DOI: 10.1137/140978168.

[SBT17] D. Sutter, M. Berta, and M. Tomamichel. “Multivariate trace inequalities”. In: *Comm. Math. Phys.* 352.1 (2017), pages 37–58. DOI: 10.1007/s00220-016-2778-5.

[Tro09] J. A. Tropp. “Column subset selection, matrix factorization, and eigenvalue optimization”. In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Philadelphia, PA, 2009, pages 978–986.

[Tro12] J. A. Tropp. “From joint convexity of quantum relative entropy to a concavity theorem of Lieb”. In: *Proc. Amer. Math. Soc.* 140.5 (2012), pages 1757–1760. DOI: 10.1090/S0002-9939-2011-11141-9.

[VB96] L. Vandenberghe and S. Boyd. “Semidefinite programming”. In: *SIAM Rev.* 38.1 (1996), pages 49–95. DOI: 10.1137/1038003.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

[Wat18] J. Watrous. *The theory of quantum information*. Cambridge University Press, 2018.

[Zha05] F. Zhang, editor. *The Schur complement and its applications*. Springer-Verlag, New York, 2005. DOI: 10.1007/b105056.

1. Means for Matrices and Norm Inequalities

Date: 14 March 2022 **Author:** Edoardo Calvello

In this project we broadly investigate means of positive matrices and norm inequalities involving them. We begin with the goal of extending the arithmetic–geometric mean inequality $\sqrt{\lambda\nu} \leq 2^{-1}(\lambda+\nu)$ for positive reals λ, ν, to the case of positive semidefinite matrices and unitarily invariant norms. In the search of a proof for the matrix arithmetic–geometric mean inequality, we present a simple method introduced by Bhatia [Bha07b]. Yet, a more general analysis of means for matrices due to Hiai and Kosaki [HK99] leads to a unified framework for comparing norms of these matrix means. This perspective thus leads to an alternative proof of the matrix arithmetic–geometric mean inequality and a variety of other notable norm inequalities. In this presentation, primarily based on the material from [Bha07b, Chap. 5] and [HK99], we outline this general analysis and the important consequences that arise.

Agenda:
1. A simple proof for matrix arithmetic–geometric mean inequality
2. From scalar means to matrix means
3. A unified analysis of means for matrices
4. Norm inequalities

1.1 A simple proof for the matrix arithmetic–geometric mean inequality

Let $\mathbb{M}_n(\mathbb{C})$ denote the space of all $n \times n$ complex matrices and $\mathbb{H}_n^+(\mathbb{C})$ the space of n-dimensional positive semidefinite Hermitian matrices over $\mathbb{C}$.

Our investigation begins with the aim of proving the following theorem.

Theorem 1.1 (Matrix arithmetic–geometric mean inequality). For any $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$ and any $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$ it holds that

$$|||\boldsymbol{H}^{1/2}\boldsymbol{X}\boldsymbol{K}^{1/2}||| \leq \frac{1}{2}|||\boldsymbol{H}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{K}||| \tag{1.1}$$

for any unitarily invariant norm $|||\cdot|||$.

Before we present a unified analysis of means for matrices and derive norm inequalities, we discuss a simple proof for the matrix arithmetic–geometric mean inequality (1.1).

We can prove Theorem 1.1 by establishing the inequality in the special case $\boldsymbol{H} = \boldsymbol{K}$, a key insight that appears in Bhatia [Bha07b]. Indeed, it is enough to prove the apparently weaker inequality

$$|||\boldsymbol{D}^{1/2}\boldsymbol{Y}\boldsymbol{D}^{1/2}||| \leq \frac{1}{2}|||\boldsymbol{D}\boldsymbol{Y} + \boldsymbol{Y}\boldsymbol{D}|||, \tag{1.2}$$

for any $\boldsymbol{D} \in \mathbb{H}_n^+(\mathbb{C}), \boldsymbol{Y} \in \mathbb{M}_n(\mathbb{C})$. In fact, by suitably choosing $\boldsymbol{D}$ and $\boldsymbol{Y}$ in inequality (1.2) as

$$\boldsymbol{D} := \begin{bmatrix} \boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{X} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

one can recover the more general inequality (1.1). Before we present a straightforward proof of Theorem 1.1, we first note an auxiliary lemma.

Lemma 1.2 (Unitarily invariant norm of a Schur product). If $\boldsymbol{A} \in \mathbb{H}_n^+(\mathbb{C})$, for any unitarily invariant norm $\|\!|\cdot|\!\|$ on $\mathbb{M}_n(\mathbb{C})$

$$\|\!|\boldsymbol{A} \odot \boldsymbol{X}|\!\| \leq \max_{1\leq i\leq n} a_{ii} \cdot \|\!|\boldsymbol{X}|\!\|, \quad \text{for any } \boldsymbol{X} \in \mathbb{M}_n(\mathbb{C}) ,$$

where $\odot$ is the Schur product.

Proof. A proof for this lemma can be found in [AHJ87, p.363]. ∎

The following proof of Theorem 1.1 is drawn from [Bha07b].

Proof of Theorem 1.1. We consider $\|\!|\boldsymbol{D}^{1/2}\boldsymbol{Y}\boldsymbol{D}^{1/2}|\!\|$ and note that since the norm in question is unitarily invariant, the expression reduces to the case where $\boldsymbol{D}$ is a diagonal matrix of the form $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. It is easily checked that in this case

$$\boldsymbol{D}^{1/2}\boldsymbol{Y}\boldsymbol{D}^{1/2} = \boldsymbol{\Lambda} \odot \left(\frac{\boldsymbol{D}\boldsymbol{Y} + \boldsymbol{Y}\boldsymbol{D}}{2}\right). \tag{1.3}$$

For each $i, j = 1, \ldots, n$, the (i, j)-th entry of $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{\Lambda}_{ij} = \frac{2\sqrt{\lambda_i\lambda_j}}{\lambda_i + \lambda_j}.$$

The matrix $\boldsymbol{\Lambda}$ is congruent to the matrix $\boldsymbol{C}$ with (i, j)th entry $c_{ij} \coloneqq (\lambda_i + \lambda_j)^{-1}$. Indeed note that $\boldsymbol{\Lambda} = 2 \cdot \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \cdot \boldsymbol{C} \cdot \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Since $(\boldsymbol{C})_{ij}$ can be represented as

$$c_{ij} = \int_0^\infty \mathrm{e}^{-(\lambda_i+\lambda_j)t}\, \mathrm{d}t,$$

we can conclude that $\boldsymbol{C}$ is positive semidefinite. Thus, by Sylvester's inertia theorem $\boldsymbol{\Lambda}$ is also positive semidefinite. Furthermore, all the diagonal entries of $\boldsymbol{\Lambda}$ are 1. Hence, from Lemma 1.2 we can observe that (1.3) implies (1.2), concluding the proof. ∎

1.2 From scalar means to matrix means

In this section we will first recall our discussion on scalar means. We will then introduce a notion of matrix mean due to [HK99] that is more general than the one due to [KA79], which was the object of Lecture 16.

1.2.1 Scalar means

We first recall the definition of a scalar mean.

Definition 1.3 (Scalar mean). A map $M : \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ is a scalar mean if for any $\lambda, \nu > 0$ we have the following properties.

$\mathbb{R}_{++} \coloneqq (0, \infty)$.

1. **Strict postivity.** It holds that $M(\lambda, \nu) > 0$.
2. **Ordering.** The inequality $\min\{\lambda, \nu\} \leq M(\lambda, \nu) \leq \max\{\lambda, \nu\}$ holds.
3. **Symmetry.** It holds that $M(\lambda, \nu) = M(\nu, \lambda)$.
4. **Monotonicity.** The functions $\lambda \mapsto M(\lambda, \nu)$ and $\nu \mapsto M(\lambda, \nu)$ are increasing.
5. **Homogeneity.** For any $c > 0$, $M(c\lambda, c\nu) = cM(\lambda, \nu)$ for all $\lambda, \nu > 0$.
6. **Continuity.** It holds that $(\lambda, \nu) \mapsto M(\lambda, \nu)$ is continuous on $\mathbb{R}_{++}^2$.

We denote by $\mathfrak{M}$ the set of all such scalar means.

We now recall two propositions that establish a bijection between the class of scalar means $\mathfrak{M}$ and what we define as representing functions.

Proposition 1.4 (Representing functions for scalar means). Let $M \in \mathfrak{M}$. Let the function f be defined as $f(t) := M(t, 1)$ for $t > 0$. Note that $f(0)$ can be defined by continuity and hence by taking limits. The function f then satisfies the following properties.

1. **Strict positivity.** For any $t > 0$ it holds that $f(t) > 0$.
2. **Normalization.** It holds that $f(1) = 1$.
3. **Symmetry and homogeneity.** For any $t > 0$ it holds that $f(t) = t \cdot f(1/t)$.
4. **Monotonicity.** The function f is increasing in t, while $t \mapsto f(t)/t$ is decreasing in t for $t > 0$.
5. **Continuity.** The function f is continuous.

We call such a function f a representing function.

Proof. The proof follows from the properties of scalar means. ∎

Proposition 1.5 (Scalar means from representing functions). Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be a function satisfying properties (1) through (5) of Proposition 1.4. Let M_f be defined by

$$M_f(\lambda, \nu) = \lambda \cdot f(\lambda/\nu) \quad \text{for} \quad \lambda, \nu > 0.$$

Then M_f is a scalar mean and we say that scalar mean M_f has representation f.

Proof. The proof of this proposition follows from direct application of the properties of the function f. ∎

Indeed, together Propositions 1.4 and 1.5 establish a bijection between scalar means and functions with the properties in Proposition 1.4, which we call representing functions.

1.2.2 A more general notion of means for matrices

We recall that in Lecture 16 we lifted the definition of scalar mean to the case of n-dimensional positive semidefinite Hermitian matrices over $\mathbb{C}$, $\mathbb{H}_n^+(\mathbb{C})$, thus introducing the definition of matrix mean due to [KA79] . Just as for scalar means, we showed that there exists a bijection between this class of matrix means and a matrix notion of representing functions.

In contrast to our previous discussion, we outline a different construction of a more general notion of matrix mean, due to Hiai and Kosaki [HK99], which will allow us to perform an analysis useful for establishing a wide range of matrix norm inequalities.

We endow $\mathbb{M}_n(\mathbb{C})$ with the trace inner product denoted by $\langle \cdot, \cdot \rangle$. For $M \in \mathfrak{M}$ and $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$, we consider the mean associated to left multiplication by $\boldsymbol{H}$ and right multiplication by $\boldsymbol{K}$, which we denote by $M(\boldsymbol{H}, \boldsymbol{K})$.

Recall that $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{tr}(\boldsymbol{X}\boldsymbol{Y}^*)$.

Definition 1.6 (Matrix mean; Hiai & Kosaki 1999). Let $M \in \mathfrak{M}$ and $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$. Let $\boldsymbol{H}$ and $\boldsymbol{K}$ have spectral decompositions $\boldsymbol{H} = \sum_{i=1}^n \lambda_i \boldsymbol{P}_i$ and $\boldsymbol{K} = \sum_{j=1}^n \nu_j \boldsymbol{Q}_j$, respectively. We define by the operator $M(\boldsymbol{H}, \boldsymbol{K}) : \mathbb{M}_n(\mathbb{C}) \to \mathbb{M}_n(\mathbb{C})$ by

$$M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} = \sum\nolimits_{i,j} M(\lambda_i, \nu_j) \boldsymbol{P}_i \boldsymbol{X} \boldsymbol{Q}_j, \quad \text{for } \boldsymbol{X} \in \mathbb{M}_n(\mathbb{C}).$$

This definition leads to the following useful result.

Proposition 1.7 **(Alternative formulation of the matrix mean).** Let $M \in \mathfrak{M}$ and $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$. Let $\boldsymbol{H} = \boldsymbol{U}\operatorname{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^*$ for unitary $\boldsymbol{U}$ be an eigenvalue decomposition. Similarly, let $\boldsymbol{K} = \boldsymbol{V}\operatorname{diag}(\nu_1, \ldots, \nu_n)\boldsymbol{V}^*$ for unitary $\boldsymbol{V}$. Then

$$M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} = \boldsymbol{U}([M(\lambda_i, \nu_j)] \odot (\boldsymbol{U}^*\boldsymbol{X}\boldsymbol{V}))\boldsymbol{V}^*,$$

for $[M(\lambda_i, \nu_j)]$ the matrix defined (i, j) entrywise and where $\odot$ denotes the usual Schur product.

Exercise 1.8 **(Alternative formulation of the matrix mean).** Starting from Definition 1.6, provide a proof for Proposition 1.7.

1.3 A unified analysis of means for matrices

We now present a unified framework for comparing norms of matrix means which is summarized by the following theorem, the formulation and proof of which are due to [HK99].

Theorem 1.9 **(Comparison of norms of matrix means).** For any $M, L \in \mathfrak{M}$ the following are equivalent.

1. There exists a symmetric probability measure μ on $\mathbb{R}$ such that
$$M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} = \int_{-\infty}^{\infty} \boldsymbol{H}^{is}(L(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X})\boldsymbol{K}^{-is}\,\mathrm{d}\mu(s),$$
for any $n \times n$ matrices $\boldsymbol{H}, \boldsymbol{K}, \boldsymbol{X}$ with $\boldsymbol{H}, \boldsymbol{K} > \boldsymbol{0}$.
2. For any $n \times n$ matrices $\boldsymbol{H}, \boldsymbol{K}, \boldsymbol{X}$ with $\boldsymbol{H}, \boldsymbol{K} \geqslant \boldsymbol{0}$ it holds that
$$|||M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}||| \leq |||L(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}|||,$$
for any unitarily invariant norm $|||\cdot|||$.
3. For any $n \times n$ matrices $\boldsymbol{H}, \boldsymbol{X}$ with $\boldsymbol{H} \geqslant \boldsymbol{0}$ it holds that
$$|||M(\boldsymbol{H}, \boldsymbol{H})\boldsymbol{X}||| \leq |||L(\boldsymbol{H}, \boldsymbol{H})\boldsymbol{X}|||,$$
4. The matrix defined entrywise by
$$\left[\frac{M(\lambda_i, \lambda_j)}{L(\lambda_i, \lambda_j)}\right]_{1 \leq i,j \leq n}$$
is positive semidefinite for any $\lambda_1, \ldots, \lambda_n > 0$, for all $n \in \mathbb{N}$.
5. The function
$$h(t) := \frac{M(\mathrm{e}^t, 1)}{L(\mathrm{e}^t, 1)}$$
is positive definite on $\mathbb{R}$.

Proof. We prove the equivalence using a step-by-step approach as presented in [HK99].
(1)⇒(2) We can observe that if (1) holds, then we have

$$\begin{aligned}|||M(\boldsymbol{H}, \boldsymbol{H})\boldsymbol{X}||| &\leq |||L(\boldsymbol{H}, \boldsymbol{H})\boldsymbol{X}||| \int_{-\infty}^{\infty} |||\boldsymbol{H}^{is}||| \cdot |||\boldsymbol{K}^{-is}|||\,\mathrm{d}\mu(s) \\ &\leq |||L(\boldsymbol{H}, \boldsymbol{H})\boldsymbol{X}|||,\end{aligned}$$

for any $n \times n$ matrices $\boldsymbol{H}, \boldsymbol{K}, \boldsymbol{X}$ with $\boldsymbol{H}, \boldsymbol{K} \succ \mathbf{0}$, where the last inequality follows from μ being a probability measure. To extend to $\boldsymbol{H}, \boldsymbol{K} \succcurlyeq \mathbf{0}$ it suffices to take the limit of the inequality for $\boldsymbol{H} + \epsilon\mathbf{I}$ and $\boldsymbol{K} + \epsilon\mathbf{I}$ for $\epsilon \to 0_+$.

(2)⇒(3) By taking $\boldsymbol{K} = \boldsymbol{H}$, we have that (2) clearly implies (3).

(3)⇒(4) Consider arbitrary $\lambda_1, \ldots, \lambda_n > 0$ and the matrix $\boldsymbol{A}$ defined entrywise by $a_{ij} = M(\lambda_i, \lambda_j)/L(\lambda_i, \lambda_j)$ for any $1 \le i, j \le n$. By the symmetry property of matrix means M and L, the matrix $\boldsymbol{A}$ is Hermitian with all diagonal entries equal to 1. Hence, it also holds that $\operatorname{tr}(\boldsymbol{A}) = n$.

Letting $\boldsymbol{H} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and applying (3), given what we know from Proposition 1.7, we obtain

$$\|[M(\lambda_i, \lambda_j)] \odot \boldsymbol{X}\| \le \|[L(\lambda_i, \lambda_j)] \odot \boldsymbol{X}\|$$

for each $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$. This implies

$$\|\boldsymbol{A} \odot \boldsymbol{X}\| \le \|\boldsymbol{X}\|,$$

for each $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$.

Now, considering that $\langle \boldsymbol{A} \odot \boldsymbol{X}, \boldsymbol{Y} \rangle = \langle \boldsymbol{X}, \boldsymbol{A} \odot \boldsymbol{Y} \rangle$, it is possible to show that from the above it holds that $\|\boldsymbol{A} \odot \boldsymbol{X}\|_1 \le \|\boldsymbol{X}\|_1$. Since this holds for any $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$, if we take $\boldsymbol{X}$ to be the matrix with all entries 1, we obtain the inequality $\|\boldsymbol{A}\|_1 \le n$.

We recall that $\|\cdot\|_1$ denotes the trace norm so $\|\boldsymbol{A}\|_1 = \operatorname{tr}(\boldsymbol{A}^*\boldsymbol{A})^{1/2}$.

If $\alpha_1, \ldots, \alpha_n$ are the real eigenvalues of $\boldsymbol{A}$, to prove (4) it suffices to show that $\alpha_1, \ldots, \alpha_n$ are all positive. We observe

$$\sum_{i=1}^n |\alpha_i| = \|\boldsymbol{A}\|_1 \le n,$$

but $n = \operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^n \alpha_i$. Hence $\sum_{i=1}^n |\alpha_i| \le \sum_{i=1}^n \alpha_i$ which proves positivity of the eigenvalues of $\boldsymbol{A}$ and thus (4).

(4)⇒(5) For any $t_i, t_j \in \mathbb{R},\ 1 \le i, j \le n$,

$$h(t_i - t_j) = \left(\frac{M(\mathrm{e}^{t_i - t_j}, 1)}{L(\mathrm{e}^{t_i - t_j}, 1)}\right) = \left(\frac{M(\mathrm{e}^{t_i}, \mathrm{e}^{t_j})}{L(\mathrm{e}^{t_i}, \mathrm{e}^{t_j})}\right),$$

which by (4) defines a positive semidefinite matrix on $\mathbb{R}$, hence by definition $h(t)$ is a positive definite function.

(5)⇒(1) Since the function $h(t) \coloneqq M(\mathrm{e}^t, 1)/L(\mathrm{e}^t, 1)$ is positive definite on $\mathbb{R}$, by Bochner's theorem we have that there exists a probability measure μ on $\mathbb{R}$ such that $h(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}ts}\, \mathrm{d}\mu(s)$ for any $t \in \mathbb{R}$ (see Lecture 18). Since by the properties of scalar means $h(t) = h(-t)$, we have that μ is a symmetric measure. Now for any $\boldsymbol{H}, \boldsymbol{K} \succ \mathbf{0}$ we compute $M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}$. Indeed, using Definition 1.6, we obtain

$$\begin{aligned}
M(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} &= \sum\nolimits_{k,l} M(\lambda_k, \nu_l) \boldsymbol{P}_k \boldsymbol{X} \boldsymbol{Q}_l \\
&= \sum\nolimits_{k,l} \nu_l M\left(\mathrm{e}^{\log(\lambda_k/\nu_l)}, 1\right) \boldsymbol{P}_k \boldsymbol{X} \boldsymbol{Q}_l \\
&= \sum\nolimits_{k,l} \nu_l L\left(\mathrm{e}^{\log(\lambda_k/\nu_l)}, 1\right) \left(\int_{-\infty}^{\infty} (\lambda_k/\nu_l)^{\mathrm{i}s}\, \mathrm{d}\mu(s)\right) \boldsymbol{P}_k \boldsymbol{X} \boldsymbol{Q}_l \\
&= \int_{-\infty}^{\infty} \sum\nolimits_{k,l} (\lambda_k/\nu_l)^{\mathrm{i}s} L(\lambda_k, \nu_l) \boldsymbol{P}_k \boldsymbol{X} \boldsymbol{Q}_l\, \mathrm{d}\mu(s) \\
&= \int_{-\infty}^{\infty} \boldsymbol{H}^{\mathrm{i}s} (L(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}) \boldsymbol{K}^{-\mathrm{i}s}\, \mathrm{d}\mu(s),
\end{aligned}$$

where the second equality follows from the properties of scalar means, the third equality from using the definition of h and the fourth equality again from the properties of scalar means. ∎

As pointed out in [HK99], Theorem 1.9 suggests that to obtain norm inequalities between matrix means it is important to establish positive definiteness of the related function on $\mathbb{R}$. To this end we note the following proposition before we continue our discussion.

Proposition 1.10 (Some positive definite functions). The functions of t,

$$\frac{\sinh at}{\sinh bt} \quad \text{and} \quad \frac{\cosh at}{\cosh bt}$$

are positive definite whenever $0 \leq a < b$. The function of t,

$$\frac{t}{\sinh t/2}$$

is also positive definite.

Proof. We refer to [Kos98] and [HK99] for a proof of these facts and in particular to [Kos98] for a more comprehensive analysis of useful positive definite functions. ∎

1.4 Norm inequalities

In this section we show how we can apply Theorem 1.9 to obtain various norm inequalities involving means for matrices, and how these lead, for example, to an alternative proof for the matrix arithmetic–geometric mean inequality in Theorem 1.1.

Definition 1.11 (α-scalar mean). For $\alpha \in \mathbb{R}$ and $\lambda, \nu > 0$ we define $M_\alpha(\lambda, \nu)$ by

$$M_\alpha(\lambda, \nu) = \begin{cases} \frac{\alpha-1}{\alpha} \cdot \frac{\lambda^\alpha - \nu^\alpha}{\lambda^{\alpha-1} - \nu^{\alpha-1}} & \text{for } \lambda \neq \nu, \\ \lambda & \text{for } \lambda = \nu. \end{cases}$$

Exercise 1.12 (α-scalar mean is a scalar mean.). Using Definition 1.3, check that the α-scalar mean as defined in Definition 1.11 is indeed a scalar mean.

In particular we have the following means.

Example 1.13 (Some α-scalar means). The following are examples of α-scalar means.

1. **Arithmetic mean** For $\alpha = 2$, it holds that
$$M_2(\lambda, \nu) = \frac{\lambda + \nu}{2} \quad \text{for any } \lambda, \nu > 0.$$
2. **Geometric mean** For $\alpha = 1/2$, it holds that
$$M_{1/2}(\lambda, \nu) = \sqrt{\lambda \nu} \quad \text{for any } \lambda, \nu > 0.$$
3. **Harmonic mean** For $\alpha = -1$, it holds that
$$M_{-1}(\lambda, \nu) = \frac{2}{\lambda^{-1} + \nu^{-1}} \quad \text{for any } \lambda, \nu > 0.$$
4. **Logarithmic mean** Taking the limit as $\alpha \to 1$, it holds that
$$M_1(\lambda, \nu) = \frac{\lambda - \nu}{\log \lambda - \log \nu}. \quad \text{for any } \lambda, \nu > 0$$

5. **Zero-scalar mean** Taking the limit as $\alpha \to 0$, it holds that

$$M_0(\lambda, \nu) = \frac{\log \lambda - \log \nu}{\nu^{-1} - \lambda^{-1}} \quad \text{for any } \lambda, \nu > 0.$$

It is interesting to note that the zero-scalar mean is the reciprocal of the logarithmic mean of the reciprocals.

▪

Exercise 1.14 (Infinity-scalar mean). Characterize the α-scalar mean as defined in Definition 1.11 when $\alpha \to \pm\infty$.

The next theorem due to Hiai and Kosaki [HK99] establishes an attractive norm inequality between matrix means.

Theorem 1.15 (α-matrix mean norm inequality). If $-\infty \le \alpha < \beta \le \infty$, then

$$|||M_\alpha(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}||| \le |||M_\beta(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X}|||,$$

for any unitarily invariant norm $|||\cdot|||$.

Proof. For $1/2 \le \alpha < \beta < \infty$, by using Definition 1.11 can write

$$\begin{aligned}
\frac{M_\alpha(\mathrm{e}^{2t}, 1)}{M_\beta(\mathrm{e}^{2t}, 1)} &= \frac{(\alpha-1)\beta}{\alpha(\beta-1)} \cdot \frac{(\mathrm{e}^{2\alpha t} - 1)(\mathrm{e}^{2(\beta-1)t} - 1)}{(\mathrm{e}^{2(\alpha-1)t} - 1)(\mathrm{e}^{2\beta t} - 1)} \\
&= \frac{(\alpha-1)\beta}{\alpha(\beta-1)} \cdot \frac{(\mathrm{e}^{\alpha t} - \mathrm{e}^{-\alpha t})(\mathrm{e}^{(\beta-1)t} - \mathrm{e}^{-(\beta-1)t})}{(\mathrm{e}^{(\alpha-1)t} - \mathrm{e}^{-(\alpha-1)t})(\mathrm{e}^{\beta t} - \mathrm{e}^{-\beta t})} \\
&= \frac{(\alpha-1)\beta}{\alpha(\beta-1)} \cdot \frac{\sinh(\alpha t)\sinh((\beta-1)t)}{\sinh((\alpha-1)t)\sinh(\beta t)},
\end{aligned}$$

where we set

$$\frac{\alpha - 1}{\sinh((\alpha-1)t)} = \frac{1}{t} \text{ at } \alpha = 1, \quad \text{and} \quad \frac{\sinh((\beta-1)t)}{\beta - 1} = t \text{ at } \beta = 1.$$

Now, if $1/2 \le \alpha < \beta \le 1$, then

$$\frac{M_\alpha(\mathrm{e}^{2t}, 1)}{M_\beta(\mathrm{e}^{2t}, 1)} = \frac{(\alpha-1)\beta}{\alpha(\beta-1)} \cdot \frac{\sinh(\alpha t)}{\sinh(\beta t)} \cdot \frac{\sinh((1-\beta)t)}{\sinh((1-\alpha)t)}$$

is positive definite as it is a product of positive definite functions, by Proposition 1.10. Therefore, by Theorem 1.9, the claim of this theorem holds when $1/2 \le \alpha < \beta \le 1$.

We now consider the case $1 < \alpha < \beta < \infty$. Through some tedious computations, we observe that

$$\begin{aligned}
&\frac{\sinh(\alpha t)\sinh((\beta-1)t)}{\sinh((\alpha-1)t)\sinh(\beta t)} - 1 \\
&= \frac{\sinh((\alpha-1)t + t)\sinh((\beta-1)t)}{\sinh((\alpha-1)t)\sinh(\beta t)} - \frac{\sinh((\alpha-1)t)\sinh((\beta-1)t + t)}{\sinh((\alpha-1)t)\sinh(\beta t)} \\
&= \frac{\sinh t}{\sinh(\beta t)} \cdot \left(\frac{\cosh((\alpha-1)t)\sinh((\beta-1)t)}{\sinh((\alpha-1)t)} - \frac{\sinh((\alpha-1)t)\cosh((\beta-1)t)}{\sinh((\alpha-1)t)} \right) \\
&= \frac{\sinh t}{\sinh(\beta t)} \cdot \frac{\sinh((\beta-\alpha)t)}{\sinh((\alpha-1)t)}.
\end{aligned}$$

If $1 < \alpha < \beta < 2\alpha - 1$ and so $0 < \beta - \alpha < \alpha - 1$, then the above computations show that by Proposition 1.10, $M_\alpha(\mathrm{e}^{2t}, 1)/M_\beta(\mathrm{e}^{2t}, 1)$ is positive definite. Therefore by

Theorem 1.9, the claim of this theorem holds. In the general case of $1 < \alpha < \beta < \infty$, it is possible to choose $\alpha = \alpha_0 < \alpha_1 < \cdots < \alpha_m = \beta$ that satisfies $\alpha_k < 2\alpha_{k-1} - 1$ for $1 \leq k \leq m$, which leads to the conclusion. The special cases $1 = \alpha < \beta < \infty$ and $1 < \alpha < \beta = \infty$ can be obtained by taking the limit as $\alpha \to 1$ and $\beta \to \infty$ respectively. We refer to [HK99] for the details on why the inequalities are preserved under taking limits. Furthermore the results can be extended straightforwardly to the case $-\infty \leq \alpha < \beta \leq 1/2$, an argument for which for conciseness we refer to [HK99]. ∎

Theorem 1.15 allows us to directly obtain a proof for the matrix arithmetic–geometric mean inequality from Theorem 1.1.

Proof of Theorem 1.1. For any $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$ and any $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$, we have

$$M_{1/2}(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} = \boldsymbol{H}^{1/2}\boldsymbol{X}\boldsymbol{K}^{1/2} \quad \text{and} \quad M_2(\boldsymbol{H}, \boldsymbol{K})\boldsymbol{X} = \frac{1}{2}(\boldsymbol{H}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{K}). \tag{1.4}$$

Therefore by directly applying Theorem 1.15 we obtain that for any $\boldsymbol{H}, \boldsymbol{K} \in \mathbb{H}_n^+(\mathbb{C})$ and any $\boldsymbol{X} \in \mathbb{M}_n(\mathbb{C})$,

$$|||\boldsymbol{H}^{1/2}\boldsymbol{X}\boldsymbol{K}^{1/2}||| \leq \frac{1}{2}|||\boldsymbol{H}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{K}|||,$$

for any unitarily invariant norm $|||\cdot|||$. ∎

Exercise 1.16 (Proof of equalities in (1.4)). By using Proposition 1.7 show that the equalities in (1.4) indeed hold.

Together Theorems 1.9 and 1.15 are very powerful and a plethora of important results follow from these. We refer the reader to [HK99] for a more in depth investigation of these topics and derivations of further norm inequalities. We also refer the reader to [ABY20] for an example of an interesting application of these results.

1.5 Conclusions

In this investigation we began with the objective of proving a matrix arithmetic–geometric mean inequality. This aim led us to systematically constructing a notion of matrix mean that led to the fundamental Theorem 1.9. This allowed a more general analysis that makes it possible to derive inequalities between norms of means for matrices and thus many different norm inequalities. These insights and the corresponding analysis is due to Hiai and Kosaki [HK99], where the interested reader can find a further exploration of these topics.

Lecture bibliography

[AHJ87] T. Ando, R. A. Horn, and C. R. Johnson. "The singular values of a Hadamard product: a basic inequality". In: *Linear and Multilinear Algebra* 21.4 (1987), pages 345–365. eprint: https://doi.org/10.1080/03081088708817810. DOI: 10.1080/03081088708817810.

[ABY20] R. Aoun, M. Banna, and P. Youssef. "Matrix Poincaré inequalities and concentration". In: *Advances in Mathematics* 371 (2020), page 107251. DOI: https://doi.org/10.1016/j.aim.2020.107251.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[HK99] F. Hiai and H. Kosaki. “Means for matrices and comparison of their norms”. In: *Indiana Univ. Math. J.* 48.3 (1999), pages 899–936. DOI: 10.1512/iumj.1999.48.1665.

[Kos98] H. Kosaki. “Arithmetic–geometric mean and related inequalities for operators”. In: *Journal of Functional Analysis* 156.2 (1998), pages 429–451.

[KA79] F. Kubo and T. Ando. “Means of positive linear operators”. In: *Math. Ann.* 246.3 (1979/80), pages 205–224. DOI: 10.1007/BF01371042.

2. The Eigenvector–Eigenvalue Identity

Date: 7 April 2022 **Author:** Ruizhi Cao

Agenda:
1. Cauchy interlacing theorem
2. First order perturbation theorem
3. Eigenvector–eigenvalue identity
4. Proof of the identity
5. Application

Eigenvectors and eigenvalues are widely used in simplifying equations, data science, and are powerful analysis tools. For a $n \times n$ Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathbb{C})$, one of its eigenpairs $(\lambda,\boldsymbol{v})$ is the solution for

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

Here, λ is an eigenvalue of $\boldsymbol{A}$, and $\boldsymbol{v}$ is an eigenvector associated with λ. Moreover, the eigenvalues of a Hermitian matrix are real, and its unit-norm eigenvectors form an orthonormal basis for $\mathbb{C}^n$. One might ask if there is an intrinsic connection for the eigenvalues and eigenvectors. And over the past few decades, an identity between components of eigenvectors for a Hermitian matrix and its eigenvalues is repeatedly rediscovered in different contexts over many different fields [Den+22]. Although it is a simple yet extremely important result, it begins to attract boarder interest with the publication of Wolchover's article [Wol19]. In this section, we will establish an identity between eigenvectors and eigenvalues of a Hermitian matrix. We will write the eigenvalues in the weak increasing order throughout the following sections: $\lambda_i = \lambda_i^{\uparrow}$.

2.1 Cauchy interlacing theorem

Before we establish the eigenvector-eigenvalue identity for a Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n$, we will need a theorem on the eigenvalues of a submatrix of $\boldsymbol{A}$. We will first give the definition of a principal submatrix of a matrix $\boldsymbol{A}$. We will then present the Cauchy interlacing theorem, which tells one the order of the eigenvalues of $\boldsymbol{A}$ and the eigenvalues of one of its principal submatrices.

Definition 2.1 (Principal submatrix). For a $n \times n$ matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, let

$$\mathsf{L} = \{j_1, j_2, \ldots, j_k\}$$

be a collection of k distinct indices such that

$$1 \le j_1 < j_2 < \ldots < j_k \le n.$$

A *principal submatrix* of $\boldsymbol{A}$ is a submatrix which is obtained by keeping the entry on the ith row and the i'th column for all $i, i' \in \mathsf{L}$ and deleting all other entries.

We know that, based on the conjugation rule on matrix, a principal submatrix of a positive-semidefinite (psd) matrix is also psd. In other words, the eigenvalues of $\boldsymbol{A}$ and the eigenvalues of a submatrix of $\boldsymbol{A}$ are "correlated". Here, we will present a stronger relation on their eigenvalues.

Recall the conjugation rule tells us that if $\boldsymbol{A} \in \mathbb{M}_n$ is psd and $\boldsymbol{X} \in \mathbb{C}^{n\times k}$, then $\boldsymbol{X}^*\boldsymbol{A}\boldsymbol{X}$ is also psd.

Theorem 2.2 (Cauchy interlacing theorem). Suppose $\boldsymbol{A} \in \mathbb{H}_n$ is a Hermitian matrix. Let $\boldsymbol{B}$ be a principal $m \times m$ submatrix of $\boldsymbol{A}$. Suppose $\boldsymbol{A}$ has eigenvalues $\lambda_1 \leq \ldots \leq \lambda_n$ and $\boldsymbol{B}$ has eigenvalues $\beta_1 \leq \ldots \leq \beta_m$, then

$$\lambda_k \leq \beta_k \leq \lambda_{k+n-m} \quad \text{for} \quad k = 1, \ldots, m.$$

If $m = n - 1$, then

$$\lambda_1 \leq \beta_1 \leq \lambda_2 \leq \beta_2 \leq \ldots \leq \beta_{n-1} \leq \lambda_n.$$

Note that $\boldsymbol{B}$ is also a Hermitian matrix, that is $\boldsymbol{B} \in \mathbb{H}_m$.

Proof. We can use the Courant–Fischer–Weyl minimax principle to prove the theorem [Bha97]. Without loss, assume

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C}^* \\ \boldsymbol{C} & \boldsymbol{Z} \end{bmatrix},$$

where $\boldsymbol{Z} \in \mathbb{H}_{n-m}$, $\boldsymbol{C} \in \mathbb{C}^{n-m \times m}$. Let $\boldsymbol{x}_i \in \mathbb{C}^n$ $(1 \leq i \leq n)$ be the unit-norm eigenvector of $\boldsymbol{A}$ associated with the ordered eigenvalue λ_i, and $\boldsymbol{y}_j \in \mathbb{C}^m$ $(1 \leq j \leq m)$ be the unit-norm eigenvector of $\boldsymbol{B}$ associated with the ordered eigenvalue β_j. Let's define the following vector spaces for some $k \in \mathbb{N}$ and $1 \leq k \leq m$:

$$\mathsf{W} = \operatorname{span}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \subseteq \mathbb{C}^m,$$
$$\mathsf{M} = \left\{ \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{0} \end{pmatrix} \in \mathbb{C}^n, \boldsymbol{w} \in \mathsf{W} \right\} \subset \mathbb{C}^n.$$

Fix $\boldsymbol{w} \in \mathsf{W}$, we can find an associated vector $\widetilde{\boldsymbol{w}} \in \mathsf{M}$ via

$$\widetilde{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{0} \end{pmatrix} \in \mathsf{M}.$$

Thus, we have

$$\widetilde{\boldsymbol{w}}^* \boldsymbol{A} \widetilde{\boldsymbol{w}} = \boldsymbol{w}^* \boldsymbol{B} \boldsymbol{w} \quad \text{and} \quad \widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{w}} = \boldsymbol{w}^* \boldsymbol{w}.$$

Using the Courant–Fischer–Weyl minimax principle, the following inequality holds

$$\beta_k = \max_{\boldsymbol{w} \in \mathsf{W}} \frac{\boldsymbol{w}^* \boldsymbol{B} \boldsymbol{w}}{\boldsymbol{w}^* \boldsymbol{w}} = \max_{\widetilde{\boldsymbol{w}} \in \mathsf{M}} \frac{\widetilde{\boldsymbol{w}}^* \boldsymbol{A} \widetilde{\boldsymbol{w}}}{\widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{w}}} \geq \min_{\mathsf{M}' \subset \mathbb{C}^n, \dim \mathsf{M}' = k} \max_{\widetilde{\boldsymbol{w}} \in \mathsf{M}'} \frac{\widetilde{\boldsymbol{w}}^* \boldsymbol{A} \widetilde{\boldsymbol{w}}}{\widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{w}}} = \lambda_k.$$

This gives one side of the inequalities in the theorem.

For the other side, we can use the similar technique and apply the minimax principle to $-\boldsymbol{A}$ and $-\boldsymbol{B}$. Define the following vector spaces

$$\mathsf{W}' = \operatorname{span}(\boldsymbol{y}_k, \ldots, \boldsymbol{y}_m) \subseteq \mathbb{C}^m,$$
$$\mathsf{N} = \left\{ \begin{pmatrix} \boldsymbol{w}' \\ \boldsymbol{0} \end{pmatrix} \in \mathbb{C}^n, \boldsymbol{w}' \in \mathsf{W}' \right\} \subset \mathbb{C}^n.$$

We have

$$\dim \mathsf{N} = \dim \widetilde{\mathsf{W}}' = m - k + 1.$$

Then, for each $\boldsymbol{w}' \in \mathsf{W}'$, we can find $\widetilde{\boldsymbol{w}}' \in \mathsf{N}$ by choosing

$$\widetilde{\boldsymbol{w}}' = \begin{pmatrix} \boldsymbol{w}' \\ \boldsymbol{0} \end{pmatrix} \in \mathsf{N}.$$

Same as above, we have

$$(\widetilde{\boldsymbol{w}}')^*(-\boldsymbol{A})\widetilde{\boldsymbol{w}}' = (\boldsymbol{w}')^*(-\boldsymbol{B})\boldsymbol{w}' \quad \text{and} \quad (\widetilde{\boldsymbol{w}}')^*\widetilde{\boldsymbol{w}}' = (\boldsymbol{w}')^*\boldsymbol{w}'.$$

Note that we have

$$\lambda^{\uparrow}_{m-k+1}(-\boldsymbol{A}) = -\lambda_{k+n-m} \quad \text{and} \quad \lambda^{\uparrow}_{m-k+1}(-\boldsymbol{B}) = -\beta_k.$$

Then, apply the minimax principle to $-\boldsymbol{A}$ and $-\boldsymbol{B}$, we have

$$\begin{aligned}
-\beta_k &= \max_{\boldsymbol{w}' \in \mathsf{W}'} \frac{(\boldsymbol{w}')^*(-\boldsymbol{B})\boldsymbol{w}'}{(\boldsymbol{w}')^*\boldsymbol{w}} \\
&= \max_{\widetilde{\boldsymbol{w}}' \in \mathsf{N}} \frac{(\widetilde{\boldsymbol{w}}')^*(-\boldsymbol{A})\widetilde{\boldsymbol{w}}'}{(\widetilde{\boldsymbol{w}}')^*\widetilde{\boldsymbol{w}}'} \\
&\geq \min_{\mathsf{N}' \subset \mathbb{C}^n, \dim \mathsf{N}' = m-k+1} \max_{\widetilde{\boldsymbol{w}}' \in \mathsf{N}'} \frac{(\widetilde{\boldsymbol{w}}')^*(-\boldsymbol{A})\widetilde{\boldsymbol{w}}'}{(\widetilde{\boldsymbol{w}}')^*\widetilde{\boldsymbol{w}}'} \\
&= -\lambda_{k+n-m}.
\end{aligned}$$

That is

$$\lambda_{k+n-m} \geq \beta_k,$$

which gives the desired inequality. ∎

The Cauchy interlacing theorem allows us to bound the eigenvalues of a principal submatrix of $\boldsymbol{A}$ by the eigenvalues of the original matrix. Here, we derived the Cauchy interlacing theorem from the Courant–Fischer–Weyl minimax principle. It is also worthwhile to note that the Cauchy interlacing theorem, the Poincaré inequality and the Courant–Fischer–Weyl minimax principle have independent proofs and they can be derived from each other [Bha97].

2.2 First-order perturbation theorem

In this section, we will derive the first-order perturbation theory for a simple eigenvalue of the Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathbb{C})$. For a simple eigenpair $(\lambda, \boldsymbol{v})$ of $\boldsymbol{A}$, both the eigenvalue λ and the unit-norm eigenvector $\boldsymbol{v}$ are continuous with respect to entries of $\boldsymbol{A}$. Thus, we might expect that the perturbed eigenpair should be "close" to the original one. In this section, we will develop a rigorous proof [Saa11a; GLO20; Kat95] for such intuition. The result will also be used to prove the eigenvector–eigenvalue identity later.

Recall that $\lambda_i(\boldsymbol{A})$ is a simple eigenvalue of $\boldsymbol{A}$ if the multiplicity of $\lambda_i(\boldsymbol{A})$ is one.

Theorem 2.3 (First-order perturbation theorem). Assume $\boldsymbol{A}$ is Hermitian, and $(\lambda, \boldsymbol{v})$ is a simple eigenpair of $\boldsymbol{A}$, where $\boldsymbol{v}$ is a unit-norm eigenvector. Let $\boldsymbol{E} \in \mathbb{H}_n$ be a small perturbation and

$$\boldsymbol{A}(t) \coloneqq \boldsymbol{A} + t\boldsymbol{E}$$

be a family of perturbed matrices. If we write $\lambda(t)$ the eigenvalue of $\boldsymbol{A}(t)$ associated with $\boldsymbol{v}(t)$, where $\lambda(0) = \lambda$, and $\boldsymbol{v}(0) = \boldsymbol{v}$. Then, for small t, the perturbed eigenvalue is given by

$$\lambda(t) = \lambda + \langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v} \rangle t + O(t^2),$$

where $\langle \cdot,\ \cdot \rangle$ is the conventional Euclidean inner product.

Proof. For a Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_n(\mathbb{C})$, the left and right eigenvectors associated with the (real) eigenvalue λ coincide. That is

Recall that $\boldsymbol{v}$ is a right eigenvector of $\boldsymbol{A}$ associated with λ if $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. Similarly, $\boldsymbol{w}$ is a left eigenvector associated with λ if $\boldsymbol{w}^*\boldsymbol{A} = \lambda\boldsymbol{w}^*$.

$$\boldsymbol{A}^*\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

When t is small, the eigenvalue $\lambda(t)$ is analytic with respect to t. As $\lambda(t)$ is the eigenvalue of $\boldsymbol{A}(t)$, by definition we have

$$\boldsymbol{A}(t)\boldsymbol{v}(t) = \lambda(t)\boldsymbol{v}(t).$$

Let's calculate the Euclidean inner product of both sides with respect to $\boldsymbol{v}$, then

$$\langle \boldsymbol{v},\ \boldsymbol{A}(t)\boldsymbol{v}(t)\rangle = \langle \boldsymbol{v},\ \lambda(t)\boldsymbol{v}(t)\rangle.$$

Write it out more explicitly, we get

$$\langle \boldsymbol{v},\ \big(\boldsymbol{A} + t\boldsymbol{E}\big)\boldsymbol{v}(t)\rangle = \lambda(t)\langle \boldsymbol{v},\ \boldsymbol{v}(t)\rangle.$$

We calculate the left-hand side (LHS) of the above equation

$$\begin{aligned}\langle \boldsymbol{v},\ \big(\boldsymbol{A} + t\boldsymbol{E}\big)\boldsymbol{v}(t)\rangle &= \langle \boldsymbol{v},\ \boldsymbol{A}\boldsymbol{v}(t)\rangle + \langle \boldsymbol{v},\ t\boldsymbol{E}\boldsymbol{v}(t)\rangle \\ &= \langle \boldsymbol{A}^*\boldsymbol{v},\ \boldsymbol{v}(t)\rangle + t\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}(t)\rangle \\ &= \lambda\langle \boldsymbol{v},\ \boldsymbol{v}(t)\rangle + t\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}(t)\rangle.\end{aligned}$$

So, we have

$$\big(\lambda(t) - \lambda\big)\langle \boldsymbol{v},\ \boldsymbol{v}(t)\rangle = t\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}(t)\rangle.$$

This is equivalent to

$$\frac{\lambda(t) - \lambda(0)}{t - 0} = \frac{\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}(t)\rangle}{\langle \boldsymbol{v},\ \boldsymbol{v}(t)\rangle}$$

for a non-zero t. Because $(\lambda, \boldsymbol{v})$ is a simple eigenpair of $\boldsymbol{A}$, the eigenvector $\boldsymbol{v}$ and the eigenvalue λ are both continuous with respect to $\boldsymbol{A}$. Thus, $\boldsymbol{v}(t)$ is continuous with respect to t. Then, we have

$$\lambda'(0) = \lim_{t\to 0}\frac{\lambda(t) - \lambda(0)}{t - 0} = \lim_{t\to 0}\frac{\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}(t)\rangle}{\langle \boldsymbol{v},\ \boldsymbol{v}(t)\rangle} = \frac{\langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}\rangle}{\langle \boldsymbol{v},\ \boldsymbol{v}\rangle} = \langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}\rangle.$$

Thus, the Taylor expansion of $\lambda(t)$ about 0 is

$$\lambda(t) = \lambda(0) + \lambda'(0)(t - 0) + O(t^2) = \lambda + \langle \boldsymbol{v},\ \boldsymbol{E}\boldsymbol{v}\rangle t + O(t^2),$$

which gives the first-order perturbation theorem. ∎

2.3 Eigenvector–eigenvalue identity

The Cauchy interlacing theorem establishes inequalities for eigenvalues of a Hermitian matrix $\boldsymbol{A}$ and the eigenvalues of any principal submatrix of $\boldsymbol{A}$. In this section, we will establish a theorem that enables us to "find" the component of any eigenvector of $\boldsymbol{A}$ using the eigenvalues of $\boldsymbol{A}$ and eigenvalues of one principle minor of $\boldsymbol{A}$ [Den+22].

Theorem 2.4 (Eigenvector–eigenvalue identity). Let $\boldsymbol{A} \in \mathbb{H}_n$ be a $n \times n$ Hermitian matrix and choose j such that $1 \le j \le n$. Let $\boldsymbol{M}_j$ be the $n-1 \times n-1$ principal submatrix of $\boldsymbol{A}$ formed by deleting the jth row and column of $\boldsymbol{A}$. We arrange the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{M}$ in the weak increasing order, that is

$$\lambda_i(\boldsymbol{A}) = \lambda_i^{\uparrow}(\boldsymbol{A}) \quad \text{for} \quad 1 \le i \le n,$$

and

$$\lambda_i(\boldsymbol{M}_j) = \lambda_i^{\uparrow}(\boldsymbol{M}_j) \quad \text{for} \quad 1 \le i \le n-1.$$

Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ be the unit-norm eigenvectors of $\boldsymbol{A}$ associated with the ordered eigenvalues $\lambda_1(\boldsymbol{A}), \dots, \lambda_n(\boldsymbol{A})$, respectively. The eigenvector–eigenvalue identity for $\boldsymbol{A}$ is

$$|v_{ij}|^2 \prod_{k=1,k\neq i}^{n} \big(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\big) = \prod_{k=1}^{n-1} \big(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{M}_j)\big),$$

where v_{ij} is the ith component of $\boldsymbol{v}_j$.

We can also rewrite this identity using characteristic polynomial, as suggested in [Den+22]. Later, we will provide a proposition which generalizes the identity based on the same paper.

Definition 2.5 (Characteristic polynomial). The characteristic polynomial of a square matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ is a function $p_{\boldsymbol{M}} : \mathbb{C} \to \mathbb{C}$ defined by

$$p_{\boldsymbol{B}}(\lambda) \coloneqq \det(\lambda \mathbf{I}_n - \boldsymbol{B}) = \prod_{k=1}^{n} \big(\lambda - \lambda_k(\boldsymbol{B})\big).$$

Let $p_{\boldsymbol{A}} : \mathbb{C} \to \mathbb{C}$ be the characteristic polynomial of $\boldsymbol{A}$ defined in the above theorem. By definition, we have

$$p_{\boldsymbol{A}}(\lambda) = \det(\lambda \mathbf{I}_n - \boldsymbol{A}) = \prod_{k=1}^{n} \big(\lambda - \lambda_k(\boldsymbol{A})\big).$$

Similarly, the characteristic polynomial $p_{\boldsymbol{M}_j} : \mathbb{C} \to \mathbb{C}$ of $\boldsymbol{M}_j$ is

$$p_{\boldsymbol{M}_j}(\lambda) = \det(\lambda \mathbf{I}_{n-1} - \boldsymbol{M}_j) = \prod_{k=1}^{n-1} \big(\lambda - \lambda_k(\boldsymbol{M}_j)\big).$$

Using the chain rule, the derivative of $p_{\boldsymbol{A}}(\lambda)$ is

$$p'_{\boldsymbol{A}}(\lambda) = \sum_{l=1}^{n} \prod_{k=1,k\neq l}^{n} \big(\lambda - \lambda_k(\boldsymbol{A})\big).$$

When we evaluate this derivative at $\lambda = \lambda_i(\boldsymbol{A})$, we get

$$\begin{aligned} p'_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big) &= \sum_{l=1}^{n} \prod_{k=1,k\neq l}^{n} \big(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\big) \\ &= \prod_{k=1,k\neq i}^{n} \big(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\big). \end{aligned}$$

Then, the identity (2.4) is equivalent to

$$|v_{ij}|^2 p'_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big) = p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big).$$

There is an off-diagonal variant of the eigenvector-eigenvalue identity [Den+22]. Apart from the off-diagonal variant, the eigenvector-eigenvalue identity can be further generalized to obtain relationship between various minors of $\boldsymbol{A}$ and the unitary matrix $\boldsymbol{Q}$ which diagonalizes $\boldsymbol{A}$ (that is $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$). These results can also be found in [Den+22]. It is also possible to extend this identity to normal matrices and even diagonalizable matrices.

Proposition 2.6 (Off-diagonal characterization of the eigenvetor–eigenvalue identity). Using the above notation, and let $(\mathbf{I}_n)_{j'j}$ and $\boldsymbol{M}_{j'j}$ be the $(n-1)\times(n-1)$ minors formed by

deleting the j'th row and jth column of the identity matrix and matrix $\boldsymbol{A}$, respectively. Then, we have the following identity

$$(-1)^{j+j'} \det\left(\lambda_i(\boldsymbol{A})(\mathbf{I}_n)_{j'j} - \boldsymbol{M}_{j'j}\right) = \left(\prod_{k=1,k\neq i}^{n} \left(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\right)\right) v_{ij} v_{ij'}^*$$

for $1 \leq j, j' \leq n$.

2.4 Proof of the identity

The logic of this proof follows from [Den+22] while this is a more elaborate and self-contained version. As the eigenvectors of $\boldsymbol{A}$ are generally not continuous with respect to the entries of the matrix $\boldsymbol{A}$, one needs to be careful when applying the limiting argument. In the following proof, we are going to establish the identity by proving it in two cases. In the first part, the case for which λ_i is a repeated eigenvalue of $\boldsymbol{A}$, we will see it is easy to verify the identity, as both sides become zero. In the second part, we will focus on the case when λ_i is a simple eigenvalue of $\boldsymbol{A}$.

Let us first consider the case when $\lambda_i(\boldsymbol{A})$ is a repeated eigenvalue for some i with $1 \leq i \leq n$. If the eigenvalue $\lambda_i(\boldsymbol{A})$ of $\boldsymbol{A}$ occurs with multiplicity greater than one, without loss, we can then assume that

$$\lambda_{i+1}(\boldsymbol{A}) = \lambda_i(\boldsymbol{A}).$$

By the Cauchy interlacing theorem, the inequality

$$\lambda_i(\boldsymbol{A}) \leq \lambda_i(\boldsymbol{M}_j) \leq \lambda_{i+1}(\boldsymbol{A})$$

holds as $\boldsymbol{M}_j \in \mathbb{H}_{n-1}$ is a principal minor of $\boldsymbol{A} \in \mathbb{H}_n$. Thus, for any j, the ith smallest eigenvalue of $\boldsymbol{M}_j$ equals $\lambda_i(\boldsymbol{A})$:

$$\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{M}_j) = 0.$$

In this case, the identity is trivial as the left-hand side of the identity (2.4) is

$$\begin{aligned}\text{LHS} &= |v_{ij}|^2 \prod_{k=1,k\neq i}^{n} \left(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\right) \\ &= |v_{ij}|^2 \left(\lambda_i(\boldsymbol{A}) - \lambda_{i+1}(\boldsymbol{A})\right) \prod_{k=1,k\neq i,i+1}^{n} \left(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{A})\right) = 0,\end{aligned}$$

and the right-hand side (RHS) of the identity (2.4) is

$$\begin{aligned}\text{RHS} &= \prod_{k=1}^{n-1} \left(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{M}_j)\right) \\ &= \left(\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{M}_j)\right) \prod_{k=1,k\neq i}^{n-1} \left(\lambda_i(\boldsymbol{A}) - \lambda_k(\boldsymbol{M}_j)\right) = 0.\end{aligned}$$

Thus the identity holds for i with multiplicity of $\lambda_i(\boldsymbol{A})$ greater than one.

In the second half of the proof, we will consider the case when $\lambda_i(\boldsymbol{A})$ is a simple eigenvalue of $\boldsymbol{A}$. As the eigenvalues are continuous with respect to $\boldsymbol{A}$, the eigenvalues of a principal minor of $\boldsymbol{A}$ should also be continuous with respect to $\boldsymbol{A}$. That is, $\lambda_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{M}_j)$ are continuous with respect to $\boldsymbol{A}$. Because $\lambda_i(\boldsymbol{A})$ is a simple eigenvalue of $\boldsymbol{A}$, the eigenprojector $\boldsymbol{P}_{\lambda_i(\boldsymbol{A})}$ associated with $\lambda_i(\boldsymbol{A})$ is given by

$$\boldsymbol{P}_{\lambda_i(\boldsymbol{A})} = \boldsymbol{v}_i \boldsymbol{v}_i^*.$$

Thus, the eigenvector $\boldsymbol{v}_i$ associated with a simple eigenvalue $\lambda_i(\boldsymbol{A})$ is continuous with respect to $\boldsymbol{A}$. As the left-hand side and the right-hand side consist of products of the

eigenvalues and v_{ij}, they are both continuous with respect to $\boldsymbol{A}$. Then, it suffices to show that the identity holds for $\boldsymbol{A}$ with simple eigenvalue λ_i. Let ε be a small parameter, and consider the rank one perturbation $\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*$, where $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n$ is the standard basis. The characteristic polynomial of $\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*$ is given by

$$p_{\boldsymbol{A}+\varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*}(\lambda) = \det\Big(\lambda \mathbf{I}_n - \big(\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*\big)\Big).$$

For the sake of simplicity, we abbreviate the matrices as

$$\widetilde{\boldsymbol{A}}^{\varepsilon} = \lambda \mathbf{I}_n - \big(\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*\big) \quad \text{and} \quad \widetilde{\boldsymbol{A}} = \lambda \mathbf{I}_n - \boldsymbol{A}.$$

Furthermore, we write $\widetilde{\boldsymbol{M}}_{i,k}^{\varepsilon}$ and $\widetilde{\boldsymbol{M}}_{i,k}$ the submatrix of $\widetilde{\boldsymbol{A}}^{\varepsilon}$ and $\widetilde{\boldsymbol{A}}$ obtained by deleting the ith row and kth column of $\widetilde{\boldsymbol{A}}^{\varepsilon}$ and $\widetilde{\boldsymbol{A}}$, respectively. Note that $\boldsymbol{\delta}_j \boldsymbol{\delta}_j^*$ is a matrix with only the jth row and jth column is non-zero, so we have

$$\widetilde{\boldsymbol{M}}_{ij}^{\varepsilon} = \widetilde{\boldsymbol{M}}_{ij} \quad \text{for any} \quad 1 \le i \le n.$$

Furthermore, the submatrix $\widetilde{\boldsymbol{M}}_{jj}$ can be expressed explicitly using $\boldsymbol{M}_j$. That is,

$$\widetilde{\boldsymbol{M}}_{jj} = \lambda \mathbf{I}_{n-1} - \boldsymbol{M}_j.$$

Use cofactor expansion on the jth column, we obtain that

$$\begin{aligned}
p_{\boldsymbol{A}+\varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*}(\lambda) - p_{\boldsymbol{A}}(\lambda) &= \det(\widetilde{\boldsymbol{A}}^{\varepsilon}) - \det(\widetilde{\boldsymbol{A}}) \\
&= \sum\nolimits_{k=1}^{n} (-1)^{k+j} \det(\widetilde{\boldsymbol{M}}_{kj}^{\varepsilon}) \tilde{a}_{kj}^{\varepsilon} - \sum\nolimits_{k=1}^{n} (-1)^{k+j} \det(\widetilde{\boldsymbol{M}}_{kj}) \tilde{a}_{kj} \\
&= (-1)^{j+j} \det(\widetilde{\boldsymbol{M}}_{jj}^{\varepsilon}) \tilde{a}_{jj}^{\varepsilon} - (-1)^{j+j} \det(\widetilde{\boldsymbol{M}}_{jj}) \tilde{a}_{jj} \\
&= (-1)^{2j} \det(\widetilde{\boldsymbol{M}}_{jj}) \tilde{a}_{jj}^{\varepsilon} - (-1)^{j+j} \det(\widetilde{\boldsymbol{M}}_{jj}) \tilde{a}_{jj} \\
&= \det(\lambda \mathbf{I}_{n-1} - \boldsymbol{M}_j)(\tilde{a}_{jj}^{\varepsilon} - \tilde{a}_{jj}) \\
&= \det(\lambda \mathbf{I}_{n-1} - \boldsymbol{M}_j)(-\varepsilon) \\
&= -\varepsilon p_{\boldsymbol{M}_j}(\lambda).
\end{aligned}$$

$\tilde{a}_{i,k}^{\varepsilon}$ and $\tilde{a}_{i,k}$ are entries in the ith row and kth column of $\widetilde{\boldsymbol{A}}^{\varepsilon}$ and $\widetilde{\boldsymbol{A}}$, respectively.

In the third line of the above calculation, we use (2.4) and the fact that

$$\tilde{a}_{i,j}^{\varepsilon} = \tilde{a}_{i,j} \quad \text{for all} \quad i \ne j.$$

Thus, we have

$$p_{\boldsymbol{A}+\varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*}(\lambda) = p_{\boldsymbol{A}}(\lambda) - \varepsilon p_{\boldsymbol{M}_j}(\lambda).$$

On the other hand, based on the first order perturbation theory, the eigenvalue $\lambda_i(\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*)$ can be also expanded as

$$\begin{aligned}
\lambda_i(\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*) &= \lambda_i(\boldsymbol{A}) + \varepsilon \langle \boldsymbol{v}_i,\ \boldsymbol{\delta}_j \boldsymbol{\delta}_j^* \boldsymbol{v}_i \rangle + O(\varepsilon^2) \\
&= \lambda_i(\boldsymbol{A}) + \varepsilon |v_{ij}|^2 + O(\varepsilon^2),
\end{aligned}$$

where v_{ij} is the jth component of the ith unit eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{A}$. If we evaluate the charateristic function of $\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*$ at its eigenvalue, we have the following identity

$$p_{\boldsymbol{A}+\varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A} + \varepsilon \boldsymbol{\delta}_j \boldsymbol{\delta}_j^*)\big) = 0.$$

We expand $p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}(\lambda)$ in a Taylor series about $\lambda_i(\boldsymbol{A})$, which gives

$$p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}(\lambda) = p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big) + p'_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big)\Delta\lambda + O\big((\Delta\lambda)^2\big),$$

where $\Delta\lambda \coloneqq \lambda - \lambda_i(\boldsymbol{A})$. As we established (2.4), we have the first order approximation

$$\begin{aligned}
p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}(\lambda) &\approx p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big) + p'_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda \\
&= p_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big) - \varepsilon p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big) + p'_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda \\
&= -\varepsilon p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big) + p'_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda \\
&= -\varepsilon p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big) + p'_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda - \varepsilon p'_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda.
\end{aligned}$$

Note that we use (2.4) again in the last step of the above calculation. Based on (2.4), the quantity $p_{\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*}(\lambda)$ is zero when evaluated at $\lambda = \lambda_i(\boldsymbol{A}+\varepsilon\boldsymbol{\delta}_j\boldsymbol{\delta}_j^*)$. Using the expansion (2.4), we get $\Delta\lambda = \varepsilon|v_{ij}|^2 + O(\varepsilon^2)$. Thus, the term $\varepsilon p'_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big)\cdot\Delta\lambda$ is a second order term with respect to ε. By matching the linear term in ε, we get

$$0 = -\varepsilon p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big) + \varepsilon|v_{ij}|^2 p'_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big)$$

holds for any sufficiently small ε. That is,

$$|v_{ij}|^2 p'_{\boldsymbol{A}}\big(\lambda_i(\boldsymbol{A})\big) = p_{\boldsymbol{M}_j}\big(\lambda_i(\boldsymbol{A})\big),$$

which gives the desired result.

Thus, we can conclude that identity holds for every eigenvalue λ_i of $\boldsymbol{A}$. ■

It is interesting that both sides of the identity (2.4) equal zero when $\lambda_i(\boldsymbol{A})$ is a repeated eigenvalue of $\boldsymbol{A}$. This also matches our intuition that if the eigenspace associated with $\boldsymbol{A}$ has a dimension that is greater than one, there are infinite number of orthogonal bases for this eigenspace. Thus, it is impossible to pinpoint each component of the eigenvector.

2.5 Application

The eigenvector-eigenvalue identity allows one to reconstruct the magnitude of each component of eigenvectors of $\boldsymbol{A} \in \mathbb{H}_n$ from eigenvalues of $\boldsymbol{A}$ and its principal minor $\boldsymbol{M}_j$ (for some $j \in \mathbb{N}$, $1 \le j \le n$). However, the phase information of each component is not given in the identity. Nonetheless, proposition 2.6 shows that the relative phase $v_{ij}v_{ij'}^*$ can be determined.

As discussed in [Den+22], for a symmetric real matrix, the only ambiguity is the sign of each component, which may be recovered by direct inspection of the eigenvector equation $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i(\boldsymbol{A})\boldsymbol{v}_i$. Given the eigenvalues and eigenvectors associated with $\boldsymbol{A}$, one can obtain the original matrix $\boldsymbol{A}$ via the spectral decomposition.

Here, we are going to briefly show the application of the eigenvector-eigenvalue identity in neutrino oscillation [DPZ20]. Here, we will drop a lot of physical background knowledge needed to understand the physics of neutrino, and simply focus on the mathematical expression that models the transition among different types of neutrinos. In neutrino oscillation, the Pontecorvo–Maki–Nakagawa–Sakata matrix (PMNS matrix) describes the "probability" of transformation between mass eigenstates (denoted by 1, 2, 3, which tells the mass of the neutrino, and is the basis for neutrinos propagate in vacuum) and flavor eigenstates (denoted by e, μ, τ, which tells the species of the

neutrino, and is basis for neutrinos propagate in matter). Although it is not clear whether or not the three-neutrino model is correct, we proceed by assuming the PMNS matrix $\boldsymbol{U}_{\text{PMNS}}$ is a 3×3 unitary matrix. In this case, the PMNS matrix has the following decomposition

$$\begin{aligned}\boldsymbol{U}_{\text{PMNS}} &= \begin{bmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{bmatrix} \\ &= \begin{bmatrix} 1 & & \\ & c_{23} & s_{23}\mathrm{e}^{\mathrm{i}\delta} \\ & -s_{23}\mathrm{e}^{-\mathrm{i}\delta} & c_{23} \end{bmatrix} \begin{bmatrix} c_{13} & & s_{13} \\ & 1 & \\ -s_{13} & & c_{13} \end{bmatrix} \begin{bmatrix} c_{12} & s_{12} & \\ -s_{12} & c_{12} & \\ & & 1 \end{bmatrix},\end{aligned}$$

where $s_{ij} = \sin\theta_{ij}$, $c_{ij} = \cos\theta_{ij}$, and $U_{\alpha i}$ is the probability of i becoming α for $\alpha \in \{e, \mu, \tau\}$ and $i \in \{1, 2, 3\}$. The full solution for entries of the PMNS matrix can be obtained by solving a hard cubic equation. However, it is possible to obtain the entries via $\hat{U}_{\alpha i}$, for $\alpha \in \{e, \mu, \tau\}$ and $i \in \{1, 2, 3\}$, which is component of eigenvectors of the Hamiltonian in flavor basis.

In matter, the oscillation among these three flavors of neutrino is given by the following matrix $\boldsymbol{H}$ (the Hamiltonian in flavor basis):

$$\begin{aligned}\boldsymbol{H} &= \begin{bmatrix} H_{ee} & H_{\mu e} & H_{\tau e} \\ H_{e\mu} & H_{\mu\mu} & H_{\tau\mu} \\ H_{e\tau} & H_{\mu\tau} & H_{\tau\tau} \end{bmatrix} \\ &= \frac{1}{2E}\left[\boldsymbol{U}_{\text{PMNS}} \begin{pmatrix} 0 & & \\ & \Delta m_{21}^2 & \\ & & \Delta m_{31}^2 \end{pmatrix} \boldsymbol{U}_{\text{PMNS}}^{*} + \begin{pmatrix} a & & \\ & 0 & \\ & & 0 \end{pmatrix}\right],\end{aligned}$$

where E, Δm_{21}, Δm_{31}, and a are constant, $H_{\beta\alpha}$ denotes the probability of α becoming β for $\alpha, \beta \in \{e, \mu, \tau\}$. It is known that $\hat{\boldsymbol{U}}_{\alpha}$ ($\alpha \in \{e, \mu, \tau\}$) is the eigenvector of $\boldsymbol{H}$. So, if the eigenvalues of $\boldsymbol{H}$ and its pricipal minors are given, then entries of $\boldsymbol{U}_{\text{PMNS}}$ can be calculated via eigenvector-eigenvalue identity. Let the eigenvalues for $\boldsymbol{H}$ be

$$\frac{\lambda_j}{2E} \quad \text{for} \quad j = 1, 2, 3.$$

For each principal minor of $\boldsymbol{H}$, let the eigenvalues for the minor obtained by deleting all columns and rows contain subscript α for $\alpha \in \{e, \mu, \tau\}$ be

$$\frac{\xi_k^{\alpha}}{2E} \quad \text{for} \quad k = 1, 2.$$

Then, the eigenvector-eigenvalue identity tells us that the square of $\hat{U}_{\alpha j}$ is given by

$$|\hat{U}_{\alpha j}|^2 = \frac{\prod_{k=1}^{2}(\lambda_j - \xi_k^{\alpha})}{\prod_{i\neq j}(\lambda_j - \lambda_i)}$$

for $\alpha \in \{e, \mu, \tau\}$, and $j \in \{1, 2, 3\}$. As one can find the entries of the PMNS matrix via $\hat{U}_{\alpha j}$, the PMNS matrix can be obtained by measuring the eigenvalues of $\boldsymbol{H}$ and its principal minors. This is often easier than measuring each entries of the PMNS matrix itself. Furthermore, if experiments show that the three-neutrino theory is incorrect, the above formula is ready to be applied for $\boldsymbol{U}_{\text{PMNS}} \in \mathbb{H}_4$ with trivial modifications.

Lecture bibliography

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[DPZ20] P. B. Denton, S. J. Parke, and X. Zhang. "Neutrino oscillations in matter via eigenvalues". In: *Phys. Rev. D* 101 (2020), page 093001. DOI: 10.1103/PhysRevD.101.093001.

[Den+22] P. B. Denton et al. "Eigenvectors from eigenvalues: a survey of a basic identity in linear algebra". In: *Bull. Amer. Math. Soc. (N.S.)* 59.1 (2022), pages 31–58. DOI: 10.1090/bull/1722.

[GLO20] A. Greenbaum, R.-C. Li, and M. L. Overton. "First-Order Perturbation Theory for Eigenvalues and Eigenvectors". In: *SIAM Review* 62.2 (2020), pages 463–482. DOI: 10.1137/19M124784X.

[Kat95] T. Kato. *Perturbation theory for linear operators*. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995.

[Saa11a] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. 2nd edition. Society for Industrial and Applied Mathematics, 2011. DOI: 10.1137/1.9781611970739.

[Wol19] N. Wolchover. "Neutrinos lead to unexpected discovery in basic math". In: *Quanta Magazine* (2019). URL: https://www.quantamagazine.org/neutrinos-lead-to-unexpected-discovery-in-basic-math-20191113.

3. Bipartite Ramanujan Graphs of Any Degree

Date: 14 March 2022 **Author:** Nico Christianson

We review the work of Marcus, Spielman, and Srivastava on the existence of infinite families of bipartite Ramanujan graphs of any degree, closely following the arguments laid forth in the series of works [MSS14; MSS15a; MSS15b]. In particular, we present a general perspective that relies on the notion of *real stability* of multivariate polynomials and the resulting implications on *interlacing* of univariate polynomials. These techniques were successfully applied by the authors of the aforementioned works to several other problems, including the Kadison–Singer problem. The results highlight the power of approaching problems in matrix theory via direct consideration of the characteristic polynomial.

Agenda:
1. Bipartite Ramanujan graphs
2. Reductions
3. Interlacing polynomials and real stability
4. Proof of Theorem 3.13

3.1 Bipartite Ramanujan graphs

In this section, we provide a very brief overview of the graph-theoretic notions that are needed to define Ramanujan graphs and to motivate the problem of constructing them. A reader unfamiliar with graph theory would be well-served by reading a more thorough introduction to basic graph theory, e.g., the first few chapters of Spielman's book [Spi19].

3.1.1 Graph-theoretic preliminaries

We first recall some basic definitions from graph theory. A *graph* $G = (\mathsf{V}, \mathsf{E})$ is a set of *vertices* V along with a collection of *edges* $\mathsf{E} \subseteq \mathsf{V} \times \mathsf{V}/\sim$ between vertices, where $\sim$ is the equivalence relation of symmetry. That is, for vertices $u, v \in \mathsf{V}$, the ordered pairs (u, v) and (v, u) denote the same *undirected* edge. We will assume that graphs contain no self-edges, so that $(v, v) \notin \mathsf{E}$ for any vertex $v \in \mathsf{V}$. Throughout, we refer to the cardinality of the vertex set as $n := |\mathsf{V}|$ and the cardinality of the edge set as $m := |\mathsf{E}|$. As the vertex naming is not important, we will simply consider the vertex set as $\mathsf{V} = \{1, \dots, n\}$. We will only consider *connected* graphs, wherein each vertex is connected to every other vertex via some sequence of edges.

Given a graph G, its associated *adjacency matrix* $\boldsymbol{A} \in \mathbb{M}_n$ is the matrix whose entries indicate the connection of two vertices via an edge. That is, $\boldsymbol{A}$ has entries

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathsf{E}; \\ 0 & \text{otherwise.} \end{cases}$$

Since edges are undirected, the adjacency matrix $\boldsymbol{A}$ is symmetric: if $(i, j) \in \mathsf{E}$, then $a_{ij} = a_{ji} = 1$. By the spectral theorem, $\boldsymbol{A}$ has n real eigenvalues. For simplicity, we will also call these the eigenvalues of the graph G.

In this lecture, we pay particularly close attention to the set of d-regular bipartite graphs. A *d-regular* graph is one whose adjacency matrix $\boldsymbol{A}$ has the all-ones vector $\mathbf{1}$ as an eigenvector, with associated eigenvalue d. More tangibly, each vertex in a d-regular graph participates in exactly d edges. We call d the *trivial* eigenvalue of a d-regular graph, since any graph in this class has d as an eigenvalue.

A *bipartite* graph is one whose vertices V can be partitioned into disjoint subsets $\mathsf{S}, \mathsf{T} \subseteq \mathsf{V}$ so that any edge bridges S and T. That is, for any edge $(u, v) \in \mathsf{E}$, either $u \in \mathsf{S}$ and $v \in \mathsf{T}$, or else $u \in \mathsf{T}$ and $v \in \mathsf{S}$. The eigenvalues of a bipartite graph are symmetric about the origin; we leave this fact as an exercise.

Exercise 3.1 (Bipartite graph spectrum). Show that if $G = (\mathsf{V}, \mathsf{E})$ is bipartite, its spectrum is symmetric about the origin. **Hint:** Argue that there is a vertex ordering under which the adjacency matrix of a bipartite graph takes the form

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}$$

for some matrix $\boldsymbol{B}$.

It follows from Exercise 3.1 that a d-regular bipartite graph has an eigenvalue of $-d$ in addition to the eigenvalue d. We call both of these the *trivial eigenvalues* of a d-regular bipartite graph.

We conclude by defining an *infinite family* of graphs having a particular property.

Definition 3.2 (Infinite family). An *infinite family* of graphs with property P is an infinite sequence of graphs $(G_i)_{i=1}^{\infty}$ such that:

- Each graph G_i has property P, and
- Each graph G_{i+1} has strictly more vertices than its predecessor G_i.

3.1.2 Ramanujan graphs

In theoretical computer science and mathematics, there is great interest in the class of *expander* graphs, which are those whose nontrivial eigenvalues are small. The interested reader can refer to the classic survey of Hoory, Linial, and Widgerson [HLW06] or the more recent book of Kowalski [Kow19] for a more thorough overview of expanders and their many applications. Of particular note within the broader family of expanders are *Ramanujan* graphs, which are defined as follows.

Definition 3.3 (Ramanujan graph). A connected d-regular graph G is called *Ramanujan* if all of its nontrivial eigenvalues have magnitude at most $2\sqrt{d-1}$. In particular, a bipartite d-regular graph G is Ramanujan if all of its eigenvalues, aside from $\pm d$, are bounded in magnitude by $2\sqrt{d-1}$.

Ramanujan graphs are notable in that they are the *ideal* expander graphs. That is, no infinite family of graphs can have (asymptotically) smaller nontrivial eigenvalues. A lower bound due to Alon and Boppana [Alo86; Nil91] establishes that for any $\varepsilon > 0$ and any infinite family $(G_i)_{i=1}^{\infty}$ of d-regular graphs, there exists some $N \in \mathbb{N}$ for which G_i has a nontrivial eigenvalue of magnitude at least $2\sqrt{d-1} - \varepsilon$ when $i \geq N$.

For fixed d, it is easy to construct d-regular Ramanujan graphs with few vertices, as demonstrated in the following example.

Example 3.4 (Complete bipartite graph). Consider the bipartite graph $K_{d,d}$ with vertex set $\mathsf{V} = \mathsf{S} \cup \mathsf{T}$, where $\mathsf{S} = \{1, \ldots, d\}$ and $\mathsf{T} = \{d+1, \ldots, 2d\}$, each vertex $s \in \mathsf{S}$ has an edge to every vertex in T, and these are the only edges. This is known as a *complete bipartite graph*. Its adjacency matrix $\boldsymbol{A}$ is of the form

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{1}$ denotes the $d \times d$ matrix of ones. Since $\boldsymbol{A}$ has rank 2, it follows that all of the nontrivial eigenvalues of $K_{d,d}$ are zero. Thus $K_{d,d}$ is Ramanujan. ∎

In general, it is not known whether there exist d-regular Ramanujan graphs with arbitrarily many vertices. Our task in this lecture will be to show that this is the case when we restrict our focus to *bipartite* Ramanujan graphs. Specifically, we prove the following result of Marcus, Spielman, and Srivastava, following the works [MSS15a; MSS15b].

Theorem 3.5 (Marcus, Spielman, and Srivastava 2015). For every $d \geq 3$, there is an infinite family of d-regular bipartite Ramanujan graphs.

Our treatment of this result draws from the papers [MSS15a; MSS15b], as well as the survey [MSS14] and Spielman's course notes [Spi18b].

3.2 Reductions

In this section, we present several reductions of Theorem 3.5. We begin by reducing the existence of infinite families of d-regular bipartite Ramanujan graphs to the existence of a *signed* adjacency matrix with a certain eigenvalue bound. Subsequently, we shift our focus to the characteristic polynomial of this matrix, and we show that we can equivalently phrase the problem as a bound relating the root of a random characteristic polynomial with the root of its expectation. In the next section, these reductions enable us to deploy techniques from [MSS15b] on interlacing families of polynomials.

3.2.1 2-lifts

To prove the existence of infinite families of bipartite Ramanujan graphs, it would be very useful if, given any bipartite Ramanujan graph on n vertices, we could use it to construct another Ramanujan graph with more than n vertices. To this end, we introduce the following definitions.

Definition 3.6 (Signing). Let G be a graph with adjacency matrix $\boldsymbol{A} \in \mathbb{M}_n$. A *signing* $\boldsymbol{S} \in \mathbb{M}_n$ of G is a matrix of the form

$$\boldsymbol{S} = \sum_{(i,j) \in \mathsf{E}} \sigma_{(i,j)} \left(\boldsymbol{\delta}_i \boldsymbol{\delta}_j^* + \boldsymbol{\delta}_j \boldsymbol{\delta}_i^* \right), \tag{3.1}$$

where the sign vector $\boldsymbol{\sigma} \in \{\pm 1\}^m$ is indexed by edges. That is, $\boldsymbol{S}$ has the same nonzero pattern as the adjacency matrix $\boldsymbol{A}$, but both entries corresponding to an edge (i, j) are given a sign $\sigma_{(i,j)} \in \{\pm 1\}$.

Recall that the indicator vector $\boldsymbol{\delta}_i \in \mathbb{R}^n$ is 1 in its ith entry, and 0 everywhere else.

A signing of a graph is symmetric, and thus its eigenvalues are real.

Definition 3.7 (2-lift). Let $\boldsymbol{S} \in \mathbb{M}_n$ be a signing of a graph $G = (\mathsf{V}, \mathsf{E})$. We define the 2-*lift* $G^{\boldsymbol{S}}$ of G associated with the signing $\boldsymbol{S}$ as follows.

- For every vertex $v \in \mathsf{V}$, the 2-lift $G^{\boldsymbol{S}}$ has two vertices, which we denote v_0, v_1.
- If $s_{uv} = 1$, then $G^{\boldsymbol{S}}$ has the two edges (u_0, v_0) and (u_1, v_1).
- If $s_{uv} = -1$, then $G^{\boldsymbol{S}}$ has the two edges (u_0, v_1) and (u_1, v_0).

2-lifts give a means of constructing larger graphs out of smaller graphs in a manner that preserves certain desirable properties. In particular, any 2-lift of a bipartite,

d-regular graph is bipartite and d-regular, and the spectrum of a 2-lift $G^{\boldsymbol{S}}$ depends only on the spectra of G and $\boldsymbol{S}$. We leave these facts as exercises.

Exercise 3.8 (2-lift: Preservation of properties). Show that the following properties are preserved by 2-lifts.

- **Bipartiteness.** Any 2-lift of a bipartite graph is bipartite.
- **d-regularity.** Any 2-lift of a d-regular graph is d-regular.

Exercise 3.9 (2-lift: Spectrum). Let $\boldsymbol{S}$ be a signing of a graph G. Show that the eigenvalues of $G^{\boldsymbol{S}}$ are exactly the union of the eigenvalues of G and the eigenvalues of $\boldsymbol{S}$.

These exercises imply that, if G is a d-regular Ramanujan graph and it has a signing $\boldsymbol{S}$ satisfying $\|\boldsymbol{S}\| \leq 2\sqrt{d-1}$, then $G^{\boldsymbol{S}}$ is also Ramanujan. Thus, if it were the case that *any* d-regular graph had a signing $\boldsymbol{S}$ with $\|\boldsymbol{S}\| \leq 2\sqrt{d-1}$, this would immediately give a means for constructing infinite families of d-regular Ramanujan graphs. It was conjectured by Bilu and Linial [BL06] that such a signing can always be found; however, we will prove the following weaker result from [MSS15a].

Theorem 3.10 (Existence of signing with spectrum lower bound). Any d-regular graph G has a signing $\boldsymbol{S}$ whose minimum eigenvalue $\lambda_n(\boldsymbol{S})$ is at least $-2\sqrt{d-1}$.

Since the eigenvalues of a bipartite graph are symmetric about the origin, Theorem 3.10 immediately implies (via the preceding argument) the existence of infinite families of d-regular *bipartite* Ramanujan graphs.

3.2.2 The expected characteristic polynomial

To prove Theorem 3.10, instead of reasoning about the eigenvalues of a signing $\boldsymbol{S}$ directly, we will instead consider the roots of the characteristic polynomial of a *random* signing. In the following, we use $\lambda_k(\cdot)$ to refer both to the kth largest eigenvalue of a matrix as well as the kth largest root of a (real-rooted) polynomial. Recall that the characteristic polynomial of a square matrix $\boldsymbol{M} \in \mathbb{M}_n$ is the univariate polynomial defined as

$$\chi(\boldsymbol{M}; t) := \det(t\mathbf{I} - \boldsymbol{M})$$

whose roots are exactly the eigenvalues of $\boldsymbol{M}$. We refer to the polynomial as $\chi(\boldsymbol{M})$, and we use the notation $\chi(\boldsymbol{M}; t)$ either when we wish to distinguish the name of its variable t, or else to refer to its value at some particular choice of t. Consider a random signing $\boldsymbol{S}$ of a graph G as in (3.1), where the signs are uniformly distributed: $\boldsymbol{\sigma} \sim \textsc{uniform}\{\pm 1\}^m$. Then Theorem 3.10 is equivalent to the statement that there is strictly positive probability of the smallest root of $\chi(\boldsymbol{S})$ having lower bound

$$\lambda_n\left(\chi(\boldsymbol{S})\right) \geq -2\sqrt{d-1}.$$

Notably, the expectation of the characteristic polynomial $\chi(\boldsymbol{S})$ under a uniformly random signing $\boldsymbol{S}$ is equal to the *matching polynomial* of the graph G, which is a generator function of matchings in a graph; the interested reader can refer to [GG78] for an introduction and proof of this fact. Moreover, it was shown by Heilmann and Lieb [HL72] that the largest root of the matching polynomial of a graph is bounded above by $2\sqrt{d-1}$. Thus to prove Theorem 3.10, it suffices to prove the following result.

Theorem 3.11 (Roots: Random signing vs. expected characteristic polynomial). Suppose G is a d-regular graph and $\boldsymbol{S}$ is a uniformly random signing. Then, with strictly positive probability,

$$\lambda_n\left(\chi(\boldsymbol{S})\right) \geq -\lambda_1\left(\mathbb{E}\left[\chi(\boldsymbol{S})\right]\right).$$

3.2.3 Sums of rank-one matrices

The definition of graph signings in (3.1) expresses a signing as a sum of rank-two matrices. It will be convenient for us to instead frame the problem in terms of rank-one matrices. To achieve this, we take inspiration from the graph Laplacian and consider Laplacianized signings

$$\boldsymbol{L}_{\boldsymbol{S}} \coloneqq d\mathbf{I} - \boldsymbol{S}$$

of a d-regular graph G with signing $\boldsymbol{S}$. Expanding the definition (3.1), we can write this Laplacianized signing as

$$\begin{aligned}\boldsymbol{L}_{\boldsymbol{S}} &= d\sum\nolimits_{i\in\mathsf{V}} \boldsymbol{\delta}_i\boldsymbol{\delta}_i^* + \sum\nolimits_{(i,j)\in\mathsf{E}} \sigma_{(i,j)}\left(\boldsymbol{\delta}_i\boldsymbol{\delta}_j^* + \boldsymbol{\delta}_j\boldsymbol{\delta}_i^*\right)\\ &= \sum\nolimits_{(i,j)\in\mathsf{E}} (\boldsymbol{\delta}_i - \sigma_{(i,j)}\boldsymbol{\delta}_j)(\boldsymbol{\delta}_i - \sigma_{(i,j)}\boldsymbol{\delta}_j)^*,\end{aligned}$$

which is a sum of rank-one matrices. We introduce the vectors $\boldsymbol{v}_e = \boldsymbol{\delta}_i - \sigma_{(i,j)}\boldsymbol{\delta}_j$ indexed by edges $e = (i,j) \in \mathsf{E}$; Theorem 3.11 can then be reformulated equivalently as follows.

The graph Laplacian of a d-regular graph G with adjacency matrix $\boldsymbol{A}$ is the matrix $\boldsymbol{L}_{\boldsymbol{A}} \coloneqq d\mathbf{I} - \boldsymbol{A}$. The Laplacian is a central object of study in the field of spectral graph theory.

Theorem 3.12 (Equivalent reformulation of Theorem 3.11). Suppose G is a d-regular graph and $\boldsymbol{S}$ is a uniformly random signing with decomposition as in (3.1). Let $\boldsymbol{v}_e = \boldsymbol{\delta}_i - \sigma_{(i,j)}\boldsymbol{\delta}_j$ for each $e = (i,j) \in \mathsf{E}$. Then, with strictly positive probability,

$$\lambda_1\left(\chi\left(\sum\nolimits_{e\in\mathsf{E}} \boldsymbol{v}_e\boldsymbol{v}_e^*\right)\right) \leq \lambda_1\left(\mathbb{E}\left[\chi\left(\sum\nolimits_{e\in\mathsf{E}} \boldsymbol{v}_e\boldsymbol{v}_e^*\right)\right]\right). \tag{3.2}$$

Proof that Theorem 3.12 implies Theorem 3.11. Suppose that the theorem's conclusion (3.2) holds. Since $d\mathbf{I} - \boldsymbol{S} = \sum_{e\in\mathsf{E}} \boldsymbol{v}_e\boldsymbol{v}_e^*$, it follows that

$$\begin{aligned} d - \lambda_n(\boldsymbol{S}) &= \lambda_1(d\mathbf{I} - \boldsymbol{S})\\ &\leq \lambda_1\left(\mathbb{E}\left[\chi\left(d\mathbf{I} - \boldsymbol{S}\right)\right]\right)\\ &= d + \lambda_1\left(\mathbb{E}\left[\chi\left(-\boldsymbol{S}\right)\right]\right) &(3.3)\\ &= d + \lambda_1\left(\mathbb{E}\left[\chi\left(\boldsymbol{S}\right)\right]\right). &(3.4)\end{aligned}$$

The equality (3.3) holds because the characteristic polynomial $\chi\left(d\mathbf{I} - \boldsymbol{S}; t\right)$ is a horizontal translation (by d) of the characteristic polynomial $\chi\left(-\boldsymbol{S}; t\right)$. The equality (3.4) follows by symmetry of the uniform distribution over signings. Subtracting d and negating gives the result of Theorem 3.11. ∎

We only prove that Theorem 3.12 implies Theorem 3.11, as this is sufficient to establish the reduction. However, the reverse implication can be seen by a nearly identical argument.

Rather than prove Theorem 3.12, we will instead prove the following more general result in the case of an arbitrary sum of independent rank-one matrices. Theorem 3.12 then clearly follows as an immediate corollary.

Theorem 3.13 (Roots: Independent rank-one sum vs. expected char. poly.). Suppose that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{C}^n$ are independent random vectors, each with finite support. Then,

with strictly positive probability,

$$\lambda_1\left(\chi\left(\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^*\right)\right) \le \lambda_1\left(\mathbb{E}\left[\chi\left(\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^*\right)\right]\right).$$

Theorem 3.13 is significantly more general than Theorem 3.12: in particular, it relates the largest root of *any* sum of (finitely-supported) independent rank-one matrices to the largest root of the expected characteristic polynomial of the sum. This more general result has numerous implications beyond the existence of bipartite Ramanujan graphs. Theorem 3.13 was established by Marcus, Spielman, and Srivastava in [MSS15b] to solve the Kadison–Singer problem, and it was used by the same authors in [MSS21] to obtain sharpened versions of the restricted invertibility principle. We focus on this more general setting to highlight the theoretical techniques unifying these results.

3.3 Interlacing polynomials and real stability

Our strategy for proving Theorem 3.13, following [MSS15b], will be to show that the family comprised of the characteristic polynomials of all realizations of the random matrix $\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^*$ has a *common interlacing*. This property, which we will soon define, allows us to compare the roots of the average of polynomials with the roots of the constituents of the average. The argument crucially relies upon the notion of *real stability* of polynomials, which generalizes real-rootedness beyond the univariate setting. In this section, we will introduce the necessary background on polynomial interlacing and real stability that will be used to prove Theorem 3.13 in the subsequent section.

3.3.1 Interlacing polynomials

We begin by defining the notion of interlacing of two polynomials.

Definition 3.14 (Interlacing polynomials). Let f be a polynomial of degree n with real roots $(\lambda_i)_{i=1}^n$, and let g be a polynomial of degree n or $n-1$ with real roots $(\mu_i)_{i=1}^{n \text{ or } n-1}$. The polynomial g is said to *interlace* f if their roots are alternating and those of f are larger; i.e.,

$$\mu_n \le \lambda_n \le \mu_{n-1} \le \cdots \le \mu_1 \le \lambda_1,$$

with the first inequality disregarded when g is of degree $n-1$.

We illustrate an example of a polynomial g interlacing another polynomial f in Figure 3.1. Next, we define what it means for a family of polynomials to have a common interlacing.

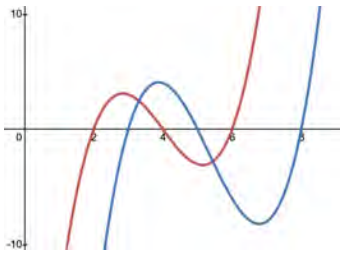

Figure 3.1 An example of interlacing polynomials plotted in Desmos. Here, the red polynomial $g(x) = (x-2)(x-4)(x-6)$ interlaces the blue polynomial $f(x) = (x-3)(x-5)(x-8)$, since their roots are alternating and the roots of f are larger.

Definition 3.15 (Common interlacing). Let $f_1, \ldots, f_m$ be a family of real-rooted polynomials, each with degree n. If there is a polynomial g that interlaces each f_i, then the family is said to have a *common interlacing*.

The existence of a common interlacing is in fact a pairwise property of a family of polynomials; we leave the proof of this fact as an exercise.

Exercise 3.16 (Pairwise common interlacing implies common interlacing of family). Let $f_1, \ldots, f_m$ be a family of real-rooted, degree-n polynomials; and suppose that, for any indices

$i, j \in \{1, \dots, m\}$, the polynomials f_i and f_j have a common interlacing. Show that $f_1, \dots, f_m$ have a common interlacing. **Hint:** This is an easy consequence of the combinatorial Helly theorem on the real line, which states that the intersection of a finite collection of non-pairwise-disjoint intervals in $\mathbb{R}$ is nonempty; for example, see [Pak10, Chapter 1].

If a family of polynomials has a common interlacing, then we can relate each root of the average with the corresponding root of a member of the family. We provide the following lemma without proof, though we recommend that the interested reader give it some thought; its source and proof can be found in [MSS15a, Lemma 4.2].

Lemma 3.17 (Roots of family with common interlacing). Let $f_1, \dots, f_m$ be a family of real-rooted, degree-n polynomials with positive leading coefficients, and let μ be a probability distribution on $\{1, \dots, m\}$. If $f_1, \dots, f_m$ have a common interlacing, then there exists some $j \in \{1, \dots, m\}$ for which

$$\lambda_1(f_j) \leq \lambda_1\left(\mathbb{E}_{k\sim\mu} f_k\right).$$

This existential result actually extends beyond the leading root to all other roots λ_i, $i \in \{2, \dots, n\}$; see [MSS14, Theorem 2.2]. For our purposes, we only need the stated result involving the leading root λ_1.

It is evident that Lemma 3.17 reduces the proof of Theorem 3.13 to the problem of showing that the characteristic polynomials $\chi\left(\sum_{i=1}^m \boldsymbol{v}_i \boldsymbol{v}_i^*\right)$ have a common interlacing (when considered as a family of polynomials under realizations of the random vectors $\boldsymbol{v}_i$). To this end, it will be useful to establish a sufficient condition for the existence of a common interlacing of a family of polynomials. As it turns out, existence of such an interlacing is equivalent to the real-rootedness of arbitrary convex combinations of polynomials within the family. This basic idea has appeared in various forms a number of times, including [Ded92, Theorem 2.1], [Fel80, Theorem 2'], and [CS07, Theorem 3.6]. Our proof follows that of [MSS14].

Theorem 3.18 (Sufficient condition for common interlacing). Let $f_1, \dots, f_m$ be a family of degree-n polynomials. Suppose that, for all probability distributions μ supported on $\{1, \dots, m\}$, the polynomial $\mathbb{E}_{k\sim\mu} f_k$ has real roots. Then the family has a common interlacing.

Proof. By Exercise 3.16, it suffices to show that any pair of polynomials f_i, f_j with $i \neq j$ has a common interlacing under the given assumptions. Define the polynomial g_t as the convex combination

$$g_t(x) \coloneqq (1-t) f_i(x) + t f_j(x) \quad \text{where } t \in [0, 1].$$

We will proceed under the assumption that f_i and f_j have no roots in common. If they do share some root $\lambda \in \mathbb{R}$, then λ will also be a root of any convex combination g_t. Then it suffices to find a common interlacing $h(x)$ for the polynomials $f_i(x)/(x-\lambda)$ and $f_j(x)/(x-\lambda)$; once we have done this, it is straightforward to see that $(x-\lambda)\cdot h(x)$ will be a common interlacing for f_i and f_j.

Let $\lambda_k(t)$ denote the kth largest root of g_t. From complex analysis, we know that the roots of a polynomial are continuous in its coefficients; thus the image of the unit interval under λ_k is a continuous curve in the complex plane that begins at the kth largest root of f_i at $t = 0$, and ends at the kth largest root of f_j at $t = 1$. In particular, this curve must reside on the real line, owing to the assumption that expectations (i.e., convex combinations) of polynomials in the family have real roots. As such, the image $\lambda_k([0, 1])$ is a closed interval in $\mathbb{R}$. Moreover, for any $t \in (0, 1)$, it cannot be the case that $\lambda_k(t)$ is a root of f_i, for if this were the case, then the assumption of no common roots would be violated:

$$0 = g_t(\lambda_k(t)) = (1-t) f_i(\lambda_k(t)) + t f_j(\lambda_k(t)) = t f_j(\lambda_k(t)).$$

Similarly, $\lambda_k(t)$ cannot be a root of f_j for any $t \in (0,1)$. It follows that the interval $\lambda_k([0,1])$ contains, for each $k = 1, \dots, n$, a single root of each of f_i and f_j. It immediately follows that the polynomial h whose kth root is the left boundary point of the interval $\lambda_k([0,1])$ is a common interlacing for f_i and f_j. ∎

3.3.2 Real-stable polynomials

In our proof of Theorem 3.13, we will exercise Theorem 3.18 by employing the notion of real stability, which is defined as follows.

Definition 3.19 (Real stability). An m-variate polynomial $p \in \mathbb{R}[z_1, \dots, z_m]$ is *real stable* either if it is identically zero or if none of its roots has all coordinates strictly in the upper half plane. That is, a nonzero p is real stable if

$$\mathrm{Im}(z_i) > 0 \quad \text{for all} \quad i = 1, \dots, m \quad \text{implies} \quad p(z_1, \dots, z_m) \neq 0.$$

Since roots of a univariate polynomial come in conjugate pairs, a real-stable univariate polynomial has only real roots.

We will need the following few results on particular real-stable polynomials, as well as closure of the class under certain operations. We begin with a proposition due to Borcea and Brändén [BB08, Proposition 2.4].

Proposition 3.20 (Determinant of linear combination of psd Hermitian matrices). Suppose that $\boldsymbol{A}_1, \dots, \boldsymbol{A}_m \in \mathbb{H}_n^+$ are positive-semidefinite Hermitian matrices. Then the m-variate polynomial $p \in \mathbb{R}[z_1, \dots, z_m]$ defined as

$$p(z_1, \dots, z_m) \coloneqq \det\left(\sum_{i=1}^m z_i \boldsymbol{A}_i\right) \quad \text{is real stable.}$$

Proof. We sketch a proof in the case that each matrix $\boldsymbol{A}_i$ is strictly positive definite; the general result then follows from complex-analytic considerations. (See the proof in [BB08] for further details.) We consider the restriction of the polynomial p to a line. Define $\boldsymbol{z}(t) = \boldsymbol{\alpha} + t\boldsymbol{\beta}$, where $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}_{++}^m$ are arbitrary, and where $t \in \mathbb{C}$. Positive-definite matrices form a convex cone, so the matrix

$$\boldsymbol{P} \coloneqq \sum_{i=1}^m \beta_i \boldsymbol{A}_i \quad \text{is positive definite.}$$

A simple calculation shows that

$$p(\boldsymbol{z}(t)) = \det(\boldsymbol{P}) \det\left(t\mathbf{I} + \boldsymbol{P}^{-1/2}\left[\sum_{i=1}^m \alpha_i \boldsymbol{A}_i\right]\boldsymbol{P}^{-1/2}\right).$$

In particular, $\det(\boldsymbol{P})$ is constant and $\det\left(t\mathbf{I} + \boldsymbol{P}^{-1/2}\left[\sum_{i=1}^m \alpha_i \boldsymbol{A}_i\right]\boldsymbol{P}^{-1/2}\right)$ is the characteristic polynomial (in the variable t) of the Hermitian matrix $-\boldsymbol{P}^{-1/2}\left[\sum_{i=1}^m a_i \boldsymbol{A}_i\right]\boldsymbol{P}^{-1/2}$. By the spectral theorem, it follows that $p(\boldsymbol{z}(t))$ has real roots, implying real stability. ∎

We next present two results on the closure of the class of real-stable polynomials under two operations: the first is the difference of the identity and a partial derivative operator, and the second is partial evaluation of the polynomial at a real number. We state these without proof, instead referring the reader to [MSS15b, Corollary 3.8 and Proposition 3.9] for a more detailed overview.

Lemma 3.21 (Closure: Identity minus derivative). Let $p \in \mathbb{R}[z_1, \dots, z_m]$ be real stable. For any $i \in \{1, \dots, m\}$, it is the case that $(1 - \partial_{z_i})p$ is real stable.

Lemma 3.22 (Closure: Partial evaluation). Let $p \in \mathbb{R}[z_1, \ldots, z_m]$ be real stable. For any fixed $a \in \mathbb{R}$, the $(m-1)$-variate polynomial $p(a, z_2, \ldots, z_m)$ is real stable.

We now define the *mixed characteristic polynomial* of a collection of matrices.

Definition 3.23 (Mixed characteristic polynomial). Let $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{M}_n$ be matrices. The *mixed characteristic polynomial* of $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ is the univariate polynomial $\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m] \in \mathbb{R}[t]$ defined as

$$\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m; t] := \left(\prod_{i=1}^m 1 - \partial_{z_i}\right) \det\left(t\mathbf{I} + \sum_{i=1}^m z_i \boldsymbol{A}_i\right)\Bigg|_{z_1,\ldots,z_m=0}. \quad (3.5)$$

Like with the characteristic polynomial, we refer to the mixed characteristic polynomial as $\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m]$, using $\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m; t]$ instead when it is useful to emphasize the name of its variable t.

Finally, we present the following result from [MSS15b, Theorem 4.1] relating the expected characteristic polynomial of the sum of independent rank-one matrices to a mixed characteristic polynomial of the corresponding covariance matrices.

Theorem 3.24 (Expected char. poly. is a mixed char. poly.). Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{C}^n$ be independent random vectors with covariance matrices $\boldsymbol{A}_i = \mathbb{E}\, \boldsymbol{v}_i \boldsymbol{v}_i^*$. Then

$$\mathbb{E}\left[\chi\left(\sum_{i=1}^m \boldsymbol{v}_i \boldsymbol{v}_i^*; t\right)\right] = \mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m; t]. \quad (3.6)$$

Moreover, $\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m]$ has real roots.

Proof sketch. For the sake of brevity, we will not prove the equality (3.6); we refer the reader to [MSS15b, Section 4] for a proof of this result. On the other hand, real-rootedness of $\mu[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m]$ follows immediately from the definition of the mixed characteristic polynomial in (3.5), Proposition 3.20, positive semidefiniteness of covariance matrices, the closure properties in Lemmas 3.21 and 3.22, and the fact that real stability is equivalent to real-rootedness in the univariate case. ∎

3.4 Proof of Theorem 3.13

Equipped with the tools of the previous section, we are ready to tackle the proof of Theorem 3.13. We proceed inductively by showing that the collection of "conditional expectation polynomials" at a certain level have a common interlacing, which establishes the desired bound at that level.

Define the conditional expectation polynomial at level ℓ with first ℓ vectors assigned as $\boldsymbol{v}_1 = \boldsymbol{u}_1, \ldots, \boldsymbol{v}_\ell = \boldsymbol{u}_\ell$ by

$$q_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell}(x) := \mathbb{E}_{\boldsymbol{v}_{\ell+1},\ldots,\boldsymbol{v}_m}\left[\chi\left(\sum_{i=1}^\ell \boldsymbol{u}_i \boldsymbol{u}_i^* + \sum_{j=\ell+1}^m \boldsymbol{v}_i \boldsymbol{v}_i^*\right)\right].$$

Write $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r\}$ for the support of $\boldsymbol{v}_{\ell+1}$ (which, by assumption, is finite). Let ν

be an arbitrary distribution on $\{1, \dots, r\}$. Then observe that

$$\begin{aligned}
&\mathbb{E}_{k\sim\nu}\left[q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell,\boldsymbol{w}_k}(x)\right] \\
&\quad= \mathbb{E}_{\boldsymbol{v}_{\ell+2},\dots,\boldsymbol{v}_m;k\sim\nu}\left[\chi\left(\sum_{i=1}^{\ell}\boldsymbol{u}_i\boldsymbol{u}_i^* + \boldsymbol{w}_k\boldsymbol{w}_k^* + \sum_{j=\ell+2}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^*\right)\right] \\
&\quad= \mu\left[\boldsymbol{u}_1\boldsymbol{u}_1^*, \dots, \boldsymbol{u}_\ell\boldsymbol{u}_\ell^*, \mathbb{E}_{k\sim\nu}\,\boldsymbol{w}_k\boldsymbol{w}_k^*, \mathbb{E}\,\boldsymbol{v}_{\ell+2}\boldsymbol{v}_{\ell+2}^*, \dots, \mathbb{E}\,\boldsymbol{v}_m\boldsymbol{v}_m^*\right].
\end{aligned}$$

The mixed characteristic polynomial in the last line has real roots, by Theorem 3.24 (which we also employed in the second equality). The distribution ν was arbitrary, so we may invoke Theorem 3.18 to discover that the family of polynomials $(q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell,\boldsymbol{w}_k})_{k=1}^r$ has a common interlacing. Then by Lemma 3.17, there is some $k \in \{1, \dots, r\}$ for which

$$\lambda_1(q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell,\boldsymbol{w}_k}) \le \lambda_1(q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell}).$$

Indeed, $q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell} = \mathbb{E}_{\boldsymbol{v}_{\ell+1}}[q_{\boldsymbol{u}_1,\dots,\boldsymbol{u}_\ell,\boldsymbol{v}_{\ell+1}}]$. Applying this argument repeatedly at each level $\ell \in \{1, \dots, m\}$ and chaining the inequalities, we eventually obtain an assignment $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ satisfying

$$\lambda_1\left(\chi\left(\sum_{i=1}^{m}\boldsymbol{u}_i\boldsymbol{u}_i^*\right)\right) \le \lambda_1\left(\mathbb{E}\left[\chi\left(\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^*\right)\right]\right),$$

which is the desired result. ∎

Lecture bibliography

[Alo86] N. Alon. "Eigenvalues and Expanders". In: *Combinatorica* 6.2 (June 1986), pages 83–96. DOI: 10.1007/BF02579166.

[BL06] Y. Bilu and N. Linial. "Lifts, Discrepancy and Nearly Optimal Spectral Gap". In: *Combinatorica* 26.5 (2006), pages 495–519. DOI: 10.1007/s00493-006-0029-7.

[BB08] J. Borcea and P. Brändén. "Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized Fischer products". In: *Duke Mathematical Journal* 143.2 (2008), pages 205 –223. DOI: 10.1215/00127094-2008-018.

[CS07] M. Chudnovsky and P. Seymour. "The roots of the independence polynomial of a clawfree graph". In: *Journal of Combinatorial Theory, Series B* 97.3 (2007), pages 350–357. DOI: https://doi.org/10.1016/j.jctb.2006.06.001.

[Ded92] J. P. Dedieu. "Obreschkoff's theorem revisited: what convex sets are contained in the set of hyperbolic polynomials?" In: *Journal of Pure and Applied Algebra* 81.3 (1992), pages 269–278. DOI: https://doi.org/10.1016/0022-4049(92)90060-S.

[Fel80] H. J. Fell. "On the zeros of convex combinations of polynomials." In: *Pacific Journal of Mathematics* 89.1 (1980), pages 43 –50. DOI: pjm/1102779366.

[GG78] C. Godsil and I. Gutman. "On the matching polynomial of a graph". In: *Algebraic Methods in Graph Theory* 25 (Jan. 1978).

[HL72] O. J. Heilmann and E. H. Lieb. "Theory of monomer-dimer systems". In: *Communications in Mathematical Physics* 25.3 (1972), pages 190 –232. DOI: cmp/1103857921.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. "Expander graphs and their applications". In: *Bulletin of the American Mathematical Society* 43.4 (2006), pages 439–561.

[Kow19] E. Kowalski. *An introduction to expander graphs*. Société mathématique de France, 2019.

[MSS14] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Ramanujan graphs and the solution of the Kadison-Singer problem". In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. III*. Kyung Moon Sa, Seoul, 2014, pages 363–386.

[MSS15a] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Interlacing families I: Bipartite Ramanujan graphs of all degrees". In: *Annals of Mathematics* 182.1 (2015), pages 307–325. URL: http://www.jstor.org/stable/24523004.

[MSS15b] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Interlacing families II: Mixed characteristic polynomials and the Kadison—Singer problem". In: *Annals of Mathematics* 182.1 (2015), pages 327–350. URL: http://www.jstor.org/stable/24523005.

[MSS21] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Interlacing families III: Sharper restricted invertibility estimates". In: *Israel Journal of Mathematics* (2021), pages 1–28.

[Nil91] A. Nilli. "On the Second Eigenvalue of a Graph". In: *Discrete Math.* 91.2 (1991), pages 207–210. DOI: 10.1016/0012-365X(91)90112-F.

[Pak10] I. Pak. *Lectures on Discrete and Polyhedral Geometry*. 2010. URL: https://www.math.ucla.edu/~pak/book.htm.

[Spi18b] D. Spielman. *Bipartite Ramanujan Graphs*. 2018. URL: http://www.cs.yale.edu/homes/spielman/561/lect25-18.pdf.

[Spi19] D. Spielman. *Spectral and Algebraic Graph Theory*. 2019. URL: http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf.

4. The Noncommutative Grothendieck Problem

Date: 14 March 2022 **Author:** Ethan N. Epperly

Grothendieck's inequality [Gro53] is a foundational result that has important uses in Banach space theory, C^* algebras, quantum information theory, and combinatorial optimization. A noncommutative analog of the inequality, conjectured by Grothendieck in 1953, was proved a quarter century later by Pisier [Pis78] and has seen significant applications in various areas of mathematics. This note will survey these inequalities, adopting a contemporary perspective centered around semidefinite programming.

4.1 Grothendieck's inequality

Let $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ be a matrix. Consider the problem of computing the $\ell_\infty \to \ell_1$ operator norm of $\boldsymbol{B}$, which can be phrased as the binary integer optimization problem

$$\|\boldsymbol{B}\|_{\ell_\infty\to\ell_1} = \max_{\substack{\boldsymbol{x}\in\{\pm 1\}^n \\ \boldsymbol{y}\in\{\pm 1\}^m}} \langle \boldsymbol{y},\ \boldsymbol{B}\boldsymbol{x}\rangle. \tag{4.1}$$

The equality (4.1) is a consequence of duality between the ℓ_∞ and ℓ_1 norms and the Bauer maximum principle:

$$\begin{aligned}\|\boldsymbol{z}\|_{\ell_1} &= \max_{\|\boldsymbol{w}\|_{\ell_\infty}\le 1} \langle \boldsymbol{w}, \boldsymbol{z}\rangle \\ &= \max_{\boldsymbol{w}\in\{\pm 1\}^n} \langle \boldsymbol{w}, \boldsymbol{z}\rangle.\end{aligned}$$

This is a combinatorial optimization problem over the 2^{m+n} assignments for $\boldsymbol{x}$ and $\boldsymbol{y}$, and it is known to be NP-hard [AN06, Prop. 2.2].

A classical technique to find approximate solutions to combinatorial optimization problems is to relax the problem to a *semidefinite program*, which can be tractably solved. Observe that the objective function of (4.1) can be written as

$$\langle \boldsymbol{B}\boldsymbol{x},\ \boldsymbol{y}\rangle = \langle \boldsymbol{y}\boldsymbol{x}^*,\ \boldsymbol{B}\rangle.$$

This formulation suggests we reframe as an optimization problem over the matrix $\boldsymbol{Z}_{12} = \boldsymbol{y}\boldsymbol{x}^*$. In fact, to capture the constraint that the entries of $\boldsymbol{x}$ and $\boldsymbol{y}$ are ± 1, we shall consider the larger matrix

$$\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix}^* = \begin{bmatrix}\boldsymbol{y}\boldsymbol{y}^* & \boldsymbol{y}\boldsymbol{x}^*\\ \boldsymbol{x}\boldsymbol{y}^* & \boldsymbol{x}\boldsymbol{x}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{Z}_{11} & \boldsymbol{Z}_{12}\\ \boldsymbol{Z}_{21} & \boldsymbol{Z}_{22}\end{bmatrix}. \tag{4.2}$$

One can verify that a matrix $\boldsymbol{Z} \in \mathbb{M}_{m+n}(\mathbb{R})$ has the form in (4.2) if and only if $\boldsymbol{Z}$ has rank one and has all ones on its diagonal. Thus, (4.1) can be *exactly* reformulated as the rank-constrained matrix optimization problem

$$\|\boldsymbol{B}\|_{\ell_\infty\to\ell_1} = \max_{\boldsymbol{Z}\in\mathbb{M}_{m+n}(\mathbb{R})} \langle \boldsymbol{Z}_{12},\ \boldsymbol{B}\rangle \text{ subject to } \operatorname{rank}\boldsymbol{Z} = 1,\ \operatorname{diag}(\boldsymbol{Z}) = \mathbb{1}. \tag{4.3}$$

We have written $\boldsymbol{Z}_{12}$ for the $(1,2)$-block of $\boldsymbol{Z}$, partitioned as in (4.2). As this is a reformulation, we have not made the problem any more tractable by phrasing it this way. To make the problem easier on ourself, we "relax" the rank-one constraint to the looser restriction that $\boldsymbol{Z}$ is positive semidefinite:

$$\mathrm{SDP}(\boldsymbol{B}) := \max_{\boldsymbol{Z}\in\mathbb{M}_{m+n}(\mathbb{R})} \langle \boldsymbol{Z}_{12},\ \boldsymbol{B}\rangle \quad \text{subject to} \quad \boldsymbol{Z} \succeq 0,\ \operatorname{diag}(\boldsymbol{Z}) = \mathbb{1}. \tag{4.4}$$

Since any feasible solution to (4.3) is also feasible for (4.4), we must have $\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1} \le \mathrm{SDP}(\boldsymbol{B})$. The content of Grothendieck's inequality is that $\mathrm{SDP}(\boldsymbol{B})$ is no larger than a modest multiple of $\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1}$. That is, the semidefinite programming relaxation (4.4) provides a *constant-factor approximation* to the Grothendieck problem (4.1).

Theorem 4.1 (Grothendieck's inequality I). There exists a constant $\mathrm{K}_{\mathrm{G}}^{\mathbb{R}}$ such that

$$\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1} \le \mathrm{SDP}(\boldsymbol{B}) \le \mathrm{K}_{\mathrm{G}}^{\mathbb{R}} \cdot \|\boldsymbol{B}\|_{\ell_\infty \to \ell_1}$$

for all real matrices $\boldsymbol{B}$ of any dimension.

We indulge in one final refomulation. Note that a matrix $\boldsymbol{Z}$ is feasible for (4.4) if and only if it is the Gram matrix of unit vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m, \boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ in a Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$. That is,

$$\boldsymbol{Z} = \begin{bmatrix} \langle \boldsymbol{u}_1, \boldsymbol{u}_1 \rangle & \cdots & \langle \boldsymbol{u}_1, \boldsymbol{u}_m \rangle & \langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle & \cdots & \langle \boldsymbol{u}_1, \boldsymbol{v}_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{u}_m, \boldsymbol{u}_1 \rangle & \cdots & \langle \boldsymbol{u}_m, \boldsymbol{u}_m \rangle & \langle \boldsymbol{u}_m, \boldsymbol{v}_1 \rangle & \cdots & \langle \boldsymbol{u}_m, \boldsymbol{v}_n \rangle \\ \langle \boldsymbol{v}_1, \boldsymbol{u}_1 \rangle & \cdots & \langle \boldsymbol{v}_1, \boldsymbol{u}_m \rangle & \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \cdots & \langle \boldsymbol{v}_1, \boldsymbol{v}_n \rangle \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{v}_n, \boldsymbol{u}_1 \rangle & \cdots & \langle \boldsymbol{v}_n, \boldsymbol{u}_m \rangle & \langle \boldsymbol{v}_n, \boldsymbol{v}_1 \rangle & \cdots & \langle \boldsymbol{v}_n, \boldsymbol{v}_n \rangle \end{bmatrix}.$$

With this observation, another equivalent statement of Grothendieck's inequality becomes evident. In premonition of what will come, we shall also generalize to the field $\mathbb{F}$ of either real or complex numbers.

Theorem 4.2 (Grothendieck's inequality II). There exists a constant $\mathrm{K}_{\mathrm{G}}^{\mathbb{F}}$ such that

$$\left| \sum_{i=1}^m \sum_{j=1}^n b_{ij} \langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle \right| \le \mathrm{K}_{\mathrm{G}}^{\mathbb{F}} \cdot \|\boldsymbol{B}\|_{\ell_\infty \to \ell_1},$$

where $\mathbb{T}$ denotes the elements of $\mathbb{F}$ with modulus one , $\boldsymbol{B}$ is an matrix over $\mathbb{F}$ of any size $m \times n$, and $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m, \boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ are unit vectors in a Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$.

For $\mathbb{F} = \mathbb{C}$, we have the following analog of (4.1):

$$\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1} = \max_{\substack{\boldsymbol{x} \in \mathbb{T}^n \\ \boldsymbol{y} \in \mathbb{T}^n}} |\langle \boldsymbol{B}\boldsymbol{x}, \boldsymbol{y} \rangle|,$$

where $\mathbb{T}$ are the complex numbers of unit modulus.

For algorithmic applications, we are not only just interested in the *optimal value* of the Grothendieck problem (4.1) but also in the *optimal solutions* $\boldsymbol{x} \in \{\pm 1\}^m$ and $\boldsymbol{y} \in \{\pm 1\}^n$. Fortunately, there are efficient ways for "rounding" the output of the semidefinite programming relaxation (4.4) to approximate optimal solutions. The following is a result of Alon and Naor [AN06].

Theorem 4.3 (Grothendieck efficient rounding). There exists a polynomial-time randomized algorithm to convert an optimal solution to the semidefinite programming problem (4.4) to random vectors $\boldsymbol{x} \in \{\pm 1\}^n$ and $\boldsymbol{y} \in \{\pm 1\}^m$ such that

$$\mathbb{E}\left[\sum_{i=1}^m \sum_{j=1}^n b_{ij} x_i y_j \right] \le \mathrm{K}_{\mathrm{G,rnd}}^{\mathbb{R}} \cdot \|\boldsymbol{B}\|_{\ell_\infty \to \ell_1},$$

where $\mathrm{K}_{\mathrm{G,rnd}}^{\mathbb{R}}$ is an absolute constant.

4.2 The noncommutative Grothendieck problem

We now turn our attention to a *noncommutative* analog of the Grothendieck problem (4.1), which we will analyze using a noncommutative analog of Grothendieck's inequality (Theorem 4.2). We shall restrict ourselves to the complex field $\mathbb{F} = \mathbb{C}$ and to the "square case" $m = n$ to streamline the presentation.

To motivate the use of the term "noncommutative", observe that we can further reformulate the Grothendieck problem (4.1) in the field $\mathbb{C}$ as an optimization problem over diagonal matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1} = \max_{\boldsymbol{X},\boldsymbol{Y} \in \{\mathrm{diag}(\boldsymbol{z}) : \boldsymbol{z} \in \mathbb{T}^n\}} |\mathcal{T}(\boldsymbol{X},\boldsymbol{Y})|$$

where

$$\mathcal{T}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^m \sum_{j=1}^m b_{ij} x_{ii} y_{jj}.$$

Thus the Grothendieck problem can be equivalently formulated as maximizing a bilinear form over two diagonal (and thus commuting) *unitary* matrices.

The noncommutative Grothendieck problem relaxes this diagonality assumption to allow for arbitrary unitary matrices. Let $\mathcal{T} : \mathbb{M}_n \times \mathbb{M}_n \to \mathbb{C}$ be a sesquilinear form, conjugate linear in its first argument. The noncommutative Grothendieck problem for $\mathcal{T}$ is to compute

$$\mathrm{OPT}(\mathcal{T}) = \max_{\boldsymbol{X},\boldsymbol{Y} \in \mathbb{U}_n(\mathbb{C})} \mathrm{Re}\, \mathcal{T}(\boldsymbol{X},\boldsymbol{Y}), \tag{4.5}$$

It leads to an equivalent problem to replace the real part in (4.5) by the modulus. We shall prefer the real part as it is $\mathbb{R}$-linear.

where $\mathbb{U}_n(\mathbb{C})$ denotes the $n \times n$ unitary matrices over the field $\mathbb{C}$.

In order to state a noncommutative analog of the Grothendieck inequality (Theorem 4.2), we must introduce some seemingly peculiar notation. Let $\mathbb{M}_n(\mathcal{H})$ denote $n \times n$ matrices taking values in a Hilbert space $\mathcal{H}$, whose inner product we denote $\langle \cdot, \cdot \rangle$. For a matrix $\boldsymbol{C} \in \mathbb{M}_n(\mathcal{H})$, we define the Gram matrices $\boldsymbol{C}\boldsymbol{C}^*, \boldsymbol{C}^*\boldsymbol{C} \in \mathbb{H}_n(\mathbb{C})$ with entries

$$(\boldsymbol{C}\boldsymbol{C}^*)_{ij} = \sum_{k=1}^n \langle \boldsymbol{c}_{ik}, \boldsymbol{c}_{jk} \rangle, \quad (\boldsymbol{C}^*\boldsymbol{C})_{ij} = \sum_{k=1}^n \langle \boldsymbol{c}_{ki}, \boldsymbol{c}_{kj} \rangle.$$

The set of matrices $\boldsymbol{U} \in \mathbb{M}_n(\mathcal{H})$ such that $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^* = \mathbf{I}$ are called *unitary* and are denoted $\mathbb{U}_n(\mathcal{H})$. For a sesquilinear form $\mathcal{T}$ on $\mathbb{M}_n(\mathbb{C})$ expressed in coordinates as

$$\mathcal{T}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i,j,k,\ell=1}^n t_{ijk\ell} \overline{x_{ij}} y_{k\ell},$$

we define $\mathcal{T}(\boldsymbol{U},\boldsymbol{V})$ for $\boldsymbol{U},\boldsymbol{V} \in \mathbb{M}_n(\mathcal{H})$ as

$$\mathcal{T}(\boldsymbol{U},\boldsymbol{V}) = \sum_{i,j,k,\ell=1}^n t_{ijk\ell} \langle \boldsymbol{u}_{ij}, \boldsymbol{v}_{k\ell} \rangle.$$

We now state a noncommutative Grothendieck inequality, analogous to our second formulation of the commutative Grothendieck inequality (Theorem 4.2).

Theorem 4.4 (Noncommutative Grothendieck Inequality I). Let $\mathcal{T}$ be a sesquilinear form on $\mathbb{M}_n(\mathbb{C})$. There exists an constant $\mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}}$ such that for any Hilbert space $\mathcal{H}$ and $\boldsymbol{U},\boldsymbol{V} \in \mathbb{U}_n(\mathcal{H})$,

$$\mathrm{Re}\, \mathcal{T}(\boldsymbol{U},\boldsymbol{V}) \le \mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}} \cdot \mathrm{OPT}(\mathcal{T}).$$

The optimal value of the constant is $\mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}} = 2$.

A result of this form was conjectured by Grothendieck and proved by Pisier [Pis78]. Haagerup obtained the sharp constant using a more streamlined proof [Haa85], and Haagerup and Itoh established the optimality of the constant [HI95].

4.2.1 Semidefinite programming formulation

By running the our argument from the commutative Grothendieck inequality in reverse, we can reformulate the noncommutative Grothendieck inequality (Theorem 4.4) as an approximation guarantee for the noncommutative Grothendieck problem (4.5).

Consider a $2n^2 \times 2n^2$ Gram matrix $\boldsymbol{G}$ whose entries are populated with the pairwise inner products of all $2n^2$ vector entries of $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{M}_n(\mathscr{H})$ for some Hilbert space $\mathscr{H}$. The unitarity constraints on $\boldsymbol{U}$ and $\boldsymbol{V}$ are all *linear equalities* in the pairwise inner products of the entries of $\boldsymbol{U}$ and $\boldsymbol{V}$, which are the entries of the Gram matrix $\boldsymbol{G}$. Further, the objective function is $\mathbb{R}$-linear in the pairwise inner products of the entries of $\boldsymbol{U}$ and $\boldsymbol{V}$. Therefore, if one picks a Hilbert space $\mathscr{H}$ of large enough dimension ($\dim \mathscr{H} \geq 2n^2$ suffices), one can optimize over $\boldsymbol{U}$ and $\boldsymbol{V}$ on the left-hand side of the noncommutative Grothendieck inequality (Theorem 4.4) to obtain a quantity

$$\mathrm{SDP}(\mathcal{T}) := \max_{\boldsymbol{U},\boldsymbol{V} \in \mathbb{U}_n(\mathbb{C}^{2n^2})} \mathrm{Re}\, \mathcal{T}(\boldsymbol{U}, \boldsymbol{V}) \tag{4.6}$$

The dimension $2n^2$ suffices since the dimension of the Hilbert space $\mathscr{H}$ must only be chosen to be large enough that all possible $2n^2 \times 2n^2$ positive semidefinite matrices can be represented as the Gram matrix of $2n^2$ vectors in $\mathscr{H}$. Cholesky factorization, for example, shows that $\mathscr{H} = \mathbb{C}^{2n^2}$ satisfies this property.

which can be solved (to a specified accuracy) in polynomial time by solving a linear semidefinite program with $O(n^4)$ entries in the matrix variable and $O(n^2)$ linear equality constraints. This optimization problem is indeed a relaxation of the noncommutative Grothendieck problem (4.5) because any $\boldsymbol{X}, \boldsymbol{Y}$ feasible for (4.5) can be converted to a pair $\boldsymbol{U}, \boldsymbol{V}$ feasible for (4.6) by setting $\boldsymbol{u}_{ij} = x_{ij}\boldsymbol{\delta}_1$ and $\boldsymbol{v}_{ij} = y_{ij}\boldsymbol{\delta}_1$ for all $i, j = 1, \ldots, n$.

The noncommutative Grothendieck inequality can be interpreted as an approximation guarantee for this semidefinite programming relaxation:

Theorem 4.5 (Noncommutative Grothendieck Inequality II). Let $\mathcal{T}$ be a sesquilinear form on $\mathbb{M}_n(\mathbb{C})$. With the optimal constant $\mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}} = 2$,

$$\mathrm{OPT}(\mathcal{T}) \leq \mathrm{SDP}(\mathcal{T}) \leq \mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}} \cdot \mathrm{OPT}(\mathcal{T}).$$

4.2.2 Efficient rounding algorithm

As with the commutative Grothendieck problem, we are often interested in obtaining an nearly optimal solution to the noncommutative Grothendieck problem (4.5) from the semidefinite programming relaxation (4.6). The following (randomized) algorithm achieves precisely this goal.

If we solve the relaxation (4.6) using the approach outlined, we obtain a Gram matrix $\boldsymbol{G}$ for the entries of a solution $\boldsymbol{U}$ and $\boldsymbol{V}$. From this Gram matrix, we can find $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{U}_n(\mathbb{C}^{2n^2})$ that solve (4.6) in $O(n^6)$ time using Cholesky factorization.

Algorithm 4.1 Efficient rounding for noncommutative Grothendieck problem

Input: (Approximate) solutions $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{U}_n(\mathbb{C}^d)$ to the SDP relaxation (4.6) of the noncommutative Grothendieck problem (4.5)
Output: Approximate solutions $\boldsymbol{X}, \boldsymbol{V} \in \mathbb{U}_n(\mathbb{C})$ to the noncommutative Grothendieck problem (4.5)

1 Choose $\boldsymbol{z} \sim \textsc{uniform}\{1, -1, \mathrm{i}, -\mathrm{i}\}^d$ and t according to a hyperbolic secant distribution:
$$\mathbb{P}\{t \in [a, b]\}\} = \int_a^b \frac{1}{2} \operatorname{sech}\left(\frac{x}{2}\right) \mathrm{d}x;$$
2 Set $\boldsymbol{U}_{\boldsymbol{z}} \leftarrow \langle \boldsymbol{z}, \boldsymbol{U} \rangle \in \mathbb{M}_n(\mathbb{C})$ and $\boldsymbol{V}_{\boldsymbol{z}} := \langle \boldsymbol{z}, \boldsymbol{V} \rangle \in \mathbb{M}_n(\mathbb{C})$, where the inner product is taken entrywise.
3 Compute polar decompositions $\boldsymbol{U}_{\boldsymbol{z}} = \boldsymbol{\Phi}_{\boldsymbol{z}}|\boldsymbol{U}_{\boldsymbol{z}}|$, $\boldsymbol{V}_{\boldsymbol{z}} = \boldsymbol{\Psi}_{\boldsymbol{z}}|\boldsymbol{V}_{\boldsymbol{z}}|$
4 **return** $\boldsymbol{X} \leftarrow \boldsymbol{\Phi}_{\boldsymbol{z}}|\boldsymbol{U}_{\boldsymbol{z}}/\sqrt{2}|^{\mathrm{i}t}$, $\boldsymbol{Y} \leftarrow \boldsymbol{\Psi}_{\boldsymbol{z}}|\boldsymbol{V}_{\boldsymbol{z}}/\sqrt{2}|^{\mathrm{i}t}$

Amazingly, this somewhat peculiar rounding algorithm produces $\boldsymbol{X}$ and $\boldsymbol{Y}$ which are, in expectation, nearly (worst-case) optimal. The preceding algorithm and the following result are due to Naor, Regev, and Vidick [NRV14, Thm. 1.4].

Theorem 4.6 (Noncommutative Grothendieck efficient rounding). Suppose $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{U}_n(\mathbb{C}^d)$ are ε-optimal solutions to (4.6):

$$\operatorname{Re} \mathcal{T}(\boldsymbol{U}, \boldsymbol{V}) \geq (1 - \varepsilon) \cdot \mathrm{SDP}(\mathcal{T}).$$

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be the outputs of Algorithm 4.1 applied to these inputs. Then

$$\mathbb{E} \operatorname{Re} \mathcal{T}(\boldsymbol{X}, \boldsymbol{Y}) \geq \left(\frac{1}{2} - \varepsilon\right) \cdot \mathrm{OPT}(\mathcal{T}).$$

Both (equivalent) versions of the noncommutative Grothendieck inequality follow directly from this rounding procedure (and the probabilistic method).

4.3 Noncommutative Grothendieck efficient rounding: Proof

We shall now work our way up to a proof. Throughout this section, we shall use the notation from the rounding procedure (Algorithm 4.1). We begin with two technical lemmas which we shall need for the proof proper.

4.3.1 Supporting lemmas

Our first lemma is an expectation bound for the "projected" matrices $\boldsymbol{U}_z$ and $\boldsymbol{V}_z$ from the rounding procedure.

Lemma 4.7 (Projection). With the prevailing notation, we have the bounds

$$\mathbb{E}\left[\boldsymbol{U}_z \boldsymbol{U}_z^*\right], \mathbb{E}\left[\boldsymbol{U}_z^* \boldsymbol{U}_z\right], \mathbb{E}\left[\boldsymbol{V}_z \boldsymbol{V}_z^*\right], \mathbb{E}\left[\boldsymbol{V}_z^* \boldsymbol{V}_z\right] \preceq 2\mathbf{I}.$$

The proof is a short and rather pedestrian calculation, which we omit. We refer the interested reader to [NRV14, Lem. 2.2].

For our second lemma, it will be helpful to "upgrade" a pair of vector-valued matrices $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{M}_n(\mathbb{C}^d)$ which are *subunitary* in the sense that

$$\|\boldsymbol{B}^* \boldsymbol{B}\|, \|\boldsymbol{B}\boldsymbol{B}^*\|, \|\boldsymbol{C}^* \boldsymbol{C}\|, \|\boldsymbol{C}\boldsymbol{C}^*\| \leq 1 \tag{4.7}$$

to full unitary vector-valued matrices $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{U}_n(\mathbb{C}^{\tilde{d}})$, possibly of a larger dimension $\tilde{d} \geq d$, while preserving the sesquilinear form $\mathcal{T}$. The next lemma shows this is possible.

Lemma 4.8 (Subunitary to unitary). Suppose $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{M}_n(\mathbb{C}^d)$ satisfy the subunitary bound (4.7). Then there exist $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{U}_n(\mathbb{C}^{d+2n^2})$ such that

$$\mathcal{T}(\boldsymbol{R}, \boldsymbol{S}) = \mathcal{T}(\boldsymbol{B}, \boldsymbol{C}).$$

Proof. Define residuals $\boldsymbol{D} = \mathbf{I} - \boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{E} = \mathbf{I} - \boldsymbol{B}^* \boldsymbol{B}$. Define $\sigma := \operatorname{tr} \boldsymbol{D} = \operatorname{tr} \boldsymbol{E}$ and consider spectral decompositions

$$\boldsymbol{D} = \sum\nolimits_{i=1}^{n} \lambda_i \boldsymbol{u}^{(i)} \boldsymbol{u}^{(i)*} \quad \text{and} \quad \boldsymbol{E} = \sum\nolimits_{i=1}^{n} \mu_i \boldsymbol{v}^{(i)} \boldsymbol{v}^{(i)*}.$$

Define $\boldsymbol{R} \in \mathbb{U}_n(\mathbb{C}^{d+2n^2})$ with (i, j)th entry

$$\boldsymbol{r}_{ij} = \boldsymbol{b}_{ij} \oplus \left(\sqrt{\frac{\lambda_k \mu_\ell}{\sigma}} u_i^{(k)} v_j^{(\ell)*} : k, \ell = 1, \dots, n \right) \oplus \mathbf{0}, \tag{4.8}$$

where $\oplus$ denotes vertical concatenation of column vectors. A short calculation confirms that $\boldsymbol{R}$ is unitary. Define $\boldsymbol{S}$ analogously, swapping the last two vectors in (4.8). It is straightforward to verify that $\mathcal{T}(\boldsymbol{R}, \boldsymbol{S}) = \mathcal{T}(\boldsymbol{B}, \boldsymbol{C})$. ∎

4.3.2 Proof of Theorem 4.6

As a final observation before the proof, note that we can identify the sesquilinear form $\mathcal{T}$ on $\mathbb{C}^n$ with a linear operator on $\mathbb{C}^n \otimes \mathbb{C}^n$ as both spaces are equidimensional and thus isomorphic. Under this identification, evaluation of $\mathcal{T}$ can be computed as an inner product

$$\mathcal{T}(\boldsymbol{X}, \boldsymbol{Y}) = \langle \boldsymbol{X} \otimes \overline{\boldsymbol{Y}},\ \mathcal{T} \rangle \quad \text{for } \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{M}_n(\mathbb{C}), \tag{4.9}$$

where $\overline{\boldsymbol{Y}}$ denotes the entrywise complex conjugate of $\boldsymbol{Y}$. With the preliminaries taken care of, we proceed with the proof in earnest.

Proof of Theorem 4.6. Let $\mathbb{E}_{\boldsymbol{z}}$, $\mathbb{E}_t$, and $\mathbb{E}_{\boldsymbol{z},t}$ denote expectations over the randomness in $\boldsymbol{z}$, t, or both $\boldsymbol{z}$ and t, respectively. Since $\mathcal{T}$ is sesquilinear and $\boldsymbol{z}$ is isotropic in the sense that $\mathbb{E}\, \boldsymbol{z}\boldsymbol{z}^* = \mathbf{I}$, we compute

$$\mathbb{E}_{\boldsymbol{z}}\, \mathcal{T}(\boldsymbol{U}_{\boldsymbol{z}}, \boldsymbol{V}_{\boldsymbol{z}}) = \mathcal{T}(\boldsymbol{U}, \boldsymbol{V}). \tag{4.10}$$

In anticipation of applying (4.9) to compute $\mathbb{E}_{\boldsymbol{z},t}\, \mathcal{T}(\boldsymbol{X}, \boldsymbol{Y})$, we begin by computing $\mathbb{E}_t[\overline{\boldsymbol{X}} \otimes \boldsymbol{Y}]$. Recall that $\boldsymbol{X}$ and $\boldsymbol{Y}$ were defined as

$$\boldsymbol{X} = \boldsymbol{\Phi}_{\boldsymbol{z}} \left(|\boldsymbol{U}_{\boldsymbol{z}}| / \sqrt{2} \right)^{\mathrm{i}t}, \quad \boldsymbol{Y} = \boldsymbol{\Psi}_{\boldsymbol{z}} \left(|\boldsymbol{V}_{\boldsymbol{z}}| / \sqrt{2} \right)^{\mathrm{i}t}$$

where $\boldsymbol{U}_{\boldsymbol{z}} = \boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|$ and $\boldsymbol{V}_{\boldsymbol{z}} = \boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|$ are polar decompositions. The log-secant distribution has the property that

$$\mathbb{E}_t\, a^{\mathrm{i}t} = 2a - \mathbb{E}_t\, a^{2+\mathrm{i}t}. \tag{4.11}$$

Apply this fact to calculate

$$\begin{aligned} \mathbb{E}_t \left[\boldsymbol{X} \otimes \overline{\boldsymbol{Y}} \right] &= (\boldsymbol{\Phi}_{\boldsymbol{z}} \otimes \overline{\boldsymbol{\Psi}}_{\boldsymbol{z}})\, \mathbb{E}_t \left(\tfrac{1}{2} |\boldsymbol{U}_{\boldsymbol{z}}| \otimes |\boldsymbol{V}_{\boldsymbol{z}}| \right)^{\mathrm{i}t} \\ &= (\boldsymbol{\Phi}_{\boldsymbol{z}} \otimes \overline{\boldsymbol{\Psi}}_{\boldsymbol{z}}) \left[|\boldsymbol{U}_{\boldsymbol{z}}| \otimes |\boldsymbol{V}_{\boldsymbol{z}}| - \mathbb{E}_t \left(\tfrac{1}{2} |\boldsymbol{U}_{\boldsymbol{z}}| \otimes |\boldsymbol{V}_{\boldsymbol{z}}| \right)^{2+\mathrm{i}t} \right] \\ &= \boldsymbol{U}_{\boldsymbol{z}} \otimes \overline{\boldsymbol{V}}_{\boldsymbol{z}} - \mathbb{E}_t \left[\frac{1}{2^{2+\mathrm{i}t}} \cdot \boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t} \otimes \overline{\boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}} \right]. \end{aligned} \tag{4.12}$$

Combining this result with (4.9) and (4.10), we conclude

$$\mathbb{E}_{\boldsymbol{z},t}\, \mathcal{T}(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{T}(\boldsymbol{U}, \boldsymbol{V}) - \mathbb{E}_{\boldsymbol{z},t}\, \mathcal{T}(\boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}).$$

By hypothesis, $\operatorname{Re} \mathcal{T}(\boldsymbol{U}, \boldsymbol{V}) \geq (1 - \varepsilon) \cdot \mathrm{SDP}(\mathcal{T})$, so it will suffice to show

$$\operatorname{Re} \mathbb{E}_{\boldsymbol{z}}\, \mathcal{T}(\boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}) \leq 2\, \mathrm{OPT}(\mathcal{T}) \quad \text{for any } t \in \mathbb{R}. \tag{4.13}$$

We shall show precisely this. For the remainder of this proof, fix $t \in \mathbb{R}$.

To show (4.13), we "derandomize" the expectation over $\boldsymbol{z}$ by passing from random scalar-valued matrices $\boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t} \in \mathbb{M}_n(\mathbb{C})$ to deterministic vector-valued matrices in $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{M}_n(\mathbb{C}^{\{1,-1,\mathrm{i},-\mathrm{i}\}^d})$, which we define as

$\mathbb{C}^{\{1,-1,\mathrm{i},-\mathrm{i}\}^d}$ denotes the Hilbert space of functions

$$f : \{1, -1, \mathrm{i}, -\mathrm{i}\}^d \to \mathbb{C}.$$

$$\boldsymbol{B}(\boldsymbol{z}) := \frac{1}{2^d} \boldsymbol{\Phi}_{\boldsymbol{z}} |\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t} \quad \text{and} \quad \boldsymbol{C}(\boldsymbol{z}) := \frac{1}{2^d} \boldsymbol{\Psi}_{\boldsymbol{z}} |\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}.$$

As promised, this allows us to remove the explicit randomness in (4.13) since

$$\begin{aligned}\mathcal{T}(\boldsymbol{B},\boldsymbol{C}) &= \frac{1}{4^d}\sum\nolimits_{\boldsymbol{z}\in\{1,-1,\mathrm{i},-\mathrm{i}\}^d} \mathcal{T}(\boldsymbol{\Phi}_{\boldsymbol{z}}|\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}}|\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}) \\ &= \mathbb{E}_{\boldsymbol{z}}\,\mathcal{T}(\boldsymbol{\Phi}_{\boldsymbol{z}}|\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}}|\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}).\end{aligned}$$

Further, $\boldsymbol{B}\boldsymbol{B}^* = \mathbb{E}_{\boldsymbol{z}}(\boldsymbol{U}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}}^*)^2$ and $\boldsymbol{B}^*\boldsymbol{B} = \mathbb{E}_{\boldsymbol{z}}(\boldsymbol{U}_{\boldsymbol{z}}^*\boldsymbol{U}_{\boldsymbol{z}})^2$, with analogous statements holding for $\boldsymbol{C}$. By Lemma 4.7, we have

$$\boldsymbol{B}\boldsymbol{B}^*,\ \boldsymbol{B}^*\boldsymbol{B},\ \boldsymbol{C}\boldsymbol{C}^*,\ \boldsymbol{C}^*\boldsymbol{C} \preceq 2\mathbf{I}.$$

Thus, $\boldsymbol{B}/\sqrt{2}$ and $\boldsymbol{C}/\sqrt{2}$ are subunitary in the sense of Lemma 4.8, so there exists $\boldsymbol{R},\boldsymbol{S} \in \mathbb{U}_n(\mathbb{C}^{\tilde{d}})$ for some dimension $\tilde{d}$ such that

$$2\,\mathcal{T}(\boldsymbol{R},\boldsymbol{S}) = \mathcal{T}(\boldsymbol{B},\boldsymbol{C}) = \mathbb{E}_{\boldsymbol{z}}\,\mathcal{T}(\boldsymbol{\Phi}_{\boldsymbol{z}}|\boldsymbol{U}_{\boldsymbol{z}}|^{2+\mathrm{i}t}, \boldsymbol{\Psi}_{\boldsymbol{z}}|\boldsymbol{V}_{\boldsymbol{z}}|^{2-\mathrm{i}t}).$$

But since $\boldsymbol{R}$ and $\boldsymbol{S}$ are unitary, $\operatorname{Re}\mathcal{T}(\boldsymbol{R},\boldsymbol{S}) \le \mathrm{SDP}(\mathcal{T})$. Plugging in and rearranging yields (4.13), completing the proof. ∎

4.3.3 Intuition for rounding procedure

The proof sheds some light on why the rounding procedure (Algorithm 4.1) works. As shown in (4.10), $\boldsymbol{U}_{\boldsymbol{z}},\boldsymbol{V}_{\boldsymbol{z}} \in \mathbb{U}_n(\mathbb{C})$ have the same $\mathcal{T}$-value in expectation as $\boldsymbol{U},\boldsymbol{V} \in \mathbb{U}_n(\mathbb{C}^d)$. Were $\boldsymbol{U}_{\boldsymbol{z}},\boldsymbol{V}_{\boldsymbol{z}}$ unitary, they would be excellent approximate solutions to the noncommutative Grothendieck problem (4.5).

We are thus interested in adding additional randomness to obtain random unitary matrices which agree with $\boldsymbol{U}_{\boldsymbol{z}}$ and $\boldsymbol{V}_{\boldsymbol{z}}$ on average, up to a small error. The forms $\boldsymbol{X} = \boldsymbol{\Phi}_{\boldsymbol{z}}(|\boldsymbol{U}_{\boldsymbol{z}}|/\sqrt{2})^{\mathrm{i}t}$ and $\boldsymbol{Y} = \boldsymbol{\Psi}_{\boldsymbol{z}}(|\boldsymbol{V}_{\boldsymbol{z}}|/\sqrt{2})^{\mathrm{i}t}$ with t having the log-secant distribution are specially engineered to achieve this goal. Specifically, the property (4.11) of the log-secant distribution shows that, on average, the *unitary* matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ produced by the rounding algorithm are close to $\boldsymbol{U}_{\boldsymbol{z}}$ and $\boldsymbol{V}_{\boldsymbol{z}}$ in the sense (4.12).

4.4 Application: Robust PCA

We conclude by presenting an application of the noncommutative Grothendieck inequality to robust principal component analysis (PCA). Our presentation will be informal and high-level, seeking only to provide a colorful illustration of this machinery to a "practical" problem.

4.4.1 ℓ_1 PCA

Consider a data set encoded in the columns of a matrix $\boldsymbol{B} \in \mathbb{C}^{m\times n}$. The PCA problem is to find a k-dimensional subspace capturing most of the "energy" or "variance" in the data matrix $\boldsymbol{B}$. In classical PCA, this can be formulated as an optimization problem

$$\text{maximize} \quad \|\boldsymbol{W}^*\boldsymbol{B}\|_{\mathrm{F}} \quad \text{subject to} \quad \boldsymbol{W} \in \mathbb{U}^{m\times k}(\mathbb{C}), \tag{4.14}$$

where $\mathbb{U}^{m\times k}(\mathbb{C})$ consists of the $m\times k$ matrices with orthonormal columns. An optimal solution to this problem is readily determined by choosing $\boldsymbol{W}$ to consist of the k dominant left singular vectors of $\boldsymbol{B}$.

To make this procedure more robust to outliers, we can replace the Frobenius norm with a different norm in (4.14) For instance, Kwak [Kwa08] proposes using the following "ℓ_1 PCA" problem

$$\text{maximize} \quad \|\boldsymbol{W}^*\boldsymbol{B}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{W} \in \mathbb{U}^{m\times k}(\mathbb{C}), \tag{4.15}$$

where $\|\cdot\|_{\ell_1}$ is the entrywise ℓ_1 norm

$$\|\boldsymbol{C}\|_{\ell_1} := \sum_{i=1}^{k} \sum_{j=1}^{n} |z_{ij}| \quad \text{for } \boldsymbol{C} \in \mathbb{C}^{k \times n}.$$

See also [MT11, §2.7].

4.4.2 Reformulation

We now reformulate the ℓ_1 PCA problem as a bilinear optimization problem over suitably defined unitary matrices. First, use the duality of ℓ_1 and ℓ_∞ to write the ℓ_1 norm as a maximization:

$$\text{maximize} \quad \operatorname{Re}\langle \boldsymbol{W}^* \boldsymbol{B}, \boldsymbol{Z} \rangle \quad \text{subject to} \quad \boldsymbol{W} \in \mathbb{U}^{m \times k}(\mathbb{C}),\ \boldsymbol{Z} \in \mathbb{T}^{k \times n}.$$

One can further encode $\boldsymbol{W}$ and $\boldsymbol{Z}$ as unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ with a specific entries set to zero:

$$\begin{aligned} \boldsymbol{U} &= \begin{bmatrix} \boldsymbol{W} & \boldsymbol{0} \end{bmatrix} \in \mathbb{U}_m(\mathbb{C}); \\ \boldsymbol{V} &= \operatorname{diag}\left(z_{ij} : 1 \le i \le k,\ 1 \le j \le n\right) \in \mathbb{U}_{kn}(\mathbb{C}). \end{aligned} \tag{4.16}$$

Thus, one can reformulate the ℓ_1 PCA problem (4.15) as a bilinear optimization problem over unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ by designing the sesquilinear form such that optimal solutions $\boldsymbol{U}$ and $\boldsymbol{V}$ take the forms (4.16); see [NRV14, §5.1] for details. The noncommutative Grothendieck inequality ensures the semidefinite programming relaxation of this problem is a constant factor approximation. We conclude the that ℓ_1 PCA problem has a polynomial-time constant-factor approximation algorithm.

Notes

A survey on various Grothendieck inequalities and their applications is given by Pisier [Pis12].

The optimal (commutative) Grothendieck constants are satisfy bounds

$$1.67696 \le \mathrm{K}_{\mathrm{G}}^{\mathbb{R}} < 1.7822\ldots = \frac{\pi}{2\log(1+\sqrt{2})}, \quad 1.338 < \mathrm{K}_{\mathrm{G}}^{\mathbb{C}} \le 1.4049.$$

Davie established the lower bounds in unpublished work; Krivine showed the upper bound on $\mathrm{K}_{\mathrm{G}}^{\mathbb{R}}$ [Kri78]; Braverman, Makarychev, Makarychev, and Naor showed the slight suboptimality of Krivine's bound [Bra+11]; and Haagerup showed the upper bound on $\mathrm{K}_{\mathrm{G}}^{\mathbb{C}}$ in [Haa85]. Vershynin provides an accessible exposition on Krivine's bound [Ver18, §3.7].

The semidefinite programming interpretation of Grothendieck's inequality was popularized by Alon and Naor [AN06]. Their rounding procedure (Theorem 4.3) is based on Krivine's bound for the Grothendieck constant. Thus, in the notation of Theorem 4.3,

$$\mathrm{K}_{\mathrm{G,rnd}}^{\mathbb{R}} \le \frac{\pi}{2\log(1+\sqrt{2})}.$$

There are many ways to round answers from semidefinite programs to approximate solutions of combinatorial optimization problems; among these, the Goemans–Williamson algorithm for MaxCut is a distinguished example [GW95]. Assuming the Unique Games Conjecture, it is NP-hard to approximate $\|\boldsymbol{B}\|_{\ell_\infty \to \ell_1}$ to within a factor of $\mathrm{K}_{\mathrm{G}}^{\mathbb{R}} - \varepsilon$ for any $\varepsilon > 0$ [RS09]; the semidefinite programming relaxation (4.4) provides an

efficient approximation algorithm with the optimal approximation factor (assuming the Unique Games Conjecture).

There are real and Hermitian analogs of the noncommutative Grothendieck problem and inequality. The rounding procedure analyzed in Theorem 4.6 generalizes to these problems, giving a constructive proof of real and Hermitian noncommutative Grothendieck inequalities with constant $2\sqrt{2}$ [NRV14, §3]. Naor, Regev, and Vidick also provide a different rounding procedure for the real noncommutative inequality [NRV14, §4] that was inspired by the proof of Kaijser [Kai83]. Briët, Regev, and Saket showed [BRS17] that it is NP-hard to approximate the solution to the (complex) noncommutative Grothendieck problem (4.5) within a factor of $\mathrm{K}_{\mathrm{NCG}}^{\mathbb{C}} - \varepsilon$ for any $\varepsilon > 0$; similar to the commutative case, the semidefinite programming relaxation (4.6) provides an efficient algorithm with the optimal approximation factor (assuming $\mathsf{P} \neq \mathsf{NP}$).

Several robust versions of principal component analysis have been proposed, including the ℓ_1 version we discussed as well as a variant where the Frobenius norm in (4.14) is replaced by an entrywise mixed ℓ_1–ℓ_2 norm [Din+06]. McCoy and Tropp [MT11] provide a polynomial-time constant-factor approximation algorithm to the ℓ_1 robust PCA problem (4.15) for $k = 1$ and a polynomial-time $O(\log m)$-factor approximation algorithm for general k that builds on work of So [So09].

Lecture bibliography

[ANo6] N. Alon and A. Naor. "Approximating the cut-norm via Grothendieck's inequality". In: *SIAM J. Comput.* 35.4 (2006), pages 787–803. DOI: 10.1137/S0097539704441629.

[Bra+11] M. Braverman et al. "The Grothendieck Constant Is Strictly Smaller than Krivine's Bound". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. Oct. 2011, pages 453–462. DOI: 10.1109/FOCS.2011.77.

[BRS17] J. Briët, O. Regev, and R. Saket. "Tight Hardness of the Non-Commutative Grothendieck Problem". In: *Theory of Computing* 13.15 (Dec. 2017), pages 1–24. DOI: 10.4086/toc.2017.v013a015.

[Din+06] C. Ding et al. "R_1-PCA: Rotational Invariant L_1-Norm Principal Component Analysis for Robust Subspace Factorization". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Association for Computing Machinery, June 2006, pages 281–288. DOI: 10.1145/1143844.1143880.

[GW95] M. X. Goemans and D. P. Williamson. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". In: *J. Assoc. Comput. Mach.* 42.6 (1995), pages 1115–1145. DOI: 10.1145/227683.227684.

[Gro53] A. Grothendieck. *Résumé de La Théorie Métrique Des Produits Tensoriels Topologiques*. Soc. de Matemática de São Paulo, 1953.

[Haa85] U. Haagerup. "The Grothendieck Inequality for Bilinear Forms on C*-Algebras". In: *Advances in Mathematics* 56.2 (May 1985), pages 93–116. DOI: 10.1016/0001-8708(85)90026-X.

[HI95] U. Haagerup and T. Itoh. "Grothendieck Type Norms For Bilinear Forms On C*-Algebras". In: *Journal of Operator Theory* 34.2 (1995), pages 263–283.

[Kai83] S. Kaijser. "A Simple-Minded Proof of the Pisier-Grothendieck Inequality". In: *Banach Spaces, Harmonic Analysis, and Probability Theory*. Springer, 1983, pages 33–55.

[Kri78] J.-L. Krivine. "Constantes de Grothendieck et Fonctions de Type Positif Sur Les Spheres". In: *Séminaire d'Analyse fonctionnelle (dit" Maurey-Schwartz")* (1978), pages 1–17.

[Kwa08] N. Kwak. "Principal Component Analysis Based on L1-Norm Maximization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.9 (Sept. 2008), pages 1672–1680. DOI: 10.1109/TPAMI.2008.114.

[MT11] M. McCoy and J. A. Tropp. "Two Proposals for Robust PCA Using Semidefinite Programming". In: *Electronic Journal of Statistics* 5.none (Jan. 2011), pages 1123–1160. DOI: 10.1214/11-EJS636.

[NRV14] A. Naor, O. Regev, and T. Vidick. "Efficient rounding for the noncommutative Grothendieck inequality". In: *Theory Comput.* 10 (2014), pages 257–295. DOI: 10.4086/toc.2014.v010a011.

[Pis78] G. Pisier. "Grothendieck's Theorem for Noncommutative C*-Algebras, with an Appendix on Grothendieck's Constants". In: *Journal of Functional Analysis* 29.3 (Sept. 1978), pages 397–415. DOI: 10.1016/0022-1236(78)90038-1.

[Pis12] G. Pisier. "Grothendieck's Theorem, Past and Present". In: *Bulletin of the American Mathematical Society* 49.2 (Apr. 2012), pages 237–323. DOI: 10.1090/S0273-0979-2011-01348-9.

[RS09] P. Raghavendra and D. Steurer. "Towards Computing the Grothendieck Constant". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2009, pages 525–534. DOI: 10.1137/1.9781611973068.58.

[So09] A. M. So. "Improved Approximation Bound for Quadratic Optimization Problems with Orthogonality Constraints". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2009, pages 1201–1209. DOI: 10.1137/1.9781611973068.130.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

5. Algebraic Riccati Equations

Date: 14 March 2022 **Scribe:** Taylan Kargin

Agenda:
1. Motivation
2. Metric Geometry of PSD Cone
3. Stability and Lyapunov Theory
4. Discrete Algebraic Riccati Equations

This manuscript is a brief introduction to algebraic Riccati equations (AREs) named after Venetian mathematician Jacopo Riccati (1676-1754) for his study of first-order quadratic differential equations [Ric24]. AREs are a certain group of nonlinear matrix equations that naturally arise in the context of control theory [Ber12], filtering and estimation theory [Kal60; KSH00]. Given a complex-valued square matrix $\boldsymbol{A} \in \mathbb{M}_n$, and positive-semidefinite (psd) matrices $\boldsymbol{Q}, \boldsymbol{S} \in \mathbb{H}_n^+$, the quadratic matrix equation

$$\text{Find } \boldsymbol{X} \in \mathbb{H}_n: \quad \boldsymbol{A}^*\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{S}\boldsymbol{X} + \boldsymbol{Q} = \boldsymbol{0},$$

is known as continuous algebraic Riccati equation (CARE) whereas the equation

$$\text{Find } \boldsymbol{X} \in \mathbb{H}_n: \quad \boldsymbol{X} = \boldsymbol{Q} + \boldsymbol{A}^*(\boldsymbol{S} + \boldsymbol{X}^{-1})^{-1}\boldsymbol{A}, \tag{DARE}$$

is known as discrete algebraic Riccati equation (DARE). These equations are classified as continuous and discrete since they naturally arise from studying continuous-time and discrete-time dynamical systems in control and filtering problems. This manuscript omits the continues version and deals with the discrete version, (DARE), only.

The rest of this manuscript is organized as follows. In Section 5.1, a central problem in control theory is introduced to motivate the study of AREs. Section 5.2 follows with a digression on the metric geometry of positive-definite (pd) matrices in order to develop the tools to study AREs later in this manuscript. In Section 5.3, stability of matrices and a closely related linear matrix equation are studied in the context of linear dynamical systems. Section 5.4 concludes the manuscript with an analysis of the solution of (DARE) by applying the tools and techniques developed in the previous sections.

5.1 Motivation

In this section, the Linear-Quadratic Regulator (LQR) problem from optimal control theory will be introduced. LQR problem is central to optimal control theory due to its wide applicability in various real-world settings as well as its rich mathematical aspects including AREs.

Example 5.1 (Discrete-time infinite-horizon LQRs). Consider the following linear dynamical system,

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t, \text{ for } t \geq 0, \\ \boldsymbol{x}_0 &= \boldsymbol{z} \end{aligned} \tag{LIN}$$

where $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}} \subset \mathbb{C}^n$ is the sequence of states, $\boldsymbol{z} \in \mathbb{C}^n$ is the initial state $\{\boldsymbol{u}_t\}_{t\in\mathbb{N}} \subset \mathbb{C}^m$ is the sequence of control inputs, $\boldsymbol{A} \in \mathbb{M}_n$ is the state evolution matrix, and $\boldsymbol{B} \in \mathbb{C}^{n\times m}$ is the input matrix.

At each time step $t \in \mathbb{N}$, a controller observes the current state $\boldsymbol{x}_t$, exerts a control input $\boldsymbol{u}_t$ to the dynamics, and then suffers from an instantaneous cost $c(\boldsymbol{x}_t, \boldsymbol{u}_t)$ as a function of the state and the control input. In many real-world applications, it

is assumed that the instantaneous cost is a quadratic function of the state and the control input, that is, $c(\boldsymbol{x}_t, \boldsymbol{u}_t) \coloneqq \boldsymbol{x}_t^* \boldsymbol{Q} \boldsymbol{x}_t + \boldsymbol{u}_t^* \boldsymbol{R} \boldsymbol{u}_t$ where $\boldsymbol{Q} \in \mathbb{H}_n^{++}$ and $\boldsymbol{R} \in \mathbb{H}_n^{++}$ are given positive-definite matrices. The objective of the controller is to reduce the cumulative cost by deploying a *control policy* to design control inputs based on past state observations. One straightforward approach is to deploy a *linear state-feedback controller* $\boldsymbol{K} \in \mathbb{C}^{m \times n}$ such that the inputs are designed as $\boldsymbol{u}_t = \boldsymbol{K} \boldsymbol{x}_t$ for all $t \in \mathbb{N}$.

In fact, the optimal control of the linear dynamical system (LIN) can be achieved solely by a state-feedback policy as long as the pair $(\boldsymbol{A}, \boldsymbol{B})$ is stabilizable (see the definition 5.11) [Ber12]. The optimal state-feedback controller can be found by minimizing the infinite-horizon cumulative cost among all state-feedback controllers subject to linear dynamical system (LIN) as formulated below.

$$\begin{aligned} \min_{\boldsymbol{K} \in \mathbb{C}^{m \times n}} \; & \left\{ V(\boldsymbol{z}; \boldsymbol{K}) \coloneqq \sum\nolimits_{t=0}^{\infty} \left(\boldsymbol{x}_t^* \boldsymbol{Q} \boldsymbol{x}_t + \boldsymbol{u}_t^* \boldsymbol{R} \boldsymbol{u}_t \right) \right\}, \\ & \boldsymbol{x}_{t+1} = \boldsymbol{A} \boldsymbol{x}_t + \boldsymbol{B} \boldsymbol{u}_t, \text{ for } t \in \mathbb{N}, \\ \text{subject to} \quad & \boldsymbol{u}_t = \boldsymbol{K} \boldsymbol{x}_t, \text{ for } t \in \mathbb{N}, \\ & \boldsymbol{x}_0 = \boldsymbol{z}. \end{aligned} \tag{OPT 1}$$

Here, $\boldsymbol{z} \in \mathbb{C}^n$ is a given initial state and $V(\boldsymbol{z}; \boldsymbol{K})$ is the *value function*. The solution to this problem is known as the infinite-horizon linear-quadratic regulator (LQR). ▪

The value function gives the total cost of controlling the linear dynamical system (LIN) with state-feedback controller $\boldsymbol{K} \in \mathbb{C}^{m \times n}$ starting from an initial state $\boldsymbol{z} \in \mathbb{C}^n$. Note that the value function is not guaranteed to take a finite value for all $\boldsymbol{K} \in \mathbb{C}^{m \times n}$. In fact, a state-feedback controller $\boldsymbol{K} \in \mathbb{C}^{m \times n}$ is called *stabilizing* if $V(\boldsymbol{z}; \boldsymbol{K})$ takes a finite value for any initial state $\boldsymbol{z} \in \mathbb{C}^n$. Assuming that a state-feedback controller $\boldsymbol{K} \in \mathbb{C}^{m \times n}$ is stabilizing, the value function can be expressed recursively as

$$\begin{aligned} V(\boldsymbol{z}; \boldsymbol{K}) &= \sum\nolimits_{t=0}^{\infty} \left\{ \boldsymbol{x}_t^* \boldsymbol{Q} \boldsymbol{x}_t + (\boldsymbol{K} \boldsymbol{x}_t)^* \boldsymbol{R} (\boldsymbol{K} \boldsymbol{x}_t) \right\}, \\ &= \sum\nolimits_{t=0}^{\infty} \boldsymbol{x}_t^* \left(\boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} \right) \boldsymbol{x}_t, \\ &= \boldsymbol{x}_0^* \left(\boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} \right) \boldsymbol{x}_0 + \sum\nolimits_{t=1}^{\infty} \boldsymbol{x}_t^* \left(\boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} \right) \boldsymbol{x}_t, \\ &= \boldsymbol{z}^* \left(\boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} \right) \boldsymbol{z} + V((\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K}) \boldsymbol{z}; \boldsymbol{K}), \end{aligned} \tag{5.1}$$

for any $\boldsymbol{z} \in \mathbb{C}^n$ where (5.1) follows from taking the next step at time $t = 1$ as the new initial state.

Furthermore, it can be argued that the value function is quadratic in the initial state $\boldsymbol{z}$ by noting that the current state at time $t \geq 0$ is a linear function of the initial state as $\boldsymbol{x}_t = (\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K})^t \boldsymbol{z}$ and the instantaneous cost is quadratic in the current state. Thus, one can write $V(\boldsymbol{z}; \boldsymbol{K}) = \boldsymbol{z}^* \boldsymbol{P} \boldsymbol{z}$ where $\boldsymbol{P} \in \mathbb{H}_n^+$ is a psd matrix which depends on the feedback gain matrix, $\boldsymbol{K}$. Inserting the quadratic form into the recursive relation in (5.1), one obtains the following condition

$$\boldsymbol{z}^* \left\{ \boldsymbol{P} - \left(\boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} \right) - (\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K})^* \boldsymbol{P} (\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K}) \right\} \boldsymbol{z} = \boldsymbol{0}, \text{ for all } \boldsymbol{z} \in \mathbb{C}^n,$$

which can be simplified by removing the $\boldsymbol{z}$-dependence as

$$(\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K})^* \boldsymbol{P} (\boldsymbol{A} + \boldsymbol{B} \boldsymbol{K}) - \boldsymbol{P} + \boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R} \boldsymbol{K} = \boldsymbol{0}. \tag{5.2}$$

The equation (5.2) is known as the *Lypaunov equation* with state-feedback controller $\boldsymbol{K}$. This relationship makes it possible to rewrite the optimal control problem (OPT 1)

in the following equivalent form as

$$\begin{aligned} \min_{\boldsymbol{K}\in\mathbb{C}^{m\times n},\,\boldsymbol{P}\in\mathbb{H}_n} \quad & \boldsymbol{z}^*\boldsymbol{P}\boldsymbol{z}, \\ \text{subject to} \quad & (\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K})^*\boldsymbol{P}(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K}) - \boldsymbol{P} + \boldsymbol{Q} + \boldsymbol{K}^*\boldsymbol{R}\boldsymbol{K} = \boldsymbol{0}, \\ & \boldsymbol{P} \succcurlyeq \boldsymbol{0}. \end{aligned} \tag{OPT 2}$$

Noticing that the Lyapunov equation (5.2) is quadratic with respect to $\boldsymbol{K}$ and $\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B} \succ \boldsymbol{0}$, one can rewrite the equation (5.2) by completion of squares as

$$\begin{aligned} \boldsymbol{P} = \boldsymbol{Q} + \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{A} - \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{B}(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{A} \\ + (\boldsymbol{K}-\boldsymbol{L})^*(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})(\boldsymbol{K}-\boldsymbol{L}). \end{aligned} \tag{5.3}$$

where $\boldsymbol{L} := -(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{A}$. The equivalence of (5.2) and (5.3) can be verified directly by expanding the right-hand side of (5.3). This relationship is further elaborated in Lemma 5.19 of Section 5.4. Since $(\boldsymbol{K}-\boldsymbol{L})^*(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})(\boldsymbol{K}-\boldsymbol{L}) \succcurlyeq \boldsymbol{0}$, it can be concluded that a pair $(\boldsymbol{K},\boldsymbol{P}) \in \mathbb{C}^{m\times n}\times\mathbb{H}_n^+$ satisfying the Lyapunov equation (5.2) satisfies the matrix inequality

$$\boldsymbol{P} \succcurlyeq \boldsymbol{Q} + \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{A} - \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{B}(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{A}. \tag{5.4}$$

Conversely, for any $\boldsymbol{P} \in \mathbb{H}_n^+$ satisfying the inequality (5.4), one can find a matrix $\boldsymbol{K} \in \mathbb{C}^{m\times n}$ such that the equation (5.3) holds. Therefore, the optimization problem (OPT 2) is equivalent to

$$\begin{aligned} \min_{\boldsymbol{P}\in\mathbb{H}_n} \quad & \boldsymbol{z}^*\boldsymbol{P}\boldsymbol{z}, \\ \text{subject to} \quad & \boldsymbol{P} \succcurlyeq \boldsymbol{Q} + \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{A} - \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{B}(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{P}\boldsymbol{A}, \\ & \boldsymbol{P} \succcurlyeq \boldsymbol{0}. \end{aligned}$$

It can be argued that an optimal solution $\boldsymbol{P}_\star \succcurlyeq \boldsymbol{0}$ to the problem above is attained when the equality is satisfied, that is,

$$\boldsymbol{P}_\star = \boldsymbol{Q} + \boldsymbol{A}^*\boldsymbol{P}_\star\boldsymbol{A} - \boldsymbol{A}^*\boldsymbol{P}_\star\boldsymbol{B}(\boldsymbol{R}+\boldsymbol{B}^*\boldsymbol{P}_\star\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{P}_\star\boldsymbol{A}$$

which is a discrete algebraic Riccati equation (DARE).

Attention will be devoted to the analysis of DARE and the related Lyapunov equations in the rest of this manuscript. For the sake of generality, the complex-valued case will be considered.

5.2 Metric Geometry of Positive-Definite Cone

In this section, the open convex cone of positive-definite matrices, denoted as $\mathbb{P}_n$, will be studied from a metric geometry perspective. The concept of geometric mean of two positive-definite matrices gives rise to a natural way of assigning a metric distance to $\mathbb{P}_n$. The results developed in this section will be used in Section 5.4 to characterize the convergence of fixed-point iterations of Riccati operator. The proofs of theorems and lemmas introduced in this section are omitted for the sake of brevity. The interested reader is referred to the references [Bhao3],[BH06], and [LW94] for detailed discussion of the subject and the proofs omitted in this manuscript.

Suppose that a *symmetric gauge function* (sgf) $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is given. Denote by $\|\cdot\|_\Phi : \mathbb{M}_n \to \mathbb{R}_+$ unitarily invariant matrix norm associated to sgf Φ. The space of

Hermitian matrices equipped with this norm, $(\mathbb{H}_n, \|\cdot\|_\Phi)$, is a real (complete) normed vector space. The exponential map defined as

$$\begin{aligned}\exp : \mathbb{H}_n &\to \mathbb{P}_n,\\ \boldsymbol{H} &\mapsto \sum_{k=0}^{\infty} \frac{1}{k!}\boldsymbol{H}^k,\end{aligned}$$

is a *diffeomorphism*, that is, a smooth bijection with smooth inverse denoted as $\log : \mathbb{P}_n \to \mathbb{H}_n$. Thus, the open convex cone of pd matrices constitute a smooth manifold. It can be shown that the tangent space of $\mathbb{P}_n$ at a point $\boldsymbol{X} \in \mathbb{P}_n$ is isomorphic to the space of Hermitian matrices, that is, $T_{\boldsymbol{X}}\mathbb{P}_n \cong \mathbb{H}_n$.

A *Riemmannian manifold* is defined as a smooth manifold whose tangent space at every point is equipped with a smooth inner product. *Finsler manifolds* generalize Riemmannian manifolds by equipping tangent spaces at every point with a smooth norm, instead. The following theorem shows that $\mathbb{P}_n$ attains a metric function, called a *Finsler metric*, induced by a symmetric gauge function (sgf) defined on its tangent space.

Theorem 5.2 (Metric induced by an sgf, [BhaO3, p. 216]). Let $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ be a symmetric gauge function (sgf). The nonnegative function defined as

$$\begin{aligned}\delta_\Phi : \mathbb{P}_n \times \mathbb{P}_n &\to \mathbb{R}_+,\\ (\boldsymbol{X},\ \boldsymbol{Y}) &\mapsto \|\log(\boldsymbol{X}^{-1/2}\boldsymbol{Y}\boldsymbol{X}^{-1/2})\|_\Phi,\end{aligned}$$

is a metric on $\mathbb{P}_n$. Equipped with this metric, $(\mathbb{P}_n,\ \delta_\Phi)$ constitutes a Finsler manifold where tangent space at any point is equipped with the norm $\|\cdot\|_\Phi$. Furthermore, δ_Φ is invariant under matrix inversion and congruence transformations, i.e.,

$$\delta_\Phi(\boldsymbol{X}^{-1}, \boldsymbol{Y}^{-1}) = \delta_\Phi(\boldsymbol{A}^*\boldsymbol{X}\boldsymbol{A}, \boldsymbol{A}^*\boldsymbol{Y}\boldsymbol{A}) = \delta_\Phi(\boldsymbol{X}, \boldsymbol{Y}),$$

for $\boldsymbol{X},\ \boldsymbol{Y} \in \mathbb{P}_n$ and $\boldsymbol{A} \in \mathrm{GL}(n)$.

Aside: $\mathrm{GL}(n)$ is the general linear group of invertible square matrices of size $n \in \mathbb{N}$.

Remark 5.3 The special case of Euclidean sgf results in a Riemmannian manifold $(\mathbb{P}_n,\ \delta_2)$ since Euclidean norm admits an inner product. The metric δ_2 induced by Euclidean norm is is called the *Thompson metric*.

The last part of Theorem 5.2 simply tells that inversion operation and group action of $\mathrm{GL}(n)$ onto $\mathbb{P}_n$ are isometries with respect to δ_Φ. This invariance property will play a role in the convergence analysis of fixed-point iterations of the Riccati operator.

The metric δ_Φ is related to geometric mean of matrices in a natural way. The following proposition shows that δ_Φ is in fact the geodesic distance on $\mathbb{P}_n$ and that the geodesics can be parameterized by geometric mean of matrices.

Aside: A geodesic is a shortest-distance curve connecting two points on a smooth manifold.

Proposition 5.4 (Geodesics, [BhaO3, p. 216]). Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{P}_n$ be pd matrices. Define the $t-$weighted geometric mean of two pd matrices as

$$\boldsymbol{X}\sharp_t\boldsymbol{Y} := \boldsymbol{X}^{1/2}(\boldsymbol{X}^{-1/2}\boldsymbol{Y}\boldsymbol{X}^{-1/2})^t\boldsymbol{X}^{1/2},$$

for $t \in [0, 1]$. The curve $\gamma : t \mapsto \boldsymbol{X}\sharp_t\boldsymbol{Y}$ starting at $\boldsymbol{X}$ and ending at $\boldsymbol{Y}$ is a geodesic wrt the metric δ_Φ for any sgf $\Phi : \mathbb{R}^n \to \mathbb{R}_+$, . In other words,

$$\begin{aligned}\delta_\Phi(\gamma(s), \gamma(t)) &= \delta_\Phi(\boldsymbol{X}\sharp_s\boldsymbol{Y}, \boldsymbol{X}\sharp_t\boldsymbol{Y}),\\ &= |s-t|\delta_\Phi(\boldsymbol{X}, \boldsymbol{Y}) = |s-t|\delta_\Phi(\gamma(0), \gamma(1)),\end{aligned}$$

for $s, t \in [0, 1]$.

This section ends with some key definitions and theorems relevant to the convergence analysis in Section 5.4.

Definition 5.5 (Contractions). Let $f : (\mathbb{P}_n, \delta_\Phi) \to (\mathbb{P}_n, \delta_\Phi)$ be a self-map on $\mathbb{P}_n$ and denote by

$$L_\Phi(f) := \sup_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{P}_n,\boldsymbol{X}\neq\boldsymbol{Y}} \frac{\delta_\Phi(f(\boldsymbol{X}), f(\boldsymbol{Y}))}{\delta_\Phi(\boldsymbol{X},\boldsymbol{Y})}$$

the Lipschitz constant of the map f. The map f is called

1. *nonexpansive* if $L_\Phi(f) \leq 1$,
2. a *strict contraction* if $L_\Phi(f) < 1$.

The following lemma states a sufficient condition for a self-map to have a unique fixed point as well as a simple iterative algorithm to approximate it.

Lemma 5.6 (Banach's fixed-point theorem). Let $f : (\mathbb{P}_n, \delta_\Phi) \to (\mathbb{P}_n, \delta_\Phi)$ be a strict contraction on $\mathbb{P}_n$. Then, there exists a unique fixed point $\boldsymbol{X}_\star = f(\boldsymbol{X}_\star)$. Furthermore, the iterates $\{\boldsymbol{X}_k\}_{k\in\mathbb{N}}$ defined recursively as $\boldsymbol{X}_{k+1} = f(\boldsymbol{X}_k)$ converge to the fixed point $\boldsymbol{X}_\star$ for any $\boldsymbol{X}_0 \in \mathbb{P}_n$.

Finally, the next lemma shows that translations do not increase the distance in the Finsler metric.

Lemma 5.7 (Translations are nonexpansive, [LL08, Prop. 4.1]). Let $\boldsymbol{P} \in \mathbb{H}_n^+$ be a psd matrix. For any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{P}_n$, one has that

$$\delta_\Phi(\boldsymbol{P}+\boldsymbol{X}, \boldsymbol{P}+\boldsymbol{Y}) \leq \frac{\max(\lambda_{\max}(\boldsymbol{X}),\, \lambda_{\max}(\boldsymbol{Y}))}{\lambda_{\min}(\boldsymbol{P}) + \max(\lambda_{\max}(\boldsymbol{X}),\, \lambda_{\max}(\boldsymbol{Y}))}\delta_\Phi(\boldsymbol{X},\boldsymbol{Y}).$$

5.3 Stability of Matrices

In this section, a notion of stability of linear dynamical systems will be investigated in relation to spectral properties of matrices. Stability plays a central role in the theory of dynamical systems as well as control theory. Since Riccati equations often arise from studying nonautonomous dynamical systems, stability and related aspects are recurrent themes in the study of Riccati equations.

This section starts with an introduction to concepts and results fundamental to the theory of stability and ends with a subsection devoted to *Lyapunov equations* which lay the foundations of algebraic Riccati equations as discussed in Section 5.1.

Example 5.8 (Linear autonomous dynamics). Let $\boldsymbol{A} \in \mathbb{M}_n$ be a complex-valued matrix. Consider the linear dynamical system,

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t \text{ for } t \in \mathbb{N}, \text{ and } \boldsymbol{x}_0 \in \mathbb{C}^n. \tag{5.5}$$

The state at time $t \geq 0$ can be expressed in closed form as $\boldsymbol{x}_t = \boldsymbol{A}^t\boldsymbol{x}_0$. It is common to study dynamical systems by investigating whether the trajectory $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ converges to an equilibrium state $\boldsymbol{x}_\star \in \mathbb{C}^n$ such that $\boldsymbol{x}_\star = \boldsymbol{A}\boldsymbol{x}_\star$ or otherwise the trajectory $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ diverges and and the state blows up as time goes to infinity. Due to linearity of the state transitions, one can take $\boldsymbol{x}_\star = \boldsymbol{0}$ without loss of generality by transforming $\boldsymbol{x}_t \mapsto \boldsymbol{x}_t - \boldsymbol{x}_*$ for all $t \in \mathbb{N}$. The dynamical system (5.5) is said to be *globally asymptotically stable* (GAS) if the trajectory $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ converges to the equilibrium

$x_\star = \mathbf{0}$ for any starting point $\boldsymbol{x}_0 \in \mathbb{C}^n$. A necessary and sufficient condition for this to happen is $\lim_{t\to\infty} \boldsymbol{A}^t = \mathbf{0}$. ▪

The limit condition for stability stated in the example above could be impractical to verify in many settings. Therefore, alternative ways of feasibly certifying stability of linear dynamical systems is needed.

Definition 5.9 (Spectral radius). For a complex-valued matrix $\boldsymbol{A} \in \mathbb{M}_n$, denote by $\mathrm{sp}(\boldsymbol{A}) \coloneqq \{\lambda \in \mathbb{C} \mid \det(\boldsymbol{A} - \lambda \mathbf{I}_n) = 0\}$ the spectrum of it. The spectral radius of $\boldsymbol{A}$ is defined as $\rho(\boldsymbol{A}) \coloneqq \sup_{\lambda \in \mathrm{sp}(\boldsymbol{A})} |\lambda|$.

Spectral radius serves as an equivalent criterion for certifying stability as stated by the following lemma.

Lemma 5.10 The limit $\lim_{t\to\infty} \boldsymbol{A}^t = \mathbf{0}$ holds if and only if all the eigenvalues of $\boldsymbol{A}$ are inside the open unit disk of the complex plane, that is, $\rho(\boldsymbol{A}) < 1$.

Proof. Taking the Jordan canonical form $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$, one gets

$$\lim_{t\to\infty} \boldsymbol{A}^t = \lim_{t\to\infty} \left(\boldsymbol{P}\boldsymbol{J}^t\boldsymbol{P}^{-1}\right) = \boldsymbol{P}\left(\lim_{t\to\infty} \boldsymbol{J}^t\right)\boldsymbol{P}^{-1}.$$

It can be shown that the limit $\lim_{t\to\infty} \boldsymbol{J}^t = \mathbf{0}$ holds if and only if for all eigenvalues $\lambda \in \mathrm{sp}(\boldsymbol{A})$, the limit $\lim_{t\to\infty} \boldsymbol{J}_\lambda^t = \mathbf{0}$ holds for each Jordan block corresponding to eigenvalue λ. Since the elements of $\boldsymbol{J}_\lambda^t$ consist only of powers of λ_i upto t, the limit condition is satisfied if and only if all the eigenvalues have modulus less than 1. ■

The notion of stability of dynamical systems can be extended to matrices by applying the spectral radius criterion.

Definition 5.11 (Stable matrices). A matrix $\boldsymbol{A} \in \mathbb{M}_n$ is called *(Schur) stable* if all the eigenvalues of $\boldsymbol{A}$ are strictly inside the unit circle, that is, $\rho(\boldsymbol{A}) < 1$. A pair of matrices $(\boldsymbol{A}, \boldsymbol{B})$ is said to be *stabilizable* if there exists a matrix $\boldsymbol{K} \in \mathbb{C}^{m\times n}$ such that $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}$ is stable.

As discussed earlier in Section 5.1 and at the beginning of this section, Lyapunov equations are foundational in the analysis of Riccati equations. Lyapunov equations are named after Russian mathematician Aleksandr Mikhailovich Lyapunov (1857-1918) who pioneered the mathematical theory of stability of dynamical systems and worked in the fields of mathematical physics, probability theory and potential theory [SMI92]. He originally proposed Lyapunov function method to certify stability of equilibrium points of ordinary differential equations.

When applied to discrete-time linear dynamical systems, Lyapunov function method leads to *Lyapunov stability theorem* for certifying stability by showing existence of solution to a Lyapunov equation as stated in Theorem 5.14. The details of Lyapunov function method is omitted in this manuscript for the sake of brevity. However, the interested reader is refereed to [Kha96] for detailed derivations.

The proof of Theorem 5.14 relies on the following lemmas.

Lemma 5.12 (Gelfand's formula, [Lax02, Thm. 17.4]). Suppose a consistent matrix norm $\|\cdot\|$ is given. For any matrix $\boldsymbol{A} \in \mathbb{M}_n$, the spectral radius is bounded from above as $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}^k\|^{1/k}$ for all $k \in \mathbb{N}$. Furthermore, $\|\boldsymbol{A}^k\|^{1/k} \downarrow \rho(\boldsymbol{A})$ as $k \to \infty$.

Gelfand's formula provides a way to approximate spectral radius from matrix norms.

Lemma 5.13 (Bounded matrix norm, [HJ13, Lem. 5.6.10]). Let $\boldsymbol{A} \in \mathbb{M}_n$ and $\epsilon > 0$ be given. There exists a submultiplicative matrix norm $\|\cdot\|_{\boldsymbol{A},\epsilon}$ such that $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}^s\|_{\boldsymbol{A},\epsilon} \leq \rho(\boldsymbol{A}) + \epsilon$.

Theorem 5.14 (Lyapunov stability theorem). Let $\boldsymbol{A} \in \mathbb{M}_n$ be given. The following are equivalent.

1. The linear dynamical system $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t$ is GAS.
2. For every pd matrix $\boldsymbol{Q} \in \mathbb{P}_n$, there exists a unique pd matrix $\boldsymbol{P} \in \mathbb{P}_n$ such that $\boldsymbol{P} - \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{A} = \boldsymbol{Q}$.

Proof. Suppose that the equation $\boldsymbol{P} - \boldsymbol{A}^*\boldsymbol{P}\boldsymbol{A} = \boldsymbol{Q}$ holds for $\boldsymbol{P} \succ \boldsymbol{0}$ and $\boldsymbol{Q} \succ \boldsymbol{0}$. Multiplying both hand sides with $\boldsymbol{P}^{-1/2}$ from left and right as

$$\mathbf{I}_n - (\boldsymbol{P}^{-1/2}\boldsymbol{A}^*\boldsymbol{P}^{1/2})(\boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}) = \boldsymbol{P}^{-1/2}\boldsymbol{Q}\boldsymbol{P}^{-1/2} \succ \boldsymbol{0},$$

one obtains the bound $\|\boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\|_2 < 1$ on the spectral norm. Since the mapping $\boldsymbol{A} \mapsto \boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}$ is a similarity transformation, the characteristic polynomial of $\boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}$ and its spectrum are unchanged. Thus, one can certify stability of $\boldsymbol{A}$ by noticing

$$\rho(\boldsymbol{A}) = \rho(\boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}) \leq \|\boldsymbol{P}^{1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\|_2 < 1,$$

where the first inequality is due to Lemma 5.12.

For the converse, suppose $\rho(\boldsymbol{A}) < 1$. Consider the sequence $\{\boldsymbol{P}_t\}_{t\in\mathbb{N}}$ generated by the iterations $\boldsymbol{P}_{t+1} = \boldsymbol{Q} + \boldsymbol{A}^*\boldsymbol{P}_t\boldsymbol{A}$ with $\boldsymbol{P}_0 = \boldsymbol{0}$ for a given pd matrix $\boldsymbol{Q} \succ \boldsymbol{0}$. The iterate at time $t \geq 0$ can be expressed in closed form as

$$\boldsymbol{P}_t = \sum\nolimits_{s=0}^{t-1} (\boldsymbol{A}^*)^s \boldsymbol{Q}\boldsymbol{A}^s \succ \boldsymbol{0}.$$

Notice that the iterates are nondecreasing, that is, $\boldsymbol{P_{t+1}} \succcurlyeq \boldsymbol{P_t} \succ \boldsymbol{0}$ for all $t \geq 0$. Let $\epsilon > 0$ be small enough such that $\rho(\boldsymbol{A}) + \epsilon < 1$. By Lemma 5.12, there exists a constant $T_\epsilon \in \mathbb{N}$ such that $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}^t\|^{1/t} < \rho(\boldsymbol{A}) + \epsilon < 1$ for all $t \geq T_\epsilon$. Therefore, for any $t \geq s \geq T_\epsilon + 1$, the distance between two iterates can be bounded as

$$\begin{aligned}
\|\boldsymbol{P}_t - \boldsymbol{P}_s\|_2 &= \left\| \sum\nolimits_{k=s}^{t-1} (\boldsymbol{A}^*)^k \boldsymbol{Q}\boldsymbol{A}^k \right\|_2, \\
&\leq \sum\nolimits_{k=s}^{t-1} \|(\boldsymbol{A}^*)^k \boldsymbol{Q}\boldsymbol{A}^k\|_2, \\
&\leq \|\boldsymbol{Q}\|_2 \sum\nolimits_{k=s}^{t-1} \|\boldsymbol{A}^k\|_2^2, \\
&\leq \|\boldsymbol{Q}\|_2 \sum\nolimits_{k=s}^{t-1} (\rho(\boldsymbol{A}) + \epsilon)^{2k}, \\
&= \|\boldsymbol{Q}\|_2 \frac{(\rho(\boldsymbol{A}) + \epsilon)^{2s} - (\rho(\boldsymbol{A}) + \epsilon)^{2t}}{1 - (\rho(\boldsymbol{A}) + \epsilon)^2} \leq \|\boldsymbol{Q}\|_2 \frac{(\rho(\boldsymbol{A}) + \epsilon)^{2s}}{1 - (\rho(\boldsymbol{A}) + \epsilon)^2},
\end{aligned}$$

where the first inequality is due to triangle inequality and the second one is due to submultiplicative property of spectral norm. Notice that the bound above is independent of t which leads to the bound

$$\sup_{t \geq s} \|\boldsymbol{P}_t - \boldsymbol{P}_s\|_2 \leq \|\boldsymbol{Q}\|_2 \frac{(\rho(\boldsymbol{A}) + \epsilon)^{2s}}{1 - (\rho(\boldsymbol{A}) + \epsilon)^2}.$$

Aside: Spectral norm $\|A\|_2$ is equal to the maximum singular value.

The right-hand side converges to zero as $s \to \infty$ since $\rho(\boldsymbol{A}) + \epsilon < 1$. Therefore, the sequence of iterates, $\{\boldsymbol{P}_t\}_{t\in\mathbb{N}}$, is Cauchy in $\mathbb{H}_n^+$ and there exists a limit $\lim_{t\to\infty} \boldsymbol{P}_t = \boldsymbol{P} \in \mathbb{H}_n^+$ such that $\boldsymbol{P} \succcurlyeq \boldsymbol{P}_t \succ \boldsymbol{0}$. The limit can be expressed as an infinite sum as $\boldsymbol{P} = \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t$. Substituting this limit in the Lyapunov equation yields

$$\begin{aligned}
\boldsymbol{P} - \boldsymbol{A}^* \boldsymbol{P} \boldsymbol{A} &= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^{t+1} \boldsymbol{Q} \boldsymbol{A}^{t+1}, \\
&= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^{t+1} \boldsymbol{Q} \boldsymbol{A}^{t+1}, \\
&= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \sum_{t=1}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t = \boldsymbol{Q}.
\end{aligned}$$

In order to show uniqueness of the solution of the Lyapunov equation, suppose $\boldsymbol{P}_1, \boldsymbol{P}_2 \succ \boldsymbol{0}$ are two different solutions. Subtracting the equation $\boldsymbol{P}_1 = \boldsymbol{A}^* \boldsymbol{P}_1 \boldsymbol{A} + \boldsymbol{Q}$ from $\boldsymbol{P}_2 = \boldsymbol{A}^* \boldsymbol{P}_2 \boldsymbol{A} + \boldsymbol{Q}$, one obtains

$$\boldsymbol{P}_1 - \boldsymbol{P}_2 = \boldsymbol{A}^* (\boldsymbol{P}_1 - \boldsymbol{P}_2) \boldsymbol{A}.$$

Observe that multiplying the equality above repeatedly with $\boldsymbol{A}^*$ from left and with $\boldsymbol{A}$ from right yields

$$\boldsymbol{P}_1 - \boldsymbol{P}_2 = \sum_{s=0}^{t-1} (\boldsymbol{A}^*)^s (\boldsymbol{P}_1 - \boldsymbol{P}_2) \boldsymbol{A}^s,$$

for any $t \geq 0$. Let $\epsilon > 0$ be given such that $\rho(\boldsymbol{A}) + \epsilon < 1$. By Lemma 5.13, there exists a submultiplicative norm $\|\cdot\|_{\boldsymbol{A},\epsilon}$ such that $\|\boldsymbol{A}^s\|_{\boldsymbol{A},\epsilon} \leq (\rho(\boldsymbol{A}) + \epsilon)^s$. By triangle inequality and submultiplicative property, the distance between the two solutions can be bounded as

$$\begin{aligned}
\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\boldsymbol{A},\epsilon} &\leq \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\boldsymbol{A},\epsilon} \sum_{s=0}^{t-1} \|\boldsymbol{A}^s\|_{\boldsymbol{A},\epsilon}^2, \\
&\leq \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\boldsymbol{A},\epsilon} \sum_{s=0}^{t-1} (\rho(\boldsymbol{A}) + \epsilon)^{2s}, \\
&= \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\boldsymbol{A},\epsilon} \frac{(\rho(\boldsymbol{A}) + \epsilon)^{2t}}{1 - (\rho(\boldsymbol{A}) + \epsilon)^2},
\end{aligned}$$

which holds for all $t \geq 0$. Taking the limit as $t \to \infty$, one gets that $\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\boldsymbol{A},\epsilon} = 0$, which is a contradiction. Hence, the solution must be unique. ■

5.3.1 Lyapunov Equation

Given a matrix $\boldsymbol{A} \in \mathbb{M}_n$, the *discrete Lyapunov operator* $\mathscr{L}_{\boldsymbol{A}} : \mathbb{H}_n \to \mathbb{H}_n$ is a linear operator defined as $\mathscr{L}_{\boldsymbol{A}}(\boldsymbol{P}) = \boldsymbol{P} - \boldsymbol{A}^* \boldsymbol{P} \boldsymbol{A}$ for any $\boldsymbol{P} \in \mathbb{H}_n$. The following proposition lists some of the equivalent properties of $\mathscr{L}_{\boldsymbol{A}}$.

Proposition 5.15 (Properties of Lyapunov operator). Suppose $\boldsymbol{A} \in \mathbb{M}_n$ is given. The following statements are equivalent.

1. $\boldsymbol{A}$ is Schur stable.
2. $\mathscr{L}_{\boldsymbol{A}}(\boldsymbol{P}) \succ \boldsymbol{0}$ for all $\boldsymbol{P} \succ \boldsymbol{0}$.
3. There exists pd matrices $\boldsymbol{P} \succ \boldsymbol{0}$, and $\boldsymbol{Q} \succ \boldsymbol{0}$ such that $\mathscr{L}_{\boldsymbol{A}}(\boldsymbol{P}) = \boldsymbol{Q}$.
4. There exists pd matrices $\boldsymbol{P} \succ \boldsymbol{0}$, and $\boldsymbol{Q} \succ \boldsymbol{0}$ such that $\mathscr{L}_{\boldsymbol{A}}(\boldsymbol{P}) \succ \boldsymbol{Q}$.
5. The inverse operator $\mathscr{L}_{\boldsymbol{A}}^{-1} : \mathbb{H}_n \to \mathbb{H}_n$ exists.

Proof. Equivalence of these statements can be shown following a direction similar to the proof of Theorem 5.14. ■

The unique solution to the Lyapunov equation is constructed in the proof of Theorem 5.14 for a pd matrix $\boldsymbol{Q} \succ \boldsymbol{0}$. The same form of the solution is valid for any Hermitian matrix $\boldsymbol{Q} \in \mathbb{H}_n$ as long as $\boldsymbol{A}$ is stable.

Theorem 5.16 (Solution operator, [Lan70, Thm. 3]). Let $\boldsymbol{A} \in \mathbb{M}_n$ be a stable matrix. The solution operator $\mathscr{L}_{\boldsymbol{A}}^{-1} : \mathbb{H}_n \to \mathbb{H}_n$ can be expressed as

$$\mathscr{L}_{\boldsymbol{A}}^{-1}(\boldsymbol{Q}) = \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t, \tag{5.6}$$

for any $\boldsymbol{Q} \in \mathbb{H}_n$.

Proof. Since $\boldsymbol{A}$ is stable, all of its eigenvalues are inside the open unit disk. Hence, the series $\sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t$ converges for any $\boldsymbol{Q} \in \mathbb{H}_n$ (see the proof of Theorem 5.14).

By evaluating the Lyapunov operator at the candidate solution (5.6), one gets

$$\begin{aligned}
\mathscr{L}_{\boldsymbol{A}} \left(\sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t \right) &= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \boldsymbol{A}^* \left(\sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t \right) \boldsymbol{A}, \\
&= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^{t+1} \boldsymbol{Q} \boldsymbol{A}^{t+1}, \\
&= \sum_{t=0}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t - \sum_{t=1}^{\infty} (\boldsymbol{A}^*)^t \boldsymbol{Q} \boldsymbol{A}^t, \\
&= \boldsymbol{Q}.
\end{aligned}$$

Uniqueness of the solution operator for stable $\boldsymbol{A}$ can be argued in a similar fashion as in the proof of Theorem 5.14.

■

Remark 5.17 (Solution in Kronecker product form). The Lyapunov equation can also be solved for general $\boldsymbol{A}, \boldsymbol{Q} \in \mathbb{M}_n$ by vectorizing and using Kronecker products as

$$(\mathbf{I} \otimes \mathbf{I} - \boldsymbol{A}^\intercal \otimes \boldsymbol{A}^*) \operatorname{vec}(\boldsymbol{P}) = \operatorname{vec}(\boldsymbol{Q}).$$

If the eigenvalues of matrix $\boldsymbol{A}$ are such that $\lambda_i \lambda_j^* \neq 1$, then the unique solution exists for any $\boldsymbol{Q}$ and is given by

$$\operatorname{vec}(\boldsymbol{P}) = (\mathbf{I} \otimes \mathbf{I} - \boldsymbol{A}^\intercal \otimes \boldsymbol{A}^*)^{-1} \operatorname{vec}(\boldsymbol{Q}).$$

Notice that the special case $\rho(\boldsymbol{A}) < 1$ satisfies the spectrum condition since all eigenvalue have modulus less than 1.

5.4 Discrete Algebraic Riccati Equations

In the previous sections, several sophisticated tools and techniques were developed. In this section, existence and uniqueness of the solution of discrete algebraic Riccati equations (DARE) will be investigated utilizing these tools and techniques.

Let $\boldsymbol{A} \in \mathbb{M}_n$, $\boldsymbol{B} \in \mathbb{C}^{n \times m}$ be complex-valued matrices and $\boldsymbol{Q} \in \mathbb{P}_n$, $\boldsymbol{R} \in \mathbb{P}_m$ be pd matrices. The Riccati operator $\mathscr{R} : \mathbb{H}_n^+ \to \mathbb{H}_n^+$ is defined as

$$\mathscr{R}(\boldsymbol{X}) = \boldsymbol{Q} + \boldsymbol{A}^* \boldsymbol{X} \boldsymbol{A} - \boldsymbol{A}^* \boldsymbol{X} \boldsymbol{B} (\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X} \boldsymbol{B})^{-1} \boldsymbol{B}^* \boldsymbol{X} \boldsymbol{A},$$

for any $\boldsymbol{X} \in \mathbb{H}_n^+$. Lemma 5.19 establishes the formal link between Riccati and Lyapunov equations as hinted in Section 5.1. The following lemma is needed in order to prove Lemma 5.19.

Lemma 5.18 (Block upper-diagonal-lower (UDL) decomposition). Let $\boldsymbol{A} \in \mathbb{M}_n$, $\boldsymbol{B} \in \mathbb{C}^{n \times m}$, $\boldsymbol{C} \in \mathbb{C}^{m \times n}$, and $\boldsymbol{D} \in \mathbb{M}_m$ be given complex matrices. The following decomposition is admitted as long as $\boldsymbol{D}$ is invertible.

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \mathbf{0}_{mn} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \mathbf{0}_{nm} \\ \mathbf{0}_{mn} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{nm} \\ \boldsymbol{D}^{-1}\boldsymbol{C} & \mathbf{I}_m \end{bmatrix}.$$

Proof. The proof follows directly by multiplying the matrices in the right-hand side. ∎

Lemma 5.19 (A useful upper bound). Consider the function $\Gamma : \mathbb{C}^{m\times n} \times \mathbb{H}_n^+ \to \mathbb{H}_n^+$ defined as

$$\Gamma(\boldsymbol{K}, \boldsymbol{X}) \coloneqq (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^* \boldsymbol{X}(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}) + \boldsymbol{Q} + \boldsymbol{K}^* \boldsymbol{R}\boldsymbol{K}.$$

For any $\boldsymbol{K} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{X} \in \mathbb{H}_n^+$, it holds that $\mathcal{R}(\boldsymbol{X}) \preccurlyeq \Gamma(\boldsymbol{K}, \boldsymbol{X})$. Furthermore, the equality $\mathcal{R}(\boldsymbol{X}) = \Gamma(\boldsymbol{K}_\star(\boldsymbol{X}), \boldsymbol{X})$ holds for

$$\boldsymbol{K}_\star(\boldsymbol{X}) \coloneqq -(\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B})^{-1} \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{A}.$$

Proof. Observe that $\Gamma(\boldsymbol{K}, \boldsymbol{X})$ is quadratic in $\boldsymbol{K}$ and linear in $\boldsymbol{X}$ as it can be expressed as

$$\Gamma(\boldsymbol{K}, \boldsymbol{X}) = \begin{bmatrix} \mathrm{I}_n & \boldsymbol{K}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{A} + \boldsymbol{Q} & \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{B} \\ \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{A} & \boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B} \end{bmatrix} \begin{bmatrix} \mathrm{I}_n \\ \boldsymbol{K} \end{bmatrix},$$

Since $\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B} \succ \boldsymbol{0}$, block UDL decomposition can be applied to the middle block matrix in (5.4) by Lemma 5.18 as

$$\begin{bmatrix} \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{A} + \boldsymbol{Q} & \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{B} \\ \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{A} & \boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B} \end{bmatrix} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*,$$

where

$$\boldsymbol{U} = \begin{bmatrix} \mathrm{I}_n & \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{B}(\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B})^{-1} \\ \boldsymbol{0}_n & \mathrm{I}_n \end{bmatrix},$$

and

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{A} + \boldsymbol{Q} - \boldsymbol{A}^* \boldsymbol{X}\boldsymbol{B}(\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B})^{-1} \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{A} & \boldsymbol{0}_n \\ \boldsymbol{0}_n & \boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B} \end{bmatrix}.$$

Notice that the first block of $\boldsymbol{D}$ is same as $\mathcal{R}(\boldsymbol{X})$ and the upper second block of $\boldsymbol{U}$ is simply $-\boldsymbol{K}_\star(\boldsymbol{X})^*$. Using this decomposition, $\Gamma(\boldsymbol{K}, \boldsymbol{X})$ can be rewritten as

$$\begin{aligned} \Gamma(\boldsymbol{K}, \boldsymbol{X}) &= \begin{bmatrix} \mathrm{I}_n & \boldsymbol{K}^* \end{bmatrix} \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^* \begin{bmatrix} \mathrm{I}_n \\ \boldsymbol{K} \end{bmatrix}, \\ &= \begin{bmatrix} \mathrm{I}_n & \boldsymbol{K}^* - \boldsymbol{K}_\star(\boldsymbol{X})^* \end{bmatrix} \begin{bmatrix} \mathcal{R}(\boldsymbol{X}) & \boldsymbol{0}_n \\ \boldsymbol{0}_n & \boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B} \end{bmatrix} \begin{bmatrix} \mathrm{I}_n \\ \boldsymbol{K} - \boldsymbol{K}_\star(\boldsymbol{X}) \end{bmatrix}, \\ &= \mathcal{R}(\boldsymbol{X}) + (\boldsymbol{K} - \boldsymbol{K}_\star(\boldsymbol{X}))^* (\boldsymbol{R} + \boldsymbol{B}^* \boldsymbol{X}\boldsymbol{B})(\boldsymbol{K} - \boldsymbol{K}_\star(\boldsymbol{X})). \end{aligned}$$

Thus, $\Gamma(\boldsymbol{K}, \boldsymbol{X}) \succcurlyeq \mathcal{R}(\boldsymbol{X})$ for any $\boldsymbol{K} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{X} \in \mathbb{H}_n^+$ with equality only if $\boldsymbol{K} = \boldsymbol{K}_\star(\boldsymbol{X})$. ∎

Proposition 5.20 (Properties of Riccati operator). Suppose $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{H}_n^+$. Then, we have that

- **Monotone.** $\mathcal{R}(\boldsymbol{X}) \preccurlyeq \mathcal{R}(\boldsymbol{X}')$ whenever $\boldsymbol{X} \preccurlyeq \boldsymbol{X}'$,
- **Concave.** $\mathcal{R}(t\boldsymbol{X} + (1-t)\boldsymbol{X}') \succcurlyeq t\mathcal{R}(\boldsymbol{X}) + (1-t)\mathcal{R}(\boldsymbol{X}')$ for $t \in [0, 1]$,
- **Continuous.** If $\{\boldsymbol{X}_k\}_{k\in\mathbb{N}} \subset \mathbb{H}_n^+$ is a sequence such that $\lim \boldsymbol{X}_k = \boldsymbol{X}$ as $k \to \infty$, then $\lim \mathcal{R}(\boldsymbol{X}_k) = \mathcal{R}(\boldsymbol{X})$ as $k \to \infty$.

Proof. To show monotonicity, suppose $\boldsymbol{X} \preccurlyeq \boldsymbol{X}'$. By Lemma 5.19, one has that $\mathcal{R}(\boldsymbol{X}) \preccurlyeq \Gamma(\boldsymbol{K}, \boldsymbol{X}) \preccurlyeq \Gamma(\boldsymbol{K}, \boldsymbol{X}')$ for any $\boldsymbol{K} \in \mathbb{C}^{m\times n}$. This holds in particular for $\boldsymbol{K}_\star(\boldsymbol{X}')$. Hence, $\mathcal{R}(\boldsymbol{X}) \preccurlyeq \Gamma(\boldsymbol{K}_\star(\boldsymbol{X}'), \boldsymbol{X}) \preccurlyeq \Gamma(\boldsymbol{K}_\star(\boldsymbol{X}'), \boldsymbol{X}') = \mathcal{R}(\boldsymbol{X}')$.

In order to show concavity, Lemma 5.19 and linearity of $\Gamma(\boldsymbol{K}, \boldsymbol{X})$ wrt $\boldsymbol{X}$ can be used to write

$$\begin{aligned}\Gamma(\boldsymbol{K}, t\boldsymbol{X} + (1-t)\boldsymbol{X}') &= t\Gamma(\boldsymbol{K}, \boldsymbol{X}) + (1-t)\Gamma(\boldsymbol{K}, \boldsymbol{X}') \\ &\succcurlyeq t\mathscr{R}(\boldsymbol{X}) + (1-t)\mathscr{R}(\boldsymbol{X}')\end{aligned}$$

for any $\boldsymbol{K} \in \mathbb{C}^{m\times n}$ and $t \in [0,1]$. Choosing $\boldsymbol{K} = \boldsymbol{K}_\star(t\boldsymbol{X} + (1-t)\boldsymbol{X}')$ yields the desired result.

Continuity of $\mathscr{R}(\cdot)$ follows trivially from continuity of matrix inversion, congruence transformations, translations and matrix multiplication.

■

Characterizing the conditions for existence and uniqueness of the solutions to DARE lies at the heart of the analysis of Riccati equations. The following theorem gives a necessary and sufficient condition for the existence and uniqueness of a solution, $\boldsymbol{X}_\star$, for which the the closed-loop matrix $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star(\boldsymbol{X}_\star)$ is stable.

Theorem 5.21 (Existence and uniqueness of stabilizing fixed point). If the pair $(\boldsymbol{A}, \boldsymbol{B})$ is stabilizable, then there exists a unique positive-definite fixed point $\boldsymbol{X}_\star \succ \boldsymbol{0}$ of the Riccati operator such that $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star(\boldsymbol{X}_\star)$ is a stable matrix. Conversely, if there exists a positive definite fixed-point $\boldsymbol{X}_\star \succ \boldsymbol{0}$, then it is uniquely determined and $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star(\boldsymbol{X}_\star)$ is a stable matrix.

Proof. For the converse direction, suppose $\boldsymbol{X}_\star = \mathscr{R}(\boldsymbol{X}_\star) \succ \boldsymbol{0}$, that is,

$$\boldsymbol{X}_\star = \Gamma(\boldsymbol{K}_\star, \boldsymbol{X}_\star) = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star)^* \boldsymbol{X}_\star (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star) + \boldsymbol{Q} + \boldsymbol{K}_\star^* \boldsymbol{R}\boldsymbol{K}_\star$$

where $\boldsymbol{K}_\star \coloneqq \boldsymbol{K}_\star(\boldsymbol{X}_\star)$. Notice that the positive-definite matrix $\boldsymbol{X}_\star \succ \boldsymbol{0}$ satisfies the Lyapunov equation with parameters $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star$ and $\boldsymbol{Q} + \boldsymbol{K}_\star^* \boldsymbol{R}\boldsymbol{K}_\star \succ \boldsymbol{0}$. Hence, $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star$ is a stable matrix by Proposition 5.15 and $\boldsymbol{X}_\star$ is uniquely determined by the unique solution to the Lyapunov equation $\Gamma(\boldsymbol{K}_\star, \boldsymbol{X}_\star) = \boldsymbol{X}_\star$. This proves the uniqueness and the stabilizing property of a positive-definite solution.

Now, suppose that $(\boldsymbol{A}, \boldsymbol{B})$ is a stabilizable pair, that is, there exists a matrix $\boldsymbol{K}_0 \in \mathbb{C}^{m\times n}$ such that $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_0$ is stable. By Proposition 5.15, there exists a unique positive-definite matrix $\boldsymbol{X}_0 \succ \boldsymbol{0}$ satisfying the Lyapunov equation $\Gamma(\boldsymbol{K}_0, \boldsymbol{X}_0) = \boldsymbol{X}_0$ with parameters $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_0$ and $\boldsymbol{Q} + \boldsymbol{K}_0^* \boldsymbol{R}\boldsymbol{K}_0 \succ \boldsymbol{0}$. Then, the inequality $\mathscr{R}(\boldsymbol{X}_0) = \Gamma(\boldsymbol{K}_\star(\boldsymbol{X}_0), \boldsymbol{X}_0) \preccurlyeq \Gamma(\boldsymbol{K}_0, \boldsymbol{X}_0) = \boldsymbol{X}_0$ holds by Lemma 5.19. Denoting $\boldsymbol{K}_1 \coloneqq \boldsymbol{K}_\star(\boldsymbol{X}_0)$, the inequality $\Gamma(\boldsymbol{K}_1, \boldsymbol{X}_0) \preccurlyeq \boldsymbol{X}_0$ implies that $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_1$ is a stable matrix by the 4^{th} item of Proposition 5.15. Hence, there exists a positive-definite matrix $\boldsymbol{X}_1 \succ \boldsymbol{0}$ satisfying Lyapunov equation $\Gamma(\boldsymbol{K}_1, \boldsymbol{X}_1) = \boldsymbol{X}_1$ with parameters $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_1$ and $\boldsymbol{Q} + \boldsymbol{K}_1^* \boldsymbol{R}\boldsymbol{K}_1 \succ \boldsymbol{0}$. In addition, the inequality $\mathscr{R}(\boldsymbol{X}_1) = \Gamma(\boldsymbol{K}_\star(\boldsymbol{X}_1), \boldsymbol{X}_1) \preccurlyeq \Gamma(\boldsymbol{K}_1, \boldsymbol{X}_1) = \boldsymbol{X}_1$ holds by Lemma 5.19. Taking the difference $\Gamma(\boldsymbol{K}_1, \boldsymbol{X}_0) - \Gamma(\boldsymbol{K}_1, \boldsymbol{X}_1) = \mathscr{R}(\boldsymbol{X}_0) - \boldsymbol{X}_1 \preccurlyeq \boldsymbol{X}_0 - \boldsymbol{X}_1$, one obtains

$$(\boldsymbol{X}_0 - \boldsymbol{X}_1) - (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_1)^* (\boldsymbol{X}_0 - \boldsymbol{X}_1)(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_1) \succcurlyeq \boldsymbol{0}.$$

By the 2^{th} item of Proposition 5.15, it can be seen that $\boldsymbol{X}_0 \succcurlyeq \boldsymbol{X}_1$ which implies $\mathscr{R}(\boldsymbol{X}_0) \succcurlyeq \mathscr{R}(\boldsymbol{X}_1)$ by monotonicity.

Recursively repeating this process, a nonincreasing sequence of pd matrices $\{\boldsymbol{X}_k\}_{k\in\mathbb{N}}$ can be constructed such that $\boldsymbol{X}_k \succcurlyeq \boldsymbol{X}_{k+1}$. The stabilizing matrix of the next step is $\boldsymbol{K}_{k+1} \coloneqq \boldsymbol{K}_\star(\boldsymbol{X}_k)$ such that $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_{k+1}$ is stable. Each iterate also satisfies the Lyapunov equation $\Gamma(\boldsymbol{K}_k, \boldsymbol{X}_k) = \boldsymbol{X}_k \succ \boldsymbol{0}$.

Since $\{\boldsymbol{X}_k\}_{k\in\mathbb{N}}$ is nonincreasing and bounded below, there exists a limit $\boldsymbol{X}_\star = \lim \boldsymbol{X}_k$ such that $\boldsymbol{X}_k \succcurlyeq \boldsymbol{X}_\star \succcurlyeq \boldsymbol{0}$ for any $k \in \mathbb{N}$. Since the mapping $\boldsymbol{K}_\star : \boldsymbol{X} \mapsto \boldsymbol{K}_\star(\boldsymbol{X})$ is continuous, the limit $\lim \boldsymbol{K}_k = \boldsymbol{K}_\star(\boldsymbol{X}_\star)$ exists. Taking the limit of both hand sides of the equation $\Gamma(\boldsymbol{K}_k, \boldsymbol{X}_k) = \boldsymbol{X}_k$ leads to the equation $\Gamma(\boldsymbol{K}_\star(\boldsymbol{X}_\star), \boldsymbol{X}_\star) = \boldsymbol{X}_\star$. In other words, $\boldsymbol{X}_\star \succcurlyeq \boldsymbol{0}$ is a fixed point $\mathcal{R}(\boldsymbol{X}_\star) = \boldsymbol{X}_\star$.

In order to show positivity and uniqueness of $\boldsymbol{X}_\star$, observe that $\boldsymbol{0} \prec \boldsymbol{Q} = \mathcal{R}(\boldsymbol{0}) \preccurlyeq \mathcal{R}(\boldsymbol{X}_\star) = \boldsymbol{X}_\star$ due to monotonicity of the operator $\mathcal{R}$. Therefore, the fact that $\boldsymbol{X}_\star \succ \boldsymbol{0}$ together with the converse theorem implies uniqueness of $\boldsymbol{X}_\star$ and stability of $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}_\star(\boldsymbol{X}_\star)$. ∎

Remark 5.22 In fact, any positive-semidefinite fixed point must necessarily be nonsingular and unique whenever $(\boldsymbol{A}, \boldsymbol{B})$ is stabilizable and $\boldsymbol{Q} \succ \boldsymbol{0}$. If the condition $\boldsymbol{Q} \succcurlyeq \boldsymbol{0}$ is relaxed, then the unique stabilizing fixed-point might be singular.

5.4.1 Convergence Analysis of Fixed-Point Iterates

In this section, strict contraction property of Riccati operator with respect to Finsler metric δ_Φ will be shown and convergence rate of fixed-point iterations $\boldsymbol{X}_{k+1} = \mathcal{R}(\boldsymbol{X}_k)$ will be analyzed. In order to simplify the analysis, the domain of Riccati operator will be restricted to the cone of pd matrices, $\mathbb{P}_n$. In this case, matrix inversion lemma can be used to rewrite the Riccati operator as

$$\begin{aligned}\mathcal{R}(\boldsymbol{X}) &= \boldsymbol{Q} + \boldsymbol{A}^* \left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}(\boldsymbol{R} + \boldsymbol{B}^*\boldsymbol{X}\boldsymbol{B})^{-1}\boldsymbol{B}^*\boldsymbol{X}\right)\boldsymbol{A},\\ &= \boldsymbol{Q} + \boldsymbol{A}^* \left(\boldsymbol{S} + \boldsymbol{X}^{-1}\right)^{-1}\boldsymbol{A}\end{aligned}$$

where $\boldsymbol{S} = \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^*$. For the sake of brevity in analysis, it will be assumed that $\boldsymbol{S} \succ \boldsymbol{0}$ in the rest of this manuscript. However, similar results can be obtained for the case of invertible $\boldsymbol{A}$ and $\boldsymbol{S} \succcurlyeq \boldsymbol{0}$. Interested reader is referred to [Bou93], [LL07], and [LL08].

Theorem 5.23 (Strict contraction of Riccati map, [LL08, Thm. 4.4]). Suppose that $\boldsymbol{Q}, \boldsymbol{S} \succ \boldsymbol{0}$ are positive definite. Let $\boldsymbol{A} \in \mathbb{M}_n$ be a complex-valued matrix. Then, the Riccati map is a strict contraction with respect to δ_Φ with the Lipschitz constant

$$L_\Phi(\mathcal{R}) \leq \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{Q}) + \lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})} < 1,$$

for any sgf Φ.

Proof. Assume that $\boldsymbol{A}$ is nonsingular. One has that $\boldsymbol{0} \prec (\boldsymbol{S}+\boldsymbol{X}^{-1})^{-1} \preccurlyeq \boldsymbol{S}^{-1}$ for $\boldsymbol{X} \in \mathbb{P}_n$ by order-reversing property of matrix inversion. Congruence transformation of both sides by $\boldsymbol{A}$ leads to $\boldsymbol{0} \prec \boldsymbol{A}^*(\boldsymbol{S} + \boldsymbol{X}^{-1})^{-1}\boldsymbol{A} \preccurlyeq \boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A}$. Weyl's monotonicity theorem implies that

$$0 < \lambda_{\max}(\boldsymbol{A}^*(\boldsymbol{S} + \boldsymbol{X}^{-1})^{-1}\boldsymbol{A}) \leq \lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})$$

for any $\boldsymbol{X} \in \mathbb{P}_n$.

Suppose $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{P}_n$. Define

$$\alpha := \max(\lambda_{\max}(\boldsymbol{A}^*(\boldsymbol{S} + \boldsymbol{X}^{-1})^{-1}\boldsymbol{A}),\ \lambda_{\max}(\boldsymbol{A}^*(\boldsymbol{S} + \boldsymbol{Y}^{-1})^{-1}\boldsymbol{A})).$$

Notice that $\alpha \leq \lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})$. Then, for any sgf Φ, the Finsler distance between

two points under the Riccati operation can be bounded as

$$\begin{aligned}\delta_\Phi(\mathcal{R}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{Y})) &\leq \frac{\alpha}{\lambda_{\min}(\boldsymbol{Q})+\alpha}\delta_\Phi(\boldsymbol{A}^*(\boldsymbol{S}+\boldsymbol{X}^{-1})^{-1}\boldsymbol{A},\ \boldsymbol{A}^*(\boldsymbol{S}+\boldsymbol{Y}^{-1})^{-1}\boldsymbol{A}),\\ &= \frac{\alpha}{\lambda_{\min}(\boldsymbol{Q})+\alpha}\delta_\Phi(\boldsymbol{S}+\boldsymbol{X}^{-1},\ \boldsymbol{S}+\boldsymbol{Y}^{-1}),\\ &\leq \frac{\alpha}{\lambda_{\min}(\boldsymbol{Q})+\alpha}\delta_\Phi(\boldsymbol{X}^{-1},\ \boldsymbol{Y}^{-1}),\\ &= \frac{\alpha}{\lambda_{\min}(\boldsymbol{Q})+\alpha}\delta_\Phi(\boldsymbol{X},\ \boldsymbol{Y}),\\ &\leq \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{Q})+\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}\delta_\Phi(\boldsymbol{X},\ \boldsymbol{Y}),\end{aligned}$$

where the first and third lines are due to Lemma 5.7, the second and fourth lines are due to invariance of δ_Φ under congruence transformation and matrix inversion, and the last line is due to the monotonically increasing map $\alpha \mapsto \frac{\alpha}{\alpha+\beta}$, $\beta > 0$. This bound implies that

$$L_\Phi(\mathcal{R}) \leq \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{Q})+\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})} < 1$$

for nonsingular $\boldsymbol{A} \in \mathbb{M}_n$ where strict inequality is due to positivity $\lambda_{\min}(\boldsymbol{Q}) > 0$.

Now, suppose that $\boldsymbol{A}$ is arbitrary. Let $\{\boldsymbol{A}_k\}_{k\in\mathbb{N}} \subset \mathrm{GL}(n)$ be a sequence of nonsingular matrices converging to $\boldsymbol{A}$. Define the sequence of Riccati maps $\mathcal{R}_k(\boldsymbol{X}) \coloneqq \boldsymbol{Q} + \boldsymbol{A}_k^*(\boldsymbol{S}+\boldsymbol{X}^{-1})^{-1}\boldsymbol{A}_k$. Then, for $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{P}_n$, one has that

$$\delta_\Phi(\mathcal{R}_k(\boldsymbol{X}),\mathcal{R}_k(\boldsymbol{Y})) \leq \frac{\lambda_{\max}(\boldsymbol{A}_k^*\boldsymbol{S}^{-1}\boldsymbol{A}_k)}{\lambda_{\min}(\boldsymbol{Q})+\lambda_{\max}(\boldsymbol{A}_k^*\boldsymbol{S}^{-1}\boldsymbol{A}_k)}\delta_\Phi(\boldsymbol{X},\ \boldsymbol{Y}).$$

Taking the limit of both sides and using the continuity of the metric $(\boldsymbol{X},\boldsymbol{Y}) \mapsto \delta_\Phi(\boldsymbol{X},\boldsymbol{Y})$ and the eigenvalue function $\boldsymbol{A} \mapsto \lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})$, one obtains the final bound

$$\delta_\Phi(\mathcal{R}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{Y})) \leq \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{Q})+\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}\delta_\Phi(\boldsymbol{X},\ \boldsymbol{Y}),$$

where the coefficient is strictly less then 1. ∎

Strict contraction of Riccati operator enables one to approximate the unique fixed point by fixed-point iterations starting from any initial point. The rate of convergence is exponential and controlled by the Lipschitz constant of the Riccati map as shown in the following corollary.

Corollary 5.24 (Convergence rate of Riccati iterates, [LL08, Thm. 4.6]). Suppose $\boldsymbol{Q},\boldsymbol{S} > \boldsymbol{0}$ and let $\boldsymbol{A} \in \mathbb{M}_n$ be arbitrary. Given an initial seed $\boldsymbol{X}_0 \in \mathbb{P}_n$, define recursions $\boldsymbol{X}_{k+1} \coloneqq \mathcal{R}(\boldsymbol{X}_k)$ for $k \in \mathbb{N}$. Then, for any sgf Φ, the distance of the iterate $\boldsymbol{X}_k$ to the unique fixed point $\boldsymbol{X}_\star = \mathcal{R}(\boldsymbol{X}_\star)$ is controlled as

$$\delta_\Phi(\boldsymbol{X}_\star,\boldsymbol{X}_k) \leq \frac{L^k}{1-L}\delta_\Phi(\boldsymbol{X}_1,\boldsymbol{X}_0),$$

where

$$L = \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{Q})+\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{S}^{-1}\boldsymbol{A})} < 1.$$

Proof. From Theorem 5.23, one has that

$$\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_k) = \delta_\Phi(\mathcal{R}(\boldsymbol{X}_\star), \mathcal{R}(\boldsymbol{X}_{k-1})) \leq L\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_{k-1}).$$

Recursively applying the first inequality, the following bound

$$\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_k) \leq L^{k-1}\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_1),$$

is obtained. In particular, one has that

$$\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_1) \leq L\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_0) \leq L(\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_1) + \delta_\Phi(\boldsymbol{X}_1, \boldsymbol{X}_0)),$$

where the second inequality is due to triangle inequality. Hence,

$$\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_1) \leq \frac{L}{1-L}\delta_\Phi(\boldsymbol{X}_1, \boldsymbol{X}_0),$$

which implies

$$\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_k) \leq L^{k-1}\delta_\Phi(\boldsymbol{X}_\star, \boldsymbol{X}_1) \leq \frac{L^k}{1-L}\delta_\Phi(\boldsymbol{X}_1, \boldsymbol{X}_0).$$

■

Lecture bibliography

[Ber12] D. Bertsekas. *Dynamic Programming and Optimal Control: Volume I*. v. 1. Athena Scientific, 2012. URL: https://books.google.com/books?id=qVBEEAAAQBAJ.

[Bha03] R. Bhatia. "On the exponential metric increasing property". In: *Linear Algebra and its Applications* 375 (2003), pages 211–220. DOI: https://doi.org/10.1016/S0024-3795(03)00647-5.

[BH06] R. Bhatia and J. Holbrook. "Riemannian geometry and matrix geometric means". In: *Linear Algebra and its Applications* 413.2 (2006). Special Issue on the 11th Conference of the International Linear Algebra Society, Coimbra, 2004, pages 594–618. DOI: https://doi.org/10.1016/j.laa.2005.08.025.

[Bou93] P. Bougerol. "Kalman Filtering with Random Coefficients and Contractions". In: *SIAM Journal on Control and Optimization* 31.4 (1993), pages 942–959. eprint: https://doi.org/10.1137/0331041. DOI: 10.1137/0331041.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[KSH00] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall, 2000.

[Kal60] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pages 35–45. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. DOI: 10.1115/1.3662552.

[Kha96] H. Khalil. "Nonlinear Systems, Printice-Hall". In: *Upper Saddle River, NJ* 3 (1996).

[Lan70] P. Lancaster. "Explicit Solutions of Linear Matrix Equations". In: *SIAM Review* 12.4 (1970), pages 544–566. URL: http://www.jstor.org/stable/2028490.

[LL07] J. Lawson and Y. Lim. "A Birkhoff Contraction Formula with Applications to Riccati Equations". In: *SIAM Journal on Control and Optimization* 46.3 (2007), pages 930–951. eprint: https://doi.org/10.1137/050637637. DOI: 10.1137/050637637.

[Lax02] P. D. Lax. *Functional analysis*. Wiley-Interscience, 2002.

[LL08] H. Lee and Y. Lim. "Invariant metrics, contractions and nonlinear matrix equations". In: *Nonlinearity* 21.4 (2008), pages 857–878. DOI: 10.1088/0951-7715/21/4/011.

[LW94] C. Liverani and M. P. Wojtkowski. "Generalization of the Hilbert metric to the space of positive definite matrices." In: *Pacific Journal of Mathematics* 166.2 (1994), pages 339 –355. DOI: pjm/1102621142.

[Ric24] J. Riccati. "Animadversiones in aequationes differentiales secundi gradus". In: *Actorum Eruditorum quae Lipsiae publicantur Supplementa*. Actorum Eruditorum quae Lipsiae publicantur Supplementa v. 8. prostant apud Joh. Grossii haeredes & J.F. Gleditschium, 1724. URL: https://books.google.com/books?id=UjTw1w7tZsEC.

[SMI92] V. I. SMIRNOV. "Biography of A. M. Lyapunov". In: *International Journal of Control* 55.3 (1992), pages 775–784. eprint: https://doi.org/10.1080/00207179208934258. DOI: 10.1080/00207179208934258.

6. Hyperbolic Polynomials

Date: 14 March 2022 **Author:** Eitan Levin

In this note, we survey the basic properties of hyperbolic polynomials and their consequences. These polynomials generalize the determinant, and the roots of their analogously-defined characteristic polynomials share remarkably many of the properties of eigenvalues. Hyperbolic polynomials were originally introduced in the study of (hyperbolic) PDEs by Gårding [Går51], and have since found applications in several other areas, most famously in the resolution of the Kadison–Singer problem by Marcus, Spielman, and Srivastava [MSS14].

Agenda:
1. Hyperbolic polynomials
2. Differentiation
3. Quadratics and Alexandrov's inequality
4. SDPs and perturbation theory
5. Compositions
6. Euclidean structure

6.1 Basic definitions and properties

Let $\mathscr{E}$ be a Euclidean vector space. Denote by $\mathbb{R}[\mathscr{E}]_d$ the space of polynomial functions on $\mathscr{E}$ which are homogeneous of degree d.

In more detail, $\mathscr{E}$ is a finite-dimensional real inner-product space equipped with the associated norm topology. The space $\mathbb{R}[\mathscr{E}]_d$ consists of functions that are homogeneous polynomials of degree d in the coordinates with respect to some basis for $\mathscr{E}$. This definition is independent of the choice of basis.

Definition 6.1 (Hyperbolic polynomials). A *homogeneous* polynomial $p \in \mathbb{R}[\mathscr{E}]_d$ is *hyperbolic* on $\mathscr{E}$ with respect to a direction $\boldsymbol{e} \in \mathscr{E}$ if $p(\boldsymbol{e}) \neq 0$ and if the univariate polynomial $t \mapsto p(\boldsymbol{x} - t\boldsymbol{e})$ has only real roots for any $\boldsymbol{x} \in \mathscr{E}$. The set of hyperbolic polynomials of degree d with respect to $\boldsymbol{e}$ is denoted $\mathsf{Hyp}_d(\boldsymbol{e})$.

The polynomial $t \mapsto p(\boldsymbol{x} - t\boldsymbol{e})$ is called the *characteristic polynomial* of $\boldsymbol{x}$, and its roots (listed with multiplicity) are called the *eigenvalues* of $\boldsymbol{x}$ and are denoted by $\lambda_{\max}(\boldsymbol{x}) := \lambda_1(\boldsymbol{x}) \geq \ldots \geq \lambda_d(\boldsymbol{x}) =: \lambda_{\min}(\boldsymbol{x})$.

Note that the characteristic polynomial always has degree d, and hence has d roots counting multiplicities, since the leading term of $t \mapsto p(\boldsymbol{x} - t\boldsymbol{e})$ is $(-1)^d p(\boldsymbol{e}) t^d$ and we assume $p(\boldsymbol{e}) \neq 0$. See also Remark 6.8 below.

It is common to normalize hyperbolic polynomials so that $p(\boldsymbol{e}) = 1$, but we do not require this.

Example 6.2 The following are three of the most basic examples of hyperbolic polynomials.

(a) The product $p(\boldsymbol{x}) = \prod_{i=1}^d x_i$ is hyperbolic on $\mathbb{R}^d$ with respect to $\mathbf{1} = (1, \ldots, 1)^\intercal$ since the roots of $p(\boldsymbol{x} - t\mathbf{1}) = \prod_{i=1}^d (x_i - t)$ are the coordinates x_i. In this case, $\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{x}^\downarrow$.

(b) The determinant $p(\boldsymbol{X}) = \det(\boldsymbol{X})$ is hyperbolic on $\mathbb{H}_d$ with respect to $\mathbf{I}$ since the roots of $p(\boldsymbol{X} - t\mathbf{I}) = \det(\boldsymbol{X} - t\mathbf{I})$ are the eigenvalues of $\boldsymbol{X}$. In this case $\boldsymbol{\lambda}(\boldsymbol{X}) = \boldsymbol{\lambda}(\boldsymbol{X})^\downarrow$ is the vector of eigenvalues of $\boldsymbol{X}$ sorted in decreasing order.

(c) We have $p(\boldsymbol{x}) = \langle \boldsymbol{e}, \boldsymbol{x} \rangle^d \in \mathsf{Hyp}_d(\boldsymbol{e})$, showing that $\mathsf{Hyp}_d(\boldsymbol{e}) \neq \emptyset$ for all $d \in \mathbb{N}$ and $\boldsymbol{e} \in \mathscr{E}$.

More sophisticated examples are given in Sections 6.3 and 6.5. ∎

Hyperbolic polynomials unify the analysis of certain families of inequalities and cones arising in optimization.

Note that the eigenvalues $\lambda_i(\boldsymbol{x})$ are continuous in $\boldsymbol{x}$, since the roots of a polynomial are continuous functions of its coefficients (which follows from the argument principle in complex analysis), and the coefficients of $t \mapsto p(\boldsymbol{x} - t\boldsymbol{e})$ are themselves continuous (in fact, polynomial) functions of $\boldsymbol{x}$. Note also that

$$\begin{aligned} &\boldsymbol{\lambda}(s\boldsymbol{x} + t\boldsymbol{e}) = s\boldsymbol{\lambda}(\boldsymbol{x}) + t\mathbf{1} \text{ for } s \geq 0 \text{ and } t \in \mathbb{R}, \\ &\lambda_i(-\boldsymbol{x}) = \lambda_{d-i}(\boldsymbol{x}), \\ &p(\boldsymbol{x}) = p(\boldsymbol{e}) \textstyle\prod_{i=1}^{d} \lambda_i(\boldsymbol{x}). \end{aligned} \tag{6.1}$$

In particular, $\boldsymbol{\lambda}(\boldsymbol{e}) = \mathbf{1}$.

As usual in this class, we begin by characterizing the geometry of the set of hyperbolic polynomials.

Proposition 6.3 (Basic geometry). If $p \in \mathsf{Hyp}_d(\boldsymbol{e})$ then $\alpha p \in \mathsf{Hyp}_d(\boldsymbol{e})$ for any nonzero $\alpha \in \mathbb{R}$. In particular, $\mathsf{Hyp}_d(\boldsymbol{e})$ is a cone, but it is not convex. If $p \in \mathsf{Hyp}_d(\boldsymbol{e})$ and $q \in \mathsf{Hyp}_m(\boldsymbol{e})$, then $pq \in \mathsf{Hyp}_{d+m}(\boldsymbol{e})$.

In this note, a *cone* is a set closed under multiplication by strictly positive numbers.

Proof. The first and last claims are easy to verify directly from Definition 6.1.

The cone $\mathsf{Hyp}_d(\boldsymbol{e})$ is not convex since if $p \in \mathsf{Hyp}_d(\boldsymbol{e})$ then $-p \in \mathsf{Hyp}_d(\boldsymbol{e})$ but $(p-p)/2 = 0 \notin \mathsf{Hyp}_d(\boldsymbol{e})$ because the zero polynomial vanishes at $\boldsymbol{e}$. ∎

We do not get convexity even if we restrict to polynomials satisfying $p(\boldsymbol{e}) = 1$, as can be seen from Proposition 6.13(d) below.

Next, we introduce the generalization of the psd cone.

Definition 6.4 The *hyperbolicity cone* for p with respect to $\boldsymbol{e}$ is the cone $\Lambda_{++} = \{\boldsymbol{x} \in \mathscr{E} : \lambda_{\min}(\boldsymbol{x}) > 0\}$. Its closure is denoted by Λ_+.

Note that Λ_{++} is indeed a cone, since $\lambda_{\min}$ is positively homogeneous by (6.1). Note also that

$$\Lambda_+ = \{\boldsymbol{x} \in \mathscr{E} : \lambda_{\min}(\boldsymbol{x}) \geq 0\}.$$

Indeed, the inclusion $\subseteq$ follows by continuity of $\lambda_{\min}$, while the reverse inclusion follows since if $\lambda_{\min}(\boldsymbol{x}) = 0$ then for $t > 0$ we have $\boldsymbol{x} + t\boldsymbol{e} \in \Lambda_{++}$ by (6.1) and $\boldsymbol{x} + t\boldsymbol{e} \to \boldsymbol{x}$ as $t \downarrow 0$. Equation (6.1) also shows that $p/p(\boldsymbol{e}) > 0$ on Λ_{++} and $p/p(\boldsymbol{e}) \geq 0$ on Λ_+.

We have $\Lambda_{++} = \mathbb{R}^d_{++}$, $\Lambda_+ = \mathbb{R}^d_+$ in Example 6.2(a) and $\Lambda_{++} = \mathbb{H}^{++}_d$, $\Lambda_+ = \mathbb{H}^+_d$ in Example 6.2(b).

We proceed to show that hyperbolicity cones are convex.

Lemma 6.5 The set Λ_{++} is the connected component containing $\boldsymbol{e}$ in $\{\boldsymbol{x} \in \mathscr{E} : p(\boldsymbol{x}) \neq 0\}$. Moreover, it is star-shaped with center $\boldsymbol{e}$.

A set S is *star-shaped* with center $s_0 \in \mathsf{S}$ if the line segment between s_0 and any $s \in \mathsf{S}$ is contained in S.

Proof. We have $\boldsymbol{e} \in \Lambda_{++}$ since $\lambda_{\min}(\boldsymbol{e}) = 1$, and $\Lambda_{++} \subseteq \{\boldsymbol{x} \in \mathscr{E} : p(\boldsymbol{x}) \neq 0\}$ since $p/p(\boldsymbol{e}) > 0$ on Λ_{++}. If $\boldsymbol{x} \in \Lambda_{++}$ then $\lambda_{\min}(\tau\boldsymbol{x} + \bar{\tau}\boldsymbol{e}) = \tau\lambda_{\min}(\boldsymbol{x}) + \bar{\tau} > 0$ for all $\tau \in [0,1]$ and $\bar{\tau} = 1 - \tau$. This shows that the line segment between $\boldsymbol{x}$ and $\boldsymbol{e}$ is contained in Λ_{++} and hence Λ_{++} is connected, and moreover start-shaped with center $\boldsymbol{e}$. Therefore, Λ_{++} is contained in the connected component of $\boldsymbol{e}$ in $\{\boldsymbol{x} \in \mathscr{E} : p(\boldsymbol{x}) \neq 0\}$. Conversely, if $\boldsymbol{x}$ is in that connected component, then $\boldsymbol{x} \in \Lambda_{++}$ because $\lambda_{\min}$ varies continuously on any path between $\boldsymbol{x}$ and $\boldsymbol{e}$ contained in $\{\boldsymbol{x} \in \mathscr{E} : p(\boldsymbol{x}) \neq 0\}$ and is never zero on it. ∎

The following theorem gives an important property of hyperbolicity cones, from which their convexity readily follows. The proof below is a slight simplification of Renegar's proof in [Ren06, Thm. 3], which in turn is a simplification of Gårding's original proof in [Går51, Lemma 2.7].

Theorem 6.6 If $\boldsymbol{x} \in \Lambda_{++}$, then p is hyperbolic with respect to $\boldsymbol{x}$. The hyperbolicity cone of p with respect to $\boldsymbol{x}$ is also Λ_{++}.

Proof. Fix arbitrary $\boldsymbol{y} \in \mathscr{C}$. For the first claim, we need to show that $t \mapsto p(\boldsymbol{y} - t\boldsymbol{x})$ has only real roots. Since this is a polynomial with real coefficients, and hence its complex roots come in conjugate pairs, it is equivalent to show that all of its roots have nonnegative imaginary parts. Suppose for contradiction that this is not the case.

Define $q(\alpha, s, t) = p(\mathrm{i}\alpha\boldsymbol{e} + s\boldsymbol{y} - t\boldsymbol{x})$. All the roots of $t \mapsto q(1, 0, t) = p(\boldsymbol{e}) \prod_j (\mathrm{i} - t\lambda_j(\boldsymbol{x}))$ have positive imaginary parts (namely, $\lambda_j(\boldsymbol{x})^{-1}$), while some root of $t \mapsto q(0, 1, t)$ has a negative imaginary part. Using the continuous dependence of the roots on the coefficients, we conclude that there must exist $\beta \in (0, 1)$ such that some root of $t \mapsto q(1 - \beta, \beta, t)$ is real, say that root is t'. Then $\boldsymbol{z} = \beta\boldsymbol{y} - t'\boldsymbol{x} \in \mathscr{C}$ is such that $t \mapsto p(\boldsymbol{z} - t\boldsymbol{e})$ has a purely imaginary root (namely, $t = -\mathrm{i}(1 - \beta)$), contradicting the hyperbolicity of p with respect to $\boldsymbol{e}$.

The second claim follows from Lemma 6.5. ∎

Corollary 6.7 The cones Λ_{++} and Λ_+ are convex.

Proof. The cone Λ_{++} is star-shaped and every $\boldsymbol{x} \in \Lambda_{++}$ is a center by Lemma 6.5, hence it is convex. The cone Λ_+ is the closure of a convex set, hence convex as well. ∎

Remark 6.8 Corollary 6.7 can fail without the requirement that $p(\boldsymbol{e}) \neq 0$ in Definition 6.1, see [Reno6].

We can now generalize the convexity of the largest eigenvalue.

Theorem 6.9 The function $\lambda_{\max}$ is convex and $\lambda_{\min}$ is concave on all of $\mathscr{C}$.

Proof. Since $\lambda_1(-\boldsymbol{x}) = -\lambda_{\min}(\boldsymbol{x})$, it suffices to show the second statement.

The superlevel sets of $\lambda_{\min}$ are convex, because $\{\boldsymbol{x} \in \mathscr{C} : \lambda_{\min}(\boldsymbol{x}) \geq \alpha\} = \Lambda_+ + \alpha\boldsymbol{e}$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{C}$ such that $\lambda_{\min}(\boldsymbol{y}) \geq \lambda_{\min}(\boldsymbol{x})$, define

$$\boldsymbol{z} = \boldsymbol{y} + [\lambda_{\min}(\boldsymbol{x}) - \lambda_{\min}(\boldsymbol{y})]\boldsymbol{e},$$

and note that $\lambda_{\min}(\boldsymbol{x}) = \lambda_{\min}(\boldsymbol{z})$. Therefore, for any $\tau \in [0, 1]$ and $\bar{\tau} = 1 - \tau$, we have $\tau\boldsymbol{x} + \bar{\tau}\boldsymbol{z} \in \{\boldsymbol{w} \in \mathscr{C} : \lambda_{\min}(\boldsymbol{w}) \geq \lambda_{\min}(\boldsymbol{x})\}$, so

$$\lambda_{\min}(\boldsymbol{x}) \leq \lambda_{\min}(\tau\boldsymbol{x} + \bar{\tau}\boldsymbol{z}) = \lambda_{\min}(\tau\boldsymbol{x} + \bar{\tau}\boldsymbol{y}) + \bar{\tau}[\lambda_{\min}(\boldsymbol{x}) - \lambda_{\min}(\boldsymbol{y})].$$

Rearranging gives $\lambda_{\min}(\tau\boldsymbol{x} + \bar{\tau}\boldsymbol{y}) \geq \tau\lambda_{\min}(\boldsymbol{x}) + \bar{\tau}\lambda_{\min}(\boldsymbol{y})$, as desired. ∎

6.2 Derivatives and multilinearization

Differentiating a hyperbolic polynomial gives another hyperbolic polynomial. This fact, which we prove below, ultimately yields Alexandrov's mixed discriminant inequality (Corollary 6.16) as an easy consequence of the theory we develop.

Definition 6.10 For a polynomial p on $\mathscr{C}$ and vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \mathscr{C}$, define induc-

tively the following polynomials:

$$p^{(1)}_{\boldsymbol{y}_1}(\boldsymbol{x}) = \left.\frac{d}{dt}\right|_{t=0} p(\boldsymbol{x} + t\boldsymbol{y}_1) = \langle \nabla p(\boldsymbol{x}), \boldsymbol{y}_1 \rangle,$$
$$p^{(k)}_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k} = \left.\frac{d}{dt}\right|_{t=0} p^{(k-1)}_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{k-1}}(\boldsymbol{x} + t\boldsymbol{y}_k).$$

Proposition 6.11 (Derivatives are hyperbolic). Suppose p is hyperbolic on $\mathscr{E}$ of degree d with hyperbolicity cone Λ_{++}. Then for any $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \Lambda_{++}$, the derivative polynomial $p^{(k)}_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k}$ is hyperbolic of degree $d-k$ with respect to any $\boldsymbol{y} \in \Lambda_{++}$.

Proof. It suffices to prove the case $k = 1$, since then the general case follows by induction. If $\boldsymbol{y}_1 \in \Lambda_{++}$, then p is hyperbolic with respect to $\boldsymbol{y}_1$ by Theorem 6.6, so $t \mapsto p(\boldsymbol{x} - t\boldsymbol{y}_1)$ has d real roots. By Rolle's theorem, the roots $\boldsymbol{\lambda}^{(1)}(\boldsymbol{x})$ of $t \mapsto p^{(1)}_{\boldsymbol{y}_1}(\boldsymbol{x} - t\boldsymbol{y}_1)$ are nested between the roots $\boldsymbol{\lambda}(\boldsymbol{x})$ of $t \mapsto p(\boldsymbol{x} - t\boldsymbol{y}_1)$, meaning that

$$\lambda_1(\boldsymbol{x}) \geq \lambda^{(1)}_1(\boldsymbol{x}) \geq \lambda_2(\boldsymbol{x}) \geq \ldots \geq \lambda_{d-1}(\boldsymbol{x}) \geq \lambda^{(1)}_{d-1}(\boldsymbol{x}) \geq \lambda_d(\boldsymbol{x}).$$

In particular, this shows that all $d-1$ roots of $t \mapsto p^{(1)}(\boldsymbol{x} - t\boldsymbol{y}_1)$ are real. By Euler's homogeneous function theorem, we have $p^{(1)}_{\boldsymbol{y}_1}(\boldsymbol{y}_1) = p(\boldsymbol{y}_1)d \neq 0$. Thus, $p^{(1)}_{\boldsymbol{y}_1}$ is hyperbolic with respect to $\boldsymbol{y}_1$. Moreover, for any $\boldsymbol{y} \in \Lambda_{++}$ we have $\lambda_d(\boldsymbol{y}) > 0$, hence $\lambda^{(1)}_{d-1}(\boldsymbol{y}) > 0$ so $\boldsymbol{y}$ is in the hyperbolicity cone of $p^{(1)}_{\boldsymbol{y}_1}$. Theorem 6.6 then shows that $p^{(1)}_{\boldsymbol{y}_1}$ is hyperbolic with respect to $\boldsymbol{y}$. ∎

An important special case of derivative polynomials is the *multilinearization* of a homogeneous polynomial. If p is a homogeneous polynomial of degree d on $\mathscr{E}$, one defines polynomials $p_{i_1,\ldots,i_d}$ by

$$p\left(\sum\nolimits_{i=1}^{d} \alpha_i \boldsymbol{x}_i\right) = \sum\nolimits_{i_1,\ldots,i_d} p_{i_1,\ldots,i_d}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)\alpha_1^{i_1} \cdots \alpha_d^{i_d},$$

where the sum is over indices $i_1, \ldots, i_d \in \{0, \ldots, d\}$ summing to d. By viewing this as a polynomial in $\alpha \in \mathbb{R}^d$ and equating coefficients of corresponding monomials, it is easy to show that $p_{i_1,\ldots,i_d}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ is homogeneous of degree i_j in $\boldsymbol{x}_j$, and that for any permutation σ on d letters, we have

$$p_{i_1,\ldots,i_d}(\boldsymbol{x}_{\sigma(1)}, \ldots, \boldsymbol{x}_{\sigma(d)}) = p_{i_{\sigma^{-1}(1)},\ldots,i_{\sigma^{-1}(d)}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d).$$

In particular, $\widetilde{p} = p_{1,\ldots,1}$ is multilinear and symmetric in its d arguments, and it is called the *multilinearization* or full polarization of p. The map $p \mapsto \widetilde{p}$ defines a linear isomorphism between $\mathbb{R}[\mathscr{E}]_d$ and symmetric multilinear polynomials on $\mathscr{E}^d$, since we have:

$$p(\boldsymbol{x}) = \frac{1}{d!}\widetilde{p}(\boldsymbol{x}, \ldots, \boldsymbol{x}). \tag{6.2}$$

Multilinearization realizes the isomorphism $\mathrm{Sym}^d(\mathscr{E}) \cong (\mathscr{E}^{\otimes d})^{S_d}$ briefly discussed in class, where the symmetric group S_d acts by permuting the factors.

This can be seen by noting that $p\left(\sum_{i=1}^{d} \alpha_i \boldsymbol{x}\right) = \left(\sum_i \alpha_i\right)^d p(\boldsymbol{x})$ and equating the coefficients of $\alpha_1 \cdots \alpha_d$. Proposition 6.11 implies the following Corollary.

Corollary 6.12 Suppose p is a hyperbolic polynomial on $\mathscr{E}$ of degree d with hyperbolicity cone Λ_{++}, and let $\widetilde{p}$ be its multilinearization. For any $1 \leq k \leq d$ and any $\boldsymbol{x}_k, \ldots, \boldsymbol{x}_d \in \Lambda_{++}$, the polynomial $\boldsymbol{x} \mapsto \widetilde{p}(\boldsymbol{x}, \ldots, \boldsymbol{x}, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_d)$ is hyperbolic with respect to any $\boldsymbol{y} \in \Lambda_{++}$.

Proof. Using (6.2) and the symmetry and multilinearity of $\widetilde{p}$, we have

$$\begin{aligned} p(\boldsymbol{x}+t\boldsymbol{x}_d) &= \frac{1}{d!}\widetilde{p}(\boldsymbol{x}+t\boldsymbol{x}_d,\dots,\boldsymbol{x}+t\boldsymbol{x}_d) \\ &= \frac{1}{d!}\left[p(\boldsymbol{x}) + td\widetilde{p}(\boldsymbol{x},\dots,\boldsymbol{x},\boldsymbol{x}_d) + O(t^2)\right]. \end{aligned}$$

Therefore, $p^{(1)}_{\boldsymbol{x}_d}(\boldsymbol{x}) = \frac{1}{(d-1)!}\widetilde{p}(\boldsymbol{x},\dots,\boldsymbol{x},\boldsymbol{x}_d)$. Continuing inductively, we get

$$\widetilde{p}(\boldsymbol{x},\dots,\boldsymbol{x},\boldsymbol{x}_k,\dots,\boldsymbol{x}_d) = (k-1)!p^{(d-k+1)}_{\boldsymbol{x}_k,\dots,\boldsymbol{x}_d}(\boldsymbol{x}),$$

so the claim follows from Proposition 6.11. ∎

6.3 Hyperbolic quadratics and Alexandrov's mixed discriminant inequality

In this section, we consider the special case of a quadratic polynomial $p(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}$. Its hyperbolicity is characterized by a "reverse Cauchy-Schwarz" inequality satisfied by its associated symmetric bilinear form. We then instantiate this inequality for the multilinearization of the determinant, yielding Alexandrov's mixed discriminant inequality. The following proposition is based on [SH19, Lemma 2.9].

Proposition 6.13 (Hyperbolic quadratics). Set $\mathscr{C} = \mathbb{R}^n$ and fix $\boldsymbol{A} \in \mathbb{H}_n$. Define $Q(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{y}$ and $p(\boldsymbol{x}) = Q(\boldsymbol{x},\boldsymbol{x})$, and suppose $\boldsymbol{e} \in \mathbb{R}^n$ satisfies $p(\boldsymbol{e}) > 0$. The following are equivalent.

(a) p is hyperbolic with respect to $\boldsymbol{e}$.
(b) $Q(\boldsymbol{x},\boldsymbol{y})^2 \geq p(\boldsymbol{x})p(\boldsymbol{y})$ for all $\boldsymbol{x},\boldsymbol{y} \in \mathscr{C}$ such that $p(\boldsymbol{y}) \geq 0$.
(c) $\boldsymbol{A}$ has at most one positive eigenvalue.
(d) There exists $\boldsymbol{w} \in \mathbb{R}^n$ such that $Q(\boldsymbol{x},\boldsymbol{w}) = 0$ implies $p(\boldsymbol{x}) \leq 0$.

The hyperbolicity cone of p with respect to any such $\boldsymbol{e}$ is

$$\Lambda_{++} = \{\boldsymbol{x} \in \mathbb{R}^n : p(\boldsymbol{x}) > 0\}.$$

Proof. We show (a) ⇔ (b) ⇒ (c) ⇒ (d) ⇒ (b). Note that $p(\boldsymbol{x}-t\boldsymbol{y}) = p(\boldsymbol{y})t^2 - 2tQ(\boldsymbol{x},\boldsymbol{y}) + p(\boldsymbol{x})$, whose roots are (assuming $p(\boldsymbol{y}) \neq 0$):

$$\frac{Q(\boldsymbol{x},\boldsymbol{y}) \pm \sqrt{Q(\boldsymbol{x},\boldsymbol{y})^2 - p(\boldsymbol{x})p(\boldsymbol{y})}}{p(\boldsymbol{y})}. \tag{6.3}$$

Suppose (a) holds. If $p(\boldsymbol{y}) = 0$ then the inequality in (b) is trivially satisfied. If $p(\boldsymbol{y}) > 0$, then the smallest root of $t \mapsto p(\boldsymbol{y}-t\boldsymbol{e})$ (given by (6.3) with $\boldsymbol{x} = \boldsymbol{e}$) is strictly positive (and conversely, showing the claimed expression for Λ_{++}), hence $\boldsymbol{y} \in \Lambda_{++}$ and p is hyperbolic with respect to $\boldsymbol{y}$. This implies that both roots in (6.3) are real, and hence the term inside the square root is positive, giving (b). Conversely, suppose (b) holds. Setting $\boldsymbol{y} = \boldsymbol{e}$ in the inequality of (b), we obtain $Q(\boldsymbol{x},\boldsymbol{e})^2 - p(\boldsymbol{x})p(\boldsymbol{e}) \geq 0$, hence (a) holds. Thus, (a) and (b) are equivalent.

Suppose (b) holds. If (c) does not hold, then there exist orthogonal eigenvectors $\boldsymbol{x},\boldsymbol{y}$ corresponding to positive eigenvalues, in which case $Q(\boldsymbol{x},\boldsymbol{y}) = 0$ but $p(\boldsymbol{x})p(\boldsymbol{y}) > 0$. Hence (b) does not hold, a contradiction, so (b) implies (c). Suppose (c) holds. If $\boldsymbol{A}$ has no positive eigenvalues, then $p \leq 0$ and (d) holds with any $\boldsymbol{w}$. If $\boldsymbol{A}$ has exactly one positive eigenvalue, then (d) holds with $\boldsymbol{w}$ equal to the corresponding eigenvector. Hence (c) implies (d). Suppose (d) holds. We claim (b) holds as well.

Indeed, if $p(\boldsymbol{y}) = 0$ there is nothing to prove, so assume $p(\boldsymbol{y}) > 0$. Then (d) implies $Q(\boldsymbol{y}, \boldsymbol{w}) \neq 0$. For each $\boldsymbol{x} \in \mathbb{R}^n$, define $\boldsymbol{z} = \boldsymbol{x} - t\boldsymbol{y}$ with $t = Q(\boldsymbol{x}, \boldsymbol{w})/Q(\boldsymbol{y}, \boldsymbol{w})$, which satisfies $Q(\boldsymbol{z}, \boldsymbol{w}) = 0$. Therefore, (d) gives

$$0 \geq p(\boldsymbol{z}) = p(\boldsymbol{x}) - 2tQ(\boldsymbol{x}, \boldsymbol{y}) + t^2 p(\boldsymbol{y}) \geq p(\boldsymbol{x}) - \frac{Q(\boldsymbol{x}, \boldsymbol{y})^2}{p(\boldsymbol{y})},$$

where last inequality is obtained by minimizing over t. This shows (b) holds. ■

Example 6.14 (The second-order cone). Let $\mathscr{C} = \mathbb{R}^{n+1}$ and $p(\boldsymbol{x}) = x_0^2 - \sum_{i=1}^n x_i^2 = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}$ for $\boldsymbol{A} = \operatorname{diag}(1, -1, \ldots, -1)$. Then p is hyperbolic with respect to $\boldsymbol{e} = (1, 0, \ldots, 0)^\mathsf{T}$ by Proposition 6.13(d), and its corresponding hyperbolicity cone is

$$\Lambda_+ = \left\{ \boldsymbol{x} \in \mathbb{R}^{n+1} : \sum\nolimits_{i=1}^n x_i^2 \leq x_0^2 \right\}.$$

This is the so-called *second-order cone*, the epigraph of the ℓ_2 norm on $\mathbb{R}^n$. ▪

Applying Proposition 6.13 to the multilinearization, we obtain the following result.

Theorem 6.15 ([Sch14, Thm. 5.5.3]). Let p be hyperbolic on $\mathscr{C}$ of degree d with hyperbolicity cone Λ_{++}, and let $\widetilde{p}\colon \mathscr{C}^d \to \mathbb{R}$ be its multilinearization. Fix $\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d \in \Lambda_{++}$.

(a) $\widetilde{p}(\boldsymbol{z}, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d) \geq 0$ for all $\boldsymbol{z} \in \Lambda_+$.
(b) For any $\boldsymbol{y} \in \Lambda_+$ and $\boldsymbol{x} \in \mathscr{C}$, we have

$$\widetilde{p}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d)^2 \geq \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d)\widetilde{p}(\boldsymbol{y}, \boldsymbol{y}, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d).$$

Proof. Corollary 6.12 shows that $\boldsymbol{x} \mapsto \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d)$ is hyperbolic with respect to any $\boldsymbol{z} \in \Lambda_{++}$, hence in particular, nonnegative on Λ_+. This shows (a).

The same corollary also shows that $\boldsymbol{x} \mapsto \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_d)$ is a hyperbolic quadratic on $\mathscr{C}$ with respect to any element in Λ_{++}, so Proposition 6.13 give part (b). ■

Corollary 6.16 (Alexandrov's mixed discriminant inequality). Let $\mathrm{D}\colon (\mathbb{H}_n)^n \to \mathbb{R}$ be the multilinearization of $p = \det$ on $\mathbb{H}_n$.

(a) $\mathrm{D}(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n) \geq 0$ whenever $\boldsymbol{C}_1 \in \mathbb{H}_n^+$ and $\boldsymbol{C}_i \in \mathbb{H}_n^{++}$ for all $i \geq 2$.
(b) For any $\boldsymbol{A} \in \mathbb{H}_n$, any $\boldsymbol{B} \in \mathbb{H}_n^+$, and any $\boldsymbol{C}_3, \ldots, \boldsymbol{C}_n \in \mathbb{H}_n^{++}$, we have:

$$\mathrm{D}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}_3, \ldots, \boldsymbol{C}_n)^2 \geq \mathrm{D}(\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{C}_3, \ldots, \boldsymbol{C}_n)\mathrm{D}(\boldsymbol{B}, \boldsymbol{B}, \boldsymbol{C}_3, \ldots, \boldsymbol{C}_n). \tag{6.4}$$

The function D is called the *mixed discriminant*, and inequality (6.4) is called Alexandrov's mixed discriminant inequality. One can also derive the equality cases for (6.4) from hyperbolicity, see [Sch14, Thm. 5.5.4].

The inequality (6.4) can be used to derive the Alexandrov–Fenchel inequality for mixed volumes, a fundamental result in convex geometry. See [SH19], who also give a more elementary proof of (6.4).

6.4 Semidefinite representability, and additive perturbation theory

An important subset of $\mathsf{Hyp}_d(\boldsymbol{e})$ for $\boldsymbol{e} \in \mathbb{R}^n$ consists of determinants of matrix pencils $p(\boldsymbol{x}) = \det\left(\sum_{i=1}^n x_i \boldsymbol{A}_i\right)$ with $\boldsymbol{A}_i \in \mathbb{H}_d$ such that $\sum_i e_i \boldsymbol{A}_i \succ \boldsymbol{0}$. The corresponding hyperbolicity cones are linear slices of the psd cone $\mathbb{H}_d^+$, namely, $\Lambda_+ = \{\boldsymbol{x} \in \mathscr{C} : \sum_i x_i \boldsymbol{A}_i \succeq \boldsymbol{0}\}$.

Such cones Λ_+ are said to be semidefinite-representable, and they are significant because optimizing a linear function over Λ_+ subject to linear equality constraints is a semidefinite program (SDP), which can be solved efficiently using interior-point methods. Those same methods can be used to efficiently solve linear optimization over any hyperbolicity cone [Gül97] (because $\log 1/p$ is a self-concordant barrier function for Λ_+), leading to so-called *hyperbolic programs*. A natural question is then whether the class of hyperbolic programs is more expressive than SDPs, that is, can every hyperbolic program be written as an SDP? This is the content of the general Lax conjecture [Reno6], which conjectures that for any hyperbolicity cone Λ_+ there exists $n \in \mathbb{N}$, a subspace $S \subset \mathbb{H}_n$, and a linear isomorphism $\mathcal{C} \xrightarrow{\sim} S$ sending Λ_+ to $S \cap \mathbb{H}_n^+$.

The special case of hyperbolicity cones in $\mathbb{R}^3$ (which was the original Lax conjecture) has been proved in [LPR05] using, according to [Sero9], a "deep result about Riemann surfaces by Helton–Vinnikov [HV07]". They show that for any hyperbolic polynomial p on $\mathbb{R}^3$ with respect to $(1,0,0)^\mathsf{T}$, there exist $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_d$ (where $d = \deg p$) satisfying

$$p(\xi_0, \xi_1, \xi_2) = p(1,0,0) \det(\xi_0 \mathbf{I} + \xi_1 \boldsymbol{A} + \xi_2 \boldsymbol{B}). \tag{6.5}$$

The corresponding statement for hyperbolic polynomials on $\mathbb{R}^n$ for $n \geq 4$ is false [LPR05] (but note that the existence of a representation of the form (6.5) is stronger than the general Lax conjecture).

The representation (6.5) can be used to show that all the additive perturbation theory for eigenvalues of matrices extend to eigenvalues of any hyperbolic polynomial. Indeed, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, consider the polynomial on $\mathbb{R}^3$ given by $q(\boldsymbol{\xi}) = p(\xi_0 \boldsymbol{e} + \xi_1 \boldsymbol{x} + \xi_2 \boldsymbol{y})$. This is hyperbolic with respect to $(1,0,0)^\mathsf{T}$, hence (6.5) yields $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_d$ satisfying

$$p(\xi_0 \boldsymbol{e} + \xi_1 \boldsymbol{x} + \xi_2 \boldsymbol{y}) = \det(\xi_0 \mathbf{I} + \xi_1 \boldsymbol{A} + \xi_2 \boldsymbol{B}).$$

This implies that $\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{\lambda}(\boldsymbol{A})$, $\boldsymbol{\lambda}(\boldsymbol{y}) = \boldsymbol{\lambda}(\boldsymbol{B})$, and $\boldsymbol{\lambda}(\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{\lambda}(\boldsymbol{A} + \boldsymbol{B})$, giving the general result:

Theorem 6.17 (Perturbation theory). Let p be a hyperbolic polynomial on $\mathcal{C}$ of degree d. Then *any* relation satisfied by the eigenvalues of arbitrary matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A}+\boldsymbol{B} \in \mathbb{H}_d$ is also satisfied by $\boldsymbol{\lambda}(\boldsymbol{x}), \boldsymbol{\lambda}(\boldsymbol{y}), \boldsymbol{\lambda}(\boldsymbol{x} + \boldsymbol{y})$, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$.

In particular, Lidskii's theorem holds: $\boldsymbol{\lambda}(\boldsymbol{x} + \boldsymbol{y}) - \boldsymbol{\lambda}(\boldsymbol{x}) \prec \boldsymbol{\lambda}(\boldsymbol{y})$.

A more elementary proof of Lidskii's theorem for hyperbolic polynomials is given in [Sero9]. We give a self-contained proof of the weaker statement $\boldsymbol{\lambda}(\boldsymbol{x} + \boldsymbol{y}) \prec \boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y})$ in Corollary 6.21 in the next section. There, we also use hyperbolicity to derive inequalities for eigenvalues of matrices, rather than the other way around.

6.5 Hyperbolicity and convexity of compositions

In this section, we consider hyperbolicity and convexity of the composition of functions and the eigenvalue map $\boldsymbol{\lambda}$. These properties enable us to prove inequalities for eigenvalues of matrices, in particular a weak form of Lidskii's theorem and Minkowski's determinant inequality.

Recall that the kth elementary symmetric polynomial on $\mathbb{R}^n$ is defined as

$$E_k(\boldsymbol{x}) = \sum\nolimits_{1 \leq i_1 < \ldots < i_k \leq n} x_{i_1} \cdots x_{i_k},$$

for $k = 1, \ldots, n$, and $E_0 = 1$. The ring of symmetric polynomials on $\mathbb{R}^n$ is generated by $E_0, \ldots, E_n$.

Lemma 6.18 If p is hyperbolic with respect to $\boldsymbol{e}$ with eigenvalue map $\boldsymbol{\lambda}$, then $E_k \circ \boldsymbol{\lambda}$ is a hyperbolic polynomial of degree k with respect to any $\boldsymbol{y} \in \Lambda_{++}$.

Proof. Note that

$$p(\boldsymbol{x} + t\boldsymbol{e}) = p(\boldsymbol{e}) \prod_{i=1}^{d} (\lambda_i(\boldsymbol{x}) + t) = p(\boldsymbol{e}) \sum_{i=0}^{d} E_i(\boldsymbol{\lambda}(\boldsymbol{x})) t^{d-i},$$

hence $E_k \circ \boldsymbol{\lambda} = \frac{1}{p(\boldsymbol{e})(d-k)!} p^{(d-k)}_{\boldsymbol{e},\dots,\boldsymbol{e}}$. The claim follows from Proposition 6.11. ∎

Corollary 6.19 The following are consequences of Lemma 6.18.

(a) E_k is hyperbolic on $\mathbb{R}^n$ with respect to any $\boldsymbol{y} \in \mathbb{R}^n_{++}$.
(b) The sum of all the eigenvalues $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda}$ is linear.

Proof. For (a), instantiate Lemma 6.18 on Example 6.2(a). For (b), note that $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda} = E_1 \circ \boldsymbol{\lambda}$ which is a polynomial of degree 1 by Lemma 6.18. ∎

The following theorem, taken from [Bau+01, Thm. 3.1], shows that compositions of symmetric hyperbolic polynomials with the eigenvalue map is hyperbolic.

Theorem 6.20 (Hyperbolicity of composition). Let p be a hyperbolic polynomial on $\mathscr{C}$ with respect to $\boldsymbol{e}$ of degree d, with eigenvalue map $\boldsymbol{\lambda}$. Let q be a *symmetric* hyperbolic polynomial on $\mathbb{R}^d$ with respect to $\mathbf{1}$ of degree k, with eigenvalue map $\boldsymbol{\mu}$. Then $q \circ \boldsymbol{\lambda}$ is a hyperbolic polynomial on $\mathscr{C}$ with respect to $\boldsymbol{e}$ of degree k, with eigenvalue map $\boldsymbol{\mu} \circ \boldsymbol{\lambda}$.

Proof. Since q is symmetric of degree k, it can be written as a polynomial in $E_1, \dots, E_k$, hence $q \circ \boldsymbol{\lambda}$ is a homogeneous polynomial of degree k by Lemma 6.18. Next, recall from (6.1) that $\boldsymbol{\lambda}(\boldsymbol{x} - t\boldsymbol{e}) = \boldsymbol{\lambda}(\boldsymbol{x}) - t\mathbf{1}$. Since q is hyperbolic with respect to $\mathbf{1}$, we have $(q \circ \boldsymbol{\lambda})(\boldsymbol{e}) = q(\mathbf{1}) \neq 0$ and

$$(q \circ \boldsymbol{\lambda})(\boldsymbol{x} - t\boldsymbol{e}) = q(\boldsymbol{\lambda}(\boldsymbol{x}) - t\mathbf{1}) = q(\mathbf{1}) \prod_j (\mu_j(\boldsymbol{\lambda}(\boldsymbol{x})) - t),$$

whose roots are $(\boldsymbol{\mu} \circ \boldsymbol{\lambda})(\boldsymbol{x})$, which are all real. ∎

We can use Theorem 6.20 to prove a weaker form of Lidskii's theorem:

Corollary 6.21 Let p be a hyperbolic polynomial on $\mathscr{C}$ with eigenvalue map $\boldsymbol{\lambda}$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{C}$, we have $\boldsymbol{\lambda}(\boldsymbol{x} + \boldsymbol{y}) \prec \boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y})$.

Proof. For each $1 \leq k \leq d$ let

$$q_k(u) = \prod_{1 \leq i_1 < \dots < i_k \leq d} \sum_{j=1}^{k} u_{i_j}.$$

Note that q_k is a symmetric polynomial, hyperbolic with respect to $\mathbf{1}$ with eigenvalues $\boldsymbol{\mu}(\boldsymbol{u}) = \left(\sum_{j=1}^{k} u_{i_j}\right)_{1 \leq i_1 < \dots < i_k \leq d}$. Theorem 6.20 then shows that $q_k \circ \boldsymbol{\lambda}$ is hyperbolic with respect to $\boldsymbol{e}$, and its largest eigenvalue is $\sum_{i=1}^{k} \lambda_i$. By (6.1) and Theorem 6.9, this largest eigenvalue is positively homogeneous and convex. Thus, $\sum_{i=1}^{k} \lambda_i$ is subadditive for all k, implying $\boldsymbol{\lambda}(\boldsymbol{x}+\boldsymbol{y}) \prec_w \boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y})$. Finally, $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda}(\boldsymbol{x}+\boldsymbol{y}) = \mathbf{1}^\mathsf{T}(\boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y}))$ by Corollary 6.19(b). ∎

Using the above weak form of Lidskii's theorem, we can prove that the composition of convex and symmetric functions with the eigenvalue map is also convex, a fact that yields in particular Minkowski's determinant inequality. The next result appears in [Bau+01, Thm. 3.9].

Theorem 6.22 (Convexity of compositions). Let p be a hyperbolic polynomial on $\mathscr{C}$ with eigenvalue map $\boldsymbol{\lambda}$. If $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex and symmetric, then $f \circ \boldsymbol{\lambda}$ is convex on $\mathscr{C}$.

Recall that $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is the extended reals.

Proof. By Corollary 6.21 and the positive homogeneity of λ, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{C}$, any $\tau \in [0,1]$ and $\bar{\tau} = 1 - \tau$, we have

$$\boldsymbol{\lambda}(\tau \boldsymbol{x} + \bar{\tau}\boldsymbol{y}) \prec \boldsymbol{\lambda}(\tau \boldsymbol{x}) + \boldsymbol{\lambda}(\bar{\tau}\boldsymbol{y}) = \tau \boldsymbol{\lambda}(\boldsymbol{x}) + \bar{\tau}\boldsymbol{\lambda}(\boldsymbol{y}).$$

Since f is convex and symmetric, it is isotone, hence

$$(f \circ \boldsymbol{\lambda})(\tau \boldsymbol{x} + \bar{\tau}\boldsymbol{y}) \leq f(\tau \boldsymbol{\lambda}(\boldsymbol{x}) + \bar{\tau}\boldsymbol{\lambda}(\boldsymbol{y})) \leq \tau (f \circ \boldsymbol{\lambda})(\boldsymbol{x}) + \bar{\tau}(f \circ \boldsymbol{\lambda})(\boldsymbol{y}),$$

where for the last inequality we used the convexity of f. ∎

We can now use hyperbolicity to derive inequalities for eigenvalues of matrices.

Corollary 6.23 (Gårding's inequality). Let p be a hyperbolic polynomial on $\mathscr{C}$ with respect to $\boldsymbol{e}$ of degree d. If $p(\boldsymbol{e}) > 0$, then $\boldsymbol{x} \mapsto p(\boldsymbol{x})^{1/d}$ is superlinear on Λ_+. In particular, we have

$$\operatorname{tr}(\wedge^d(\boldsymbol{A} + \boldsymbol{B}))^{1/d} \geq \operatorname{tr}(\wedge^d \boldsymbol{A})^{1/d} + \operatorname{tr}(\wedge^d \boldsymbol{B})^{1/d}, \tag{6.6}$$

A function is *superlinear* if it is superadditive and positively homogeneous.

for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_n^+$ and any $n \in \mathbb{N}$.

Setting $n = d$ in (6.6), we get Minkowski's determinant inequality $\det(\boldsymbol{A}+\boldsymbol{B})^{1/d} \geq \det(\boldsymbol{A})^{1/d} + \det(\boldsymbol{B})^{1/d}$, valid for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_d^+$.

Proof. Note that $p(\boldsymbol{x}) = p(\boldsymbol{e}) \prod_i \lambda_i(\boldsymbol{x}) = p(\boldsymbol{e})(E_d \circ \boldsymbol{\lambda})(\boldsymbol{x})$. The function $\boldsymbol{x} \mapsto E_d(\boldsymbol{x})^{1/d}$ is the geometric mean, which is concave on $\mathbb{R}_+^d$ (exercise) and symmetric. Hence, $p^{1/d}$ is concave on Λ_+ by Theorem 6.22 (with $f = E_d^{1/d}$ on $\mathbb{R}_+^d$ and $+\infty$ otherwise). Since p is homogeneous of degree d, we conclude that $p^{1/d}$ is positively homogeneous. Thus, it is superlinear on Λ_+.

For the second claim, Lemma 6.18 applied to $p = \det$ and $\mathscr{C} = \mathbb{H}_n$ shows that $E_d \circ \boldsymbol{\lambda}$ is hyperbolic of degree d with respect to $\mathbf{I}_n$. Therefore, the map $\boldsymbol{X} \mapsto (E_d(\boldsymbol{\lambda}(\boldsymbol{X})))^{1/d} = \operatorname{tr}\left(\wedge^d \boldsymbol{X}\right)^{1/d}$ is superlinear on $\Lambda_+ = \mathbb{H}_n^+$, as desired. ∎

Exercise 6.24 Show that $\boldsymbol{x} \mapsto \left(\prod_{i=1}^d x_i\right)^{1/d}$ is concave on $\mathbb{R}_+^d$.

6.6 Euclidean structure

The eigenvalue map λ induces a psd symmetric bilinear form, which is an inner product for all our examples. It generalizes the Frobenius inner product on symmetric matrices, and satisfies a sharpened Cauchy–Schwarz inequality generalizing von Neumann's trace inequality.

Definition 6.25 Define $\|\cdot\| : \mathscr{C} \to \mathbb{R}_+$ by $\|\boldsymbol{x}\| = \|\boldsymbol{\lambda}(\boldsymbol{x})\|_2$, and define $\langle \cdot, \cdot \rangle : \mathscr{C}^2 \to \mathbb{R}$ by the polarization identity

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \frac{1}{4}\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \frac{1}{4}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Proposition 6.26 ([Bau+01, Thm. 4.2]). The function $\|\cdot\|$ is a seminorm on $\mathscr{C}$, and $\langle \cdot, \cdot \rangle$ is a positive-semidefinite bilinear form on $\mathscr{C}$.

A seminorm is sublinear but may be zero on non-zero vectors.

Proof. The function $\|\cdot\|$ is clearly nonnegative, and it is absolutely homogeneous by (6.1). It is convex by Theorem 6.22 as it is the composition of the convex and symmetric ℓ_2 norm on $\mathbb{R}^d$ with the eigenvalue map λ. Thus, $\|\cdot\|$ is subadditive. Moreover, $\|\boldsymbol{x}\|^2 = \sum_i \lambda_i(\boldsymbol{x})^2 = (E_1(\boldsymbol{\lambda}(\boldsymbol{x})))^2 - 2E_2(\boldsymbol{\lambda}(\boldsymbol{x}))$, which is a homogeneous quadratic polynomial on $\mathscr{C}$ by Lemma 6.18. Any quadratic seminorm is induced by a psd symmetric bilinear form given by the polarization identity. ∎

In order to make $\|\cdot\|$ a norm and $\langle\cdot,\cdot\rangle$ an inner-product, we must require an additional property from p.

Definition 6.27 A hyperbolic polynomial p with respect to $\boldsymbol{e}$ is *complete* if $\{\boldsymbol{x} \in \mathscr{C} : \boldsymbol{\lambda}(\boldsymbol{x}) = 0\} = \{\boldsymbol{0}\}$.

Proposition 6.28 The following are equivalent.

(a) p is complete.
(b) Λ_+ is pointed.
(c) $\|\boldsymbol{x}\| = \|\boldsymbol{\lambda}(\boldsymbol{x})\|_2$ is a norm.

A closed cone K is *pointed* if $\mathsf{K} \cap (-\mathsf{K}) = \{\boldsymbol{0}\}$, or equivalently, if K contains no lines.

Proof. If $\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{0}$ then $\boldsymbol{\lambda}(-\boldsymbol{x}) = \boldsymbol{0}$ as well by (6.1), so $\boldsymbol{x} \in \Lambda_+ \cap (-\Lambda_+)$. Conversely, if $\boldsymbol{x} \in \Lambda_+ \cap (-\Lambda_+)$ then $\lambda_{\min}(\boldsymbol{x}) \geq 0$ and $\lambda_{\min}(-\boldsymbol{x}) = -\lambda_{\max}(\boldsymbol{x}) \geq 0$, hence $0 \leq \lambda_{\min}(\boldsymbol{x}) \leq \lambda_{\max}(\boldsymbol{x}) \leq 0$ and $\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{0}$. This shows (a) and (b) are equivalent. Finally, (a) and (c) are equivalent because $\|\boldsymbol{x}\|$ is always a seminorm, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{0}$. ∎

The polynomials in Example 6.2 and Example 6.14 are complete, since $\mathbb{R}^n_+$, $\mathbb{H}^+_n$, and the second-order cone are pointed. The above norm is also the Hessian norm used in interior-point methods with the barrier $\log 1/p$ [Bau+01, Rmk. 4.3].

The inner-product in Definition 6.25 satisfies a sharpened Cauchy–Schwarz inequality [Bau+01, Prop. 4.4].

Proposition 6.29 (Refined Cauchy–Schwarz). For any $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{C}$,

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq \langle \boldsymbol{\lambda}(\boldsymbol{x}), \boldsymbol{\lambda}(\boldsymbol{y})\rangle_{\ell_2} \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|.$$

Proof. Corollary 6.21 shows that $\boldsymbol{\lambda}(\boldsymbol{x}+\boldsymbol{y}) \prec \boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y})$. Since the ℓ_2 norm is convex and symmetric, it is isotone, hence $\|\boldsymbol{\lambda}(\boldsymbol{x}+\boldsymbol{y})\|_{\ell_2} \leq \|\boldsymbol{\lambda}(\boldsymbol{x}) + \boldsymbol{\lambda}(\boldsymbol{y})\|_{\ell_2}$. Squaring both sides, this is equivalent to

$$\begin{aligned} 2\langle \boldsymbol{\lambda}(\boldsymbol{x}), \boldsymbol{\lambda}(\boldsymbol{y})\rangle_{\ell_2} &\geq \|\boldsymbol{\lambda}(\boldsymbol{x}+\boldsymbol{y})\|^2_{\ell_2} - \|\boldsymbol{\lambda}(\boldsymbol{x})\|^2_{\ell_2} - \|\boldsymbol{\lambda}(\boldsymbol{y})\|^2_{\ell_2} \\ &= \|\boldsymbol{x}+\boldsymbol{y}\|^2 - \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 = 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle, \end{aligned}$$

giving the first claimed inequality. The second is Cauchy–Schwarz for the ℓ_2 norm. ∎

In Example 6.2(a), the sharpened Cauchy–Schwarz inequality reduces to Chebyshev's rearrangement inequality, and in Example 6.2(b) it reduces to von Neumann's trace inequality. Note also that $\langle \boldsymbol{x}, \boldsymbol{e}\rangle = \sum_i \lambda_i(\boldsymbol{x})$, generalizing the trace $\langle \boldsymbol{X}, \mathbf{I}_d\rangle_{\mathrm{F}}$ on $\mathbb{H}_d$ and making its linearity (proved separately in Corollary 6.19(b)) apparent.

Lecture bibliography

[Bau+01] H. H. Bauschke et al. "Hyperbolic polynomials and convex analysis". In: *Canad. J. Math.* 53.3 (2001), pages 470–488. DOI: 10.4153/CJM-2001-020-6.

[Går51] L. Gårding. "Linear hyperbolic partial differential equations with constant coefficients". In: *Acta Mathematica* 85.none (1951), pages 1 –62. DOI: 10.1007/BF02395740.

[Gül97] O. Güler. "Hyperbolic Polynomials and Interior Point Methods for Convex Programming". In: *Mathematics of Operations Research* 22.2 (1997), pages 350–377.

[HV07] J. W. Helton and V. Vinnikov. "Linear matrix inequality representation of sets". In: *Communications on Pure and Applied Mathematics* 60.5 (2007), pages 654–674. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20155. DOI: https://doi.org/10.1002/cpa.20155.

[LPR05] A. Lewis, P. Parrilo, and M. Ramana. "The Lax conjecture is true". In: *Proceedings of the American Mathematical Society* 133.9 (2005), pages 2495–2499.

[MSS14] A. W. Marcus, D. A. Spielman, and N. Srivastava. "Ramanujan graphs and the solution of the Kadison-Singer problem". In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. III*. Kyung Moon Sa, Seoul, 2014, pages 363–386.

[Ren06] J. Renegar. "Hyperbolic Programs, and Their Derivative Relaxations". In: *Found. Comput. Math.* 6.1 (2006), pages 59–79. DOI: 10.1007/s10208-004-0136-z.

[Sch14] R. Schneider. *Convex bodies: the Brunn–Minkowski theory*. 151. Cambridge university press, 2014. DOI: 10.1017/CBO9781139003858.

[Ser09] D. Serre. "Weyl and Lidskiĭ inequalities for general hyperbolic polynomials". In: *Chinese Annals of Mathematics, Series B* 30.6 (2009), pages 785–802.

[SH19] Y. Shenfeld and R. van Handel. "Mixed volumes and the Bochner method". In: *Proc. Amer. Math. Soc.* 147.12 (2019), pages 5385–5402. DOI: 10.1090/proc/14651.

7. The Laplace Transform Method for Matrices

Date: 9 March 2022 **Author:** Elvira Moreno

Agenda:
1. The Laplace transform method
2. The Laplace transform bound for sums of random matrices
3. The matrix Chernoff bound
4. Sparsification via random sampling

In previous lectures, we have observed how mathematical theory concerning scalars can be extended to the realm of matrices. In this lecture, we explore how concepts in probability, originally defined to study large-deviation behavior of sequences of random variables and their sums, can be paralleled to study the behavior of the extremal eigenvalues of independent sums of random matrices.

We begin our discussion with a matrix version of the Laplace transform method, which has proven to be an invaluable tool for producing tail bounds for the sums of random variables. We then present a Laplace-tranform-like bound for the sum of independent random matrices due to Tropp [Tro11], who builds on Lieb's [Lie73b] work on convex trace functions to extend the classical result to the matrix setting. As a testament to the power of Tropp's result, we show how it can be employed to prove a matrix version of the Chernoff bound. Finally, we discuss an application of the matrix Chernoff bound in spectral graph theory, where it is used to show that any undirected, connected graph can be well approximated by a sparse graph with high probability.

7.1 The Laplace transform method

The Laplace transform method is an invaluable tool for producing bounds on the tail probabilities of random variables and their sums. In this section, we recall the statement of the Laplace transform bound and we show an extension of this result to the matrix setting due to Ahlswede and Winter [AW01].

7.1.1 The Laplace transform method: Scalar case

We begin by recalling the classical definitions for the moment generating function and the cumulant generating function of a random variable.

Definition 7.1 (Mgf and cgf). Let X be a real random variable. The *moment generating function (mgf)* of X is defined by

$$m_X(\theta) := \mathbb{E}\exp(\theta X) \quad \text{for each } \theta \in \mathbb{R}.$$

The *cumulant generating function (cgf)* of X is defined by

$$\xi_X(\theta) := \log \mathbb{E}\exp(\theta X) \quad \text{for each } \theta \in \mathbb{R}.$$

Next, we state the Laplace transform method, which illustrates how the cgf of a real random variable can be used to establish bounds on its tail probabilities.

Theorem 7.2 (Laplace transform method). Let X be a real random variable. Then, for

each $t \in \mathbb{R}$,

$$\mathbb{P}\{X \geq t\} \leq \inf_{\theta>0} \exp\left(-\theta t + \xi_X(\theta)\right);$$
$$\mathbb{P}\{X \leq t\} \leq \inf_{\theta<0} \exp\left(-\theta t + \xi_X(\theta)\right).$$

Exercise 7.3 Provide a proof for Theorem 7.2. **Hint:** Markov's inequality.

The previous result can be readily employed to produce tail bounds for the sum of independent random variables.

Corollary 7.4 (Laplace transform method for the independent sum). Let $\{X_i\}_{i=1}^k$ be an independent family of real random variables in L_∞. Then

$$\mathbb{P}\left\{\sum_{i=1}^k X_i \geq t\right\} \leq \inf_{\theta>0} \exp\left(-\theta t + \sum_{i=1}^k \xi_{X_i}(\theta)\right);$$
$$\mathbb{P}\left\{\sum_{i=1}^k X_i \leq t\right\} \leq \inf_{\theta<0} \exp\left(-\theta t + \sum_{i=1}^k \xi_{X_i}(\theta)\right).$$

Proof. Let $Y = \sum_{i=1}^k X_i$. By independence of the random variables X_i, we have that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \sum_{i=1}^k X_i} = \log \mathbb{E}\left(\prod_{i=1}^k \mathrm{e}^{\theta X_i}\right) = \log \prod_{i=1}^k \mathbb{E}\, \mathrm{e}^{\theta X_i} = \sum_{i=1}^k \log \mathbb{E}\, \mathrm{e}^{\theta X_i}$$

for all $\theta \in \mathbb{R}$. So $\xi_Y = \sum_{i=1}^k \xi_{X_i}$, and the result follows by replacing Y and ξ_Y in Theorem 7.2. ∎

In the next section, we extend these concepts and results to the matrix case.

7.1.2 The Laplace transform method: Matrix case

We begin by defining the concepts of moment generating function and cumulant generating function of a random matrix in analogy to their classical definitions for real random variables.

Definition 7.5 (Matrix mgf and cgf). Let $\boldsymbol{X} \in \mathbb{M}_n$ be a random self-adjoint matrix. Define the *moment generating function (mgf)* of $\boldsymbol{X}$ by

$$M_{\boldsymbol{X}}(\theta) := \mathbb{E} \exp(\theta \boldsymbol{X}) \quad \text{for } \theta \in \mathbb{R}.$$

Define the *cumulant generating function (cgf)* of $\boldsymbol{X}$ by

$$\Xi_{\boldsymbol{X}}(\theta) := \log \mathbb{E} \exp(\theta \boldsymbol{X}) \quad \text{for } \theta \in \mathbb{R}.$$

Recall from Lecture 13, that the matrix exponential is neither monotone nor convex. This means that, unlike the scalar mgf, the matrix mgf does not enjoy these properties.

Note that when $n = 1$, the definitions for the mgf and the cgf of a random matrix $\boldsymbol{X} \in \mathbb{M}_n$ coincide with their classical definition in the setting of real random variables.

As in the scalar case, the cgf of a random matrix can be employed to provide bounds on the tail probabilities of its largest eigenvalue. An extension of Theorem 7.2 to the matrix setting was first provided by Ahlswede and Winter in [AW01]. We will instead consider the following variant by Oliveira [Oli10]. The proof we present is that by Tropp [Tro11].

Theorem 7.6 (Laplace transform method: Matrix case). Let $\boldsymbol{X}$ be a random self-adjoint

matrix and let $t \in \mathbb{R}$. Then,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \geq t\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta t}\, \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \boldsymbol{X}};$$
$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{X}) \leq t\} \leq \inf_{\theta<0} \mathrm{e}^{-\theta t}\, \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \boldsymbol{X}}.$$

Proof. (Upper Bound) For $\theta > 0$, Markov's inequality yields

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \geq t\} = \mathbb{P}\{\lambda_{\max}(\theta\boldsymbol{X}) \geq \theta t\} = \mathbb{P}\left\{\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})} \leq \mathrm{e}^{\theta t}\right\} \leq \frac{\mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})}}{\mathrm{e}^{\theta t}}.$$

The first identity follows because the eigenvalue map is positively homogeneous. The second is due to the fact that the scalar exponential function is monotone increasing.

Next, recall from Definition 13.2 in Lecture 13, that matrix functions are defined via the spectral resolution. This means that the eigenvalues of the matrix $\mathrm{e}^{\theta\boldsymbol{X}}$ correspond to the exponentials of the eigenvalues of $\theta\boldsymbol{X}$, and we have that

$$\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})} = \lambda_{\max}(\mathrm{e}^{\theta\boldsymbol{X}}) \leq \mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{X}}.$$

The inequality is due to the fact that all eigenvalues of the matrix $\mathrm{e}^{\theta\boldsymbol{X}}$ are strictly positive, so its trace dominates its largest eigenvalue. Hence, for $\theta > 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \geq t\} \leq \frac{\mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})}}{\mathrm{e}^{\theta t}} \leq \frac{\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{X}}}{\mathrm{e}^{\theta\boldsymbol{X}}}.$$

The result is obtained by taking the infimum over all strictly positive θ. ∎

Exercise 7.7 (Laplace transform method: Lower bound). Provide a proof for the lower bound of Theorem 7.2.

Paralleling the scalar case, we would like to employ the Laplace transform method for random matrices, Theorem 7.6, to establish tail bounds for the largest eigenvalue of the sum of independent random matrices. It is here where we encounter the biggest challenge. Indeed, as is evidenced in the proof of Corollary 7.4, the natural step from the Laplace transform bound to a Laplace-transform-like statement regarding the tail decay of the sum of random variables, strongly depends on the additivity of the cgf. This property of the scalar cgf is not transferred into the matrix case. The reason for this is that, unlike the scalar exponential, the matrix exponential function does not convert sums of matrices into products. That is, the identity

$$\mathrm{e}^{\boldsymbol{X}_1+\boldsymbol{X}_2} = \mathrm{e}^{\boldsymbol{X}_1}\mathrm{e}^{\boldsymbol{X}_2}$$

does not always hold.

In the next section, we show how a result on the concavity of a particular trace function can be leveraged to circumvent this issue and establish the desired extension of Corollary 7.4 to the matrix setting.

7.2 Laplace transform tail bound for sums of random matrices

To move forward in our aim of extending the Laplace transform method to the matrix setting, we require a result by Lieb on the concavity of a particular trace function. This result is one of his many insights concerning the convexity of trace functions. We refer the interested reader to [Lie73b], where Theorem 7.8 appears as Theorem 6. We present a proof of the result by Tropp [Tro12], who derives it from joint convexity of the quantum relative entropy function.

Theorem 7.8 (Lieb). Fix a self-adjoint matrix $\boldsymbol{H}$. The function

$$\boldsymbol{A} \longmapsto \operatorname{tr} \exp(\boldsymbol{H} + \log(\boldsymbol{A}))$$

is concave on the PSD cone.

Before proceeding with the proof of Lieb's Theorem, it is important to recall the definition of the quantum relative entropy function.

Definition 7.9 (Quantum relative entropy). Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be positive-definite matrices. The *quantum relative entropy* of $\boldsymbol{X}$ with respect to $\boldsymbol{Y}$ is defined by

$$\mathrm{D}(\boldsymbol{X}; \boldsymbol{Y}) := \operatorname{tr}(\boldsymbol{X} \log \boldsymbol{X} - \boldsymbol{X} \log \boldsymbol{Y} - (\boldsymbol{X} - \boldsymbol{Y})).$$

In Lecture 17, we defined the (Umegaki) quantum relative entropy of two density matrices $\boldsymbol{\varrho}, \boldsymbol{\nu} \in \boldsymbol{\Delta}_n$ as

$$\mathrm{S}(\boldsymbol{\varrho}; \boldsymbol{\nu}) := \operatorname{tr}(\boldsymbol{\varrho} \log \boldsymbol{\varrho} - \boldsymbol{\varrho} \log \boldsymbol{\nu}),$$

and we proved that it is both nonnegative and jointly convex. These two properties also hold for the quantum relative entropy function as defined by 7.9 and play a central role in the proof of Theorem 7.8.

Proof. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be positive definite. By nonnegativity of the quantum relative entropy function, we have that

$$\operatorname{tr}(\boldsymbol{Y}) \geq \operatorname{tr}(\boldsymbol{X} \log \boldsymbol{Y} - \boldsymbol{X} \log \boldsymbol{X} + \boldsymbol{X}).$$

Since equality is attained when $\boldsymbol{X} = \boldsymbol{Y}$, we deduce that

$$\operatorname{tr}(\boldsymbol{Y}) = \max_{\boldsymbol{X} > \mathbf{0}} \operatorname{tr}(\boldsymbol{X} \log \boldsymbol{Y} - \boldsymbol{X} \log \boldsymbol{X} + \boldsymbol{X}).$$

Replacing $\boldsymbol{Y} = \exp(\boldsymbol{H} + \log \boldsymbol{A})$ in the identity above yields

$$\begin{aligned} \operatorname{tr}(\exp(\boldsymbol{H} + \log \boldsymbol{A})) &= \max_{\boldsymbol{X} > \mathbf{0}} \operatorname{tr}(\boldsymbol{X} \log \exp(\boldsymbol{H} + \log \boldsymbol{A}) - \boldsymbol{X} \log \boldsymbol{X} + \boldsymbol{X}) \\ &= \max_{\boldsymbol{X} > \mathbf{0}} \operatorname{tr}(\boldsymbol{X}\boldsymbol{H}) - \operatorname{tr}(\boldsymbol{X} \log \boldsymbol{X} - \boldsymbol{X} \operatorname{tr}(\boldsymbol{A}) - \boldsymbol{X} + \boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A}) \\ &= \max_{\boldsymbol{X} > \mathbf{0}} \operatorname{tr}(\boldsymbol{X}\boldsymbol{H}) - \mathrm{D}(\boldsymbol{X}; \boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A}). \end{aligned}$$

Observe that $\boldsymbol{Y} > \mathbf{0}$, as matrix functions are defined via the spectral resolution and $\mathrm{e}^{\lambda} > 0$ for all $\lambda \in \mathbb{R}$. So the substitution is valid. Furthermore, joint convexity of D and linearity of the trace imply joint concavity of the function $(\boldsymbol{X}, \boldsymbol{A}) \mapsto \operatorname{tr}(\boldsymbol{X}\boldsymbol{H}) - \mathrm{D}(\boldsymbol{X}; \boldsymbol{A}) - \operatorname{tr}(\boldsymbol{A})$. The result follows from the fact that the partial maximization of a jointly concave function is concave (see Exercise 7.10). ∎

Exercise 7.10 (Partial maximization of a jointly concave function). Let $f(\cdot, \cdot)$ be a jointly concave function. Then the map $y \mapsto \sup_x f(x, y)$ is concave.

Corollary 7.11 (Expectation of the trace exponential). Let $\boldsymbol{H}$ be a fixed self-adjoint matrix. Then, for any random self-adjoint matrix $\boldsymbol{X}$,

$$\mathbb{E} \operatorname{tr} \exp(\boldsymbol{H} + \boldsymbol{X}) \leq \operatorname{tr} \exp(\boldsymbol{H} + \log(\mathbb{E}\, \mathrm{e}^{\boldsymbol{X}})).$$

Exercise 7.12 Provide a proof for Corollary 7.11. **Hint:** Employ Theorem 7.8 with $\boldsymbol{A} = \mathrm{e}^{\boldsymbol{X}}$ and invoke Jensen's inequality.

Recall from the end of Section 1 that nonadditivity of the cgf was the mayor obstacle for establishing a matrix parallel to Corollary 7.4. Tropp gives rest to this issue by proving the following subadditivity result for matrix cgfs [Tro11].

Lemma 7.13 (Subadditivity of Matrix cgfs). Consider an independent sequence $\{\boldsymbol{X}_i\}_{i=1}^k$ of random, self-adjoint matrices. Then

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^k \theta\boldsymbol{X}_i\right) \le \mathrm{tr}\exp\left(\sum\nolimits_{i=1}^k \log\mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{X}_i}\right) \quad \text{for } \theta\in\mathbb{R}.$$

Proof. Assume without loss of generality that $\theta = 1$. For $i = 1,\dots,k$, we adopt the conventions

$$\mathbb{E}_i \coloneqq \mathbb{E}\left[\cdot \mid \boldsymbol{X}_1,\dots,\boldsymbol{X}_i\right] \quad \text{and} \quad \Xi_i \coloneqq \log\mathbb{E}\,\mathrm{e}^{\boldsymbol{X}_i}.$$

Using the tower property of the conditional expectation we can write

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^k \boldsymbol{X}_i\right) = \mathbb{E}_0\,\mathbb{E}_1\cdots\mathbb{E}_{k-1}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-1}\boldsymbol{X}_i + \boldsymbol{X}_k\right).$$

Invoking Corollary 7.11 with $\boldsymbol{H} = \sum_{i=1}^{k-1}\boldsymbol{X}_i$ yields

$$\begin{aligned}
&\mathbb{E}_0\cdots\mathbb{E}_{k-1}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-1}\boldsymbol{X}_i + \boldsymbol{X}_k\right)\\
&\qquad\le \mathbb{E}_0\cdots\mathbb{E}_{k-2}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-1}\boldsymbol{X}_i + \log\mathbb{E}_{k-1}\,\mathrm{e}^{\boldsymbol{X}_k}\right)\\
&\qquad= \mathbb{E}_0\cdots\mathbb{E}_{k-2}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-1}\boldsymbol{X}_i + \Xi_k\right).\\
&\qquad= \mathbb{E}_0\cdots\mathbb{E}_{k-2}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-2}\boldsymbol{X}_i + \boldsymbol{X}_{k-1} + \Xi_k\right).
\end{aligned}$$

Taking $\boldsymbol{H} = \sum_{i=1}^{k-2}\boldsymbol{X}_i + \Xi_k$ and invoking Corollary 7.11 once more, we obtain

$$\begin{aligned}
&\mathbb{E}_0\cdots\mathbb{E}_{k-2}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-2}\boldsymbol{X}_i + \boldsymbol{X}_{k-1} + \Xi_k\right)\\
&\qquad\le \mathbb{E}_0\cdots\mathbb{E}_{k-3}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-2}\boldsymbol{X}_i + \Xi_{k-1} + \Xi_k\right)\\
&\qquad= \mathbb{E}_0\cdots\mathbb{E}_{k-3}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{k-3}\boldsymbol{X}_i + \boldsymbol{X}_{\boldsymbol{k-2}} + \Xi_{k-1} + \Xi_k\right).
\end{aligned}$$

Repeating this procedure recursively with $\boldsymbol{H} = \sum_{i=1}^{k-2}\boldsymbol{X}_i + \sum_{j=m+1}^{k}\Xi_k$ for $m = k, k-1,\dots,1$ yields

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^k \boldsymbol{X}_i\right) \le \mathrm{tr}\sum\nolimits_{i=1}^k \Xi_k,$$

from which the result follows. ∎

Having established subadditivity of the matrix cgf, the following tail bound can be derived from the first bound in Theorem 7.6 [Tro11].

Theorem 7.14 (Tail bounds for independent sums of random matrices). Consider an independent sequence $\{\boldsymbol{X}_i\}_{i=1}^k$ of random, self-adjoint matrices. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \geq t\right\} \leq \inf_{\theta>0}\left\{\mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\sum_{i=1}^k \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_i}\right)\right\};$$
$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \leq t\right\} \leq \inf_{\theta<0}\left\{\mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\sum_{i=1}^k \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_i}\right)\right\}.$$

Proof. Follows directly from subadditivity of matrix cgfs, Lemma 7.13, and Theorem 7.6. ∎

We also consider the following Corollary, as it will be useful in the next section when we prove an extension of the Chernoff bound to the matrix setting. We refer the reader to [Tro11, Corollary 3.9] for a proof.

Corollary 7.15 Consider an independent sequence $\{\boldsymbol{X}_i\}_{i=1}^k$ of random, self-adjoint matrices in $\mathbb{M}_n$. For all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \geq t\right\} \leq n \inf_{\theta>0} \exp\left(-\theta t + k \log \lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^k \mathbb{E}\,\mathrm{e}^{\boldsymbol{X}_i}\right)\right).$$

It is worth mentioning that these results can be generalized to inequalities involving the singular values of random rectangular matrices by means of their *self-adjoint dilation*. Refer to [Tro11] for more details.

7.3 The matrix Chernoff bound

Recall Chernoff's inequality for bounded, positive random variables.

Theorem 7.16 (Chernoff's inequality). Consider an independent sequence $\{X_i\}_{i=1}^k$ of random variables satisfying $X_i \in [0,1]$ for $i = 1,\dots,k$, and let $\mu := \sum_{i=1}^k \mathbb{E}\,X_i$. Then,

$$\mathbb{P}\left\{\sum_{i=1}^k X_i \geq (1+\delta)\mu\right\} \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \text{ for } \delta \geq 0;$$
$$\mathbb{P}\left\{\sum_{i=1}^k X_i \leq (1-\delta)\mu\right\} \leq \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \text{ for } \delta \in [0,1].$$

Next, we establish a similar result for controlling the extremal eigenvalues of the sum of independent random matrices that satisfy some additional properties. For this purpose, we first need a simple lemma.

Lemma 7.17 (Chernoff mgf). Let $\boldsymbol{X}$ be a random positive-definite matrix such that $\lambda_{\max}(\boldsymbol{X}) \leq 1$. Then,

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + (\mathrm{e}^{\theta} - 1)(\mathbb{E}\,\boldsymbol{X}).$$

Exercise 7.18 Provide a proof for Lemma 7.17. **Hint:** First establish the result for the scalar case by bounding the exponential on $[0,1]$ by a straight line.

The following theorem is presented as Corollary 5.2 in [Tro11], where the reader can find a stronger version of the Chernoff bound [Tro11, Theorem 5.1].

Theorem 7.19 (Matrix Chernoff bound). Consider an independent sequence $\{\boldsymbol{X}_i\}_{i=1}^k$ of random, self-adjoint matrices in $\mathbb{M}_n$ satisfying $\boldsymbol{X}_i \succcurlyeq \mathbf{0}$ and $\lambda_{\max}(\boldsymbol{X}_i) \le R$ almost surely, for each $i \in \{1,\dots,k\}$. Define

$$\mu_{\min} := \lambda_{\min}\left(\sum_{i=1}^k \mathbb{E}\,\boldsymbol{X}_i\right) \quad \text{and} \quad \mu_{\max} := \lambda_{\max}\left(\sum_{i=1}^k \mathbb{E}\,\boldsymbol{X}_i\right).$$

Then,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \ge (1+\delta)\mu_{\max}\right\} \le n\left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \text{for } \delta \ge 0;$$

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \le (1-\delta)\mu_{\min}\right\} \le n\left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1].$$

Proof. For each $i = 1,\dots,k$, let $\boldsymbol{Y}_i = \frac{1}{R}\boldsymbol{X}_i$. Then $\lambda_{\max}(\boldsymbol{Y}_i) \le 1$ for every $i \in \{1,\dots,k\}$, and for $t > 0$ we have that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) \ge t\right\} = \mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{Y}_i\right) \ge \bar{t}\right\}$$

where $\bar{t} = \frac{t}{R}$. By Lemma 7.17, $\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{Y}_i} \preccurlyeq \mathbf{I} + (\mathrm{e}^{\theta} - 1)(\mathbb{E}\,\boldsymbol{Y}_i)$ for all i. So Weyl's monotonicity principle implies that $\lambda_{\max}(\mathrm{e}^{\theta \boldsymbol{Y}_i}) \le \lambda_{\max}(\mathbf{I} + (\mathrm{e}^{\theta} - 1)(\mathbb{E}\,\boldsymbol{Y}_i))$ for all i, and Corollary 7.15 yields

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^k \boldsymbol{Y}_i\right) \ge \bar{t}\right\} &\le n \exp\left(-\theta\bar{t} + k \log \lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^k \mathbb{E}\,\mathrm{e}^{\boldsymbol{Y}_i}\right)\right) \\
&\le n \exp\left(-\theta\bar{t} + k \log \lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^k \mathbf{I} + (\mathrm{e}^{\theta} - 1)(\mathbb{E}\,\boldsymbol{Y}_i)\right)\right) \\
&\le n \exp\left(-\theta\bar{t} + k \log \lambda_{\max}\left(\mathbf{I} + \frac{(\mathrm{e}^{\theta} - 1)}{k}\sum_{i=1}^k \mathbb{E}\,\boldsymbol{Y}_i\right)\right) \\
&\le n \exp\left(-\theta\bar{t} + k \log\left(1 + \frac{(\mathrm{e}^{\theta} - 1)}{k}\left(\frac{\mu_{\max}}{R}\right)\right)\right) \\
&\le n \exp\left(-\theta\bar{t} + k\frac{(\mathrm{e}^{\theta} - 1)\mu_{\max}}{kR}\right).
\end{aligned}$$

The last inequality is due to the fact that $\log(1+x) \le x$ for $x > -1$. The result is obtained by replacing $\bar{t} = (1+\delta)\frac{\mu_{\max}}{R}$ and selecting $\theta = \log(1+\delta)$.
The lower bound follows from a similar argument. ∎

It is worth mentioning that the Chernoff bound is only one of many matrix inequalities that can be deduced from the Laplace transform tail bounds for independent sums of random matrices, Theorem 7.14. In fact, Theorem 7.14 is used in [Tro11] to produce matrix versions of the bounds known by the names of Azuma, Bennett, Bernstein, Chernoff, Hoeffding, and McDiarmid.

7.4 Application : Sparsification via random sampling

In this section we show how the matrix Chernoff bound can be employed to prove that any graph can be well approximated by a sparse graph with high probability. With this purpose in mind, we recall a couple of key definitions from graph theory.

Definition 7.20 (Adjacency matrix). Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a graph. The adjacency matrix $\boldsymbol{A}$ of G is the $|\mathsf{V}| \times |\mathsf{V}|$ matrix defined by

$$A_{(u,v)} \coloneqq \begin{cases} 1 & \text{if } (u,v) \in \mathsf{E} \\ 0 & \text{otherwise.} \end{cases}$$

In words, the adjacency matrix of G is the $(0,1)$-matrix whose (u,v)-th entry indicates whether there is an edge from node u to node v.

For a set V, we denote by $|\mathsf{V}|$ its cardinality.
In this section we will only be dealing with undirected graphs G. In this case, the adjacency matrix $\boldsymbol{A}$ is always symmetric.

Definition 7.21 (Laplacian matrix). Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an undirected graph. The Laplacian matrix of G, denoted by $\boldsymbol{L}_{\mathsf{G}}$, is defined by

$$\boldsymbol{L}_{\mathsf{G}} \coloneqq \boldsymbol{D} - \boldsymbol{A},$$

where $\boldsymbol{D}$ is the diagonal matrix whose entries are the degrees of the nodes in V. The Laplacian of a graph is often understood as its matrix representation.

Recall that the *degree* of node $u \in \mathsf{V}$ is the number of edges that are attached to u.

An important property of the Laplacian of any graph is that it is positive semidefinite. Indeed, for $\boldsymbol{x} \in \mathbb{R}^n$, we have that

$$\boldsymbol{x}^\intercal \boldsymbol{L}\boldsymbol{x} = \sum_u \deg(x_u) {x_u}^2 - \sum_{(u,v)\in\mathsf{E}} x_u x_v = \sum_{(u,v)\in\mathsf{E}} (x_u - x_v)^2.$$

So far we have mentioned that any graph can be approximated by a sparse graph with high probability. Nonetheless, we have not yet specified what we mean by an approximation. The notion of "closeness" between two graphs that we will consider is based on the spectrum of their Laplacian matrices.

Definition 7.22 (Spectral ε-approximation). Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a connected, undirected graph, and let H be a graph that is defined over the vertex set V. Given $\varepsilon > 0$, we say that H is a *spectral ε-approximation* of G if

$$(1-\varepsilon)\boldsymbol{L}_{\mathsf{G}} \preccurlyeq \boldsymbol{L}_{\mathsf{H}} \preccurlyeq (1+\varepsilon)\boldsymbol{L}_{\mathsf{G}}.$$

Our main objective is to show, by means of the probabilistic method, that every undirected, connected graph has a sparse spectral ε-approximation. This result is due to Spielman and Srivastava [SS11] and the analysis presented in this section was based on that by Spielman in his lecture notes [Spi19]. The argument consists of two steps. First, we describe an algorithm that, given $\varepsilon > 0$ and a graph G, produces a sparse graph H whose Laplacian coincides in expectation to that of G. Then, we invoke the matrix Chernoff bound, Theorem 7.19, to show that H, the graph resulting from the algorithm, is a spectral ε-approximation of G with high probability.

For the remainder of this section, we denote by $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ an undirected, connected graph with $|\mathsf{V}| = n$ nodes. We also denote by $w_{(u,v)}$ the weight of the node $(u,v) \in \mathsf{E}$. Finally, for $(u,v) \in \mathsf{E}$, we denote by $\boldsymbol{L}_{(u,v)}$ the Laplacian matrix of the graph with n nodes and a single edge connecting u and v. Note that $\boldsymbol{L}_{(u,v)}$ can alternatively be written as $(\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)(\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^\intercal$, where $\boldsymbol{\delta}_u$ is the standard basis vector associated to the position of u.

7.4.1 The algorithm

We begin by presenting the randomized algorithm that Spielman proposed for the construction of sparse approximations of a graph [Spi19]. Given G and a parameter $R > 0$, the algorithm produces a graph H whose Laplacian matrix coincides in expectation with $\boldsymbol{L}_{\mathsf{G}}$, and whose edge density is inversely proportional to R.

Algorithm 7.1

Input: G, $R \geq 0$
Ouput: H
For each $(u, v) \in \mathsf{E}$ do:

1. Compute the probability

$$p_{(u,v)} := \frac{1}{R} w_{(u,v)} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^{\mathsf{T}} \boldsymbol{L}_{\mathsf{G}}^{\dagger} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v),$$

   where $\boldsymbol{L}_{\mathsf{G}}^{\dagger}$ denotes the generalized inverse of $\boldsymbol{L}_{\mathsf{G}}$.
2. Add the edge (u, v) with weight $w_{(u,v)}/p_{(u,v)}$ to H with probability $p_{(u,v)}$.

For $(u, v) \in \mathsf{E}$ the quantities $r_{(u,v)} = (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^{\mathsf{T}} \boldsymbol{L}_{\mathsf{G}}^{\dagger} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)$ and $l_{(u,v)} = w_{(u,v)} r_{(u,v)}$ are called the *effective resistance* and the *leverage score* of the edge (u, v), respectively, and are of interest on their own.

Therefore, to create the approximating graph H, we first assign a probability $p_{(u,v)}$ to each edge $(u, v) \in \mathsf{E}$ of G and then add the edge (u, v) with weight $w_{(u,v)}/p_{(u,v)}$ to H with probability $p_{(u,v)}$. As expressed in the following proposition, this choice of weights allows for the expectation of $\boldsymbol{L}_{\mathsf{H}}$ to be exactly what we want it to be.

Proposition 7.23 (Expectation of the $\boldsymbol{L}_{\mathsf{H}}$). Given a graph G and $R > 0$, the graph H returned by Algorithm 7.1 is such that $\mathbb{E}\, \boldsymbol{L}_{\mathsf{H}} = \boldsymbol{L}_{\mathsf{G}}$.

Proof. Computing the expectation of $\boldsymbol{L}_{\mathsf{H}}$ yields

$$\mathbb{E}\, \boldsymbol{L}_{\mathsf{H}} = \sum\nolimits_{(u,v) \in \mathsf{E}} p_{(u,v)} \left(\frac{w_{(u,v)}}{p_{(u,v)}} \right) \boldsymbol{L}_{(u,v)} = \sum\nolimits_{(u,v) \in \mathsf{E}} w_{(u,v)} \boldsymbol{L}_{(u,v)} = \boldsymbol{L}_{\mathsf{G}},$$

as desired. ∎

The substance of Algorithm 7.1 lies in the choice of the probabilities $p_{(u,v)}$, which have been cleverly selected so as to produce a sparse approximator H.

Proposition 7.24 (Sparsity of H). Given a graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, the expected number of edges of the graph H returned by Algorithm 7.1 is $\frac{n-1}{R}$.

Proof. The expected number of edges in H is:

$$\begin{aligned}
\sum\nolimits_{(u,v) \in \mathsf{E}} p_{(u,v)} &= \frac{1}{R} \sum\nolimits_{(u,v) \in \mathsf{E}} w_{(u,v)} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^{\mathsf{T}} \boldsymbol{L}_{\mathsf{G}}^{\dagger} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v) \\
&= \frac{1}{R} \sum\nolimits_{(u,v) \in \mathsf{E}} w_{(u,v)} \operatorname{tr}(\boldsymbol{L}_{\mathsf{G}}^{\dagger} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)(\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^{\mathsf{T}}) \\
&= \frac{1}{R} \operatorname{tr}\left(\sum\nolimits_{(u,v) \in \mathsf{E}} \boldsymbol{L}_{\mathsf{G}}^{\dagger} w_{(u,v)} (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)(\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^{\mathsf{T}} \right) \\
&= \frac{1}{R} \operatorname{tr}\left(\boldsymbol{L}_{\mathsf{G}}^{\dagger} \sum\nolimits_{(u,v) \in \mathsf{E}} w_{(u,v)} \boldsymbol{L}_{(v,u)} \right) \\
&= \frac{1}{R} \operatorname{tr}(\boldsymbol{L}_{\mathsf{G}}^{\dagger} \boldsymbol{L}_{\mathsf{G}}^{\dagger}) \\
&= \frac{n-1}{R}.
\end{aligned}$$

The last identity is due to the fact that G is connected, so L_G has 0 as an eigenvalue with multiplicity exactly one. For more information on the relationship between connectivity of a graph and the eigenvalues of its Laplacian matrix we refer the reader to Chapter 2 of [MP93]. ■

Exercise 7.25 (Sparsity with high probability). Show that choosing $R = \Omega(\varepsilon^{-2} \ln n)$ in Algorithm 7.1 produces a graph H with expected number of edges $O(\varepsilon^{-2} n \ln n)$ with high probability.

7.4.2 Analysis of the algorithm

We have described an algorithm by Spielman and Srivastava that, given an undirected, connected graph G, produces a sparse graph H whose Laplacian coincides with the Laplacian of G in expectation. It remains to show that H is a spectral ε-approximation of G with high probability. We begin with a useful transformation, which will later allow us to employ the matrix Chernoff bound.

Lemma 7.26 Let $\varepsilon > 0$ and let G and H be as in Definition 7.22. Then

$$(1-\varepsilon)\boldsymbol{L}_G \preccurlyeq \boldsymbol{L}_H \preccurlyeq (1+\varepsilon)\boldsymbol{L}_G; \text{ if and only if}$$

$$(1-\varepsilon)\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_G\boldsymbol{L}_G^{\dagger/2} \preccurlyeq \boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_H\boldsymbol{L}_G^{\dagger/2} \preccurlyeq (1+\varepsilon)\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_G\boldsymbol{L}_G^{\dagger/2},$$

where $\boldsymbol{L}_G^{\dagger/2}$ denotes the square root of the generalized inverse of $\boldsymbol{L}_G$.

Exercise 7.27 Provide a proof for Lemma 7.26.

We are now ready to prove the main theorem of this section.

Theorem 7.28 (Spielman and Srivastava (2011)). Let $G = (V, E)$ be an undirected, connected graph with $|V| = n$, and let $\varepsilon > 0$. Then there exists a graph H on V, with at most $4\varepsilon^{-2} n \log n$ edges, that is a spectral ε-approximation of G with high probability.

Proof. Let $\boldsymbol{\Pi} := \boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_G\boldsymbol{L}_G^{\dagger/2}$ and define

$$\boldsymbol{X}_{(u,v)} := \begin{cases} \frac{w_{(u,v)}}{p_{(u,v)}}\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_{(u,v)}\boldsymbol{L}_G^{\dagger/2}, & \text{if } (u,v) \text{ is added to H} \\ 0, & \text{otherwise.} \end{cases}$$

By Lemma 7.26, we have that H is a spectral ε-approximation of G if and only if the graph whose Laplacian is $\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_H\boldsymbol{L}_G^{\dagger/2}$ is a spectral ε-approximation of the graph with Laplacian $\boldsymbol{\Pi}$. Applying the transformation to the identity in the proof of Proposition 7.23, we obtain

$$\begin{aligned} \mathbb{E}\,\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_H\boldsymbol{L}_G^{\dagger/2} &= \sum\nolimits_{(u,v)\in E} p_{(u,v)} \frac{w_{(u,v)}}{p_{(u,v)}} \boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_{(u,v)}\boldsymbol{L}_G^{\dagger/2} \\ &= \mathbb{E}\left(\sum\nolimits_{(u,v)\in E} \boldsymbol{X}_{(u,v)}\right) \\ &= \boldsymbol{\Pi}. \end{aligned}$$

Therefore, it suffices to show that the extremal eigenvalues of the sum $\sum_{(u,v)\in E} \boldsymbol{X}_{(u,v)}$ stay within a factor of $(1 \pm \varepsilon)$ of that of $\boldsymbol{\Pi}$ with high probability. It is easy to check that

$$\lambda_{\max}(\boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_{(u,v)}\boldsymbol{L}_G^{\dagger/2}) = (\boldsymbol{\delta}_u - \boldsymbol{\delta}_v)^\top \boldsymbol{L}_G^{\dagger}(\boldsymbol{\delta}_u - \boldsymbol{\delta}_v).$$

So $\lambda_{\max}(\boldsymbol{X}_{(v,u)}) \leq R$ for all $(u,v) \in \mathsf{E}$. Furthermore, the equality above implies that $\mu_{\max}(\boldsymbol{X}_{(u,v)}) = \mu_{\min}(\boldsymbol{X}_{(u,v)}) = 1$. Therefore, choosing $R = \frac{\varepsilon^2}{3.5 \ln n}$ and applying the matrix Chernoff bound yields

$$\begin{aligned}\mathbb{P}\left\{\sum\nolimits_{(u,v)\in\mathsf{E}} \boldsymbol{X}_{(u,v)} \geq (1+\varepsilon)\boldsymbol{\Pi}\right\} &\leq n\left(\frac{\mathrm{e}^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{1/R} \\ &\leq n\left(\mathrm{e}^{-\varepsilon^2/3}\right)^{\ln n/(3.5\varepsilon^2)} \\ &= n^{-1/6}.\end{aligned}$$

The second inequality uses the relation $\mathrm{e}^{\varepsilon}(1+\varepsilon)^{-1-\varepsilon} \leq \mathrm{e}^{-\varepsilon^2/3}$ for $\varepsilon \in [0,1]$. Similarly, applying the matrix Chernoff lower bound and using the fact that $\mathrm{e}^{-\varepsilon}(1-\varepsilon)^{\varepsilon-1} \leq \mathrm{e}^{-\varepsilon^2/2}$ for $\varepsilon \in [0,1]$ we obtain

$$\mathbb{P}\left\{\sum\nolimits_{(u,v)\in\mathsf{E}} \boldsymbol{X}_{(u,v)} \leq (1-\varepsilon)\boldsymbol{\Pi}\right\} \leq n^{-3/2}.$$

We conclude that $\boldsymbol{L}_{\mathsf{G}}^{\dagger/2}\boldsymbol{L}_{\mathsf{H}}\boldsymbol{L}_{\mathsf{G}}^{\dagger/2}$ approximates $\boldsymbol{\Pi}$, hence $\boldsymbol{L}_{\mathsf{H}}$ approximates $\boldsymbol{L}_{\mathsf{G}}$, sufficiently well with high probability. Finally, observe that in light of Proposition 7.24, and given the choice of R, the expected number of edges of the resulting approximating graph H is $3.5(n-1)\ln n\varepsilon^{-2}$. ■

To understand the significance of this result, first recall that the spectrum of a graph contains relevant information about the graph's connectivity, underlying modularity, and other invariants. As indicated by their name, spectral approximations of a graph G have eigenvalues that are similar to those of G, and thus preserve many of G's structural properties. Furthermore, the solutions to linear systems of equations associated to the Laplacian matrices of graphs that approximate each other are similar [Spi19]. This property allows for more efficient computation of the solutions of Laplacian linear systems. A well known testament of the utility of graph sparsification are expander graphs, which are sparse spectral approximations of complete graphs.

The greatest potential of application of matrix sparsification lies in its utility for reducing the computational burden of otherwise complex tasks. The perspicacious reader may thus wonder about the computational feasibility of implementing the algorithm, and particularly of computing the probabilities $p_{(u,v)}$ for each $(u,v) \in \mathsf{E}$. In [SS11], Spielman and Srivastava provide an efficient way of estimating the effective resistances of every edge on a graph, hence the probabilities $p_{(u,v)}$.

7.5 Conclusion

In this lecture we extended the Laplace transform method for random variables to the matrix setting, obtaining powerful tail bounds for the extremal eigenvalues of random matrices and their independent sums. As an example of the results that can be deduced from the matrix extension of the Laplace transform method, we established a Chernoff inequality for random matrices and exhibited an application of the latter in spectral graph theory.

Lecture bibliography

[AW01] R. Ahlswede and A. Winter. *Strong Converse for Identification via Quantum Channels*. 2001. arXiv: quant-ph/0012127 [quant-ph].

[Lie73b] E. H. Lieb. "Convex trace functions and the Wigner-Yanase-Dyson conjecture". In: *Advances in Mathematics* 11.3 (1973), pages 267–288. DOI: https://doi.org/10.1016/0001-8708(73)90011-X.

[MP93] B. Mohar and S. Poljak. "Eigenvalues in Combinatorial Optimization". In: *Combinatorial and Graph-Theoretical Problems in Linear Algebra*. Springer New York, 1993, pages 107–151.

[Oli10] R. I. Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *Electronic Communications in Probability* 15 (2010), pages 203–212.

[Spi19] D. Spielman. *Spectral and Algebraic Graph Theory*. 2019. URL: http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf.

[SS11] D. A. Spielman and N. Srivastava. "Graph Sparsification by Effective Resistances". In: *SIAM Journal on Computing* 40.6 (2011), pages 1913–1926. eprint: https://doi.org/10.1137/080734029. DOI: 10.1137/080734029.

[Tro11] J. A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4 (2011), pages 389–434. DOI: 10.1007/s10208-011-9099-z.

[Tro12] J. A. Tropp. "From joint convexity of quantum relative entropy to a concavity theorem of Lieb". In: *Proc. Amer. Math. Soc.* 140.5 (2012), pages 1757–1760. DOI: 10.1090/S0002-9939-2011-11141-9.

8. Operator-Valued Kernels

Date: 14 March 2022 **Author:** Nicholas H. Nelsen

In this lecture, we given an overview of the theory of operator-valued kernels. These objects generalize scalar-valued kernels to vector-valued output data and underlie statistically analyzable algorithms for vector-valued learning. Motivated by machine learning considerations, we take a function space perspective and introduce reproducing kernel Hilbert spaces for scalar- and operator-valued kernels. Both finite-dimensional (matrix-valued) and infinite-dimensional (operator-valued) examples of positive-definite kernels are presented. After developing some of the properties of these kernels, we state vector-valued extensions of Bochner's and Schoenberg's characterization theorems. We conclude with a connection to vector-valued Gaussian processes and regression.

Agenda:
1. Scalar kernels and RKHS
2. Operator-valued kernels
3. Examples
4. Vector-valued GPs

Throughout this lecture, we write $\mathscr{L}(\mathcal{U};\mathcal{V})$ for the Banach space of bounded linear operators taking normed space $\mathcal{U}$ to Banach space $\mathcal{V}$, $\mathscr{L}(\mathcal{V})$ when $\mathcal{U} = \mathcal{V}$, and $\mathscr{L}_+(\mathcal{H}) \subset \mathscr{L}(\mathcal{H})$ for the set of bounded positive-semidefinite operators on a Hilbert space $\mathcal{H}$.

8.1 Scalar kernels and reproducing kernel Hilbert space

In Lectures 18 and 19, we studied some basic properties of positive-definite kernels and positive-definite functions. Two characterization theorems were developed for translation invariant and radial kernels. Bochner's theorem characterized translation invariant kernels in terms of the Fourier transform of a measure, and Schoenberg's theorem characterized radial kernels in terms of the Laplace transform of a measure. Some of these ideas will be generalized in today's lecture.

Recall that a measurable bivariate function $k\colon \mathbb{F}^d \times \mathbb{F}^d \to \mathbb{F}$ over the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$ is a *positive-definite kernel* if it is continuous and satisfies

1. (**Symmetry**) $k(\boldsymbol{x},\boldsymbol{x}') = \overline{k(\boldsymbol{x}',\boldsymbol{x})}$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathbb{F}^d$,
2. (**psd**) For all $n \in \mathbb{N}$ and any set $\{\boldsymbol{x}_i\}_{i=1,\dots,n} \subset \mathbb{F}^d$, the matrix

$$\boldsymbol{K} \coloneqq [k(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{ij} \succcurlyeq \boldsymbol{0}\,.$$

These kernels naturally model scalar-valued data. Hence, they may be referred to as *scalar-valued kernels*. Kernels implicitly define functions from $\mathbb{F}^d$ into $\mathbb{F}$, and it is often the goal in machine learning, statistics, and approximation theory to exploit this space of functions for modeling, inference, prediction, and other tasks. We now briefly introduce the notion of reproducing kernel Hilbert space in the scalar setting.

8.1.1 Reproducing kernel Hilbert space

The vector space of all functions mapping an input space $\mathcal{X}$ to the field $\mathbb{F}$ is a messy infinite-dimensional linear space. Instead of working here, functional analysts often study smaller infinite-dimensional spaces that have more structure and better properties.

For example, the function space $\mathsf{L}_2(\mathbb{F}^d;\mathbb{F})$ of all square integrable scalar-valued functions on $\mathbb{F}^d$ is a natural generalization of finite-dimensional Euclidean space

to infinite dimensions. It is a Hilbert space whenever its elements are viewed as equivalence classes of functions equal Lebesgue almost everywhere. However, this construction means that evaluation of an element of $\mathsf{L}_2(\mathbb{F}^d; \mathbb{F})$ at a single point in $\mathbb{F}^d$ is not even defined!

Reproducing kernel Hilbert spaces (RKHS) come to the rescue.

Definition 8.1 (Reproducing kernel Hilbert space). A Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$ of functions mapping $\mathcal{X}$ to $\mathbb{F}$ is a *reproducing kernel Hilbert space* if for every $\boldsymbol{x} \in \mathcal{X}$, the linear functional $\Phi_{\boldsymbol{x}} \colon \mathscr{H} \to \mathbb{F}$ defined by

$$f \mapsto \Phi_{\boldsymbol{x}} f \coloneqq f(\boldsymbol{x}) \quad \text{is continuous.}$$

This definition ensures that functions in an RKHS have a pointwise meaning. Intuitively, it follows that functions in RKHS are smoother than those in L_2. However, this abstract property does not explain the terminology "reproducing" or "kernel" in the name.

The following alternative definition of a RKHS makes clear the connection to positive-definite kernels.

Definition 8.2 (RKHS). A Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$ of functions mapping the input space $\mathcal{X}$ to $\mathbb{F}$ is said to be the *reproducing kernel Hilbert space* of the positive-definite kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ if

1. (**Inclusion**) $k(\cdot, \boldsymbol{x}) \in \mathscr{H}$ for every $\boldsymbol{x} \in \mathcal{X}$.
2. (**Reproducing property**) $f(\boldsymbol{x}) = \langle k(\cdot, \boldsymbol{x}),\ f \rangle$ for every $\boldsymbol{x} \in \mathcal{X}$ and $f \in \mathscr{H}$.

The RKHS inner-product *is not* easily expressed in closed form for most kernels.

Item (2) in Definition 8.2 shows that the pointwise values of the function are "reproduced" by the RKHS inner-product of the kernel with the function itself. Since most real-world data can be modeled as pointwise values of some underlying scalar function, it seems that RKHSs are natural candidates for hypothesis spaces in learning theory.

Today, we will go beyond the scalar function setting by answering the question:

Can we use kernels to model vector-valued data?

The key idea is to generalize scalar kernels to *operator-valued kernels*.

8.2 Operator-valued kernels

Consider the space of functions mapping $\mathbb{F}^d$ to $\mathbb{F}^p$, where $d, p \in \mathbb{N}$. For example, vector-valued functions arise as the vector fields of dynamical systems or as the parameter-to-image map in computer graphics. One could treat a vector-valued function in this space as a family of p scalar-valued functions and apply the previous RKHS formulation componentwise. This is commonplace in many fields. However, this approach may ignore correlations between the p components that could be essential to fully describe the function. The field of *multitask learning* [Cap+08; Car97; Evg+05] has developed to explicitly model the correlation in vector-valued output data. At the core of this effort is the operator-valued kernel.

8.2.1 Kernels valued as linear maps

We consider a setting in which the input space $\mathcal{X}$ is a separable Banach space but now the output space $\mathcal{Y}$ is a separable Hilbert space instead of the scalar field $\mathbb{F}$. This

framework is powerful because it allows for both $\mathcal{X}$ and $\mathcal{Y}$ to be infinite-dimensional, in which case maps between them are often called *operators*.

The following definition generalizes what it means to be a kernel.

Definition 8.3 (Positive-definite operator-valued kernel). An *operator-valued kernel* on $\mathcal{X}$ is a Borel-$\mathcal{X}$ measurable function $\boldsymbol{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$. The kernel is *positive-definite* if

1. (**Symmetry**) $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x})^*$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,
2. (**psd**) For all $n \in \mathbb{N}$ and any $\{\boldsymbol{x}_i\}_{i=1,\dots,n} \subset \mathcal{X}$ and $\{\boldsymbol{y}_j\}_{j=1,\dots,n} \subset \mathcal{Y}$, the quadratic form
$$\sum\nolimits_{i,j=1}^{n} \langle \boldsymbol{y}_i,\ \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{y}_j \rangle_{\mathcal{Y}} \geq 0\,. \tag{8.1}$$

Here * denotes adjoint with respect to the Hilbert space $\mathcal{Y}$.

That is, an operator-valued kernel takes a pair of vectors in the input space into the space of bounded linear maps between the output space and itself. Not only does an operator-valued kernel measure similarity between inputs, but it also accounts for relationships between the outputs. When $\mathcal{Y}$ is finite-dimensional, we use the term *matrix-valued kernel*.

Notice that the second condition (8.1) in Definition 8.3 generalizes the corresponding scalar definition condition from n-by-n psd kernel matrices to n-by-n psd *block kernel operators* $\boldsymbol{K}(\mathsf{X}, \mathsf{X}) := [\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j} \in \mathcal{L}(\mathcal{Y}^n)$. Here $\mathsf{X} := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ and $\mathcal{Y}^n := \mathcal{Y} \times \cdots \times \mathcal{Y}$ is the n-fold product of $\mathcal{Y}$ (which is itself a Hilbert space).

8.2.2 Vector-valued RKHS

Since scalar-valued kernels on $\mathcal{X}$ lead to functions taking $\mathcal{X}$ to $\mathbb{F}$, it is reasonable to expect that operator-valued kernels induce vector-valued functions taking $\mathcal{X}$ to $\mathcal{Y}$. This is indeed the case. The basic theory of *vector-valued RKHS* can be developed in a parallel manner to the scalar case, albeit with a few nontrivial generalizations such as the vector-valued reproducing property.

We begin this endeavor by considering an arbitrary Hilbert space $(\mathcal{H}, \langle \cdot,\ \cdot \rangle_{\mathcal{H}})$ of maps taking $\mathcal{X}$ to $\mathcal{Y}$.

Definition 8.4 (Vector-valued RKHS [MPO5]). A Hilbert space $(\mathcal{H}, \langle \cdot,\ \cdot \rangle_{\mathcal{H}})$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$ is a *reproducing kernel Hilbert space* if for every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, the linear functional $\varphi_{\boldsymbol{x},\boldsymbol{y}} : \mathcal{H} \to \mathbb{F}$ defined by
$$\boldsymbol{f} \mapsto \Phi_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{f} := \langle \boldsymbol{y},\ \boldsymbol{f}(\boldsymbol{x}) \rangle_{\mathcal{Y}} \quad \text{is continuous.}$$

While seemingly benign, the implications of Definition 8.4 are quite profound. Indeed, let $\mathcal{H}$ be a vector-valued RKHS. For $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, the linear functional $\Phi_{\boldsymbol{x},\boldsymbol{y}}$ is bounded because it is continuous. Hence, the Riesz representation theorem ensures the existence of a unique element $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{y}} \in \mathcal{H}$, parametrized by $\boldsymbol{x}$ and $\boldsymbol{y}$, such that
$$\Phi_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{f} = \langle \boldsymbol{y},\ \boldsymbol{f}(\boldsymbol{x}) \rangle_{\mathcal{Y}} = \langle \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{y}},\ \boldsymbol{f} \rangle_{\mathcal{H}} \quad \text{for every} \quad \boldsymbol{f} \in \mathcal{H}\,. \tag{8.2}$$

The left-hand side of (8.2) is linear in $\boldsymbol{y}$. It follows that $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{y}}$ on the right-hand side is also linear. We can define a linear operator $\boldsymbol{K}_{\boldsymbol{x}} : \mathcal{Y} \to \mathcal{H}$ via the rule
$$\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{y}} =: \boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{y}\,.$$

Furthermore, we can define an operator-valued kernel $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{Y} \to \mathcal{Y}$ by
$$\boldsymbol{y} \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{y} := (\boldsymbol{K}_{\boldsymbol{x}'}\boldsymbol{y})(\boldsymbol{x})\,. \tag{8.3}$$

We sometimes write $\boldsymbol{K}_{\boldsymbol{x}} = \boldsymbol{K}(\cdot, \boldsymbol{x})$ to emphasize that the first argument of the kernel is open. Provided that $\boldsymbol{K}_{\boldsymbol{x}} \colon \mathcal{Y} \to \mathcal{H}$ is bounded (which we prove in Proposition 8.5), it follows from (8.2) and (8.3) that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{K}_{\boldsymbol{x}}^* \boldsymbol{f} \quad \text{and} \quad \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}_{\boldsymbol{x}}^* \boldsymbol{K}_{\boldsymbol{x}'} \quad \text{for each} \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$$

The second equality parallels the well known scalar identity $k(\boldsymbol{x}, \boldsymbol{x}') = \langle k(\cdot, \boldsymbol{x}),\ k(\cdot, \boldsymbol{x}') \rangle$.

because $\mathcal{Y}$ is a Hilbert space. This is a form of the *vector-valued reproducing property* (cf. Definition 8.2).

Using some basic functional analysis, it is possible to deduce the following facts about the kernel $\boldsymbol{K}$ in (8.3).

Proposition 8.5 (Operator-valued kernel: Properties). The operator-valued kernel $\boldsymbol{K}$ on $\mathcal{X}$ from (8.3) satisfies the following properties for all $\boldsymbol{x}$ and $\boldsymbol{x}' \in \mathcal{X}$.

1. (**Kernel**) $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{L}(\mathcal{Y})$, $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')^* \in \mathcal{L}(\mathcal{Y})$, and $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \in \mathcal{L}_+(\mathcal{Y})$.
2. (**psd**) $\boldsymbol{K}$ is positive-definite.
3. (**Kernel sections**) $\boldsymbol{K}_{\boldsymbol{x}} \in \mathcal{L}(\mathcal{Y}; \mathcal{H})$ with $\|\boldsymbol{K}_{\boldsymbol{x}}\|_{\mathcal{L}(\mathcal{Y};\mathcal{H})} = \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\|_{\mathcal{L}(\mathcal{Y})}^{1/2}$.
4. (**Cauchy–Schwarz**) $\|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\|_{\mathcal{L}(\mathcal{Y})} \le \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\|_{\mathcal{L}(\mathcal{Y})}^{1/2} \|\boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x}')\|_{\mathcal{L}(\mathcal{Y})}^{1/2}$.

Proof. The only hard item is (1), which is left as an exercise; see Problem 8.8. We will prove (3), noting that the remaining items are similar or trivial given (1). Applying the vector-valued reproducing property and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \|\boldsymbol{K}_{\boldsymbol{x}}\|^2 = \sup_{\|\boldsymbol{y}\|_{\mathcal{Y}}=1} \|\boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{y}\|_{\mathcal{H}}^2 &= \sup_{\|\boldsymbol{y}\|_{\mathcal{Y}}=1} \langle \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{y},\ \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{y} \rangle_{\mathcal{H}} \\ &= \sup_{\|\boldsymbol{y}\|_{\mathcal{Y}}=1} \langle \boldsymbol{y},\ \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y} \rangle_{\mathcal{Y}} \\ &\le \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\|, \end{aligned}$$

where all unlabeled norms are the induced operator norms. For the other direction,

$$\begin{aligned} \langle \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y},\ \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y} \rangle_{\mathcal{Y}} &= \langle \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y},\ \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{y} \rangle_{\mathcal{H}} \\ &\le \|\boldsymbol{K}_{\boldsymbol{x}}\| \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y}\|_{\mathcal{Y}} \|\boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{y}\|_{\mathcal{H}} \\ &\le \|\boldsymbol{K}_{\boldsymbol{x}}\|^2 \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y}\|_{\mathcal{Y}} \|\boldsymbol{y}\|_{\mathcal{Y}}. \end{aligned}$$

These two bounds imply that $\|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \boldsymbol{y}\|_{\mathcal{Y}} \le \|\boldsymbol{K}_{\boldsymbol{x}}\|^2 \|\boldsymbol{y}\|_{\mathcal{Y}}$ as desired. ∎

> **Aside:** Further assumptions on $\boldsymbol{K}$ (such as $\operatorname{tr}(\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})) < \infty$ for all $\boldsymbol{x}$) beyond just the default properties in Proposition 8.5 are made in the paper [CDV07] of Caponnetto and de Vito. They derive optimal convergence rates for the statistical learning of vector-valued maps taking $\mathcal{X}$ to $\mathcal{Y}$ using RKHS-penalized least-squares (i.e., kernel ridge) regression.

Exercise 8.6 (Matrix coordinates). For $p \in \mathbb{N}$, let $\boldsymbol{K} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{F}^{p \times p}$ be the operator-valued kernel of RKHS $\mathcal{H}_{\boldsymbol{K}}$. For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $i, j = 1, \dots, p$, show that

$$\big[\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\big]_{ij} = \langle \boldsymbol{K}_{\boldsymbol{x}} \boldsymbol{\delta}_i,\ \boldsymbol{K}_{\boldsymbol{x}'} \boldsymbol{\delta}_j \rangle_{\mathcal{H}_{\boldsymbol{K}}},$$

where $(\boldsymbol{\delta}_i)_{i=1,\dots,p}$ are the standard basis vectors of $\mathbb{F}^p$.

Exercise 8.7 (Point evaluation). From Proposition 8.5 deduce that the point evaluation map is a bounded linear operator. For every $\boldsymbol{x} \in \mathcal{X}$, there exists $C_{\boldsymbol{x}} > 0$ such that

$$\|\boldsymbol{f}(\boldsymbol{x})\|_{\mathcal{Y}} \le C_{\boldsymbol{x}} \|\boldsymbol{f}\|_{\mathcal{H}} \quad \text{for all} \quad \boldsymbol{f} \in \mathcal{H}.$$

Explicitly determine a choice of the constant $C_{\boldsymbol{x}}$.

Problem 8.8 (Kernel bounds). Complete the proof of Proposition 8.5. **Hint:** For item (1), use the adjoint and apply the uniform boundedness principle.

Warning 8.9 (Continuity). Although point evaluation is well-defined, functions $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ in the RKHS $\mathcal{H}_{\boldsymbol{K}}$ of a positive-definite operator-valued reproducing kernel $\boldsymbol{K}$ need not be continuous. To *guarantee continuity*, that is, $\boldsymbol{f} \in \mathsf{C}(\mathcal{X}; \mathcal{Y})$, it is necessary and sufficient for the kernel to be *Mercer* [Car+10, Proposition 2]. More precisely,

1. $\boldsymbol{x} \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$ is locally bounded,
2. $\boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{y} \in \mathsf{C}(\mathcal{X}; \mathcal{Y})$ for all $\boldsymbol{x} \in \mathcal{X}$ and all $\boldsymbol{y} \in \mathcal{Y}$. ∎

Item (2) is natural from the characterization of $\mathcal{H}_{\boldsymbol{K}}$ as the closure of the span of $\boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{y}$ over all $\boldsymbol{x}$ and $\boldsymbol{y}$ [ARL+12].

The following is an alternative definition of vector-valued RKHS that parallels Definition 8.2 in the scalar case.

Definition 8.10 (Vector-valued RKHS [Kad+16]). A Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ of functions mapping the input space $\mathcal{X}$ to the output Hilbert space $(\mathcal{Y}, \langle \cdot, \cdot \rangle_{\mathcal{Y}})$ is said to be the *vector-valued reproducing kernel Hilbert space* of the positive-definite operator-valued kernel $\boldsymbol{K} : \mathcal{X} \times \mathcal{X} \to \mathscr{L}(\mathcal{Y})$ if

1. (**Inclusion**) $\boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{y} \in \mathcal{H}$ for every $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$,
2. (**Reproducing property**) $\langle \boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x}) \rangle_{\mathcal{Y}} = \langle \boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{y}, \boldsymbol{f} \rangle$ for every $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{f} \in \mathcal{H}$.

We conclude this section with a converse result [Sch64]. In the scalar setting, the analogous result is called the Moore–Aronszajn theorem [Aro50].

Theorem 8.11 (Schwartz 1964). Every positive-definite operator-valued kernel $\boldsymbol{K}$ is the reproducing kernel of a uniquely defined RKHS $\mathcal{H}_{\boldsymbol{K}}$.

This one-to-one correspondence between operator-valued kernels and RKHS justifies our notation of indexing the RKHS by its kernel, $\boldsymbol{K} \leftrightarrow \mathcal{H}_{\boldsymbol{K}}$.

8.2.3 Vector-valued Bochner and Schoenberg theorems

Last, we present vector-valued extensions of Bochner's theorem for translation-invariant kernels and Schoenberg's theorem for radial kernels as seen in Lectures 18 and 19. We refer the reader to [MP05, Section 5] and [Car+10] for more details and references.

Theorem 8.12 (Vector-valued Bochner; Berberian 1966, Fillmore 1970). A map $\boldsymbol{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathscr{L}(\mathcal{Y})$ is a positive-definite operator-valued translation-invariant kernel if and only if it takes the form

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x} - \boldsymbol{x}', \, \boldsymbol{\xi} \rangle_{\mathbb{R}^d}} \, \boldsymbol{\mu}(\mathrm{d}\boldsymbol{\xi}) \,,$$

where $\mathcal{Y}$ is a separable Hilbert space and $\boldsymbol{\mu}$ is a $\mathscr{L}_+(\mathcal{Y})$ operator-valued Borel measure on $\mathbb{R}^d$.

Theorem 8.13 (Vector-valued Schoenberg; Berberian 1966, Fillmore 1970). A function $\boldsymbol{\varphi} : \mathbb{R}_+ \to \mathscr{L}(\mathcal{Y})$ generates a positive-definite radial operator-valued kernel

$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \varphi(\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}^2)$ on the Hilbert space $\mathcal{X}$ if and only if

$$s \mapsto \boldsymbol{\varphi}(s) = \int_0^\infty \mathrm{e}^{-su}\, \boldsymbol{\nu}(\mathrm{d}u)\,,$$

where $\mathcal{Y}$ is a separable Hilbert space and $\boldsymbol{\nu}$ is a $\mathcal{L}_+(\mathcal{Y})$ operator-valued Borel measure on $\mathbb{R}_+$.

8.3 Examples

Operator-valued kernels are much more complex than their scalar counterparts. This abstraction may make it difficult to intuit canonical ways to map vectors into operators in an expressive way. To build this intuition, we develop examples of operator-valued kernels that are related to scalar formulations. Catalogs of explicit matrix-valued and operator-valued kernels may be found in [ARL+12; BHB16; Car+10; MP05].

The most common examples are the separable scalar-type kernels in the next two examples. This is not surprising given the far-reaching implications of the vector-valued Bochner and Schoenberg characterization theorems of Section 8.2.3.

Example 8.14 (Separable scalar operator-valued kernel). Consider the function $\boldsymbol{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ given by

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{T} \tag{8.4}$$

for some scalar-valued kernel k and some self-adjoint $\boldsymbol{T} \in \mathcal{L}_+(\mathcal{Y})$. Then $\boldsymbol{K}$ is called a *separable scalar operator-valued kernel*. In supervised learning applications, the map $\boldsymbol{T}$ can be learned jointly as a hyperparameter along with the regression function [ARL+12]. Alternatively, $\boldsymbol{T}$ can be chosen as the empirical covariance operator of the labeled data (output space PCA), for example.

Notice that the input space is decoupled from the output space, which may be a strong limitation in practical settings. However, the block kernel matrix for (8.4) has a simple tensor product structure that allows for its efficient inversion.

Here is a concrete instantiation of (8.4) in infinite dimensions. For any invertible and self-adjoint $\boldsymbol{C} \in \mathcal{L}_+(\mathcal{X})$, define $\boldsymbol{K}$ by

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \exp\left(\tfrac{1}{2}\|\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{x}')\|_{\mathcal{X}}^2\right)\boldsymbol{T}, \quad \text{where}$$
$$(\boldsymbol{T}\boldsymbol{y})(\boldsymbol{t}) = \int_{\mathsf{D}} e^{-\|\boldsymbol{t}-\boldsymbol{s}\|_{\mathbb{R}^d}}\, \boldsymbol{y}(\boldsymbol{s})\, \mathrm{d}\boldsymbol{s}$$

for $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{t} \in \mathsf{D}$, where $\mathcal{Y} = \mathsf{L}_2(\mathsf{D}; \mathbb{R}^p)$ and $\mathsf{D} \subset \mathbb{R}^d$ [Kad+16]. ▪

Exercise 8.15 (Separable). Show that separable kernels $\boldsymbol{K}$ of the form (8.4) are positive-definite whenever k is positive-definite.

In Lecture 19 we encountered inner-product kernels on $\mathbb{F}^d$. These kernels take the form

$$k(\boldsymbol{x}, \boldsymbol{x}') = \varphi(\langle \boldsymbol{x},\ \boldsymbol{x}' \rangle_{\mathbb{F}^d})$$

for a scalar function $\varphi : \mathbb{F} \to \mathbb{F}$. We saw that these objects played a crucial role in the theory of entrywise psd preservers. Moreover, φ sometimes admits a convergent power series expansion via Vasudeva's theorem and its consequences.

We now demonstrate a dramatic extension of inner-product kernels to the operator-valued setting.

Example 8.16 (Kernel inner-product). Let $\mathcal{X} = \mathbb{C}^d$ for some $d \in \mathbb{N}$. Let $\boldsymbol{\alpha} \in \mathbb{Z}_+^d$ denote a multi-index (which is a succinct notation to deal with polynomials in arbitrary input dimension; for example, see [Eva10]). Let $\boldsymbol{\varphi} : \mathbb{C}^d \to \mathcal{L}(\mathcal{Y})$ be an operator-valued

entire function, which means that

$$\boldsymbol{z} \mapsto \boldsymbol{\varphi}(\boldsymbol{z}) = \sum\nolimits_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d} \boldsymbol{A}_{\boldsymbol{\alpha}} \boldsymbol{z}^{\boldsymbol{\alpha}} \quad \text{for some} \quad \{\boldsymbol{A}_{\boldsymbol{\alpha}}\} \subset \mathscr{L}_+(\mathscr{Y}) \,. \tag{8.5}$$

This series converges everywhere on $\mathbb{C}^d$ in an appropriate sense ([MP05]). Here $\boldsymbol{z}^{\boldsymbol{\alpha}} \coloneqq z_1^{\alpha_1} z_2^{\alpha_2} \cdots z_d^{\alpha_d}$ and $\mathscr{Y}$ is allowed to be infinite-dimensional. Let d scalar kernels $k_i \colon \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$ be given. These represent "inner-products. Denote their vectorized concatenation by $\boldsymbol{k} \coloneqq (k_1, \ldots, k_d) \colon \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d$. Then the composition

$$\begin{aligned} \boldsymbol{\varphi} \circ \boldsymbol{k} \colon \mathbb{C}^d \times \mathbb{C}^d &\to \mathscr{L}(\mathscr{Y}) \\ (\boldsymbol{x}, \boldsymbol{x}') &\mapsto \sum\nolimits_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d} k_1(\boldsymbol{x}, \boldsymbol{x}')^{\alpha_1} \cdots k_d(\boldsymbol{x}, \boldsymbol{x}')^{\alpha_d} \boldsymbol{A}_{\boldsymbol{\alpha}} \end{aligned} \tag{8.6}$$

is an operator-valued kernel. ▪

We now show that such an inner-product kernel is positive-definite.

Proposition 8.17 (Positive-definite inner-product kernel). Let $\boldsymbol{\varphi} \colon \mathbb{C}^d \to \mathscr{L}(\mathscr{Y})$ be an entire function of the form (8.5), where each $\boldsymbol{A}_{\boldsymbol{\alpha}}$ is self-adjoint. Let $\{k_i\}_{i=1,\ldots,d}$ be a family of positive-definite scalar kernels. Then $\boldsymbol{K} \coloneqq \boldsymbol{\varphi} \circ \boldsymbol{k}$ as defined in (8.6) is a positive-definite operator-valued kernel.

Proof. We must verify Definition 8.3. Clearly $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x})^*$ since each k_i is positive-definite and $\boldsymbol{A}_{\boldsymbol{\alpha}} = \boldsymbol{A}_{\boldsymbol{\alpha}}^*$. It remains to show the psd property. Note

$$\begin{aligned} \sum_{i,j=1}^n \langle \boldsymbol{y}_i, \ \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \boldsymbol{y}_j \rangle_{\mathscr{Y}} &= \sum_{i,j=1}^n \sum_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d} \prod_{\ell=1}^d k_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j)^{\alpha_\ell} \langle \boldsymbol{y}_i, \ \boldsymbol{A}_{\boldsymbol{\alpha}} \boldsymbol{y}_j \rangle_{\mathscr{Y}} \\ &= \sum_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d} \left[\sum_{i,j=1}^n \mathbf{1}_i \left(\prod_{\ell=1}^d k_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j)^{\alpha_\ell} \langle \boldsymbol{y}_i, \ \boldsymbol{A}_{\boldsymbol{\alpha}} \boldsymbol{y}_j \rangle_{\mathscr{Y}} \right) \mathbf{1}_j \right] \end{aligned}$$

for any $\{\boldsymbol{x}_i\}_{i=1,\ldots,n}$, $\{\boldsymbol{y}_j\}_{j=1,\ldots,n}$, and $n \in \mathbb{N}$. Since each k_ℓ is positive-definite, the function $k_\ell^{\alpha_\ell}$ by entrywise psd preservation of the power function (as seen in Lecture 19). By the Schur product theorem,

$$\left[\prod_{\ell=1}^d k_\ell(\boldsymbol{x}_i, \boldsymbol{x}_j)^{\alpha_\ell} \right]_{ij} \succcurlyeq \mathbf{0} \,.$$

Since $\boldsymbol{A}_{\boldsymbol{\alpha}} \in \mathscr{L}_+(\mathscr{Y})$ is psd, it defines a scalar positive-definite inner-product kernel $(\boldsymbol{y}, \boldsymbol{y}') \mapsto \langle \boldsymbol{y}, \ \boldsymbol{A}_{\boldsymbol{\alpha}} \boldsymbol{y}' \rangle_{\mathscr{Y}}$ on $\mathscr{Y}$. Thus, for all multi-indices $\boldsymbol{\alpha}$,

$$[\langle \boldsymbol{y}_i, \ \boldsymbol{A}_{\boldsymbol{\alpha}} \boldsymbol{y}_j \rangle_{\mathscr{Y}}]_{ij} \succcurlyeq \mathbf{0} \,,$$

One final application of the Schur product theorem shows that all the terms in the first set of large brackets above are greater than or equal to zero. This yields the desired result. ■

We conclude this example by remarking that there is a converse result to Proposition 8.17 [MP05, Proposition 3] that mimics Vasudeva's and Schoenberg's scalar characterization of entrywise psd preservers (Lecture 19) as everywhere convergent power series.

The previous two examples concerned separable kernels or sums thereof. It is possible to go beyond the separable case using special feature map factorizations, as the next example demonstrates.

Example 8.18 (Random features). Nonseparable kernels can be constructed with *vector-valued random features*. In their most general form [NS21], they are defined as a pair $(\boldsymbol{\varphi}, \mu)$, where $\boldsymbol{\varphi} : \mathcal{X} \times \Theta \to \mathcal{Y}$ is jointly square Bochner integrable and μ is a Borel probability measure on Θ. It follows that

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}_{\boldsymbol{\theta} \sim \mu}[\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{\theta}) \otimes_{\mathcal{Y}} \boldsymbol{\varphi}(\boldsymbol{x}', \boldsymbol{\theta})] \tag{8.7}$$

is a positive-definite operator-valued kernel on $\mathcal{X}$. A key insight is that the expectation can be approximated empirically with Monte Carlo. For some $m \in \mathbb{N}$, this leads to a low rank approximation $\boldsymbol{K}_m$ to $\boldsymbol{K}$ given by

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \boldsymbol{K}_m(\boldsymbol{x}, \boldsymbol{x}') := \frac{1}{m} \sum_{\ell=1}^{m} \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{\theta}_\ell) \otimes_{\mathcal{Y}} \boldsymbol{\varphi}(\boldsymbol{x}', \boldsymbol{\theta}_\ell), \quad \text{where } \boldsymbol{\theta}_\ell \overset{\text{iid}}{\sim} \mu .$$

Operator-valued kernel regression in, say, $\mathcal{Y} = \mathbb{F}^p$ with p large, using $\boldsymbol{K}_m$ is *much more efficient* than that using $\boldsymbol{K}$, even when $m \ll np$ is relatively large. This is because all the linear algebra is performed in $\mathbb{F}^m$ for the former and in $\mathbb{F}^{np}$ for the latter, where n is the sample size of the dataset.

One can further exploit the vector-valued Bochner characterization (Theorem 8.12) to design *random Fourier feature* [RR08] approximations to translation-invariant operator-valued kernels, as demonstrated in [BHB16; Min16]. ▪

Exercise 8.19 (Random feature). Prove that the operator-valued kernel (8.7) defined in terms of random features is positive-definite.

8.4 Vector-valued Gaussian processes

In the context of learning models from data, kernel methods are closely linked to Gaussian process (GP) regression in the scalar setting. It should come as no surprise that this remains true in the vector-valued setting. Indeed, there is a correspondence between a vector-valued RKHS $\mathcal{H}_{\boldsymbol{K}}$ and a *vector-valued GP*, which is a Gaussian distribution over vector-valued functions, given by

$$\begin{aligned} \boldsymbol{f}_{\boldsymbol{K}} &\sim \mathcal{GP}(\boldsymbol{m}, \boldsymbol{K}), \quad \text{where} \\ \boldsymbol{f}_{\boldsymbol{K}}(\boldsymbol{x}) &\sim \text{NORMAL}(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})) \ \text{ for } \ \boldsymbol{x} \in \mathcal{X}, \end{aligned}$$

barring some measurability issues in infinite dimensions. GPs are fully specified by their finite-dimensional distributions. Here, $\boldsymbol{f}_{\boldsymbol{K}} \sim \mathcal{GP}(\boldsymbol{m}, \boldsymbol{K})$ signifies that $\boldsymbol{f}_{\boldsymbol{K}}$ has mean function $\boldsymbol{m} = \mathbb{E}\, \boldsymbol{f}_{\boldsymbol{K}} \in \mathcal{H}_{\boldsymbol{K}}$ and covariance function

$$(\boldsymbol{x}, \boldsymbol{x}') \mapsto \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[\boldsymbol{f}_{\boldsymbol{K}}(\boldsymbol{x}) \otimes_{\mathcal{Y}} \boldsymbol{f}_{\boldsymbol{K}}(\boldsymbol{x}')] ,$$

which is just the operator-valued kernel! The covariance kernel is in the random feature form (8.7) with the feature map given by the GP itself. Therefore, the kernel is positive-definite.

Warning 8.20 (Sample inclusion). The GP satisfies $\mathbb{P}\{\boldsymbol{f}_{\boldsymbol{K}} \in \mathcal{H}_{\boldsymbol{K}}\} = 0$. ▪

Although this GP connection is conceptually useful, it is not advisable to use such covariance kernels in practice unless they are known in closed form or the GP is fast to sample from and evaluate.

8.4.1 GP examples

We now present some examples of operator-valued kernels defined from a single scalar kernel [NS21, Example 2.9] and their associated RKHSs. We explicitly make use of the GP perspective and will see how output space correlations (or lack thereof) influence sample paths of the corresponding GP. For simplicity in all that follows, the input space is set to $\mathcal{X} := (0, 1)$.

Example 8.21 (Brownian bridge: Scalar). Take $\mathcal{Y} := \mathbb{R}$ and define the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, x') := \min\{x, x'\} - xx' . \tag{8.8}$$

Clearly k is symmetric. It is positive-definite because k is the covariance function of the *Brownian bridge* GP $\{b(x)\}_{x\in\mathcal{X}} \subset \mathbb{R}$, which is a Brownian motion constrained to zero at locations $x = 0$ and $x = 1$. That is,

$$b \sim \mathcal{GP}(0, k) \quad \text{and} \quad b(x) \sim \text{NORMAL}(0, k(x,x)) \ \text{ for every } \ x \in \mathcal{X}\,.$$

The RKHS $\mathcal{H}_k$ is given by the function space

$$\mathsf{H}_0^1(\mathcal{X};\mathbb{R}) = \big\{f \in \mathsf{L}_2(\mathcal{X};\mathbb{R}) \colon f' \in \mathsf{L}_2(\mathcal{X};\mathbb{R}),\ f(0) = f(1) = 0\big\}.$$

This is the Hilbert space of real-valued functions on $(0,1)$ that vanish at the endpoints with one weak derivative and equipped with the inner-product

$$\langle f,\ g\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R})} = \langle f',\ g'\rangle_{\mathsf{L}_2(\mathcal{X};\mathbb{R})} = \int_0^1 f'(t)g'(t)\,\mathrm{d}t\,.$$

Aside: Without appealing to the RKHS formalism, one can deduce that functions in $\mathsf{H}_0^1(\mathcal{X};\mathbb{R})$ are pointwise well defined by the *Sobolev embedding theorem* since here $1 > d/2 = 1/2$.

To see this connection, we perform a direct calculation in the spirit of Definition 8.2. For any $x \in \mathcal{X}$, the function $k(\cdot,x) \in \mathsf{L}_2(\mathcal{X};\mathbb{R})$ because it is bounded, vanishes at the endpoints, and has a bounded weak derivative $t \mapsto k'(t,x) = \mathbb{1}\{t < x\} - x$. Hence $k(\cdot,x) \in \mathcal{H}_k$. Last, we verify the reproducing property:

$$\begin{aligned}\langle k(\cdot,x),\ f\rangle_{(\mathcal{X};\mathbb{R})} &= \int_0^1 k'(\cdot,x)(t)f'(t)\,\mathrm{d}t \\ &= \int_0^x f'(t)\,\mathrm{d}t - x\int_0^1 f'(t)dt \\ &= f(x)\,.\end{aligned}$$

Therefore, $\mathcal{H}_k = \mathsf{H}_0^1(\mathcal{X};\mathbb{R})$. ∎

We can lift to a matrix-valued kernel by considering a finite *vector* of Brownian bridges.

Example 8.22 (Brownian bridge: Matrix). For some $p \in \mathbb{N}$, let $\mathcal{Y} = \mathbb{R}^p$. Define the function $\boldsymbol{K} \colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{p\times p}$ by

$$\boldsymbol{K}(x,x') \coloneqq k(x,x')\mathbf{I}\,,$$

where k is as in (8.8). Then $\boldsymbol{K}$ is the covariance function of $\boldsymbol{b} \coloneqq (b_1,\dots,b_p)$, where each $b_i(\cdot)$ is an independent Brownian bridge:

$$\boldsymbol{b} \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{K}) \quad \text{and} \quad \boldsymbol{b}(x) \sim \text{NORMAL}(\mathbf{0}, \boldsymbol{K}(x,x)) \ \text{ for every } \ x \in \mathcal{X}\,.$$

This is because the covariance function of the process satisfies

$$\mathbb{E}[\boldsymbol{b}(x)\boldsymbol{b}(x')^\mathsf{T}]_{ij} = \mathbb{E}[b_i(x)b_j(x')] = \delta_{ij}k(x,x') = [\boldsymbol{K}(x,x')]_{ij}\,.$$

By an argument similar to the previous example, $\boldsymbol{K}$ is a positive-definite matrix-valued kernel. Its RKHS $\mathcal{H}_{\boldsymbol{K}}$ is

$$\mathsf{H}_0^1(\mathcal{X};\mathbb{R}^p) = \big\{\boldsymbol{f} \in \mathsf{L}_2(\mathcal{X};\mathbb{R}^p) \colon \boldsymbol{f}' \in \mathsf{L}_2(\mathcal{X};\mathbb{R}^p), \boldsymbol{f}(0) = \boldsymbol{f}(1) = \mathbf{0}\big\}.$$

This is the Hilbert space of $\mathbb{R}^p$-valued functions on $(0,1)$ that vanish component-wise at the endpoints, with one weak derivative in each component, and equipped with the inner-product

$$\langle \boldsymbol{f},\ \boldsymbol{g}\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R}^p)} = \int_0^1 \langle \boldsymbol{f}'(t),\ \boldsymbol{g}'(t)\rangle_{\mathbb{R}^p}\,\mathrm{d}t = \sum\nolimits_{i=1}^p \int_0^1 f_i'(t)g_i'(t)\,\mathrm{d}t\,.$$

Indeed, verifying the two defining properties (Definition 8.10) of a vector-valued RKHS shows that $[\boldsymbol{K}(\cdot,x)\boldsymbol{y}]_i = k(\cdot,x)y_i \in \mathsf{H}_0^1(\mathcal{X};\mathbb{R})$ for all $x \in \mathcal{X}$, $\boldsymbol{y} \in \mathbb{R}^p$, and $i = 1,\dots,p$. By the inner-product definition above, we obtain $\boldsymbol{K}(\cdot,x)\boldsymbol{y} \in \mathsf{H}_0^1(\mathcal{X};\mathbb{R}^p)$. As a consequence, $f_i(x) = \langle k(\cdot,x),\ f_i\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R})}$ so that

$$\begin{aligned}\langle \boldsymbol{y},\ \boldsymbol{f}(x)\rangle_{\mathbb{R}^p} = \sum\nolimits_{i=1}^p y_i f_i(x) &= \sum\nolimits_{i=1}^p \langle k(\cdot,x)y_i,\ f_i\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R})} \\ &= \sum\nolimits_{i=1}^p \langle [\boldsymbol{K}(\cdot,x)\boldsymbol{y}]_i,\ f_i\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R})} \\ &= \langle \boldsymbol{K}(\cdot,x)\boldsymbol{y},\ \boldsymbol{f}\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathbb{R}^p)}\,.\end{aligned}$$

This is the required statement. ∎

Lifting once again, this time to an *infinite-dimensional* $\mathcal{Y}$, is straightforward given the last example and corresponds to a *function-valued* GP.

Example 8.23 (Brownian bridge: Operator). Take $\mathcal{Y} = \mathsf{L}_2(\mathsf{D};\mathbb{R})$ for some finite-dimensional, bounded domain D in Euclidean space. Define the function $\boldsymbol{K}\colon \mathcal{X}\times\mathcal{X} \to \mathscr{L}(\mathcal{Y})$ by

$$\boldsymbol{K}(x,x') \coloneqq k(x,x')\mathbf{I}\,, \tag{8.9}$$

where again k is as in (8.8) and $\mathbf{I} \in \mathscr{L}(\mathcal{Y})$ is the identity operator on $\mathsf{L}_2(\mathsf{D};\mathbb{R})$. Then $\boldsymbol{K}$ is the covariance function of GP $\boldsymbol{b} \coloneqq \{b_{\boldsymbol{s}}\}_{\boldsymbol{s}\in\mathsf{D}}$,

$$\boldsymbol{b} \sim \mathcal{GP}(\mathbf{0},\boldsymbol{K}) \quad\text{and}\quad \boldsymbol{b}(x) \sim \textsc{normal}(\mathbf{0},\boldsymbol{K}(x,x)) \ \text{ for every } \ x \in \mathcal{X}\,,$$

where each $b_{\boldsymbol{s}}(\cdot)$ is independent Brownian bridge. This is because for any $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{s} \in \mathsf{D}$, the covariance operator satisfies

$$\begin{aligned}\big(\mathbb{E}[\boldsymbol{b}(x)\otimes_{\mathcal{Y}}\boldsymbol{b}(x')]\boldsymbol{y}\big)(\boldsymbol{s}) &= \mathbb{E}[\langle \boldsymbol{b}(x'),\ \boldsymbol{y}\rangle_{\mathcal{Y}} b_{\boldsymbol{s}}(x)] \\ &= \int_{\mathsf{D}} \mathbb{E}[b_{\boldsymbol{s}}(x)b_{\boldsymbol{s}'}(x')]y(\boldsymbol{s}')\,\mathrm{d}\boldsymbol{s}' \\ &= \int_{\mathsf{D}} k(x,x')\delta(\boldsymbol{s}-\boldsymbol{s}')y(\boldsymbol{s}')\,\mathrm{d}\boldsymbol{s}' \\ &= k(x,x')y(\boldsymbol{s})\,.\end{aligned}$$

These calculations with white noise, though formal, can be made rigorous with the *Itô isometry* from stochastic analysis.

Calculations similar to those in the previous two examples show that $\boldsymbol{K}$ is a positive-definite operator-valued kernel with RKHS $\mathscr{H}_{\boldsymbol{K}}$ given by

$$\mathsf{H}_0^1(\mathcal{X};\mathcal{Y}) = \big\{\boldsymbol{f} \in \mathsf{L}_2(\mathcal{X};\mathcal{Y})\colon \boldsymbol{f}' \in \mathsf{L}_2(\mathcal{X};\mathcal{Y}), \boldsymbol{f}(0) = \boldsymbol{f}(1) = \mathbf{0}\big\}. \tag{8.10}$$

This is the Hilbert space of $\mathsf{L}_2(\mathsf{D};\mathbb{R})$-valued functions on $(0,1)$ that vanish in the L_2-sense at the endpoints with one weak Fréchet derivative, and equipped with the inner-product

$$\begin{aligned}\langle \boldsymbol{f},\ \boldsymbol{g}\rangle_{\mathsf{H}_0^1(\mathcal{X};\mathcal{Y})} &= \int_0^1 \langle \boldsymbol{f}'(t),\ \boldsymbol{g}'(t)\rangle_{\mathcal{Y}}\,\mathrm{d}t \\ &= \int_{\mathsf{D}}\int_0^1 [f'(t)](\boldsymbol{s})[g'(t)](\boldsymbol{s})\mathrm{d}t\mathrm{d}\boldsymbol{s}\,.\end{aligned}$$

Aside: *Lebesgue–Bochner spaces* such as $\mathsf{L}_2((0,1);\mathsf{L}_2)$ here arise frequently in the analysis of time-dependent partial differential equations.

∎

Exercise 8.24 (Lebesgue–Bochner). Complete the calculations in Example 8.23 that prove $\boldsymbol{K}$ in (8.9) is positive-definite with RKHS (8.10).

Warning 8.25 (Sample path regularity). Using a separable operator-valued kernel with isotropic output operator such as $\mathbf{I} \in \mathscr{L}(\mathcal{Y})$ in (8.9) can be *detrimental in infinite dimensions*. Not only does this completely uncorrelate the outputs, but also $\mathbf{I}$ does not have finite trace on $\mathcal{Y}$ whenever $\mathcal{Y}$ is infinite-dimensional. Thus, the function-valued GP associated to the operator-valued kernel does not take its values in $\mathcal{Y}$ with probability one! In the language of GPs, this is a bad prior distribution.

Indeed, following Example 8.23, for $x \in (0,1)$,

$$\boldsymbol{b}(x) \sim \text{NORMAL}(\mathbf{0}, \boldsymbol{K}(x,x)) = k(x,x)\text{NORMAL}(\mathbf{0}, \mathbf{I}) .$$

This is not a Gaussian random function in $\mathcal{Y}$ but instead Gaussian white noise on $\mathcal{Y}$; it belongs to a much "rougher" space than $\mathcal{Y}$ itself. Instead, taking $\boldsymbol{K} : (x,x') \mapsto k(x,x')\boldsymbol{T}$ with self-adjoint and psd trace-class $\boldsymbol{T} \in \mathsf{S}_1(\mathcal{Y})$ addresses this issue. ∎

Here S_1 denotes the Schatten-1 class operators.

8.4.2 Implications for learning theory

We end this lecture with a brief remark on kernel methods in machine learning. Our emphasis on the RKHS model stems partly from its role in *kernel ridge regression* (KRR) [CDV07], which is a regularization procedure to recover a vector-valued function from a finite number of noisy input–output data pairs. The direct link to vector-valued GPs is the fact that the KRR estimator is the posterior mean of a GP conditioned on these data pairs.

We now state a core result in this field.

Theorem 8.26 (Kernel ridge regression). Let $\mathcal{H}_{\boldsymbol{K}}$ be a vector-valued RKHS with positive-definite operator-valued kernel $\boldsymbol{K} : \mathcal{X} \times \mathcal{X} \to \mathscr{L}(\mathcal{Y})$. For $n \in \mathbb{N}$, let $(\mathsf{X}, \mathsf{Y}) := \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1,\dots,n} \subset \mathcal{X} \times \mathcal{Y}$ be a dataset of n input–output pairs. Write $\boldsymbol{y} := (\boldsymbol{y}_i)_i \in \mathcal{Y}^n$ for the vector of concatenated outputs. Then for $\lambda > 0$, the solution to

$$\operatorname{minimize}\left\{ \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{x}_i)\|_{\mathcal{Y}}^2 + \frac{\lambda}{n} \|\boldsymbol{f}\|_{\mathcal{H}_{\boldsymbol{K}}}^2 : \boldsymbol{f} \in \mathcal{H}_{\boldsymbol{K}} \right\}$$

is attained by

$$\hat{\boldsymbol{f}} = \sum_{i=1}^n \boldsymbol{K}(\cdot, \boldsymbol{x}_i)\boldsymbol{\beta}_i ,$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}_i)_i \in \mathcal{Y}^n$ solves the block linear operator equation

$$(\boldsymbol{K}(\mathsf{X},\mathsf{X}) + \lambda \mathbf{I})\boldsymbol{\beta} = \boldsymbol{y} . \tag{8.11}$$

A major shortcoming of vector-valued kernel ridge regression is its lack of scalability. Taking $\mathcal{Y} = \mathbb{F}^p$, the block kernel matrix $\boldsymbol{K}(\mathsf{X},\mathsf{X})$ requires $O(n^2p^2)$ space complexity and the linear solve in (8.11) has $O(n^3p^3)$ time complexity. Yet, there is much research activity attempting to improve scalability, e.g., with random features.

Proof. The proof is a standard exercise in orthogonality, see Problem 8.27. ∎

Problem 8.27 (Representer). Prove Theorem 8.26.

Notes

The presentation of operator-valued kernels in this lecture is new. The theory of RKHS is largely drawn from [MP05] while many of the exercises scattered throughout are original. The examples of separable kernels are from [Kad+16] and [MP05]. The random feature and GP Brownian bridge examples in the operator-valued setting are new, as is the characterization of their RKHSs.

A complete theory of operator-valued kernels in a much more general setting than that presented here was developed in the 1960s [Sch64]. Since then, many refinements

have been made, especially with learning theoretic considerations in mind [CDV07; Cap+08; Car+10; MP05]. There is a rich but technical literature on *universal kernels* in both the scalar- and operator-valued setting [Cap+08; Car+10; MXZ06]. These are kernels whose associated RKHS is dense in the space of continuous functions equipped with the topology of uniform convergence over compact subsets. This universal approximation property is of direct relevance to well-posed learning algorithms. Under fairly weak assumptions the translation-invariant and radial kernels characterized by Bochner's and Schoenberg's theorems are universal. Random feature developments along similar lines in the operator-valued setting may be found in [BHB16; Min16].

Lecture bibliography

[ARL+12] M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3 (2012), pages 195–266.

[Aro50] N. Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pages 337–404.

[BHB16] R. Brault, M. Heinonen, and F. Buc. "Random fourier features for operator-valued kernels". In: *Asian Conference on Machine Learning*. PMLR. 2016, pages 110–125.

[CDV07] A. Caponnetto and E. De Vito. "Optimal rates for the regularized least-squares algorithm". In: *Foundations of Computational Mathematics* 7.3 (2007), pages 331–368.

[Cap+08] A. Caponnetto et al. "Universal multi-task kernels". In: *The Journal of Machine Learning Research* 9 (2008), pages 1615–1646.

[Car+10] C. Carmeli et al. "Vector valued reproducing kernel Hilbert spaces and universality". In: *Analysis and Applications* 8.01 (2010), pages 19–61.

[Car97] R. Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pages 41–75.

[Eva10] L. C. Evans. *Partial differential equations*. American Mathematical Soc., 2010.

[Evg+05] T. Evgeniou et al. "Learning multiple tasks with kernel methods." In: *Journal of machine learning research* 6.4 (2005).

[Kad+16] H. Kadri et al. "Operator-valued kernels for learning from functional response data". In: *Journal of Machine Learning Research* 17.20 (2016), pages 1–54.

[MP05] C. A. Micchelli and M. Pontil. "On learning vector-valued functions". In: *Neural computation* 17.1 (2005), pages 177–204.

[MXZ06] C. A. Micchelli, Y. Xu, and H. Zhang. "Universal Kernels." In: *Journal of Machine Learning Research* 7.12 (2006).

[Min16] H. Q. Minh. "Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning". In: *arXiv preprint arXiv:1608.05639* (2016).

[NS21] N. H. Nelsen and A. M. Stuart. "The random feature model for input-output maps between banach spaces". In: *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243.

[RR08] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., 2008, pages 1177–1184. URL: http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf.

[Sch64] L. Schwartz. "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants)". In: *Journal d'analyse mathématique* 13.1 (1964), pages 115–256.

9. Spectral Radius Perturbation and Stability

Date: 15 March 2022 **Author:** Jing Yu

Agenda:
1. System Stability and Spectral Radius
2. Geršgorin Disks
3. Bounding Spectral Radius

Eigenvalues of a matrix are continuous functions of the entries of the matrix. We have studied how eigenvalues of a Hermitian matrix change with respect to additive Hermitian perturbations in Lecture 9. Beyond the Hermitian setting, there are many perturbation results. The eigenvalue perturbation problem arises in many applications. In particular, the stability of a linear dynamical system is characterized by the location of the eigenvalues of the system matrices.

In this project, we introduce an eigenvalue perturbation problem that is intimately related to the robustness of a given linear controller for a liner dynamical system. We will survey some perturbation results relevant to our problem setting, and demonstrate their usage for the control problem at hand.

9.1 System stability and spectral radius

A central question in linear control theory is whether a given linear time-invariant (LTI) control system is stable under a chosen feedback gain. An LTI system is represented as

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t, \tag{9.1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$ is the state indexed by discrete time t, vector $\boldsymbol{u}_t \in \mathbb{R}^m$ is the control input, while $\boldsymbol{A} \in \mathbb{M}_n$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ are called system matrices. Generally, one wants to design a feedback gain matrix $\boldsymbol{K} \in \mathbb{R}^{m\times n}$ such that control inputs are chosen as $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{x}_t$ for the system (9.1). In general, feedback gain $\boldsymbol{K}$ is designed with respect to the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Under a fixed feedback gain $\boldsymbol{K}$, the *closed-loop dynamics* of (9.1) is

$$\boldsymbol{x}_{t+1} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}_t. \tag{9.2}$$

A key property for studying LTI systems is the *spectral radius*.

Definition 9.1 (Spectral radius). The spectral radius of a matrix $\boldsymbol{A} \in \mathbb{M}_n$ is denoted as $\rho(\boldsymbol{A})$ and defined as

$$\rho(\boldsymbol{A}) := \max_i |\lambda_i(\boldsymbol{A})|$$

where $\lambda_i(\boldsymbol{A})$'s are the eigenvalues of $\boldsymbol{A}$.

The next definition connects *stability* of (9.2) with the spectral radius of the system matrices.

Definition 9.2 (Stability). We say the closed-loop dynamics is stable under feedback gain $\boldsymbol{K}$ if $\rho(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}) < 1$.

A natural question in control design is the following.

Aside: Another common stability definition states that (9.2) is stable if for all $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{x}_t \to 0$ as $t \to \infty$. Using Jordan decomposition of A, it can be readily checked that the two definitions of the stability are equivalent.

For a feedback gain $\boldsymbol{K}$ designed for $\boldsymbol{A}$ and $\boldsymbol{B}$, how much can we perturb $\boldsymbol{A}$ with $\boldsymbol{\Delta} \in \mathbb{M}_n$ such that $\rho(\boldsymbol{A} + \boldsymbol{\Delta} - \boldsymbol{B}\boldsymbol{K}) < 1$?

Depending on how $\boldsymbol{K}$ is synthesized according to $\boldsymbol{A}$, the answer to this question can be quite different. The property that a feedback gain $\boldsymbol{K}$ maintains closed-loop stability when $\boldsymbol{A}$ varies is called *robustness* in control theory literature. In this project, we will investigate the robustness of a specific feedback gain that is synthesized from the discrete-time algebraic Riccati equation (DARE) for a class of LTI systems that has block diagonal $\boldsymbol{A}$ and $\boldsymbol{B}$.

Consider dynamics (9.1) with diagonal $\boldsymbol{A}$ and $\boldsymbol{B}$. A linear quadratic (LQ) optimal control gain $\boldsymbol{K}^\star$ is the solution to the following optimization,

$$\begin{aligned} &\text{minimize}_{\boldsymbol{K} \in \mathbb{R}^{m \times n}} \sum_{t=1}^{\infty} \|\boldsymbol{x}_t\|_{\boldsymbol{Q}}^2 + \|\boldsymbol{u}_t\|_{\boldsymbol{R}}^2 \\ &\text{subject to} \quad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t \\ &\qquad\qquad\ \ \boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{x}_t, \end{aligned}$$

where $\|\boldsymbol{x}_t\|_{\boldsymbol{Q}}^2 = \boldsymbol{x}_t^\mathsf{T}\boldsymbol{Q}\boldsymbol{x}_t$ and $\|\boldsymbol{u}_t\|_{\boldsymbol{R}}^2 = \boldsymbol{u}_t^\mathsf{T}\boldsymbol{R}\boldsymbol{u}_t$ with block diagonal matrices $\boldsymbol{Q}, \boldsymbol{R} \succ 0$. Note that the objective is finite if and only if the closed loop is stable.

It is well known that the optimal solution is given by the solution $\boldsymbol{P} = \boldsymbol{P}^\mathsf{T} \succ 0$ to DARE shown below,

$$\boldsymbol{P} = \boldsymbol{A}^\mathsf{T}\boldsymbol{P}\boldsymbol{A} - \left(\boldsymbol{A}^\mathsf{T}\boldsymbol{P}\boldsymbol{B}\right)\left(\boldsymbol{R} + \boldsymbol{B}^\mathsf{T}\boldsymbol{P}\boldsymbol{B}\right)^{-1}\left(\boldsymbol{B}^\mathsf{T}\boldsymbol{P}\boldsymbol{A}\right) + \boldsymbol{Q}. \tag{9.4}$$

The optimal control gain $\boldsymbol{K}^\star$ can be computed as

$$\boldsymbol{K}^\star = \left(\boldsymbol{R} + \boldsymbol{B}^\mathsf{T}\boldsymbol{P}\boldsymbol{B}\right)^{-1}\left(\boldsymbol{B}^\mathsf{T}\boldsymbol{P}\boldsymbol{A}\right). \tag{9.5}$$

When $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{Q}$, $\boldsymbol{R}$ are all diagonal, it can be shown that the $\boldsymbol{K}^\star$ is diagonal as well. These diagonal structures arise when one considers dynamical systems that are made up of multiple independent scalar subsystems. The generalized setting considers block-diagonal matrices instead of diagonal matrices. Example of such systems are networks of swarm robots.

Now suppose that $\boldsymbol{A}$ is not perfectly diagonal but instead is nonzero at the off-diagonal entries. How much of these nonzero off-diagonal entries can $\boldsymbol{K}^\star$ tolerate before the closed-loop stability is lost under $\boldsymbol{K}^\star$? This question corresponds to the case where we design an optimal controller for each subsystem in the network independently, while in reality the subsystems are dynamically coupled loosely. This type of analysis quantifies how dynamically coupled the subsystems can be before $\boldsymbol{K}^\star$ (designed assuming no dynamical coupling) no longer stabilizes the closed loop.

9.2 Geršgorin disks

Given a matrix $\boldsymbol{A} \in \mathbb{M}_n$, we can always write $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{B}$ where $\boldsymbol{D}$ is a diagonal matrix containing the main diagonal of $\boldsymbol{A}$ and $\boldsymbol{B}$ is a zero-diagonal matrix containing the off-diagonal entries of $\boldsymbol{A}$. Now consider the matrix $\boldsymbol{A}_t = \boldsymbol{D} + t\boldsymbol{B}$, then we have $\boldsymbol{A}_0 = \boldsymbol{D}$ and $\boldsymbol{A}_1 = \boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}_0$ are immediately the diagonal elements of the matrix. On the other hand, if t is small enough, the eigenvalues of $\boldsymbol{A}_t$ will lie in a neighborhood of the diagonal entries of $\boldsymbol{A}$ due to continuity. The Geršgorin disk theorem formalizes this intuition about the locations of the eigenvalues. In particular, it provides an explicit bound for the eigenvalues using the diagonal entries of a matrix. Let us first define the Geršgorin discs of a matrix.

Definition 9.3 (Geršgorin disc). Let $\boldsymbol{A} = [a_{ij}] \in \mathbb{M}_n$. The Geršgorin discs of $\boldsymbol{A}$ are

$$\mathsf{D}_i(\boldsymbol{A}) = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum\nolimits_{j \neq i} |a_{ij}| \right\},$$

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

In other words, the Geršgorin discs are balls in the complex plane with a_{ii} as the center and $\sum_{j \neq i} |a_{ij}|$ as the radius. It turns out the Geršgorin discs can be used to locate the eigenvalues of a matrix.

Aside: Both Definition 9.3 and Theorem 9.4 are valid for complex matrices as well.

Theorem 9.4 (Geršgorin [HJ13]). The eigenvalues of $\boldsymbol{A} \in \mathbb{M}_n$ are in the union of its Geršgorin discs $\bigcup_{i=1}^{n} \mathsf{D}_i(\boldsymbol{A})$. Furthermore, if the union of k discs is disjoint from the remaining $n - k$ discs, then the union contains exactly k eigenvalues of $\boldsymbol{A}$, counted according to their algebraic multiplicities.

Proof. Let $\lambda, \boldsymbol{v} = [v_j] \in \mathbb{R}^n$ be an eigenpair of $\boldsymbol{A}$ where $\boldsymbol{v} \neq 0$. Let $i \in \{1, 2, \ldots, n\}$ be an index such that $\|\boldsymbol{v}\|_\infty = |v_i| \neq 0$. Rewriting $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$ for the ith entry, we have

$$\sum\nolimits_j a_{ij} v_j = \lambda v_i.$$

Subtracting $a_{ii} v_i$ from both sides,

$$\sum\nolimits_{j \neq i} a_{ij} v_j = (\lambda - a_{ii})\, v_i.$$

Therefore, using triangle inequality and $\|\boldsymbol{v}\|_\infty = |v_i|$,

$$|\lambda - a_{ii}||v_i| = |(\lambda - a_{ii})\, v_i| = \left| \sum\nolimits_{j \neq i} a_{ij} v_j \right| \leq \sum\nolimits_{j \neq i} |a_{ij}||v_i|.$$

Dividing both sides by $|v_i|$, we conclude $\lambda \in \mathsf{D}_i(\boldsymbol{A})$.

Now suppose there are k Geršgorin discs such that the union is disjoint from the rest of the $n - k$ discs. Without loss of generality, we assume the first k Geršgorin discs are such discs, and we denote their union as $G_k(\boldsymbol{A}) = \bigcup_{i=1}^{k} \mathsf{D}_i(\boldsymbol{A})$. We write $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{B}$ where $\boldsymbol{D}$ is a diagonal matrix containing the diagonal entries of $\boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{D}$. Consider $\boldsymbol{A}_t = \boldsymbol{D} + t\boldsymbol{B}$ with $t \in [0, 1]$. Then $\boldsymbol{A_0} = \boldsymbol{D}$ and $\boldsymbol{A}_1 = \boldsymbol{A}$. In particular, observe that the radii of each Geršgorin disc of $\boldsymbol{A}_t$, $t\boldsymbol{B}$, and $t\boldsymbol{A}$ are equal. Therefore, each Geršgorin disc of $\boldsymbol{A}_t$ is contained in the corresponding Geršgorin disc of $\boldsymbol{A}$. Consequently, the corresponding union of k discs of $\boldsymbol{A_t}$, denoted as $G_k(\boldsymbol{A}_t)$, is contained in $G_k(\boldsymbol{A})$. This in turn means that $G_k(\boldsymbol{A}_t)$ is disjoint from the rest of $n - k$ Geršgorin discs of $\boldsymbol{A}_t$. The above argument holds for all $t \in [0, 1]$.

Our goal now is to show that the number of eigenvalues contained inside of any curve surrounding $G_k(\boldsymbol{A}_t)$ remains constant for all $t \in [0, 1]$. This, combined with the fact that $G_k(\boldsymbol{A}_0)$ contains exactly k eigenvalues of $\boldsymbol{A}_0$ ($a_{11}, \ldots, a_{kk}$) assert that $G_k(\boldsymbol{A})$ contains k eigenvalues of $\boldsymbol{A}$. Specifically, let Γ be a simple closed finite-length curve in the complex plane that surrounds $G_k(\boldsymbol{A}_t)$ and is disjoint from the rest of the $n - k$ Geršgorin discs of $\boldsymbol{A}$. Curve Γ does not pass through any eigenvalue of any $\boldsymbol{A}_t$. Let $p_t(z)$ denote the characteristic polynomial of $\boldsymbol{A}_t$, that is, $p_t(z) = \det(z\mathbf{I} - \boldsymbol{A}_t)$. Observe that the function $p_t(z)$ is a polynomial in t. Further, $p_t(z)$ has no poles and has zeros that are the eigenvalues of $\boldsymbol{A}_t$ inside of Γ. Therefore, we apply Cauchy's argument principle to $p_t(z)$, which states

$$N(t) = \frac{1}{2\pi i} \oint_\Gamma \frac{p_t'(z)}{p_t(z)} \mathrm{d}z,$$

where $N(t)$ is an integer function of t and denotes the number of the zeros of $p_t(z)$ inside of Γ. Since $N(t)$ is polynomial on t and thus a continuous function of t on the bounded domain $[0,1]$, we deduce that $N(t)$ is a constant function on $t \in [0,1]$ by Liouville's theorem. This means that the number of zeros of $p_t(z)$ (eigenvalues of $\boldsymbol{A}_t$) inside of Γ remain the same regardless of our choice of t. Since $N(0) = k$, we know $N(1) = k$. Therefore, there are k eigenvalues of $\boldsymbol{A}$ inside Γ. By the first part of the theorem, we know that any eigenvalues of $\boldsymbol{A}$ has to lie on a Geršgorin disc. Therefore, we conclude that there are exactly k eigenvalues of $\boldsymbol{A}$ in $G_k(\boldsymbol{A})$. ■

Theorem 9.4 provides a nice quantitative bound on the eigenvalues. One may wonder if there are refinement or alternatives that can help improve our estimation of the locations of the eigenvalues. The answer is affirmative. Below we list a few such results.

Corollary 9.5 (Transpose discs). The eigenvalues of a matrix $\boldsymbol{A}$ are in the insersections of the Geršgorin discs of $\boldsymbol{A}^{\mathsf{T}}$.

Proof. Observe that $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathsf{T}}$ have the same eigenvalues. Apply Theorem 9.4 to $\boldsymbol{A}^{\mathsf{T}}$ to obtain the result. ■

Corollary 9.6 (Scaling discs). Let $\boldsymbol{A} = [a_{ij}] \in \mathbb{M}_n$ and let $\boldsymbol{D} \in \mathbb{M}_n$ be a diagonal matrix with positive real numbers $p_1, p_2, \ldots, p_n$ on the main diagonal. The eigenvalues λ_i of $\boldsymbol{A}$ are in the union of the Geršgorin discs of $\boldsymbol{D^{-1}AD}$, that is,

$$\lambda_i(\boldsymbol{A}) \in \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \frac{1}{p_i} \sum\nolimits_{j \neq i} p_j |a_{ij}| \right\},$$

for all $i = 1, \ldots, n$.

Together, Theorem 9.4, Corollary 9.5, and 9.6 provide three different approximations of the eigenvalues of $\boldsymbol{A}$, with 9.6 being especially useful in some cases.

Example 9.7 Consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 10 & 5 & 8 \\ 0 & -11 & 1 \\ 0 & -1 & 13 \end{bmatrix}.$$

The eigenvalues of $\boldsymbol{A}$ are $-10, 10, 12.96$. The Geršgorin discs of $\boldsymbol{A}$ are plotted in Figure 9.1. In particular, the union of the blue and the yellow disc is disjoint from the red disc. Corroborating Theorem 9.4, we see that there are exactly 2 eigenvalues of $\boldsymbol{A}$ inside the union of the blue and the yellow disc.

Now consider a scaling matrix $\boldsymbol{D} = \operatorname{diag}(13, 1, 1)$. The Geršgorin Disks for the matrix $\boldsymbol{D}^{-1}\boldsymbol{AD}$ is shown in Figure 9.2. This particular choice of scaling significantly improves the estimation for eigenvalue 10. For this particular example, we can actually compute the optimal scaling matrix. To see this, note that the radii of the scaled discs will be $(5p_2 + 8p_3)/p_1$, p_3/p_2, p_2/p_3. Therefore, the best we could do without making a worse estimation than the original discs is to set $p_2 = p_3$, and this will mean that we can scale the radius of the first disc $13p_2/p_1$ arbitrarily by making p_1 large to improve our estimation of the eigenvalue 10. We can achieve such improvement all while maintaining the same estimation for the eigenvalue -10 and 12.96. ▪

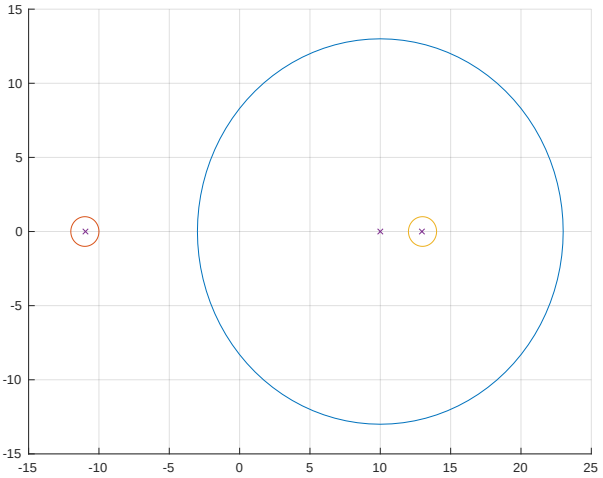

Figure 9.1 The eigenvalues and Geršgorin Disks of a matrix plotted on the complex plane. The crosses are the eigenvalues of the matrix while the circles denotes the Geršgorin Disks of the matrix.

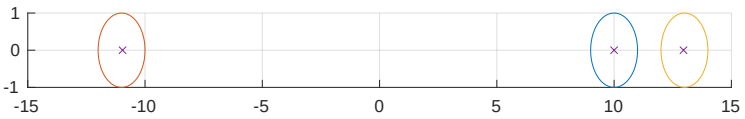

Figure 9.2 The scaled Geršgorin Disks of the same matrix plotted on the complex plane. The crosses are the eigenvalues of the matrix while the circles are the Geršgorin Disks of the matrix. Compared to the original Geršgorin Disks in Figure 9.1, the estimation using a scaling matrix significantly improves for the blue circle disc, while maintaining the same estimation radius for the other circles.

9.3 Bounding spectral radius

Recall our original problem in Section 9.1. We would like to understand how sensitive the stability of the closed-loop dynamics (9.2) is to the change of the dynamical coupling among subsystems. With tools from the previous section, we can, for example, directly apply Theorem 9.4 and conclude the following.

Corollary 9.8 (Sufficient condition for robust stability). Let $\boldsymbol{\Delta} = [\delta_{ij}]$ be a zero-diagonal matrix denoting the unaccounted dynamical coupling among subsystems. Given an LQ optimal feedback gain $\boldsymbol{K}^\star$ designed for diagonal system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the closed-loop dynamics under $\boldsymbol{K}^\star$ applied to $\boldsymbol{A}+\boldsymbol{\Delta}$ and $\boldsymbol{B}$ is stable if the coupling matrix $\boldsymbol{\Delta}$ satisfies

$$\sum_{j=1}^{n} |\delta_{ij}| < 1 - |(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}^\star)_{ii}| \quad \text{for } i = 1, \dots, n, \tag{9.6}$$

where $(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}^\star)_{ii}$ denotes the ith diagonal entry of $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}^\star$.

From this, we can also see that the more stable (eigenvalues are smaller) the LQ controller is for the ideal diagonal system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the more margin we have for the dynamically coupled true dynamics $\boldsymbol{A}+\boldsymbol{\Delta}$ and $\boldsymbol{B}$ to vary. Controllers with more margin to handle perturbation in the system matrices are said to be more robust.

Example 9.9 (of Corollary 9.8). Consider the following nominal dynamics

$$\boldsymbol{x}_{t+1} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\boldsymbol{A}} \boldsymbol{x}_t + \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\boldsymbol{B}} \boldsymbol{u}_t,$$

for two independently evolving scalar states $x_t^1 := \boldsymbol{x}_t(1)$ and $x_t^2 := \boldsymbol{x}_t(2)$ with their own control input $u_t^1 := \boldsymbol{u}_t(1)$ and $u_t^2 := \boldsymbol{u}_t(2)$ respectively. Now suppose that we have used the LQ optimal control theory to design the control gain $\boldsymbol{K} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$ via (9.4) and (9.5) for qudratic cost $\boldsymbol{Q} = \mathbf{I}$ and $\boldsymbol{R} = 20 \cdot \mathbf{I}$. Let $\boldsymbol{\Delta} = \begin{bmatrix} 0 & \alpha \\ \beta & 0 \end{bmatrix}$ be the unaccounted dynamics coupling between x^1 and x^2 such that the true system matrices are $\boldsymbol{A}+\boldsymbol{\Delta}$, $\boldsymbol{B}$ instead of $\boldsymbol{A}$, $\boldsymbol{B}$.

A simple calculation reveals that $(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}) = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$. According to the sufficient condition (9.6), as long as α and β individually does not exceed 0.2, then $\boldsymbol{K}$ designed for diagonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ continues to stabilize $\boldsymbol{A}+\boldsymbol{\Delta}$ and $\boldsymbol{B}$. We can numerically check how tight this sufficient condition is in this example for varying values of α and β. In Figure 9.3, each horizontal curve on the surface corresponds to a fixed β value as we sweep over different α values. Any points on the surface that has y axis value of more than 1 mean that the corresponding closed-loop system is unstable. Sufficient condition (9.6) is loose, because for the top curve (top boundary of the surface) which corresponds to $\beta = 0.3$, there are clearly α values small enough such that the spectral radius of the resulting closed loop remains below 1. On the other hand, the sufficient condition (9.6) is tight in the following sense. Although not explicitly shown in the plot, when $\alpha = 0.2$ and $\beta = 0.2$, i.e., (9.6) holds in equality, then the spectral radius of the closed loop is 1, which is the cross-over point between stability and instability. Combined with the plot, we see that if we fix one of the two variable to be 0.2, then the closed loop system becomes unstable as soon as the other variable exceeds 0.2. ∎

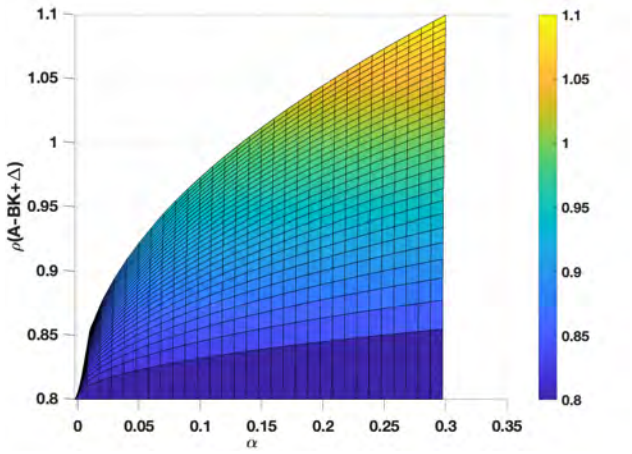

Figure 9.3 The spectral radius of the true closed loop $\rho\left(\boldsymbol{A}-\boldsymbol{BK}+\boldsymbol{\Delta}\right)$ for fixed β values varying from 0 to 0.3.

Since stability of a linear dynamical system is entirely characterized by the spectral radius of the closed-loop system, it is relevant to study various ways to bound the spectral radius of a matrix. We first present and prove the following variational characterization of the spectral radius of a generic real matrix.

Lemma 9.10 (Variational characterization of spectral radius). For $\boldsymbol{A} \in \mathbb{M}_n$, the following holds true:

$$\rho(\boldsymbol{A}) = \inf_{D \in \mathbb{M}_n \text{ invertible}} \|\boldsymbol{DAD}^{-1}\|_2.$$

Proof. Let $\boldsymbol{D} \in \mathbb{M}_n$ be an invertible matrix. Then $\rho(\boldsymbol{A}) = \rho(\boldsymbol{DAD}^{-1}) \leq \|\boldsymbol{DAD}^{-1}\|_2$, where the first equality comes from the fact that similarity transformations do not affect eigenvalues. This equality holds for any $\boldsymbol{D} \in \mathbb{M}_n$, therefore $\rho(\boldsymbol{A}) \leq \inf_{D \in \mathbb{M}_n \text{ invertible}} \|\boldsymbol{DAD}^{-1}\|_2$.

Now suppose $\boldsymbol{A}$ has Jordan decomposition $\boldsymbol{A} = \boldsymbol{T}^{-1}\boldsymbol{JT}$ where $\boldsymbol{J}$ is in the Jordan normal form with eigenvalues of $\boldsymbol{A}$ on the diagonal and 1's on the super diagonal. Let $\boldsymbol{P} = \text{diag}(1, k, k^2, \ldots, k^{n-1})$ where $k > 0$. Observe that the matrix $\boldsymbol{PJP}^{-1}$ is a matrix with eigenvalues of $\boldsymbol{A}$ on the diagonal and $\frac{1}{k}$ on the super diagonal. Therefore, as $k \to 0$, $\boldsymbol{PJP}^{-1} \to \text{diag}(\boldsymbol{\lambda}(\boldsymbol{A}))$ where $\boldsymbol{\lambda}(\boldsymbol{A})$ is the vector of eigenvalues of $\boldsymbol{A}$ repeating with algebraic multiplicity. Therefore, $\|\boldsymbol{PJP}^{-1}\|_2 \to \rho(\boldsymbol{A})$. Letting $\boldsymbol{D} = \boldsymbol{PT}$, we have $\rho(\boldsymbol{A}) = \boldsymbol{DAD}^{-1} \geq \inf_{\boldsymbol{D} \in \mathbb{M}_n \text{ invertible}} \|\boldsymbol{DAD}^{-1}\|_2$. ∎

Aside: Here we use $\|\cdot\|_2$ to denote the spectral norm.

The variational characterization of the spectral radius reminds us of Corollary 9.6, where we bounded the eigenvalues of a matrix $\boldsymbol{A}$ with the entries of $\boldsymbol{DAD}^{-1}$ for $\boldsymbol{D}$ a diagonal matrix with positive real diagonal entries. Indeed, we can bound the spectral radius of $\boldsymbol{A}$ also using the entries of $\boldsymbol{DAD}^{-1}$.

Corollary 9.11 (Variational bound on spectral radius via Geršgorin). Let $\boldsymbol{A} = [a_{ij}] \in \mathbb{M}_n$. Then

$$\rho(\boldsymbol{A}) \leq \min_{p_1,\ldots,p_n>0} \max_{1 \leq i \leq n} \frac{1}{p_i} \sum_{j=1}^{n} p_j |a_{ij}|$$

and

$$\rho(\boldsymbol{A}) \leq \min_{p_1,\ldots,p_n>0} \max_{1 \leq j \leq n} \frac{1}{p_j} \sum_{i=1}^{n} p_i |a_{ij}|. \tag{9.7}$$

Proof. By Corollary 9.6, all eigenvalues of $\boldsymbol{A}$ belong to the union of the Geršgorin discs, $\bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \frac{1}{p_i} \sum_{j \neq i} p_j |a_{ij}| \right\}$. For each of the Geršgorin disc of the scaled matrix $\boldsymbol{DAD}^{-1}$ for $\boldsymbol{D} = \operatorname{diag}(p_1, \ldots, p_n)$ with $p_1, \ldots, p_n > 0$, the furthest point away from the origin on the complex plane is $\frac{1}{p_i} \sum_{j=1}^{n} p_j |a_{ij}|$. Therefore, all eigenvalues of $\boldsymbol{A}$, including the largest eigenvalue, has to be less than the maximum furthest point over all Geršgorin discs of $\boldsymbol{DAD}^{-1}$. Now apply the same reasoning to $\boldsymbol{A}^{\mathsf{T}}$ using Corollary 9.5, we obtain (9.7). ∎

9.4 Notes

The majority of the results presented here is from chapter 6 of [HJ13]. Some corollaries are exercises from [HJ13]. The examples are inspired and tweaked from [Mar+16]. The variational characterization of the spectral radius is presented in [Hua+21] as a fact and proved here by JY.

Lecture bibliography

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[Hua+21] D. Huang et al. "Matrix Concentration for Products". In: *Foundations of Computational Mathematics* (2021), pages 1–33.

[Mar+16] D. Marquis et al. *'Gershgorin's Circle Theorem for Estimating the Eigenvalues of a Matrix with Known Error Bounds"*. 2016.

IV.

back matter

Bibliography

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. For sale by the Superintendent of Documents. U. S. Government Printing Office, Washington, D.C., 1964.

[Ahl66] L. V. Ahlfors. *Complex analysis: An introduction of the theory of analytic functions of one complex variable*. Second. McGraw-Hill Book Co., New York-Toronto-London, 1966.

[AW01] R. Ahlswede and A. Winter. *Strong Converse for Identification via Quantum Channels*. 2001. arXiv: quant-ph/0012127 [quant-ph].

[Alo86] N. Alon. "Eigenvalues and Expanders". In: *Combinatorica* 6.2 (June 1986), pages 83–96. DOI: 10.1007/BF02579166.

[ANo6] N. Alon and A. Naor. "Approximating the cut-norm via Grothendieck's inequality". In: *SIAM J. Comput.* 35.4 (2006), pages 787–803. DOI: 10.1137/S0097539704441629.

[AP20] J. Altschuler and P. Parrilo. "Approximating Min-Mean-Cycle for low-diameter graphs in near-optimal time and memory". Available at https://arxiv.org/abs/2004.03114. 2020.

[AP22] J. Altschuler and P. Parrilo. "Near-linear convergence of the Random Osborne algorithm for Matrix Balancing". In: *Math. Programming* (2022). To appear.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017.

[Alt+21] J. Altschuler et al. "Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 2021.

[ARL+12] M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3 (2012), pages 195–266.

[AGV21] N. Anari, S. O. Gharan, and C. Vinzant. "Log-concave polynomials, I: entropy and a deterministic approximation algorithm for counting bases of matroids". In: *Duke Math. J.* 170.16 (2021), pages 3459–3504. DOI: 10.1215/00127094-2020-0091.

[AJD69] W. N. Anderson Jr and R. J. Duffin. "Series and parallel addition of matrices". In: *Journal of Mathematical Analysis and Applications* 26.3 (1969), pages 576–594.

[AHJ87] T. Ando, R. A. Horn, and C. R. Johnson. "The singular values of a Hadamard product: a basic inequality". In: *Linear and Multilinear Algebra* 21.4 (1987), pages 345–365. eprint: https://doi.org/10.1080/03081088708817810. DOI: 10.1080/03081088708817810.

[And79] T. Ando. "Concavity of certain maps on positive definite matrices and applications to Hadamard products". In: *Linear Algebra and its Applications* 26 (1979), pages 203–241.

[And89] T. Ando. "Majorization, doubly stochastic matrices, and comparison of eigenvalues". In: *Linear Algebra and its Applications* 118 (1989), pages 163–248.

[ABY20] R. Aoun, M. Banna, and P. Youssef. "Matrix Poincaré inequalities and concentration". In: *Advances in Mathematics* 371 (2020), page 107251. DOI: https://doi.org/10.1016/j.aim.2020.107251.

[ABo8] T. Arbogast and J. L. Bona. *Methods of applied mathematics*. ICES Report. UT-Austin, 2008.

[Aro50] N. Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pages 337–404.

[Art11] M. Artin. *Algebra*. Pearson Prentice Hall, 2011.

[AV95] J. S. Aujla and H. Vasudeva. "Convex and monotone operator functions". In: *Annales Polonici Mathematici*. Volume 62. 1. 1995, pages 1–11.

[BCL94] K. Ball, E. A. Carlen, and E. H. Lieb. "Sharp Uniform Convexity and Smoothness Inequalities for Trace Norms". In: *Invent Math* 115.1 (Dec. 1994), pages 463–482. DOI: 10.1007/BF01231769.

[BBv21] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel. "Matrix Concentration Inequalities and Free Probability". In: *arXiv:2108.06312 [math]* (Aug. 2021). arXiv: 2108.06312 [math].

[BR97] R. B. Bapat and T. E. S. Raghavan. *Nonnegative matrices and applications*. Cambridge University Press, Cambridge, 1997. DOI: 10.1017/CBO9780511529979.

[Bar02] A. Barvinok. *A course in convexity*. American Mathematical Society, Providence, RI, 2002. DOI: 10.1090/gsm/054.

[Bar16] A. Barvinok. *Combinatorics and complexity of partition functions*. Springer, Cham, 2016. DOI: 10.1007/978-3-319-51829-9.

[Bau+01] H. H. Bauschke et al. "Hyperbolic polynomials and convex analysis". In: *Canad. J. Math.* 53.3 (2001), pages 470–488. DOI: 10.4153/CJM-2001-020-6.

[BT10] A. Beck and M. Teboulle. "On minimizing quadratically constrained ratio of two quadratic functions". In: *J. Convex Anal.* 17.3-4 (2010), pages 789–804.

[Bel+18] A. Belton et al. "A panorama of positivity". Available at https://arXiv.org/abs/1812.05482. 2018.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. Analysis, algorithms, and engineering applications. Society for Industrial and Applied Mathematics (SIAM), 2001. DOI: 10.1137/1.9780898718829.

[Ben+17] P. Benner et al. *Model reduction and approximation: theory and algorithms*. SIAM, 2017.

[Ber08] C. Berg. "Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity". In: *Positive Definite Functions: From Schoenberg to Space-Time Challenges*. Castellón de la Plana: University Jaume I, 2008, pages 15–45.

[BCR84] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups*. Theory of positive definite and related functions. Springer-Verlag, New York, 1984. DOI: 10.1007/978-1-4612-1128-0.

[BP94] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Revised reprint of the 1979 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. DOI: 10.1137/1.9781611971262.

[Ber12] D. Bertsekas. *Dynamic Programming and Optimal Control: Volume I*. v. 1. Athena Scientific, 2012. URL: https://books.google.com/books?id=qVBEEAAAQBAJ.

[Bha97] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997. DOI: 10.1007/978-1-4612-0653-8.

[Bha03] R. Bhatia. "On the exponential metric increasing property". In: *Linear Algebra and its Applications* 375 (2003), pages 211–220. DOI: https://doi.org/10.1016/S0024-3795(03)00647-5.

[Bha07a] R. Bhatia. *Perturbation bounds for matrix eigenvalues*. Reprint of the 1987 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. DOI: 10.1137/1.9780898719079.

[Bha07b] R. Bhatia. *Positive definite matrices*. Princeton University Press, Princeton, NJ, 2007.

[BH06] R. Bhatia and J. Holbrook. "Riemannian geometry and matrix geometric means". In: *Linear Algebra and its Applications* 413.2 (2006). Special Issue on the 11th Conference of the International Linear Algebra Society, Coimbra, 2004, pages 594–618. DOI: https://doi.org/10.1016/j.laa.2005.08.025.

[BJL19] R. Bhatia, T. Jain, and Y. Lim. "On the Bures-Wasserstein distance between positive definite matrices". In: *Expo. Math.* 37.2 (2019), pages 165–191. DOI: 10.1016/j.exmath.2018.01.002.

[BL06] Y. Bilu and N. Linial. "Lifts, Discrepancy and Nearly Optimal Spectral Gap". In: *Combinatorica* 26.5 (2006), pages 495–519. DOI: 10.1007/s00493-006-0029-7.

[Bir46] G. Birkhoff. "Three observations on linear algebra". In: *Univ. Nac. Tacuman, Rev. Ser. A* 5 (1946), pages 147–151.

[Boc33] S. Bochner. "Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse". In: *Math. Ann.* 108.1 (1933), pages 378–410. DOI: 10.1007/BF01452844.

[BB08] J. Borcea and P. Brändén. "Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized Fischer products". In: *Duke Mathematical Journal* 143.2 (2008), pages 205 –223. DOI: 10.1215/00127094-2008-018.

[Bou93] P. Bougerol. "Kalman Filtering with Random Coefficients and Contractions". In: *SIAM Journal on Control and Optimization* 31.4 (1993), pages 942–959. eprint: https://doi.org/10.1137/0331041. DOI: 10.1137/0331041.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. DOI: 10.1017/CBO9780511804441.

[BHB16] R. Brault, M. Heinonen, and F. Buc. "Random fourier features for operator-valued kernels". In: *Asian Conference on Machine Learning*. PMLR. 2016, pages 110–125.

[Bra+11] M. Braverman et al. "The Grothendieck Constant Is Strictly Smaller than Krivine's Bound". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. Oct. 2011, pages 453–462. DOI: 10.1109/FOCS.2011.77.

[BRS17] J. Briët, O. Regev, and R. Saket. "Tight Hardness of the Non-Commutative Grothendieck Problem". In: *Theory of Computing* 13.15 (Dec. 2017), pages 1–24. DOI: 10.4086/toc.2017.v013a015.

[CDV07] A. Caponnetto and E. De Vito. "Optimal rates for the regularized least-squares algorithm". In: *Foundations of Computational Mathematics* 7.3 (2007), pages 331–368.

[Cap+08] A. Caponnetto et al. "Universal multi-task kernels". In: *The Journal of Machine Learning Research* 9 (2008), pages 1615–1646.

[Car10] E. Carlen. "Trace inequalities and quantum entropy: an introductory course". In: *Entropy and the quantum*. Volume 529. Contemp. Math. Amer. Math. Soc., Providence, RI, 2010, pages 73–140. DOI: 10.1090/conm/529/10428.

[Car+10] C. Carmeli et al. "Vector valued reproducing kernel Hilbert spaces and universality". In: *Analysis and Applications* 8.01 (2010), pages 19–61.

[Car97] R. Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pages 41–75.

[Cha13] D. Chafaï. *A probabilistic proof of the Schoenberg theorem*. 2013. URL: https://djalil.chafai.net/blog/2013/02/09/a-probabilistic-proof-of-the-schoenberg-theorem/.

[CB21] C.-F. Chen and F. G. S. L. Brandão. "Concentration for Trotter Error". Available at https://arXiv.org/abs/2111.05324. Nov. 2021. arXiv: 2111.05324 [math-ph, physics:quant-ph].

[Cho74] M.-D. Choi. "A Schwarz inequality for positive linear maps on C^*-algebras". In: *Illinois Journal of Mathematics* 18.4 (1974), pages 565 –574. DOI: 10.1215/ijm/1256051007.

[CS07] M. Chudnovsky and P. Seymour. "The roots of the independence polynomial of a clawfree graph". In: *Journal of Combinatorial Theory, Series B* 97.3 (2007), pages 350–357. DOI: https://doi.org/10.1016/j.jctb.2006.06.001.

[Chu97] F. R. K. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[Col+21] S. Cole et al. "Quantum optimal transport". Available at https://arXiv.org/abs/2105.06922. 2021.

[CP77] "CHAPTER 6. Calculus in Banach Spaces". In: *Functional Analysis in Modern Applied Mathematics*. Volume 132. Mathematics in Science and Engineering. Elsevier, 1977, pages 87–105. DOI: https://doi.org/10.1016/S0076-5392(08)61248-5.

[Dav57] C. Davis. "A Schwarz inequality for convex operator functions". In: *Proc. Amer. Math. Soc.* 8 (1957), pages 42–44. DOI: 10.2307/2032808.

[DK70] C. Davis and W. M. Kahan. "The Rotation of Eigenvectors by a Perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pages 1–46. eprint: https://doi.org/10.1137/0707001. DOI: 10.1137/0707001.

[DKW82] C. Davis, W. M. Kahan, and H. F. Weinberger. "Norm-preserving dilations and their applications to optimal error bounds". In: *SIAM J. Numer. Anal.* 19.3 (1982), pages 445–469. DOI: 10.1137/0719029.

[Ded92] J. P. Dedieu. "Obreschkoff's theorem revisited: what convex sets are contained in the set of hyperbolic polynomials?" In: *Journal of Pure and Applied Algebra* 81.3 (1992), pages 269–278. DOI: https://doi.org/10.1016/0022-4049(92)90060-S.

[DPZ20] P. B. Denton, S. J. Parke, and X. Zhang. "Neutrino oscillations in matter via eigenvalues". In: *Phys. Rev. D* 101 (2020), page 093001. DOI: 10.1103/PhysRevD.101.093001.

[Den+22] P. B. Denton et al. "Eigenvectors from eigenvalues: a survey of a basic identity in linear algebra". In: *Bull. Amer. Math. Soc. (N.S.)* 59.1 (2022), pages 31–58. DOI: 10.1090/bull/1722.

[DT07] I. S. Dhillon and J. A. Tropp. "Matrix nearness problems with Bregman divergences". In: *SIAM J. Matrix Anal. Appl.* 29.4 (2007), pages 1120–1146. DOI: 10.1137/060649021.

[Din+06] C. Ding et al. "R_1-PCA: Rotational Invariant L_1-Norm Principal Component Analysis for Robust Subspace Factorization". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Association for Computing Machinery, June 2006, pages 281–288. DOI: 10.1145/1143844.1143880.

[ENG11] A. Ebadian, I. Nikoufar, and M. E. Gordji. "Perspectives of matrix convex functions". In: *Proceedings of the National Academy of Sciences* 108.18 (2011), pages 7313–7314. DOI: 10.1073/pnas.1102518108.

[EY39] C. Eckart and G. Young. "A principal axis transformation for non-hermitian matrices". In: *Bull. Amer. Math. Soc.* 45.2 (1939), pages 118–121. DOI: 10.1090/S0002-9904-1939-06910-3.

[Eff09] E. G. Effros. "A matrix convexity approach to some celebrated quantum inequalities". In: *Proceedings of the National Academy of Sciences* 106.4 (2009), pages 1006–1008. DOI: 10.1073/pnas.0807965106.

[Eva10] L. C. Evans. *Partial differential equations*. American Mathematical Soc., 2010.

[Evg+05] T. Evgeniou et al. “Learning multiple tasks with kernel methods.” In: *Journal of machine learning research* 6.4 (2005).

[Fel80] H. J. Fell. “On the zeros of convex combinations of polynomials.” In: *Pacific Journal of Mathematics* 89.1 (1980), pages 43 –50. DOI: pjm/1102779366.

[Går51] L. Gårding. “Linear hyperbolic partial differential equations with constant coefficients”. In: *Acta Mathematica* 85.none (1951), pages 1 –62. DOI: 10.1007/BF02395740.

[Gar+20] A. Garg et al. “Operator scaling: theory and applications”. In: *Found. Comput. Math.* 20.2 (2020), pages 223–290. DOI: 10.1007/s10208-019-09417-z.

[Gar07] D. J. H. Garling. *Inequalities: a journey into linear analysis*. Cambridge University Press, Cambridge, 2007. DOI: 10.1017/CBO9780511755217.

[GG78] C. Godsil and I. Gutman. “On the matching polynomial of a graph”. In: *Algebraic Methods in Graph Theory* 25 (Jan. 1978).

[GR01] C. Godsil and G. Royle. *Algebraic graph theory*. Springer-Verlag, New York, 2001. DOI: 10.1007/978-1-4613-0163-9.

[GW95] M. X. Goemans and D. P. Williamson. “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. In: *J. Assoc. Comput. Mach.* 42.6 (1995), pages 1115–1145. DOI: 10.1145/227683.227684.

[GM10] G. H. Golub and G. Meurant. *Matrices, moments and quadrature with applications*. Princeton University Press, Princeton, NJ, 2010.

[GVL13] G. H. Golub and C. F. Van Loan. *Matrix computations*. Fourth. Johns Hopkins University Press, Baltimore, MD, 2013.

[GR07] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Seventh. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX). Elsevier/Academic Press, Amsterdam, 2007.

[GLO20] A. Greenbaum, R.-C. Li, and M. L. Overton. “First-Order Perturbation Theory for Eigenvalues and Eigenvectors”. In: *SIAM Review* 62.2 (2020), pages 463–482. DOI: 10.1137/19M124784X.

[Gro53] A. Grothendieck. *Résumé de La Théorie Métrique Des Produits Tensoriels Topologiques*. Soc. de Matemática de São Paulo, 1953.

[Gu15] M. Gu. “Subspace iteration randomization and singular value problems”. In: *SIAM J. Sci. Comput.* 37.3 (2015), A1139–A1173. DOI: 10.1137/130938700.

[Gül97] O. Güler. “Hyperbolic Polynomials and Interior Point Methods for Convex Programming”. In: *Mathematics of Operations Research* 22.2 (1997), pages 350–377.

[Gur04] L. Gurvits. “Classical complexity and quantum entanglement”. In: *J. Comput. System Sci.* 69.3 (2004), pages 448–484. DOI: 10.1016/j.jcss.2004.06.003.

[Haa85] U. Haagerup. “The Grothendieck Inequality for Bilinear Forms on C*-Algebras”. In: *Advances in Mathematics* 56.2 (May 1985), pages 93–116. DOI: 10.1016/0001-8708(85)90026-X.

[HI95] U. Haagerup and T. Itoh. “Grothendieck Type Norms For Bilinear Forms On C*-Algebras”. In: *Journal of Operator Theory* 34.2 (1995), pages 263–283.

[Hal74] P. R. Halmos. *Finite-dimensional vector spaces*. 2nd ed. Springer-Verlag, New York-Heidelberg, 1974.

[HP82] F. Hansen and G. K. Pedersen. “Jensen’s Inequality for Operators and Löwner’s Theorem.” In: *Mathematische Annalen* 258 (1982), pages 229–241.

[HP03] F. Hansen and G. K. Pedersen. “Jensen’s Operator Inequality”. In: *Bulletin of the London Mathematical Society* 35.4 (2003), pages 553–564.

[HLP88] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Reprint of the 1952 edition. Cambridge University Press, Cambridge, 1988.

[HL72] O. J. Heilmann and E. H. Lieb. "Theory of monomer-dimer systems". In: *Communications in Mathematical Physics* 25.3 (1972), pages 190 –232. DOI: cmp/1103857921.

[HV07] J. W. Helton and V. Vinnikov. "Linear matrix inequality representation of sets". In: *Communications on Pure and Applied Mathematics* 60.5 (2007), pages 654–674. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20155. DOI: https://doi.org/10.1002/cpa.20155.

[Her63] C. S. Herz. "Fonctions opérant sur les fonctions définies-positives". In: *Ann. Inst. Fourier (Grenoble)* 13 (1963), pages 161–180. URL: http://aif.cedram.org/item?id=AIF_1963__13__161_0.

[Hia10] F. Hiai. "Matrix analysis: matrix monotone functions, matrix means, and majorization". In: *Interdiscip. Inform. Sci.* 16.2 (2010), pages 139–248. DOI: 10.4036/iis.2010.139.

[HK99] F. Hiai and H. Kosaki. "Means for matrices and comparison of their norms". In: *Indiana Univ. Math. J.* 48.3 (1999), pages 899–936. DOI: 10.1512/iumj.1999.48.1665.

[HP91] F. Hiai and D. Petz. "The proper formula for relative entropy and its asymptotics in quantum probability". In: *Communications in Mathematical Physics* 143 (1991), pages 99–114.

[HP14] F. Hiai and D. Petz. *Introduction to matrix analysis and applications*. Springer, Cham; Hindustan Book Agency, New Delhi, 2014. DOI: 10.1007/978-3-319-04150-6.

[Hig89] N. J. Higham. "Matrix nearness problems and applications". In: *Applications of matrix theory (Bradford, 1988)*. Volume 22. Inst. Math. Appl. Conf. Ser. New Ser. Oxford Univ. Press, New York, 1989, pages 1–27. DOI: 10.1093/imamat/22.1.1.

[Hig08] N. J. Higham. *Functions of matrices*. Theory and computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. DOI: 10.1137/1.9780898717778.

[HW53] A. J. Hoffman and H. W. Wielandt. "The variation of the spectrum of a normal matrix". In: *Duke J. Math.* 20 (1953), pages 37–39.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. "Expander graphs and their applications". In: *Bulletin of the American Mathematical Society* 43.4 (2006), pages 439–561.

[Hör94] L. Hörmander. *Notions of convexity*. Birkhäuser Boston, Inc., 1994.

[Hor69] R. A. Horn. "The theory of infinitely divisible matrices and kernels". In: *Trans. Amer. Math. Soc.* 136 (1969), pages 269–286. DOI: 10.2307/1994714.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Corrected reprint of the 1991 original. Cambridge University Press, Cambridge, 1994.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[Hua19] D. Huang. "Improvement on a Generalized Lieb's Concavity Theorem". Available at https://arXiv.org/abs/1902.02194. 2019.

[Hua+21] D. Huang et al. "Matrix Concentration for Products". In: *Foundations of Computational Mathematics* (2021), pages 1–33.

[Hur10] G. H. Hurlbert. *Linear optimization*. The simplex workbook. Springer, New York, 2010. DOI: 10.1007/978-0-387-79148-7.

[Jor75] C. Jordan. "Essai sur la géométrie à n dimensions". In: *Bulletin de la Société mathématique de France* 3 (1875), pages 103–174.

[Kad+16] H. Kadri et al. "Operator-valued kernels for learning from functional response data". In: *Journal of Machine Learning Research* 17.20 (2016), pages 1–54.

[Kai83] S. Kaijser. “A Simple-Minded Proof of the Pisier-Grothendieck Inequality”. In: *Banach Spaces, Harmonic Analysis, and Probability Theory*. Springer, 1983, pages 33–55.

[KSH00] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall, 2000.

[Kal60] R. E. Kalman. “A New Approach to Linear Filtering and Prediction Problems”. In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pages 35–45. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. DOI: 10.1115/1.3662552.

[Kat95] T. Kato. *Perturbation theory for linear operators*. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995.

[Kha96] H. Khalil. “Nonlinear Systems, Printice-Hall”. In: *Upper Saddle River, NJ* 3 (1996).

[Kos98] H. Kosaki. “Arithmetic–geometric mean and related inequalities for operators”. In: *Journal of Functional Analysis* 156.2 (1998), pages 429–451.

[Kow19] E. Kowalski. *An introduction to expander graphs*. Société mathématique de France, 2019.

[Kra36] F. Kraus. “Über konvexe Matrixfunktionen.” ger. In: *Mathematische Zeitschrift* 41 (1936), pages 18–42. URL: http://eudml.org/doc/168648.

[Kri78] J.-L. Krivine. “Constantes de Grothendieck et Fonctions de Type Positif Sur Les Spheres”. In: *Séminaire d’Analyse fonctionnelle (dit" Maurey-Schwartz")* (1978), pages 1–17.

[Kru37] R. Kruithof. “Telefoonverkeersrekening”. In: *De Ingenieur* 52 (1937), pp. E15–E25.

[KA79] F. Kubo and T. Ando. “Means of positive linear operators”. In: *Math. Ann.* 246.3 (1979/80), pages 205–224. DOI: 10.1007/BF01371042.

[Kwa08] N. Kwak. “Principal Component Analysis Based on L1-Norm Maximization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.9 (Sept. 2008), pages 1672–1680. DOI: 10.1109/TPAMI.2008.114.

[Lan70] P. Lancaster. “Explicit Solutions of Linear Matrix Equations”. In: *SIAM Review* 12.4 (1970), pages 544–566. URL: http://www.jstor.org/stable/2028490.

[LL07] J. Lawson and Y. Lim. “A Birkhoff Contraction Formula with Applications to Riccati Equations”. In: *SIAM Journal on Control and Optimization* 46.3 (2007), pages 930–951. eprint: https://doi.org/10.1137/050637637. DOI: 10.1137/050637637.

[Lax02] P. D. Lax. *Functional analysis*. Wiley-Interscience, 2002.

[Lax07] P. D. Lax. *Linear algebra and its applications*. second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007.

[LL08] H. Lee and Y. Lim. “Invariant metrics, contractions and nonlinear matrix equations”. In: *Nonlinearity* 21.4 (2008), pages 857–878. DOI: 10.1088/0951-7715/21/4/011.

[Lew96] A. S. Lewis. “Derivatives of spectral functions”. In: *Math. Oper. Res.* 21.3 (1996), pages 576–588. DOI: 10.1287/moor.21.3.576.

[LPR05] A. Lewis, P. Parrilo, and M. Ramana. “The Lax conjecture is true”. In: *Proceedings of the American Mathematical Society* 133.9 (2005), pages 2495–2499.

[LM99] C.-K. Li and R. Mathias. “The Lidskii-Mirsky-Wielandt theorem–additive and multiplicative versions”. In: *Numerische Mathematik* 81.3 (1999), pages 377–413.

[Lie73a] E. H. Lieb. “Convex trace functions and the Wigner-Yanase-Dyson conjecture”. In: *Advances in Math.* 11 (1973), pages 267–288. DOI: 10.1016/0001-8708(73)90011-X.

[Lie73b] E. H. Lieb. “Convex trace functions and the Wigner-Yanase-Dyson conjecture”. In: *Advances in Mathematics* 11.3 (1973), pages 267–288. DOI: https://doi.org/10.1016/0001-8708(73)90011-X.

[LR73] E. H. Lieb and M. B. Ruskai. “Proof of the strong subadditivity of quantum-mechanical entropy”. In: *J. Mathematical Phys.* 14 (1973). With an appendix by B. Simon, pages 1938–1941. DOI: 10.1063/1.1666274.

[LS13] J. Liesen and Z. Strakoš. *Krylov subspace methods*. Principles and analysis. Oxford University Press, Oxford, 2013.

[Lin73] G. Lindblad. “Entropy, information and quantum measurements”. In: *Communications in Mathematical Physics* 33 (1973), pages 305–322.

[Lin63] J. Lindenstrauss. “On the Modulus of Smoothness and Divergent Series in Banach Spaces.” In: *Michigan Mathematical Journal* 10.3 (1963), pages 241–252.

[LW94] C. Liverani and M. P. Wojtkowski. “Generalization of the Hilbert metric to the space of positive definite matrices.” In: *Pacific Journal of Mathematics* 166.2 (1994), pages 339 –355. DOI: pjm/1102621142.

[Lov79] L. Lovász. “On the Shannon capacity of a graph”. In: *IEEE Trans. Inform. Theory* 25.1 (1979), pages 1–7.

[Löw34] K. Löwner. “Über monotone Matrixfunktionen”. In: *Mathematische Zeitschrift* 38 (1934), pages 177–216. URL: http://eudml.org/doc/168495.

[Luo+10] Z. Q. Luo et al. “Semidefinite relaxation of quadratic optimization problems”. In: *IEEE Signal Process. Mag.* 27.3 (2010), pages 20–34.

[MSS14] A. W. Marcus, D. A. Spielman, and N. Srivastava. “Ramanujan graphs and the solution of the Kadison-Singer problem”. In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. III*. Kyung Moon Sa, Seoul, 2014, pages 363–386.

[MSS15a] A. W. Marcus, D. A. Spielman, and N. Srivastava. “Interlacing families I: Bipartite Ramanujan graphs of all degrees”. In: *Annals of Mathematics* 182.1 (2015), pages 307–325. URL: http://www.jstor.org/stable/24523004.

[MSS15b] A. W. Marcus, D. A. Spielman, and N. Srivastava. “Interlacing families II: Mixed characteristic polynomials and the Kadison—Singer problem”. In: *Annals of Mathematics* 182.1 (2015), pages 327–350. URL: http://www.jstor.org/stable/24523005.

[MSS21] A. W. Marcus, D. A. Spielman, and N. Srivastava. “Interlacing families III: Sharper restricted invertibility estimates”. In: *Israel Journal of Mathematics* (2021), pages 1–28.

[Mar+16] D. Marquis et al. *‘Gershgorin’s Circle Theorem for Estimating the Eigenvalues of a Matrix with Known Error Bounds”*. 2016.

[MOA11] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: theory of majorization and its applications*. Second. Springer, New York, 2011. DOI: 10.1007/978-0-387-68276-1.

[MT11] M. McCoy and J. A. Tropp. “Two Proposals for Robust PCA Using Semidefinite Programming”. In: *Electronic Journal of Statistics* 5.none (Jan. 2011), pages 1123–1160. DOI: 10.1214/11-EJS636.

[MP05] C. A. Micchelli and M. Pontil. “On learning vector-valued functions”. In: *Neural computation* 17.1 (2005), pages 177–204.

[MXZ06] C. A. Micchelli, Y. Xu, and H. Zhang. “Universal Kernels.” In: *Journal of Machine Learning Research* 7.12 (2006).

[Min88] H. Minc. *Nonnegative matrices*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1988.

[Min16] H. Q. Minh. "Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning". In: *arXiv preprint arXiv:1608.05639* (2016).

[Mir60] L. Mirsky. "Symmetric gauge functions and unitarily invariant norms". In: *Quart. J. Math. Oxford Ser. (2)* 11 (1960), pages 50–59. DOI: 10.1093/qmath/11.1.50.

[Mir59] L. Mirsky. "On the trace of matrix products". In: *Mathematische Nachrichten* 20.3-6 (1959), pages 171–174. DOI: 10.1002/mana.19590200306.

[MP93] B. Mohar and S. Poljak. "Eigenvalues in Combinatorial Optimization". In: *Combinatorial and Graph-Theoretical Problems in Linear Algebra*. Springer New York, 1993, pages 107–151.

[Nao12] A. Naor. "On the Banach-Space-Valued Azuma Inequality and Small-Set Isoperimetry of Alon–Roichman Graphs". In: *Combinatorics, Probability and Computing* 21.4 (July 2012), pages 623–634. DOI: 10.1017/S0963548311000757.

[NRV14] A. Naor, O. Regev, and T. Vidick. "Efficient rounding for the noncommutative Grothendieck inequality". In: *Theory Comput.* 10 (2014), pages 257–295. DOI: 10.4086/toc.2014.v010a011.

[NS21] N. H. Nelsen and A. M. Stuart. "The random feature model for input-output maps between banach spaces". In: *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243.

[NC00] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

[Nil91] A. Nilli. "On the Second Eigenvalue of a Graph". In: *Discrete Math.* 91.2 (1991), pages 207–210. DOI: 10.1016/0012-365X(91)90112-F.

[ON00] T. Ogawa and H. Nagaoka. "Strong converse and Stein's lemma in quantum hypothesis testing". In: *IEEE Transactions on Information Theory* 46.7 (2000), pages 2428–2433. DOI: 10.1109/18.887855.

[Oli10] R. I. Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *Electronic Communications in Probability* 15 (2010), pages 203–212.

[Olv+10] F. W. J. Olver et al., editors. *NIST handbook of mathematical functions*. With 1 CD-ROM (Windows, Macintosh and UNIX). National Institute of Standards and Technology, 2010.

[Ost59] A. M. Ostrowski. "A quantitative formulation of Sylvester's law of inertia". In: *Proceedings of the National Academy of Sciences of the United States of America* 45.5 (1959), page 740.

[Pak10] I. Pak. *Lectures on Discrete and Polyhedral Geometry*. 2010. URL: https://www.math.ucla.edu/~pak/book.htm.

[Par98] B. N. Parlett. *The symmetric eigenvalue problem*. Corrected reprint of the 1980 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. DOI: 10.1137/1.9781611971163.

[Par78] S. Parrott. "On a quotient norm and the Sz.-Nagy-Foiaş lifting theorem". In: *J. Functional Analysis* 30.3 (1978), pages 311–328. DOI: 10.1016/0022-1236(78)90060-5.

[Pin10] A. Pinkus. *Totally positive matrices*. Cambridge University Press, Cambridge, 2010.

[Pis78] G. Pisier. "Grothendieck's Theorem for Noncommutative C*-Algebras, with an Appendix on Grothendieck's Constants". In: *Journal of Functional Analysis* 29.3 (Sept. 1978), pages 397–415. DOI: 10.1016/0022-1236(78)90038-1.

[Pis12] G. Pisier. "Grothendieck's Theorem, Past and Present". In: *Bulletin of the American Mathematical Society* 49.2 (Apr. 2012), pages 237–323. DOI: 10.1090/S0273-0979-2011-01348-9.

[PX97] G. Pisier and Q. Xu. "Non-commutative martingale inequalities". In: *Comm. Math. Phys.* 189.3 (1997), pages 667–698. DOI: 10.1007/s002200050224.

[RS09] P. Raghavendra and D. Steurer. "Towards Computing the Grothendieck Constant". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2009, pages 525–534. DOI: 10.1137/1.9781611973068.58.

[RR08] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., 2008, pages 1177–1184. URL: http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.pdf.

[Ren06] J. Renegar. "Hyperbolic Programs, and Their Derivative Relaxations". In: *Found. Comput. Math.* 6.1 (2006), pages 59–79. DOI: 10.1007/s10208-004-0136-z.

[RX16] É. Ricard and Q. Xu. "A Noncommutative Martingale Convexity Inequality". In: *The Annals of Probability* 44.2 (Mar. 2016), pages 867–882. DOI: 10.1214/14-AOP990.

[Ric24] J. Riccati. "Animadversiones in aequationes differentiales secundi gradus". In: *Actorum Eruditorum quae Lipsiae publicantur Supplementa*. Actorum Eruditorum quae Lipsiae publicantur Supplementa v. 8. prostant apud Joh. Grossii haeredes & J.F. Gleditschium, 1724. URL: https://books.google.com/books?id=UjTw1w7tZsEC.

[Rud59] W. Rudin. "Positive definite sequences and absolutely monotonic functions". In: *Duke Math. J.* 26 (1959), pages 617–622. URL: http://projecteuclid.org/euclid.dmj/1077468771.

[Rud90] W. Rudin. *Fourier analysis on groups*. Reprint of the 1962 original, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1990. DOI: 10.1002/9781118165621.

[Rud91] W. Rudin. *Functional analysis*. Second. McGraw-Hill, Inc., New York, 1991.

[Saa11a] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. 2nd edition. Society for Industrial and Applied Mathematics, 2011. DOI: 10.1137/1.9781611970739.

[Saa11b] Y. Saad. *Numerical methods for large eigenvalue problems*. Revised edition of the 1992 original [1177405]. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. DOI: 10.1137/1.9781611970739.ch1.

[SSB18] B. Schlkopf, A. J. Smola, and F. Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

[Sch14] R. Schneider. *Convex bodies: the Brunn–Minkowski theory*. 151. Cambridge university press, 2014. DOI: 10.1017/CBO9781139003858.

[Sch38] I. J. Schoenberg. "Metric spaces and positive definite functions". In: *Trans. Amer. Math. Soc.* 44.3 (1938), pages 522–536. DOI: 10.2307/1989894.

[SS17] L. J. Schulman and A. Sinclair. "Analysis of a Classical Matrix Preconditioning Algorithm". In: *J. Assoc. Comput. Mach. (JACM)* 64.2 (2017), 9:1–9:23.

[Sch64] L. Schwartz. "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants)". In: *Journal d'analyse mathématique* 13.1 (1964), pages 115–256.

[Sen81] E. Seneta. *Nonnegative matrices and Markov chains*. Second. Springer-Verlag, New York, 1981. DOI: 10.1007/0-387-32792-4.

[Ser09] D. Serre. "Weyl and Lidskiĭ inequalities for general hyperbolic polynomials". In: *Chinese Annals of Mathematics, Series B* 30.6 (2009), pages 785–802.

[Sha48] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pages 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[SH19] Y. Shenfeld and R. van Handel. "Mixed volumes and the Bochner method". In: *Proc. Amer. Math. Soc.* 147.12 (2019), pages 5385–5402. DOI: 10.1090/proc/14651.

[Sim11] B. Simon. *Convexity*. An analytic viewpoint. Cambridge University Press, Cambridge, 2011. DOI: 10.1017/CBO9780511910135.

[Sim15] B. Simon. *Real analysis*. With a 68 page companion booklet. American Mathematical Society, Providence, RI, 2015. DOI: 10.1090/simon/001.

[Sim19] B. Simon. *Loewner's theorem on monotone matrix functions*. Springer, Cham, 2019. DOI: 10.1007/978-3-030-22422-6.

[Sin64] R. Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *Ann. Math. Statist.* 35 (1964), pages 876–879. DOI: 10.1214/aoms/1177703591.

[SMI92] V. I. SMIRNOV. "Biography of A. M. Lyapunov". In: *International Journal of Control* 55.3 (1992), pages 775–784. eprint: https://doi.org/10.1080/00207179208934258. DOI: 10.1080/00207179208934258.

[So09] A. M. So. "Improved Approximation Bound for Quadratic Optimization Problems with Orthogonality Constraints". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2009, pages 1201–1209. DOI: 10.1137/1.9781611973068.130.

[Spi18a] D. Spielman. "Yale CPSC 662/AMTH 561 : Spectral Graph Theory". Available at http://www.cs.yale.edu/homes/spielman/561/561schedule.html. 2018.

[Spi18b] D. Spielman. *Bipartite Ramanujan Graphs*. 2018. URL: http://www.cs.yale.edu/homes/spielman/561/lect25-18.pdf.

[Spi19] D. Spielman. *Spectral and Algebraic Graph Theory*. 2019. URL: http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf.

[SS11] D. A. Spielman and N. Srivastava. "Graph Sparsification by Effective Resistances". In: *SIAM Journal on Computing* 40.6 (2011), pages 1913–1926. eprint: https://doi.org/10.1137/080734029. DOI: 10.1137/080734029.

[SH15] S. Sra and R. Hosseini. "Conic geometric optimization on the manifold of positive definite matrices". In: *SIAM J. Optim.* 25.1 (2015), pages 713–739. DOI: 10.1137/140978168.

[Ste56] E. M. Stein. "Interpolation of linear operators". In: *Transactions of the American Mathematical Society* 83.2 (1956), pages 482–492.

[SS90] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. 1st Edition. Academic Press, 1990.

[SBT17] D. Sutter, M. Berta, and M. Tomamichel. "Multivariate trace inequalities". In: *Comm. Math. Phys.* 352.1 (2017), pages 37–58. DOI: 10.1007/s00220-016-2778-5.

[Syl84] J. J. Sylvester. "Sur l'équation en matrices px= xq". In: *CR Acad. Sci. Paris* 99.2 (1884), pages 67–71.

[Syl52] J. Sylvester. "A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.23 (1852), pages 138–142.

[Tom74] N. Tomczak-Jaegermann. "The Moduli of Smoothness and Convexity and the Rademacher Averages of the Trace Classes S_p $(1 \leq p < \infty)$". In: *Studia Mathematica* 50.2 (1974), pages 163–182.

[Tro09] J. A. Tropp. "Column subset selection, matrix factorization, and eigenvalue optimization". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Philadelphia, PA, 2009, pages 978–986.

[Tro11] J. A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4 (2011), pages 389–434. DOI: 10.1007/s10208-011-9099-z.

[Tro12] J. A. Tropp. "From joint convexity of quantum relative entropy to a concavity theorem of Lieb". In: *Proc. Amer. Math. Soc.* 140.5 (2012), pages 1757–1760. DOI: 10.1090/S0002-9939-2011-11141-9.

[Tro15] J. A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pages 1–230.

[Tro18] J. A. Tropp. "Second-order matrix concentration inequalities". In: *Appl. Comput. Harmon. Anal.* 44.3 (2018), pages 700–736. DOI: 10.1016/j.acha.2016.07.005.

[van14] R. van Handel. *Probability in High Dimension*. Technical report. Princeton University, June 2014. DOI: 10.21236/ADA623999.

[VB96] L. Vandenberghe and S. Boyd. "Semidefinite programming". In: *SIAM Rev.* 38.1 (1996), pages 49–95. DOI: 10.1137/1038003.

[Vas79] H. Vasudeva. "Positive definite matrices and absolutely monotonic functions". In: *Indian J. Pure Appl. Math.* 10.7 (1979), pages 854–858.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

[Wat18] J. Watrous. *The theory of quantum information*. Cambridge University Press, 2018.

[Wid41] D. V. Widder. *The Laplace Transform*. Princeton University Press, Princeton, N. J., 1941.

[Wol19] N. Wolchover. "Neutrinos lead to unexpected discovery in basic math". In: *Quanta Magazine* (2019). URL: https://www.quantamagazine.org/neutrinos-lead-to-unexpected-discovery-in-basic-math-20191113.

[Zha05] F. Zhang, editor. *The Schur complement and its applications*. Springer-Verlag, New York, 2005. DOI: 10.1007/b105056.