Matrix monotone and convex functions

Otte Heinävaara

January 13, 2017

Contents

Chapter 1

Introduction

1.1 Foreword

This master's thesis is about matrix monotone and convex functions. Matrix monotonicity and convexity are generalizations of standard monotonicity and convexity of real functions: now we are just having functions mapping matrices to matrices. Formally, f is *matrix monotone* if for any two matrices A and B such that

$$A \leq B \tag{1.1}$$

we should also have

$$f(A) \leq f(B). \tag{1.2}$$

This kind of function might be more properly called *matrix increasing* but we will mostly stick to the monotonicity for couple of reasons:

- For some reason, that is what people have been doing in the field.
- It doesn't make much difference whether we talk about increasing or decreasing functions, so we might just ignore the latter but try to symmetrize our thinking by choice of words.
- Somehow I can't satisfactorily fill the following table:

monotonic	monotonicity
increasing	?

  How very inconvenient.

Matrix convexity, as you might have guessed by now, is defined as follows. A function f is *matrix convex* if for any two matrices A and B and $0 \leq t \leq 1$ we have

$$f(tA + (1-t)B) \leq tf(A) + (1-t)f(B). \tag{1.3}$$

Of course, it's not really obvious how one should make any sense of these "definitions". One quickly realizes that there two things to understand.

- How should matrices be ordered?
- How should functions act on matrices?

Both of these questions can be (of course) answered in many ways, but for both of them, there's in a way very natural answer. In both cases we can get something more general: instead of comparing matrices we can compare linear maps, and we can apply function to linear mapping.

Just to give a short glimpse of how these things might be defined, we should first fix our ground field (for matrices): let's say it's $\mathbb{R}$, at least for now.

For matrix ordering we should first understand which matrices are *positive*, which here, a bit confusingly maybe, means "at least zero". We say that matrix is positive if all it's eigenvalues are non-negative. Having done this, we immediately restrict ourselves to (symmetric) diagonalizable matrices with real eigenvalues, but we will later see that we can't do much "better". Also, since sum of positive matrices should be positive, we should further restrict ourselves to even stricter class of matrices, called Hermitian matrices, which correspond self-adjoint linear maps. Now everything works nicely but we still preserve non-trivial non-commutative structure.

Matrix functions, i.e. "how to apply function to matrix" is bit simpler to explain. Instead of doing something arbitrary the idea is to take real function (a function $f : \mathbb{R} \to \mathbb{R}$, say) and intepret it as function $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, *matrix function*. Polynomials extend rather naturally, and similarly analytic functions, or at least entire. Now, a perverse definition for matrix function for continuous functions would be some kind of a limit when function is uniformly approximated by polynomials (using Weierstrass approximation theorem). This works for Hermitian matrices, but one can do better: apply the function to the eigenvalues of the mapping to get another linear map.

As it turns out, much of the study of matrix monotone and convex functions is all about understanding these definitions of positive maps and matrix functions.

Lastly, one might wonder why should one be interested in the whole business of monotone and convex functions? It's all about point of view. Let's consider a very simple inequality:

For any real numbers $0 < x \leq y$ we have

$$y^{-1} \leq x^{-1}.$$

Of course, this is quite close to the axioms of the real numbers, but there's a rather fruitful interpretation. The function $(x \mapsto \frac{1}{x})$ is decreasing.

Now there's this matrix version of the previous inequality:

For any two matrices $0 < A \leq B$ we have

$$B^{-1} \leq A^{-1}.$$

This is already not trivial, and with previous interpretation in mind, could this be interpreted as the functions $(x \mapsto \frac{1}{x})$ could be *matrix decreasing*? And is this just a special case of something bigger? Yes, and that's exactly what this thesis is about.

1.2 Plan of attack

This master's thesis is a comprehensive review of the rich theory of matrix monotone and convex functions.

Master's thesis is to be structured roughly as follows.

1. Introduction
    - Introduction to the problem, motivation
    - Brief definition of the matrix monotonicity and convexity
    - Past and present
        - Loewner's original work, Loewner-Heinz -inequality
        - Students: Dobsch' and Krauss'
        - Subsequent simplifications and further results: Bendat-Sherman, Wigner-Neumann, Koranyi, etc.
        - Donoghue's work
        - Later proofs: Krein-Milman, general spectral theorem, interpolation spaces, short proofs etc.
        - Development of the convex case
        - Recent simplifications, integral representations
        - Operator inequalities
        - Multivariate case, other variants
        - Further open problems?
    - Scope of the thesis
2. Preliminaries (partially to be dumped to appendix?)

- Divided differences: basic definition, properties, representations and smoothness
- Positive matrices
  - Setup: finite (vs infinite) dimensional inner product spaces over $\mathbb{C}$ (vs $\mathbb{R}$), basic facts
  - Linear maps, adjoint, congruence, self-adjoint maps, spectral theorem: finite and infinite dimensional
  - Good properties of spectrum
  - Positive maps: basic properties (cone structure, Sylvester's criterion etc.)
- Matrix functions
  - Several definitions
  - Derivatives of matrix functions
- Divided differences: basic definition, properties, representations and smoothness
- Pick functions
  - Basic definitions and properties
  - Pick-Nevanlinna representations theorem
  - Pick matrices/ determinants
  - Compactness
  - Pick-Nevanlinna interpolation theorem
- Regularizations
  - Basic properties
  - Lemmas needed for some of the proofs

3. Monotonic and convex matrix functions

- Basics
  - Basic definitions and properties (cone structure, pointwise limits, compositions etc.)
  - Classes P_n, K_n and their properties
  - $-1/x$
  - One directions of Loewner's theorem
  - Examples and non-examples
- Pick matrices/determinants vs matrix monotone and convex functions

  - Proofs for (sufficiently) smooth functions
- Smoothness properties
  - Ideas, simple cases
  - General case by induction and regularizations
- Global characterizations
  - Putting everything together: we get original characterization of Loewner and determinant characterization

4. Local characterizations

   - Dobsch (Hankel) matrix: basic properties, easy direction (original and new proof)
   - Integral representations
     - Introducing the general weight functions for monotonicity and convexity (and beyond?)
     - Non-negativity of the weights
     - Proof of integral representations
   - Proof of local characterizations

5. Structure of the classes P_n and K_n, interpolating properties (?)

   - Strict inclusions, strict smoothness conditions
   - Strictly increasing functions
   - Extreme values
   - Interpolating properties

6. Loewner's theorem

   - Preliminary discussion, relation to operator monotone functions
   - Loewner's original proof
   - Pick-Nevanlinna proof
   - Bendat-Sherman proof
   - Krein-Milman proof
   - Koranyi proof
   - Discussion of the proofs

- Convex case

7. Alternative characterizations (?)
    - Some discussion, maybe proofs
8. Bounded variations (?)
    - Dobsch' definition, basic properties
    - Decomposition, Dobsch' theorems

1.3 Some random ideas

1. It's easy to see that [Something]. Actually, it's so so easy that we have no excuse for not doing it.
2. When is matrix of the form $f(a_i + a_j)$ positive: f is completely monotone (?).
3. Polynomial regression...
4. TODO: Maximum of two matrices (at least as big), (a + b)/2 + abs(a - b)/2
5. If $\langle Ax, y\rangle = 0$ implies $\langle x, Ay\rangle = 0$, then A is constant times hermitian.
6. Angularity preserving functions
7. If subspace of linear maps are diagonalizable with real eigenvalues, is there a inner product such that subspace consists of only Hermitian maps

Chapter 2

Positive matrices

This chapter is titled "positive matrices", although "positive maps" might be more appropriate title. We are mostly going to deal with finite-dimensional objects, but many of the ideas could be generalized infinite-dimensional settings, where matrices lose their edge. Also, one should always ask whether it really clarifies the situation to introduce concrete matrices: matrices are good at hiding the truly important properties of linear mappings. The words "matrix" and "linear map" are used somewhat synonymously, although one should always remember that the former are just special representations for the latter.

2.1 Motivation

How should one order matrices? What should we require from ordering anyway?

I could just give you the answer, but instead I try to explain why it is standard in the first place.

We would definitely like to have natural total order on the space of matrices, but it turns out that are no natural choices for that. Partial order is next best thing. Recall that a partial order on a set X is a binary relation $\leq$ on such that

1. $x \leq x$ for any $x \in X$.
2. For any $x, y \in X$ for which $x \leq y$ and $y \leq x$, necessarily $x = y$.
3. If for some $x, y, z \in X$ we have both $x \leq y$ and $y \leq z$, also $x \leq z$.

The third point is the main point, the first two are just there preventing us from doing something crazy. But we can do better: this partial order on matrices should also respect addition.

4. For any $x, y, z \in X$ such that $x \leq y$, we should also have $x + z \leq y + z$.

There's another way to think about this last point. Instead of specifying order among all the pairs, we just say which matrices are positive: matrix is positive if and only it's at least 0.

If we know all the positive matrices, we know all the "orderings". To figure out whether $x \leq y$, we just check whether $0 = x - x \leq y - x$, i.e. whether $y - x$ is positive. Also, positive matrices are just differences of the form $y - x$ where $x \leq y$. Now, conditions on the partial order are reflected to the set of positive matrices.

1'. 0 (zero matrix) is positive.

2'. If both x and $-x$ are positive, then $x = 0$.

3'. If both x and y are positive, so is their sum $x + y$.

Here $3'$ is kind of combination of 3 and 4.

The terminology here is rather unfortunate. Natural ordering of the reals satisfies all of the above with obvious interpretation of positive numbers, which however differs from the standard definition: 0 is itself positive in our above definition. This is undoubtedly confusing, but what can you do? For real numbers we have total order, so every number is either zero, strictly positive or strictly negative, so when we say non-negative, it literally means "not negative": we get all the positive numbers and zero. But with partial orders we might get more. So the main reasons why we are using this terminology are

1. It's short.

Also, now that we have decided to preserve the word "positive" for "at least zero" one might be tempted to preserve "strictly positive" for "at least zero, but not zero". We won't do that, we save that phrase for something more important.

Back to business. When we try to define ordering of the matrices, everything of course depends on the ground field. It hardly makes any sense to order matrices over $\mathbb{F}_p$: even 1×1 matrices, namely (canonically) the elements of $\mathbb{F}_p$ defy reasonable ordering. But real numbers, for instance, have ordering, so there's a serious change that all real matrices could be ordered.

We will first try to order all real square matrices. (Actually, we won't even try to order non-square matrices.) 1×1 matrices are easy to order, but as soon one moves to larger matrices, one faces difficult decisions:

$$\text{Is } \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \leq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \geq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} ?$$

That's okay, we don't necessarily have to order all pairs of matrices. But there are other problems. We would like the ordering of the matrices to be independent of the

choice matrix representation. If we flip basis vectors we change the rows and colums of the matrix, so change it's sign! So if we want

- Positivity is readable from the matrix with respect to an arbitrary basis, alone.
- Positivity doesn't depend on the chosen basis.

there is no way to assign sign to the previous matrix.

Let's consider diagonalizable matrices: by change of basis these can be written as diagonal matrices, but these diagonal values, eigenvalues, need not be real anymore. One could try to ignore the complex bases, but how often really is ignoring the complex structure satisfactory. Another approach would be only to consider matrices with real eigenvalues. There's problem here though: sum of two matrices with real eigenvalues need not have real eigenvalues, even if the former are diagonalizable! Sad as it may sound, it shouldn't be too surprising since it's not very clear that eigenvalues should have any reasonable behaviour with respect to addition. Of course, one shouldn't just take my word for it: here are congrete examples:

$$\begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & -1 \\ -4 & 0 \end{bmatrix}$$

The two matrices have both distinct eigenvalues -2 and 2 and are hence diagonlizable, but their sum has characteristic polynomial x^2+9, which most definitely has no real zeros.

There is however very special class of matrices, correponding to special class of linear mappings, which satisfy our requirements.

- They have real eigenvalues.
- They are all diagonalizable.
- They are closed under sum, and (real) scalar multiplication.

2.2 Hermitian maps

Hermitian mappings will be our non-commutative playing ground in which we can define rather natural partial order. Fix any finite-dimensional inner-product space $(V, \langle \cdot, \cdot \rangle)$ over $\mathbb{R}$ or $\mathbb{C}$. For any $v \in V \setminus \{0\}$ we may define the corresponding projection, denoted by P_v by setting $P_v(x) = \langle x, v \rangle / \langle v, v \rangle v$.

Definition 2.1. Let V be an finite-dimensional innerproduct space over $\mathbb{R}$ or $\mathbb{C}$. Now set of *Hermitian maps* of V, denoted by $\mathcal{H}(V)$, is defined as

$$\operatorname{span}\{P_v \mid v \in V \setminus \{0\}\},$$

where the span is $\mathbb{R}$-linear, i.e. Hermitian maps are the maps of the form

$$\sum_{i=1}^{m} \lambda_i P_{v_i},$$

for some positive integer m, $v_i \in V \setminus \{0\}$ and $\lambda_i \in \mathbb{R}$.

Also note that since for any non-zero α and $v \in V \setminus \{0\}$ we have $P_v = P_{\alpha v}$, we could just as well only allow vectors of norm one in our definition of Hermitian maps. This is of course just to simplify notation. This is not the standard definition, and in a way it's horrible, but it very directly answers our needs. Remember that we want to have diagonalizable maps with real eigenvalues. Projections are very prototypical examples of such maps and since we want real eigenvalues that is simply what we require in the definition (span over $\mathbb{R}$). Lastly, one of course wants to take the span to have any linear structure in the first place.

It is however very surprising that such construction works: basis vectors (projections) satisfy our requirements, but there are no reasons to expect that these would be preserved in addition. This is guaranteed by the following theorem.

Theorem 2.2 (Spectral theorem)**.** *Let V be an n-dimensional inner-product space over $\mathbb{R}$ or $\mathbb{C}$, and $A \in \mathcal{H}(V)$. Then there exists real numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ and orthonormal vectors $v_1, v_2, \ldots, v_n$ such that*

$$A = \sum_{i=1}^{n} \lambda_i P_{v_i}. \tag{2.3}$$

It follows that A is diagonalizable, vectors $v_1, v_2, \ldots, v_n$ are eigenvectors of A with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Indeed, for any v_j we have

$$Av_j = \sum_{i=1}^{n} \lambda_i P_{v_i} v_j = \sum_{i=1}^{n} \lambda_i \langle v_j, v_i \rangle v_i = \sum_{i=1}^{n} \lambda_i \delta_{ij} v_i = \lambda_j v_j. \tag{2.4}$$

Moreover, the rank of A is simply the number of non-zero eigenvalues of A, and the all the eigenvalues appear in the representation exactly as many times the characteristic polynomial allows.

Proof (of the Spectral theorem). We prove the statement by induction on n the dimension of the space V. The case $n = 0$ is trivial: the span itself is trivial, and it's very easy to express 0-mapping as a empty sum.

Now fix a positive integer n.

Step 1. First note A has at least one eigenvector over $\mathbb{C}$, as every other linear mapping. Let v_1 be that eigenvector (of lenght one) with corresponding eigenvalue λ_1.

Step 2. Write $A = \sum_{i=1}^{m} c_i P_{u_i}$ for $c_1, c_2, \ldots, c_m \in \mathbb{R}$ and $u_1, u_2, \ldots, u_m \in V \setminus \{0\}$ are of norm 1. Now the definition of the eigenvector and eigenvalue rewrites to

$$Av_1 = \sum_{i=1}^{m} c_i \langle v_1, u_i \rangle u_i = \lambda_i v_1.$$

Take inner product with v_1 from both sides to get

$$\sum_{i=1}^{m} c_i \langle v_1, u_i \rangle \langle u_i, v_1 \rangle = \lambda_1 \langle v_1, v_1 \rangle.$$

Since $\langle v_1, u_i \rangle \langle u_i, v_1 \rangle = |\langle v_1, u_i \rangle|^2$, the left-hand side is real, so is right-hand side, and finally so is λ_1.

Step 3. Now we have found λ_1 and v_1 as in the theorem statement. The idea is then to factorize A to two parts: projection corresponding to v_1 and an orthogonal part living in $v_1^\perp$. If our choices for λ_1 and v_1 work, anything orthogonal to v_1 maps to something orthogonal to v_1. We verify this.

Take any $v \perp v_1$. Now,

$$\begin{aligned}
\langle Av, v_1 \rangle &= \langle \sum_{i=1}^{m} c_i \langle v, u_i \rangle u_i, v_1 \rangle \\
= \overline{\sum_{i=1}^{m} c_i \langle v, u_i \rangle \langle u_i, v_1 \rangle} &= \overline{\sum_{i=1}^{m} c_i \langle v_1, u_i \rangle \langle u_i, v \rangle} \\
= \overline{\langle \sum_{i=1}^{m} c_i \langle v_1, u_i \rangle u_i, v \rangle} &= \overline{\langle Av_1, v \rangle} \\
= \overline{\langle \lambda_1 v_1, v \rangle} &= 0,
\end{aligned}$$

so also $Av \perp v_1$. It follows that we get a map $A' : v_1^\perp \to v_1^\perp$ which extends to A.

Step 4. We would naturally like to use our induction hypothesis for this map, but for that we need it to be of special form: although the map is well defined, there is no reason to expect that we could just pick some of the u_i:s for our representation. But we can do something else.

Write $u_i = u_i' + \alpha_i v_1 = P_{v_1^\perp} + P_{v_1}$. Now if $v \perp v_1$, we have

$$\begin{aligned} Av &= \sum_{i=1}^{m} c_i \langle v, u_i \rangle u_i \\ &= \sum_{i=1}^{m} c_i \langle v, u_i' + \alpha_i v_1 \rangle (u_i' + \alpha_i v_1) \\ &= \sum_{i=1}^{m} c_i \langle v, u_i' \rangle u_i' + v_1 \sum_{i=1}^{m} c_i \langle v, u_i' \rangle \alpha_i. \end{aligned}$$

Now since $u_i' \perp v_1$, the latter sum vanishes, and

$$A' = \sum_{i=1}^{m} c_i P_{u_i'}.$$

This representation meets our requirements, so we can write $A' = \sum_{i=2}^{n} \lambda_i P_{v_i}$, for some $\lambda_2, \ldots, \lambda_n \in \mathbb{R}$ and $v_2, \ldots, v_n \in v_1^\perp$, and adjoining v_1 and λ_1, we get the required representation. □

Representation 2.3, which we will call *spectral representation* is by no means unique. If $A = I$ for instance, we could choose the vectors v_i rather arbitrarily. It also should be made clear why such representation is useful. First of all, calculating with spectral representation is easy. If $A = \sum_{i=1} \lambda_i P_{v_i}$,

$$A^2 = \left(\sum_{i=1} \lambda_i P_{v_i} \right) \left(\sum_{j=1} \lambda_j P_{v_j} \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j P_{v_i} P_{v_j} = \sum_{i=1}^{n} \lambda_i^2 P_{v_i},$$

since $P_v P_w$ for any orthogonal $v, w \in V$. There is an obvious generalization for higher powers, and polynomials too. Similar identities work for more general diagonalizable maps, but now we have additionally projection representation. Spectral representation has however many other nice properties not shared with the more general diagonalizable maps: operator norm of map is determined by its eigenvalues only. More precisely $\|A\| = \max_{i=1}^{n} |\lambda_i|$. We will come back to this in a minute.

It should be noted that we only used one special property of ($\mathbb{R}$ linear combination of) projections A in the proof:

$$\langle Av, w \rangle = \overline{\langle Aw, v \rangle} = \langle v, Aw \rangle,$$

for any $v, w \in V$. In the first step we didn't use any properties of A. In the step 3 this is exactly what we do in the manipulation. Also in the step 2 the idea is to show that

$$\lambda_1 \langle v_1, v_1 \rangle = \langle Av_1, v_1 \rangle \in \mathbb{R},$$

but $\langle Av_1, v_1\rangle = \overline{\langle Av_1, v_1\rangle}$ so $\langle Av_1, v_1\rangle \in \mathbb{R}$. Of course here we didn't need the fact that v_1 is an eigenvector in any way. Step 4 is just simpler: once we have constructed A', it will obviously have the same property as A.

In all of the above it doesn't make much difference whether we have $\mathbb{R}$ or $\mathbb{C}$ as the ground field, since the eigenvalues will be anyway real. Especially when one is working over $\mathbb{R}$, one might whether there is a more straightforward way to get through step 1: get rid of the complex numbers. There is. Another way to find the eigenvector and eigenvalue is to look at the so called *Rayleigh quotient*

$$R(A, v) = \frac{\langle Av, v\rangle}{\langle v, v\rangle},$$

when $v \in V \setminus \{0\}$. This is scale-invariant bounded real quantity so it attains maximum somewhere outside zero, say at v. We claim that v is an eigenvector of A. Looking at $v_t = v + tw$ and differentiating at zero yields

$$\begin{aligned} 0 = \frac{d}{dt} R(A, v_t)\Big|_{t=0} &= \frac{(\langle Av, w\rangle + \langle Aw, v\rangle)\langle v, v\rangle - (\langle v, w\rangle + \langle w, v\rangle)\langle Av, v\rangle}{\langle v, v\rangle^2} \\ &= \frac{\Re\left(\langle Av, w\rangle\langle v, v\rangle - \langle v, w\rangle\langle Av, v\rangle\right)}{\langle v, v\rangle^2}. \end{aligned}$$

We write $Av = \lambda v + v'$ where $v \perp v'$ and set $w = v'$. Now

$$\begin{aligned} 0 &= \frac{\Re\left(\langle \lambda v + v', v'\rangle\langle v, v\rangle - \langle v, v'\rangle\langle \lambda v + v', v\rangle\right)}{\langle v, v\rangle^2} \\ &= \frac{\Re\langle v', v'\rangle\langle v, v\rangle}{\langle v, v\rangle^2}, \end{aligned}$$

so $v' = 0$ and v is an eigenvector.

2.3 Self-adjoint maps

Definition 2.5. Let $(V, \langle\cdot,\cdot\rangle)$ be an finite-dimensional inner-product space. *Adjoint* of a linear map $A : V \to V$ is a linear map $A^* : V \to V$ such that

$$\langle Av, w\rangle = \langle v, A^* w\rangle$$

for any $v, w \in V$.

Definition 2.6. Let $(V, \langle\cdot,\cdot\rangle)$ be an finite-dimensional inner-product space. A linear map $A : V \to V$ is *self-adjoint* if it is its own adjoint, i.e.

$$\langle Av, w\rangle = \langle v, Aw\rangle$$

for any $v, w \in V$.

The spectral theorem shows that Hermitian maps correspond exactly to the self-adjoint maps. It is the good property of projections that made the Spectral theorem work: we got to move mappings from one side of inner-product to another. Generally the mapping changes, and adjoint is what comes out. We gather here many of the important properties of adjoint.

Theorem 2.7. *Let $(V, \langle\cdot,\cdot\rangle)$ be an n-dimensional inner-product space, $A, B \in \mathcal{L}(V)$ linear, and $\lambda \in \mathbb{C}$. Then*

i) A has an unique adjoint

ii) Matrix of A^* with respect to any orthonormal basis is conjugate transpose of matrix of A, i.e. $A^*_{i,j} = \overline{A_{j,i}}$.

iii) $(A^*)^* = A$

iv) $(A+B)^* = A^* + B^*$

v) $(\lambda I)^* = \overline{\lambda} I$

vi) $(AB)^* = B^*A^*$.

vii) $\ker(A) = \mathrm{im}(A)^\perp$

Proof. i) Fix any orthonormal basis of V, $(e_i)_{i=1}^n$. Now for any $v, w \in V$

$$\begin{aligned}\langle Av, w\rangle &= \langle A\sum_{i=1}^n \langle v, e_i\rangle e_i, w\rangle \\ &= \sum_{i=1}^n \langle Ae_i, w\rangle\langle v, e_i\rangle \\ &= \langle v, \sum_{i=1}^n \langle w, Ae_i\rangle e_i\rangle,\end{aligned}$$

so we should set $A^*w = \sum_{i=1}^n \langle w, Ae_i\rangle e_i$. This clearly defines an linear mapping A^*. Also, if A has two adjoints, their difference C is linear map such that $\langle v, Cw\rangle = 0$ for any $v, w \in V$. Setting $v = Cw$ we get that $Cw = 0$ for any $w \in V$ so the adjoints are equal.

ii) Observe that $A^*_{i,j} = \langle A^*e_j, e_i\rangle = \overline{\langle e_i, A^*e_j\rangle} = \overline{\langle Ae_i, e_j\rangle} = \overline{A_{j,i}}$.

iii) We have for any $v, w \in V$ that $\langle A^*v, w\rangle = \overline{\langle w, A^*v\rangle} = \overline{\langle Aw, v\rangle} = \langle v, Aw\rangle$, so A is adjoint of A^*.

iv) Since $\langle (A+B)v, w\rangle = \langle Av, w\rangle + \langle Bv, w\rangle = \langle v, A^*w\rangle + \langle v, B^*w\rangle = \langle v, (A^*+B^*)w\rangle$, A^*+B^* is adjoint of $A+B$.

v) We simply calculate that $\langle \lambda v, w\rangle = \lambda\langle v, w\rangle = \langle v, \overline{\lambda}v\rangle$.

vi) Note that $\langle ABv, w\rangle = \langle Bv, A^*w\rangle = \langle v, B^*A^*w\rangle$.

vii) We have $v \in \ker(A)$ i.e. $Av = 0$ if and only $\langle Av, w\rangle = \langle v, Aw\rangle$ for any $w \in V$, which is to say that $v \perp \mathrm{im}(A)$.

□

2.4 Operator norm

As we noticed, if $A \in \mathcal{L}(V)$ is self-adjoint $\langle Av, v\rangle \in \mathbb{R}$ for any $v \in V$. This motivates us to define the quadratic form of A, a map $Q_A : V \to \mathbb{R}$, by setting $Q_A(v) = \langle Av, v\rangle$. Similarly, if Q_A is real, A is self-adjoint. Indeed, looking at $Q_A(v+tw) = Q_A(v) + |t|^2 Q_A(w) + \bar{t}\langle Av, w\rangle + t\langle Aw, v\rangle$ we see that $\bar{t}\langle Av, w\rangle + t\langle Aw, v\rangle \in \mathbb{R}$ for any $v, w \in V$ and $t \in \mathbb{C}$. Still $\bar{t}\langle Av, w\rangle + t\langle Aw, v\rangle = \bar{t}\langle Av, w\rangle + t\overline{\langle Av, w\rangle} - t\langle w, Av\rangle + t\langle Aw, v\rangle = \Re(\bar{t}\langle Av, w\rangle) + t(\langle Aw, v\rangle - \langle w, Av\rangle)$ so we must have $\langle Aw, v\rangle = \langle w, Av\rangle$.

Rayleigh quotient is sort of a scale-invariant version of the quadratic form, and that's why it captures many of the properties of self-adjoint maps. It is also a good way to think about the operator norm of a linear map. Take any $\mathcal{H}(V) \ni A = \sum_{i=1}^n \lambda_i P_{v_i}$ and also any $V \setminus \{0\} \ni x = \sum_{i=1}^n x_i v_i$. Now

$$R(A, x) = \frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\sum_{i=1}^n \lambda_i |x_i|^2}{\sum_{i=1}^n |x_i|^2},$$

the Rayleigh quotient is a weighted average of the eigenvalues of A. Now if A is any linear map, and $x \in V \setminus \{0\}$

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle = R(A^*A, x)\|x\|^2$$

Now A^*A is self-adjoint. Also its eigenvalues must be non-negative, since the Rayleigh quotient is: square roots of these values are called the *singular values* of A. These considerations make it also clear that the largest singular value of A is the operator norm of A. If A is self-adjoint itself, singular values are simply the absolute values of the eigenvalues, and the operator norm is the maximum of these absolute values, as claimed before. More generally, if A is normal, i.e. we can write $A = \sum_{i=1}^n \lambda_i P_{v_i}$ with $\lambda_i \in \mathbb{C}$ and v_i's orthonormal, we have

$$A^*A = \left(\sum_{i=1}^n \overline{\lambda_i} P_{v_i}\right)\left(\sum_{j=1}^n \lambda_j P_{v_j}\right) = \sum_{i=1}^n |\lambda_i|^2 P_{v_i},$$

so again singular values are the absulute values of the eigenvalues and $\|A\| = \max_{i=1}^{n} |\lambda_i|$. Rayleigh quotient also has interpratation for normal maps, it is again weighted average of the eigenvalues. When x ranges over non-zero vectors, or equivalently over some sphere, $R(A, x)$ ranges over the convex full of the eigenvalues.

Many of the previous properties don't hold for general linear maps. Also, singular values are absolute values of eigenvalues exactly for the normal maps.

2.5 Intuition

One way to understand adjoints is to look at a bit more general case. Fix any two finite-dimensional inner-product spaces, $(V, \langle\cdot,\cdot\rangle_V)$ and $(W, \langle\cdot,\cdot\rangle_W)$, not necessarily of the same dimension, over, say, $\mathbb{C}$. Now given any linear $A \in \mathcal{L}(V, W)$, adjoint is a linear map $A^* \in \mathcal{L}(W, V)$ such that for any $v \in V$ and $w \in W$ we have

$$\langle Av, w\rangle_W = \langle v, A^* w\rangle_V.$$

Again, adjoint exists and it's unique. Fix any $w \in W$. Vector w induces a linear mapping $W \to \mathbb{C}$ by $w' \mapsto \langle w', w\rangle_W$. Hence we get a map $\Phi_W : W \to W^*$ where V^* is the dual of V. It turns out that Φ_W is anti-linear bijection. Anti-linearity and injectivity are easy to check, and since both spaces have same dimension, map has to be also surjection.

Composing with A, Φ_W induces a map $W \to V^*$ given by $w \mapsto \Phi_W(w) \circ A$. Now we finally compose this with Φ_V^{-1} to get a map $W \to V$, the adjoint. So $A^*(w) = \Phi_V^{-1}(\Phi_W(w) \circ A)$.

We can still clean the definition a bit. Any linear map $A \in \mathcal{L}(V, W)$ induces a mapping between the correponding duals (with the order reversed) ${}^tA : W^* \to V^*$, given by $({}^tA)(\phi) = \phi \circ A$ for any $\phi \in W^*$. With this interpretation in mind we rewrite $w \mapsto \Phi_W(w) \circ A = {}^tA \circ \Phi_W$, and finally

$$A^* = \Phi_V^{-1} \circ {}^tA \circ \Phi_W.$$

So the point is: from map $A \in \mathcal{L}(V, W)$ we (canonically) get a map ${}^tA : W^* \to V^*$ and since we can, thanks to the inner product, identify inner-product spaces with their duals, this gives us a map $A^* \in \mathcal{L}(W, V)$.

It should be noted that all parts of the theorem 2.7 carry directly to this more general setup with obvious modifications.

Inner products give also other canonical isomorphism: one between linear maps and sesquilinear forms. Sesquilinear forms are complex generalizations of bilinear forms. Sesquilinear form $B : V \times W \to \mathbb{C}$ is a mapping linear in its first entry and anti-linear in the second. The vector space of all sesquilinear forms on $V \times W$ is denoted by $\mathcal{B}(V, W)$.

Any linear mapping $A \in \mathcal{L}(V, W)$ gives rise to a sesquilinear form given by $\langle A\cdot, \cdot\rangle_W$. This gives us a map $\Psi_{V,W} : \mathcal{L}(V, W) \to \mathcal{B}(V, W)$, and we also denote $\Psi_{V,W}A = B_A$. This map is an isomorphism: it can be easily seen that is linear, and injection, and since both spaces are of dimension $\dim(V) \cdot \dim(W)$, the map is also an surjection.

We also have a natural map $(\cdot)^H : \mathcal{B}(V, W) \to \mathcal{B}(W, V)$, *conjugate transpose* given by $B^H(w, v) = \overline{B(v, w)}$: we simply exchange the arguments, and conjugate to preserve the sesquilinearity. Now it's not hard to check that $(\cdot)^* = \Psi^{-1}_{W,V} \circ (\cdot)^H \circ \Psi_{V,W}$, so we get an another route to adjoint.

Self-adjoint maps of course only make sense when $V = W$. When $A \in \mathcal{L}(V)$ is a self-adjoint, $B^H_A = B_A$: such sesquilinear forms are called Hermitian. The diagonal of B is the of course just the quadratic form, i.e. $B(v, v) = Q_A(v)$ for any $v \in V$.

Adjoint also behaves in many ways like (complex) conjugate.

- Notions agree in 1-dimensional spaces.
- Taking adjoint is an anti-linear involution.
- Eigenvalues of the adjoint are conjugates of the eigenvalues of the original map.
- If $A \in \mathcal{L}(V)$ has an orthonormal eigenbasis, A^* has the same eigenbasis with conjugated eigenvalues.
- Self-adjoint maps correspond to real numbers: they have real eigenvalues.

Maps for which the fourth point holds are called *normal*. The class of normal maps on V is denoted by $\mathcal{N}(V)$

One of the very good properties of conjugation is that it commutes with many operations, taking inverse, for instance. The same carries to adjoints:

$$I = (A^{-1}A)^* = A^*(A^{-1})^*,$$

so $(A^{-1})^* = (A^*)^{-1}$. There's a natural generalization we'll come back to later.

2.6 Commuting self-adjoint maps

Warning! Composition of self-adjoint maps need not be self-adjoint!

It's rather surprising that they are closed under sums in the first place. Of course both Hermitian and self-adjoint maps are by definition, but it's rather surprising that such class has such nice properties (most of which are left to discuss).

The problem is that even if $A, B \in \mathcal{H}(V)$, $(AB)^* = B^*A^* = BA$, which is in general different from AB: two self-adjoint maps need not commute. Again, one should have a

counterexample and almost anything works for that. We need to have $n > 1$. Also we should be able to find counterexamples that are projetions, since if any two projections commute, so do maps in their span, Hermitian maps. Finally we should not take parallel or orthogonal vectors: projections are idemponent and if one applies two orthogonal projections, result if the zero map. Actually anything else works as a counterexample. Indeed, when v and w are not parallel or orthogonal. Now $P_w P_v v = P_w v =$ something non-zero parallel to w, while $P_v P_w$ is something non-zero parallel to v or at least not parallel to w. Hence P_v and P_w do not commute.

From the previous we see that projections commute exactly when the corresponding vectors are parallel or orthogonal. More generally one would like to determine exactly which products of self-adjoint maps are self-adjoint, i.e. which products commute. Large family of examples is given by pairs of self-adjoint maps with the same eigenvectors. Namely if $A = \sum_{i=1}^n \lambda_i P_{v_i}$ and $B = \sum_{i=1}^n \lambda_i' P_{v_i}$,

$$AB = \left(\sum_{i=1}^n \lambda_i P_{v_i}\right)\left(\sum_{j=1}^n \lambda_j' P_{v_j}\right) = \sum_{i=1}^n \sum_{j=1}^n \left(\lambda_i P_{v_i}\right)\left(\lambda_j' P_{v_j}\right) = \sum_{i=1}^n \lambda_i \lambda_i' P_{v_i} = BA.$$

In this case eigenvalues of the product are also products of the eigenvalus of A and B, but not necessarily all of them. It turns out that that's all. Before going through the proof, let's get back to the spectral theorem. One of the unfortunate aspects of it is that the spectral representation is not unique. We can however make it unique. Essentially only choices we get to make in the representation are the choices for the orthonormal bases of the eigenspaces of the map, so if we replace the projections to vectors with projections to the eigenspaces, we have the uniqueness.

Recall that for arbitrary subspace of V, say T, (orthogonal) projection to T, denoted by P_T is the unique linear map such that $P_T^2 = P_T$, $\mathrm{im}(P_T) = T$, $\ker(P_T) = \mathrm{im}(P_T)^\perp$. The first condition is the projection part, the second projetion being to T and the third is the orthogonality. Orthogonality can be also replaced by requiring P_T to be self-adjoint. Equivalently for any $v \in V$ $P_T v$ is the unique vector in T, such that $v - P_T v \perp T$. Or still equivalently, $P_T v$ is the unique vector in T closest to v. Equivalence of these definitions is easy to check, and also the fact that projections to vectors used before are just projections to corresponding subspaces. Finally, if $e_1, e_2, \ldots, e_m$ is an orthonormal basis for an subspace T, $P_T = \sum_{i=1}^m P_{e_i}$.

Theorem 2.8 (Spectral theorem)**.** *Let V be an n-dimensional inner-product space and $A \in \mathcal{H}(V)$. Then there exists unique non-negative integer m, real numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$ and non-trivial orthonormal subspace of V, $E_{\lambda_1}, E_{\lambda_2}, \ldots E_{\lambda_m}$ with $E_{\lambda_1} + E_{\lambda_2} + \ldots + E_{\lambda_m} = V$, such that*

$$A = \sum_{i=1}^m \lambda_i P_{E_{\lambda_i}}.$$

Moreover, this representation is unique.

Here m is the number of distinct eigenvalues of A, λ_i's are the eigenvalues, and E_{λ_i}'s are the corresponding eigenspaces: representation is necessarily unique. Existence of such representation is immediate from the previous version of the spectral theorem.

We can also reinterpret the original proof a bit further. One of the key facts we used was that if v is an eigenvector, $v^\perp$ is mapped to itself. This fact has rather natural generalization: if $A \in \mathcal{H}(V)$ and T is a subspace of V which maps to itself, so does its orthocomplement. Now we can decompose A to A_T and $A_{T^\perp}$. Put a bit differently: every self-adjoint map is canonically a product of real multiplication maps.

Theorem 2.9. *Let V be an n-dimensional inner-product space and $A_j \in \mathcal{H}(V)$ for $j \in J$ be an arbitrary family of (pairwise) commuting self-adjoint maps on V. Then there exists on-negative integer m, non-trivial orthogonal subspaces $T_1, T_2, \ldots, T_m$ of V with $T_1 + T_2 + \ldots + T_m = V$, and sequences of real numbers $(\lambda_{i,j})_{i=1}^m$ for $j \in J$ such that for any $j \in J$ we have*

$$A_j = \sum_{i=1}^{m} \lambda_{i,j} P_{T_i}$$

and if $i \neq i'$, for some $j \in J$ we have $\lambda_{i,j} \neq \lambda_{i',j}$. Moreover, this representation is unique.

When such a representation exists, we call the family $\{A_j | j \in J\}$ *simultaneously diagonalizable*. We again see that the $\lambda_{i,j}$'s are eigenvalues of the corresponding A_j's. The last requirement on the subspaces and $\lambda_{i,j}$'s is just technical minimality condition to ensure uniqueness. If we have any representation ignoring this constraint, we can just combine the subspaces if all the eigenvalues are equal.

Proof. Let us first check the uniqueness. It is clear that if two representations share subspaces, also the corresponding $\lambda_{i,j}$'s are equal. So we may assume to the contrary that there are two such representations with subspaces $T_1, T_2, \ldots, T_m$ and $T'_1, T'_2, \ldots, T'_{m'}$ and $T_1 \not\subset T'_1$, but T_1 and T'_1 are not orthogonal. Take $v \in T'_1 \setminus (T_1 \cup T_1^\perp)$ and write it in the form $v_1 + v'_1$ where $v_1 \in T_1$ and $v'_1 \in T_1^\perp$. For any A_j both v and v'_1 are eigenvalues not orthogonal, so they correspond to the same eigenvalue. Hence v'_1 is also a eigenvector for any A_j corresponding to the same eigenvalue. Now v'_1 can't be orthogonal to everything in T'_1 so it is not orthogonal to some vector v_2 in T_2, say. As before we see that for any A_j the eigenvalue corresponding to T_1 is same as the eigenvalue corresponding to v_2, and so T_2. This means that the first representation is not minimal, we could have combined the T_1 and T_2, a contradiction.

We proceed by induction on n, the dimension of the space. The case $n = 0$ is trivial. When $n > 0$ if we look at the spectral representations of the maps A_j. If all of them

simple representation ($m = 1$), i.e. they are multiples of identity map, we can also choose $m = 1$, let $\lambda_{1,j}$'s be the corresponding unique eigenvalue of A_j and let $T_1 = V$.

Assume then that some A_j has non-trivial decomposition and let T be an subspace in the decomposition, eigenspace of A_j, with eigenvalue λ. We know that A_j itself can be decomposed into maps living in T and $T^\perp$: we shall prove that same is true about all the other maps $A_{j'}$.

Take any $j' \in J$. We want to prove that $A_{j'}T \subset T$ Since $A_{j'}$ and A_j, for any $v \in T$ we have

$$\lambda A_{j'}v = A_{j'}A_jv = A_jA_{j'}v.$$

If $A_{j'}v = 0$, everything's fine. If not, $A_{j'}v$ is an eigenvector of A_j with eigenvalue λ: it lives by definition in T. It follows that also $A_{j'}T^\perp \subset T^\perp$, and $A_{j'}$ factorizes to maps in T and $T^\perp$.

Once we perform this factorization for all the maps, we have reduced our problem to smaller case: T was a proper non-trivial subspace of V. It follows by induction that we can find factorizations as in the statement of the theorem for both sides, and we simply put them together to get final representation. This representation is also minimal, as is rather is easy to check, put even if it wasn't, we could cook up such as explained before the proof. □

From the proof we also see that any representation ignoring the minimality condition is just a subdivision of a minimal one: the subspaces are further divided to subspaces.

Previous theorem also sheds some light to the normal maps. Take any normal map $A = \sum_{i=1}^n \lambda_i P_{v_i}$. When the eigenvalues are split to real and imaginary parts $\lambda_i = a_i + ib_i$, we can write $A = \sum_{i=1}^n a_i P_{v_i} + i\sum_{i=1}^n b_i P_{v_i}$, where now $\sum_{i=1}^n a_i P_{v_i}$ and $\sum_{i=1}^n b_i P_{v_i}$ are self-adjoint: we have decomposed normal map itself to its real and imaginary part. This construction can be done in general too. We will denote $\Re(A) = \frac{1}{2}(A + A^*)$ and $\Im(A) = \frac{1}{2i}(A - A^*)$. For general linear map A, we can write $A = \Re(A) + i\Im(A)$. While again both $\Re(A)$ and $\Im(A) = \Re(A/i)$ are self-adjoint, $i\Im(A)$ is "purely imaginary": such maps are called *skew-Hermitian*: they are maps satisfying $A^* = -A$. We see that if A is normal, the eigenvalues of $\Re(A)$ and $\Im(A)$ are exactly the real and imaginary parts of the eigenvalues of A, respectively.

For any linear map A we can definitely find spectral representation for its real and imaginary parts. If they are simultaneously diagonalizable, they give rise to the spectral representation of the original map. By the previous theorem this is equivalent to real and imaginary parts commuting, which by short computation is equivalent to $A^*A = AA^*$: A commutes with its adjoint. It follows that if A commutes with its adjoint, A is normal. Also if A is normal, A definitely commutes with its adjoint, so these conditions are equivalent. The previous is actually a very common definition of normal map.

2.7 Unitary maps

Besides self-adjoint maps, there is other very nice subfamily of normal maps: unitary maps. Unitary maps are normal maps with eigenvalues on unit circle. Such maps satisfy

$$A^*A = \sum_{i=1}^{n} \overline{\lambda_i} P_{v_i} \sum_{i=1}^{n} \lambda_i P_{v_i} = \sum_{i=1}^{n} |\lambda_i|^2 P_{v_i} = I,$$

On the other hand, if $A \in \mathcal{L}(V)$ with $A^*A = I$, then A^* is inverse of A^{-1}, and A^* commutes with A: A is normal. By previous computation we see that all the eigenvalues of A are on unit circle, hence map is unitary, if and only if $A^*A = I$. Finally if A is unitary, for any $v, w \in V$ we have

$$\langle Av, Aw \rangle = \langle v, A^*Aw \rangle = \langle v, w \rangle,$$

so A preserves innerproduct. Again, if A preserves innerproducts, for any $v, w \in V$ we have

$$\langle v, w \rangle = \langle Av, Aw \rangle = \langle v, A^*Aw \rangle,$$

so $A^*A = I$: we have third equivalent definition for unitary maps. This definition might be also most enlightening: unitary maps are exactly the automorphisms of the inner-product space V.

There is also fourth equivalent definition for unitary maps: they are the isometries of V. It's clear that any unitary map is isometry, but the other direction might not be entirely obvious. The innerproduct can however be recovered from the norm. If V is over $\mathbb{R}$, we have

$$\langle v, w \rangle = \frac{1}{4} \left(\|v + w\|^2 - \|v - w\|^2 \right).$$

It follows that if A is an isometry, for any $v, w \in V$ we have

$$\langle Av, Aw \rangle = \frac{1}{4} \left(\|Av + Aw\|^2 - \|Av - Aw\|^2 \right) = \frac{1}{2} \left(\|v + w\|^2 - \|v - w\|^2 \right) = \langle v, w \rangle.$$

If V is over $\mathbb{C}$ instead, we have to modify our polarization identity a little:

$$\langle v, w \rangle = \frac{1}{4} \left(\|v + w\|^2 - \|v - w\|^2 + i\|v + iw\|^2 - i\|v - iw\|^2 \right).$$

Both the third and the fourth definitions make it clear that the unitary maps on V form a group. This group is often denoted by $\mathrm{Hilb}(V)$, but we will use $\mathcal{U}(V)$, for brevity, and because $\mathcal{H}$ is preserved.

TODO: sesquilinear forms vs. quadratic form correspondence.

TODO: Unitary maps and matrices

TODO: Cayley transform

2.8 Positive maps

Definition 2.10. Hermitian map is said to be *positive* if all of its eigenvalues are non-negative.

Such maps are usually called *positive semi-definite*: the term *positive definite* is usually preserved for the following class.

Definition 2.11. Hermitian map is said to be *strictly positive* if all of its eigenvalues are positive.

"A is positive" is denoted by $A \geq 0$, and "A is stricly positive" by $A > 0$. One could adopt analogous notation for (strictly) negative maps, definition of which should be clear. The set of positive maps is denoted by $\mathcal{H}_+(V)$, shortened to $\mathcal{H}_+$. Here is a big list of important facts and equivalent definitions for the previous.

- A map is positive if and only if its quadratic form, or equivalently the Rayleigh quotient is non-negative. This is clear since we noticed the Rayleigh is just a weighted average of the eigenvalues, and quadratic form is just a positively scaled version of Rayleigh quotient. Similarly, a map is strictly positive if and only if the Rayleigh quotient is positive, of the quadratic form is positive, except at 0. In both instances one could replace the non-negativity/positivity on everywhere/outside zero by requiring it only on unit sphere, for instance.

- Previous point makes it clear that sum of (strictly) positive maps is (stricly) positive. Analogous claim is true for positive scalar multiple.

- A map is strictly positive if and only if the respective sesquilinear form is inner-product. Inner-products can hence also be naturally called stricly positive.

  Positive map don't necessarily lead to inner-products, but to a so called semi-definite sesquilinear forms, which are here also called positive (sesquilinear) forms.

- Composition of positive maps need not be positive, but this is just because it need not be even Hermitian. If it is, however, by our previous results, the maps commute, are simultaneosly diagonalizable and they're composition is positive: its eigenvalues are (some of the) products of the eigenvalues of the original maps. Similar statement holds for striclty positive maps.

- The class of positive maps is topological closure (with operator norm topology) of the class of stricly positive maps. This is clear, and the closedness is very useful. Also, the class of stricly positive maps is the interior the class of positive maps. The

reason is the following: if $A \geq 0$, $Q_A \geq \delta_K > 0$ for any compact set $K \subset V$ not containing 0, so in particular on unit sphere. Now if $B \in \mathcal{L}(V)$ has operator norm $< \delta$, $Q_{A+B}(v) \geq Q(A)(v) - |Q_B(v)| > \delta - \delta$, so Q is positive on unit circle, and thus outside zero.

- Squares are positive, i.e. $A^2 \geq 0$ for any Hermitian A. Also, any positive map is a square, and has an unique positive square root. These claims follows easily from the better version of the spectral theorem. Positive square root of $A \geq 0$ is denoted by $A^{\frac{1}{2}}$, as one would hope.

- Projections are positive: their eigenvalues belong to $\{0, 1\}$.

- We already noticed that maps of the form A^*A are positive: their Rayleigh quotient is non-negative. Every positive map is also of this form (square root fits the bill), but not uniquely.

- Strictly positive maps are central at the study of the local maxima and mimima of multivariate functions. Let V be real inner-product space and $f : V \to \mathbb{R}$ be a twice continuously differentiable. Now me write $f(x + y) = f(x) + G_x(y) + B_x(y, y) + o(\|y\|^2)$, where G_x is the derivative of f at x, and H_x is the hessian, the bilinear form correponding to the second derivative. Now if $G_x = 0$ and the Hessian H_x is strictly positive, the function f has a local minimum at x. If H_x is merely positive, such conclusions can't be drawn. If, however, $H_\cdot$ is positive in a neighbourhood of x, f has again a local minimum at x.

- One way to think about positive maps is that they don't turn vectors too much. If we are working in real inner product space, the angle between two non-zero vectors v and w is the unique real number $\alpha \in [0, \pi]$ such that

$$\cos(\alpha) = \frac{\langle v, w \rangle}{\|v\|\|w\|}.$$

  The angle is less than $\frac{\pi}{2}$ if and only if $\cos(\alpha) > 0$, or equivalently if $\langle v, w \rangle > 0$. This means that a Hermitian map is positive if and only if the angle between v and Av is less than $\frac{\pi}{2}$: positive map doesn't turn a vector too much. This intuition can be somewhat pushed to complex spaces, but then the concept of an angle is harder to grasp.

  FIGURE: this intuition visualized in $\mathbb{R}^2$ or $\mathbb{R}^3$

2.9 Congruence and Sylvesters criterion

2.9.1 $*$-conjugation

There is one very important way to produce positive maps from others, called congruence. Given any two positive maps A and B, their composition need not be positive, but the map BAB is. First of all, it is self-adjoint, as $(BAB)^* = B^*A^*B^* = BAB$. Also $Q_{BAB}(v) = \langle BABv, v\rangle = \langle A(Bv), (Bv)\rangle \geq 0$ for any $v \in V$. We didn't really need the assumption on the positivity of B, but self-adjointness was not that important either. Namely for arbitrary linear B we could consider the product B^*AB instead: this is positive whenever A is. If $C = B^*AB$ for some $B \in \mathcal{L}(V)$, we say that C is $*$-conjugate of A.

We also see that $Q_{B^*AB} = Q_A \circ B$: conjugation is a change of basis in the quadratic form. Similar statement is evidently true for the respective sesquilinear form. This is the main motivation for the definition of the $*$-conjugation. We have already seen that the quadratic form of a map is a good way to characterize many of its good properties, so to some extent to undestand maps, we just to need to understand structure of their quadratic forms. By change of basis of the quadratic form we have a good control of what happens. We might however lose some information: if $B = 0$, for instance, the quadratic form after $*$-conjugation by B doesn't tell much about A. But if B is invertible, or equivalently if C and B are $*$-conjugates of each other, we shouldn't lose any information. If this is the case, we say that A and C are congruent. It is easily verified that congruence is a equivalence relation.

The construction of $*$-conjugation also sense for general linear map A, i.e. we could just as well $*$-conjugate non-positive, or even non-self-adjoint maps. The result then need not be positive or self-adjoint, and in general, $*$-conjugation loses its usefulness. Also, even if A is normal, $*$-conjugate of A need not be normal. TODO

The previous construction can be also performed between two spaces V and W: given any map $B \in \mathcal{L}(V, W)$ and $A \in \mathcal{H}_+(W)/\mathcal{H}(W)/\mathcal{L}(W)$, we note that $B^*AB \in \mathcal{H}_+(V)/\mathcal{H}(W)/\mathcal{L}(W)$. For self-adjoint maps we can say a lot more: while congruence doesn't in general preserve eigenvalues, it preserves their signs.

Theorem 2.12 (Sylvester's Law of Inertia)**.** *$A, B \in \mathcal{H}(V)$ are congruent, if and only if A and B have equally many positive, negative and zero eigenvalues, counted with multiplicity.*

Proof. Let's start with the "if" part. Let's denote the eigenvalues of A and B by $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ and $\lambda'_1 \leq \lambda'_2 \leq \ldots \leq \lambda'_n$, respectively, and the corresponding eigenvectors with $v_1, v_2, \ldots, v_n$ and $v'_1, v'_2, \ldots, v'_n$. By assumption λ_i and λ'_i have the same sign (or are both zero) for any $1 \leq i \leq n$, so we may find non-zero real numbers $t_1, t_2, \ldots, t_n$ such that $\lambda_i = \lambda'_i t_i^2$. Now consider a linear map C with $Cv_i = t_i v'_i$. C is clearly a surjection and hence a bijection. Also if $v = \sum_{i=1}^n x_i v_i$ $(Q_B \circ C)(v) = Q_B(\sum_{i=1}^n x_i t_i v'_i) = \sum_{i=1}^n |x_i|^2 t_i^2 \lambda'_i =$

$\sum_{i=1}^n |x_i|^2 \lambda_i = Q_A(v)$ so $Q_{C^*BC} = Q_B \circ C = Q_A$. It follows that $C^*BC = A$ and hence A and B are congruent.

The "only if" - part is a bit trickier. The idea is to find a good description for the number of positive non-negative eigenvalues. We noticed before that we can write quadratic forms in the form $Q_A(v) = \sum_{i=1}^n \lambda_i |x_i|^2$ if $v = \sum_{i=1}^n x_i v_i$, and v_i are the eigenvectos of A with $\lambda_i' s$ as the corresponding eigenvectors. In particular if say first k eigenvalues are negative, Q_A will be negative on span$\{v_i | 1 \leq i \leq k\}$, a k-dimensional subspace, minus zero. Similarly, now $n-k$ of the eigenvalues are non-negative, so the quadratic form is non-negative on a subspace of dimension of at least $n-k$. But the dimensions can't be any bigger: if Q_A were for instance negative on some $k+1$ dimesional subspace, this subspace would necessarily intersect a subspace where Q_A is non-negative, which is non-sense.

Congruence preserves the previous notion: if Q_B is negative on a subspace of dimension k, so is $Q_B \circ C$ for any invertible C; namely in the inverse image. Same reasoning holds for the the subspace on which Q_B is non-negative, so again, $Q_B \circ C$ has to have similar structure. We are done. □

If n_0, n_- and n_+ denote the number of zero, negative and positive eigenvalues of A, *inertia* of A is the triplet $\{n_0, n_-, n_+\}$. The previous theorem can be hence restated, that inertia is invariant under congruence.

It gets better: if two normal maps are congruent, we can vastly generalize the previous result: for normal maps, congruence preserves the number of eigenvalues on any open ray.

For any $A \in \mathcal{N}(V)$ define the mapping $\theta[A] : S^1 \to \mathbb{N}$ by setting $\theta[A](z)$ be the number of eigenvalues (counted with multiplicity) on ray $\{rz | r > 0\}$. This map is called the *angularity* of A.

TODO: figure of change of eigenvalues on congruence.

Theorem 2.13 (Generalized Sylvester's Law of Inertia)**.** *$A, B \in \mathcal{N}(V)$ are congruent, if and only if $\theta[A] = \theta[B]$.*

Proof. The "if"-part can be proven almost identically to proof of the self-adjoint case. For the "only if" -part, note that if $A = C^*BC$, the also $\Re(A) = C^*\Re(B)C$. Now both sides are self-adjoint, and eigenvalues of $\Re(A)$ and $\Re(B)$ are just the real parts of the eigenvalues of A and B. Applying the previous version Law of Inertia for these maps, we see that A and B have equally many eigenvalues on closed right half-plane.

Carrying out the previous reasoning for the maps $e^{it}A$ and $e^{it}B$ for real t we see that these maps have equally many eigenvalues on any closed half plane containing origin. If one considers function sending t to the number of eigenvalues of $e^{it}A$ on closed right half-plane, we see that this function has jump discontinuitys, and the value, and the left and the right limits at these point reveals number of eigenvalues at any open ray.

TODO: figure of functions $t \mapsto$ number of eigenvalues of $e^{it}A$ on closed right half-plane. □

Sylvester's Law of inertia gives another proof of the fact that strictly positive maps are exactly the maps congruent to the identity, and positive maps are the maps congruent to some projection. More precisely, the positive maps are partitioned to $n+1$ congruene classes depending on their rank, k:th congruence class containing the projections to k-dimensional subspaces. 0:th class contains only the zero map, the only rank 0 (positive) map, and the n:th class is the class of strictly positive maps.

If one $*$-conjugates with non-invertible, the angularity may change, but in quite obvious way only: some eigenvalues may move to 0. In particular, we have the following supplement even a bit more general version of the law.

Theorem 2.14 (General Sylvester's Law of Inertia)**.** *For $A, B \in \mathcal{N}(V)$ and A is $*$-conjugate of B, if and only if $\theta[A] \leq \theta[B]$.*

Proof. TODO □

This extension draws a picture about the relation of previously mentioned congcruence classes. We can move to the congruence classes of lower indeces by $*$-conjugation, but cannot move up the ladder: the complexity of quadratic forms cannot increase.

The important property of $*$-conjugates we have used repeatedly is that it commutes with taking adjoint (and hence with real and imaginary parts). This is not of course immediate from the definition but is also clear since taking adjoint corresponds to taking (complex) conjugate of quadratic form, and complex conjugation commutes with pretty much anything anyway.

2.9.2 Unitary congruence

There is a very important class of congruence called unitary congruence: congruence by unitary map. Since adjoint of an unitary map is just its inverse, $*$-conjugation by unitary map, or unitary conjugation is also conjugation in the usual sense. Unitary congruence hence also preserves eigenvalues.

TODO

2.10 Fake products

2.11 Absolute value and Polar decompositions

2.12 Matrices and computation

2.13 Positivity without inner-product

TODO:

- (canonical, löwdin) orthogonalization, polar decomposition and orthogonal Procrustes problem
- projection matrices
- Hilbert-Schmidt norm ($\rightarrow$ matrix functions?) and inner product
- Real vs. complex

Chapter 3

Divided differences

Divided differences are derivatives without limits.

Consider a function $f : \mathbb{R} \to \mathbb{R}$. It's (first) divided difference is defined as

$$[\cdot, \cdot]_f : \mathbb{R}^2 \setminus \{x \neq y\} \to \mathbb{R}$$
$$[x, y]_f = \frac{f(x) - f(y)}{x - y}.$$

If f is sufficiently smooth, we should also define $[x, x]_f = f'(x)$: if $f \in C^1(\mathbb{R})$, this gives continuous extension to $[\cdot, \cdot]_f$. Much of the power of divided differences comes however from the fact that they conveniently carry same information even if we do not do such extension.

Consider the case of increasing f. This information is exactly carried by the inequality $[x, y]_f \geq 0$. Again, if f is differentiable, this is equivalent to $f'(x) = [x, x]_f \geq 0$. There are many ways to see this fact, one of the more standard being the mean value theorem: If $x \neq y$, for some ξ between x and y we have

$$\frac{f(x) - f(y)}{x - y} = f'(\xi).$$

Now if the derivative is everywhere non-negative, so are divided differences. Also divided differences are sort of approximations for derivative.

It gets better: consider case of convex f. Convexity is usually defined using a real parameter $0 \leq t \leq 1$: we require the inequality $tf(x) + (1 - t)f(y) \geq f(tx + (1 - t)y)$ to hold for any x and y. For real two times differentiable functions this equivalent to second derivative being at least zero. There is however more symmetric form for convexity. First divided difference corresponds somewhat to first derivative, so second divided difference, defined for three distinct x, y, z as

$$[x, y, z]_f = \frac{[x, y]_f - [y, z]_f}{x - z} = \frac{f(x)}{(x - y)(x - z)} + \frac{f(y)}{(y - z)(y - x)} + \frac{f(z)}{(z - x)(z - y)},$$

should correspond to second derivative. And indeed it does. Assume first that $x < y < z$. We have $f'(\xi_x) = [x, y]_f$ for some $x < \xi_x < y$ and $f'(\xi_z) = [y, z]_f$ for some $y < \xi_z < z$. Now

$$[x, y, z]_f = \frac{[x, y]_f - [y, z]_f}{x - z} = \frac{f'(\xi_x) - f'(\xi_z)}{x - z} = \frac{\xi_x - \xi_z}{x - z} \frac{f'(\xi_x) - f'(\xi_z)}{\xi_x - \xi_z} = \frac{\xi_x - \xi_z}{x - z} [\xi_x, \xi_z]_{f'}$$

Applying mean value theorem again for f', we get that $[\xi_x, \xi_z]_{f'} = f''(\xi)$ for some $\xi_x < \xi < \xi_z$. Finally

$$[x, y, z]_f = \frac{\xi_x - \xi_z}{x - z} f''(\xi).$$

Since second divided difference is evidently symmetric in its entries, we can get rid of the assumption on the order of x, y and z. This isn't quite as good as the mean value theorem anymore, but it is sufficient to connect second derivative to second divided differences: if the second derivative is always non-negative, so is the second divided difference. Implication goes the other way too, but that's not really clear from our computation. Non-negativity of the second divided difference however corresponds exactly to the standard definition for convexity by change of variables. Fix $x \neq y$, choose $0 < t < 1$ and set $z = tx + (1 - t)y$. We compute that

$$\begin{aligned}
[x, y, z]_f &= \frac{f(x)}{(x - y)(x - z)} + \frac{f(y)}{(y - z)(y - x)} + \frac{f(z)}{(z - x)(z - y)} \\
&= \frac{f(x)}{(x - y)(x - (tx + (1 - t)y))} + \frac{f(y)}{(y - (tx + (1 - t)y))(y - x)} \\
&\quad + \frac{f(tx + (1 - t)y)}{((tx + (1 - t)y) - x)((tx + (1 - t)y) - y)} \\
&= \frac{f(x)}{(1 - t)(y - x)^2} + \frac{f(y)}{t(y - x)^2} - \frac{f(tx + (1 - t)y)}{t(1 - t)(y - x)^2} \\
&= \frac{tf(x) + (1 - t)f(y) - f(tx + (1 - t)y)}{t(1 - t)(y - x)^2}.
\end{aligned}$$

As expected, the general n:th order divided differences, defined for distinct real numbers $x_0, x_1, \ldots, x_n$ is defined recursively as

$$[x_0, x_1, \ldots, x_n]_f = \frac{[x_0, x_1, \ldots, x_{n-1}]_f - [x_1, x_2, \ldots, x_n]_f}{x_0 - x_n}.$$

Again, n:th divided difference corresponds to n:th derivative. TODO

Chapter 4

Matrix functions

Chapter 5

Pick-Nevanlinna functions

Pick-Nevanlinna function is an analytic function defined in upper half-plane with a non-negative real part. Such functions are sometimes also called Herglotz or $\mathbb{R}$ functions but we will often call them just *Pick functions*. The class of Pick functions is denoted by $\mathcal{P}$.

Pick functions have many interesting properties related to positive matrices and that is why they are central objects to the theory of matrix monotone functions.

5.1 Basic properties and examples

Most obvious examples of Pick functions might be functions of the form $\alpha z + \beta$ where $\alpha, \beta \in \mathbb{R}$ and $\alpha \geq 0$. Of course one could also take $\beta \in \overline{\mathbb{H}}_+$. Actually real constants are the only Pick functions failing to map $\mathbb{H}_+ \to \mathbb{H}_+$: non-constant analytic functions are open mappings.

Sum of two Pick functions is a Pick function and one can multiply a Pick function by non-negative constant to get a new Pick function. Same is true for composition.

The map $z \mapsto -\frac{1}{z}$ is evidently a Pick function. Hence are also all functions of the form

$$\alpha z + \beta + \sum_{i=1}^{N} \frac{m_i}{\lambda_i - z},$$

where N is non-negative integer, $\alpha, m_1, m_2, \ldots, m_N \geq 0$, $\beta \in \mathbb{H}_+$ and $\lambda_1, \ldots, \lambda_N \in \mathbb{H}_+$. So far we have constructed our function by adding simple poles to the closure of lower half-plane. We could further add poles of higher order at lower half plane, and change residues and so on, but then we have to be a bit more careful.

There are (luckily) more interesting examples: all the functions of the form x^p where $0 < p < 1$ are Pick functions. To be precise, one should choose branch for the previous so that they are real at positive real axis. Also log yields Pick function when branch

is chosen properly i.e. naturally again. Another classic example is *tan*. Indeed, by the addition formula

$$\begin{aligned}\tan(x+iy) &= \frac{\tan(x)+\tan(iy)}{1-\tan(x)\tan(iy)} = \frac{\tan(x)+i\tanh(y)}{1-i\tan(x)\tanh(y)} \\ &= \frac{\tan(x)(1+\tanh^2(y))}{1+\tan^2(x)\tanh^2(y)} + i\frac{(1+\tan^2(x))\tanh(y)}{1+\tan^2(x)\tanh^2(y)},\end{aligned}$$

and y and $\tanh(y)$ have the same sign.

We observe the following useful fact.

Proposition 5.1. *If $(\varphi_i)_{i=1}^n$ is a sequence of Pick functions converging locally uniformly, the limit function is also a Pick function.*

Proof. Locally uniform limits of analytic functions are analytic. Also the limit function has evidently non-negative imaginary part. □

This is one of the main reasons we include real constants to Pick functions, although they are exceptional in many ways. Note that for any $z \in \mathbb{H}_+$ we have $\log(z) = \lim_{p\to 0^+}(z^p-1)/p$: log can be understood as a limit of Pick functions. There's actually a considerable strengthening of the previous result.

Proposition 5.2. *If $(\varphi_i)_{i=1}^n$ is a sequence of Pick functions converging pointwise, the limit function is also a Pick function.*

We will not prove this quite surprising result yet.

5.2 Schur functions

As we have noticed, Pick functions need not be injections or surjections. Some are both: simple examples are functions of the form $\alpha z+\beta$ and $\frac{\alpha}{\lambda - z}+\beta$ for $\alpha > 0$ and $\beta, \lambda \in \mathbb{R}$. And that's all.

Before trying to understand why is that, we have to change the point of view. All the previous functions are rational functions, but even more is true: they are all Möbius transformations. Möbius transformations are analytic bijections of extended complex plane i.e. Riemann sphere, to itself. Our examples all exactly those Möbius transformation which map the extended real axis to itself, and don't change the orientation, so the upper half-plane is mapped to itself and not to the lower half-plane. When viewed as a part of the Riemann sphere, upper half-plane is just a hemisphere. Of course it shouldn't matter too much which hemisphere we are looking at, so we could also consider mappings from unit disc to itself (or closed unit disc, to be precise). These mappings are called *Schur*

functions and class of Schur functions is denoted by $\mathcal{S}$. It's then natural to conjecture that bijective Schur functions are exactly the Möbius transformations which map unit circle to unit circle, and don't change the orientation so that the inside is mapped to the inside.

These claims are easily derivable from each other as follows. Consider the pair of Möbius transformations

$$\xi : \mathbb{D} \to \mathbb{H}_+ \qquad \xi(z) = i\frac{1-z}{1+z}$$
$$\eta : \mathbb{H}_+ \to \mathbb{D} \qquad \eta(z) = \frac{i-z}{i+z}.$$

They are inverses of each other and map the (open) unit disc to upper half-plane and back, respectively. Now take any bijective Schur function $\psi : \mathbb{D} \to \mathbb{D}$. Then $\varphi = \xi \circ \psi \circ \eta$ is bijective Pick function. Similarly one could invert $\psi = \eta \circ \varphi \circ \xi$. This means that bijections can be paired: if all bijective Pick functions are Möbius transformations, so are all bijective Schur functions, since non-Möbiusness on one side would give rise to non-Möbiusness on the other side.

Still before proving anything we should think about this relation a bit further. We noticed that every bijective Pick function has a corresponding Schur function pair. This correspondence is however by no means unique, it was merely our choice to choose such ξ and η. Still, there is need to restrict ourselves to bijections anymore. If one takes **any** Schur function $\psi : \mathbb{D} \to \mathbb{D}$ we can form the corresponding Pick function by taking $\varphi = \xi \circ \psi \circ \eta$. This gives rise to bijection $\mathcal{S} \to \mathcal{P}$, and the inverse should be rather obvious by now. Of course, it's not a big surprise that there would be such bijection, that is to say that the sets are equal in size, but our bijection preserves composition of functions. All this is to say that in some sense these classes are almost the same.

One should be a bit more careful here though: we have included also real constant functions to our class $\mathcal{P}$ and we should also add unimodular constants to $\mathcal{S}$. For these the bijection doesn't quite work; we can mostly do a natural extension, but then one would be forced to map the constant function -1 to the constant ∞. This means that we should add the constant infinity function to our Pick functions. We will not do this, as it would change the whole business to Riemann sphere, since it will bring other technical problems, but we will try to indicate when you should think about this extension.

If one only thinks about composition one can of course do lot more. Take any simply connected domain in $U \subset \mathbb{C}$. By Riemann mapping theorem there's a analytic bijection $\xi_U \mathbb{D} \to U$. For the domain U we could define similar class of functions, and via ξ_U and it's inverse we could connect the classes. Again, one should be a bit careful with the boundary.

In many ways Pick and Schur functions are most natural of these classes: they are

closed under addition and multiplication, respectively. Also, they both contain the identity of the respective operations, so these properties are barely true.

5.3 Schwarz lemma

Chapter 6

Monotone and Convex matrix functions

We already introduced monotone and convex matrix functions in the introduction, but now that we have properly defined and discussed underlying structures we should take a deeper look. As mentioned, monotone and convex matrix functions are sort of generalizations for the standard properties of reals, and this is why we should undestand which of the phenomena for the real functions carry to matrix functions and which do not.

We will start with the matrix monotone functions; much of the discussion carries quite directly to the convex case.

6.1 Basic properties of the matrix monotone functions

We first state the definition.

Definition 6.1. Let $(a, b) \subset \mathbb{R}$ be an open, possibly unbounded interval and n positive integer. We say that $f : (a, b) \to \mathbb{R}$ is n-monotone or matrix monotone of order n, if for any $A, B \in \mathcal{H}_{(a,b)}$, such that $A \leq B$ we have $f(A) \leq f(B)$.

We will denote the space of n-monotone functions on open interval (a, b) by $P_n(a, b)$. One immediately sees that that all the matrix monotone functions are monotone as real functions.

Proposition 6.2. *If $f \in P_n(a, b)$, f is increasing.*

Proof. Take any $a < x \leq y < b$. Now for $xI, yI \in \mathcal{H}^n_{(a,b)}$ we have $xI \leq yI$ so by definition

$$f(x)I = f(xI) \leq f(yI) = f(y)I,$$

from which it follows that $f(x) \leq f(y)$. This is what we wanted. □

Actually, increasing functions have simple and expected role in n-monotone matrices.

Proposition 6.3. *Let (a, b) be an open interval and $f : (a, b) \to \mathbb{R}$. Then the following are equivalent:*

(i) f is increasing.

(ii) $f \in P_1(a, b)$.

(iii) For any positive integer n and commuting $A, B \in \mathcal{H}^n_{(a,b)}$ such that $A \leq B$ we have $f(A) \leq f(B)$.

Proof. TODO □

The equivalence of the first two is almost obvious and from this point on we shall identify 1-monotone and increasing functions. But the third point is very important: it is exactly the non-commutative nature which makes the classes of higher order interesting.

Let us then have some examples.

Proposition 6.4. *For any positive integer n, open interval (a, b) and $\alpha, \beta \in \mathbb{R}$ such that $\alpha \geq 0$ we have that $(x \mapsto \alpha x + \beta) \in P_n(a, b)$.*

Proof. Assume that for $A, B \in \mathcal{H}_{(a,b)}$ we have $A \leq B$. Now

$$f(B) - f(A) = (\alpha B + \beta I) - (\alpha A + \beta I) = \alpha(B - A).$$

Since by assumption $B - A \geq$ and $\alpha \geq 0$, also $\alpha(B - A) \geq 0$, so by definition $f(B) \geq f(A)$. This is exactly what we wanted. □

That was easy. It's not very easy to come up with other examples, though. Most of the common monotone functions fail to be matrix monotone. Let's try some non-examples.

Proposition 6.5. *Function $(x \mapsto x^2)$ is not n-monotone for any $n \geq 2$ and any open interval $(a, b) \subset \mathbb{R}$.*

Proof. Let us first think what goes wrong with the standard proof for the case $n = 1$.

Note that if $A \leq B$,

$$B^2 - A^2 = (B - A)(B + A)$$

is positive as a product of two positive matrices (real numbers).

There are two fatal flaws here when $n > 1$.

- $(B - A)(B + A) = B^2 - A^2 + (BA - AB)$, not $B^2 - A^2$.

- Product of two positive matrices need not be positive.

Note that both of these objections result from the non-commutativity and indeed, both would be fixed should A and B commute.

Let's write $B = A + H$ $(H \geq 0)$. Now we are to investigate

$$(A + H)^2 - A^2 = AH + HA + H^2.$$

Note that $H^2 \geq 0$, but as we have seen in TODO, $AH + HA$ need not be positive! Also, if H is small enough, H^2 is negligible compared to $AH + HA$. We are ready to formulate our proof strategy: find $A \in \mathcal{H}^n_{a,b}$ and $\mathbb{H}^n_+$ such that $AH + HA \ngeq 0$. Then choose parameter $t > 0$ so small that $A + tH \in \mathcal{H}^n(a, b)$ and

$$(A + tH)^2 - A^2 = t(AH + HA + tH^2) \ngeq 0$$

and set the pair $(A, A + tH)$ as the counterexample.

TODO □

In a similar manner one could show the similar statement for the functions $(x \mapsto x^k)$.

At this point several other important properties of the matrix monotone functions should be clear.

Proposition 6.6. *For any positive integer n and open interval (a, b) the set $P_n(a, b)$ is a convex cone, i.e. it is closed under taking summation and multiplication by non-negative scalars.*

Proof. This is easy: closedness under summation and scalar multiplication with non-negative scalars correspond exaclty to the same property of positive matrices. □

We should be a bit careful though. As we saw with the square function example, product of two n-monotone functions need not be n-monotone in general, even if they are both positive functions; similar statement holds for increasing functions. Similarly, taking maximums doesn't preserve monotonicity.

Proposition 6.7. *Maximum of two n-monotone functions need not be n-monotone for $n \geq 2$.*

Proof. Again, let's think what goes wrong with the standard proof for $n = 1$.

Fix open interval (a, b), positive integer $n \geq 2$ and two functions $f, g \in P^n(a, b)$. Take any two $A, B \in \mathcal{H}^n_{(a,b)}$ with $A \leq B$. Now $f(A) \leq f(B) \leq \max(f, g)(B)$ and Now $f(A) \leq f(B) \leq \max(f, g)(B)$. It follows that

$$\max(f, g)(A) = \max(f(A), g(A)) \leq \max(f, g)(B),$$

as we wanted.

Here the flaw is in the expression $\max(f(A), g(A))$: what is maximum of two matrices? This is an interesting question and we will come back to it a bit later, but it turns out that however you try to define it, you can't satisfy the above inequality.

We still need proper counterexamples though. Let's try $f \equiv 0$ and $g = \mathrm{id}$. So far the only n-monotone functions we know are affine functions so that's essentially our only hope for counterexamples.

TODO □

Similarly we have composition and pointwise limits.

Proposition 6.8. *If $f : (a, b) \to (c, d)$ and $g : (c, d) \to \mathbb{R}$ are n-monotone, so is $g \circ f : (a, b) \to \mathbb{R}$.*

Proof. Fix any $A, B \in \mathcal{H}^n_{(a,b)}$ with $A \leq B$. By assumption $f(A) \leq f(B)$ and $f(A), f(B) \in \mathcal{H}^n_{(c,d)}$ so again by assumption, $g(f(A)) \leq g(f(B))$, our claim. □

Proposition 6.9. *If n-monotone functions $f_i : (a, b) \to \mathbb{R}$ converge pointwise to $f : (a, b) \to \mathbb{R}$ as $i \to \infty$, also f is n-monotone.*

Proof. As always, fix $A, B \in \mathcal{H}^n_{(a,b)}$ with $A \leq B$. Now by assumption

$$f(B) - f(A) = \lim_{i\to\infty} f_i(B) - \lim_{i\to\infty} f_i(A) = \lim_{i\to\infty} (f_i(B) - f_i(A)) \geq 0,$$

so also $f \in P_n(a, b)$. □

We shall be using especially the previous result a lot.

One of the main properties of the classes of matrix monotone functions has still avoided our discussion, namely the relationship between classes of different orders. We already noticed that matrix monotone functions of all orders all monotonic, or $P_n(a, b) \subset P_1(a, b)$ for any $n \geq 1$. It should not be very surprising that we can make much more precise inclusions.

Proposition 6.10. *For any open interval (a, b) and positive integer n we have $P_{n+1}(a, b) \subset P_n(a, b)$.*

Proof. TODO □

One might ask whether these inclusions are strict. It turns out they are, as long as our interval is not the whole $\mathbb{R}$. We will come back to this.

There are also more trivial inclusions: $P_n(a, b) \subset P_n(c, d)$ for any $(a, b) \supset (c, d)$. More interval, more matrices, more restrictions, less functions. To be precise, we only allowed functions with domain (a, b) to the class $P_n(a, b)$, so maybe one should say instead something like: if $(a, b) \supset (c, d)$ and $f \in P_n(a, b)$, then also $f\big|_{(c,d)} \in P_n(c, d)$. We will try not to worry too much about these technicalities.