Part I : Linear Algebra

2024/25

(pages with a grey background were not treated this year and are not part of the exam ; but the videos still exist if you are interested)

Groups and fields

Definition of a group

Def A set G of elements with an operation $+ : G \times G \to G$ is called a group if the following properties hold:

(G1) Associativity: $\forall a, b, c \in G: \quad (a+b)+c = a+(b+c)$

(G2) Identity element: $\exists e \in G \ \forall g \in G: \quad e+g = g+e = g$

(G3) Inverse elements: $\forall a \in G \ \exists b \in G: \quad a+b = b+a = e$

The group is called a commutative group (Abelian group) if we have additionally that

(G4) $\forall a, b \in G: \quad a+b = b+a$

Examples of groups

- $(\mathbb{R}^n, +)$, $(\mathbb{R}^+, \cdot)$ are groups

- $(\mathbb{R}^-, \cdot)$ is <u>not</u> a group.

- $S_n := \{\pi : \{1, \dots, n\} \to \{1, \dots, n\} \mid \pi \text{ is bijective}\}$

  $\circ : S_n \times S_n \to S_n$, $\quad \pi_1 \circ \pi_2 (i) = \pi_1 (\pi_2 (i))$

  $(S_n, \circ)$ is a group.

Definition of a field

Def A set F with two operations $+ , \cdot : F \times F \to F$ is called a field if the following properties hold:

(F1) $(F, +)$ is a commutative group, with identity element 0.

(F2) $(F \setminus \{0\}, \cdot)$ is a commutative group with id. el. 1

(F3) Distributivity: $\forall a, b, c \in F : \quad a \cdot (b+c) = a \cdot b + a \cdot c$

Examples of fields

- $(\mathbb{R}, +, \cdot)$
- $(\mathbb{C}, +, \cdot)$
- $n \in \mathbb{Z}$, Consider $\mathbb{Z}_n := \{0, 1, \dots, n-1\}$

  $a +_n b := (a+b) \bmod n$

  $a \cdot_n b := (a \cdot b) \bmod n$

  Then $(\mathbb{Z}_n, +_n, \cdot_n)$ is a field if and only if n is prime.

Complex numbers

2024: Note: I realized in hindsight that it would have been good to introduce complex numbers. I now prepared few slides, but they are **NOT** part of the exam!

Motivation

In machine learning, our data is often represented by real numbers: $\mathbb{R}$

In linear algebra, however, it often helps if we extend the real numbers to complex numbers: $\mathbb{C}$

We are not going to use a lot of maths of complex numbers, but at least need them to factorize polynomials.

Here are the very basics:

Quadratic equations over $\mathbb{R}$

Quadratic equation: for given parameters $a, b, c \in \mathbb{R}$, want to find $x \in \mathbb{R}$ that satisfies the quadratic equation:

$$ax^2 + bx + c = 0$$

In school you learned the formula:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

It has solutions in $\mathbb{R}$ if $b^2 - 4ac \geq 0$, otherwise it doesn't.

Annoying!

Imagine ...

Imagine that $\sqrt{-1}$ exists, give it a name: $i := \sqrt{-1}$
(i for "imaginary number).

__Def__ A complex number is a number of the form $a+bi$ where $a, b \in \mathbb{R}$. We call a the real part and b the imaginary part of the number. Write $\mathbb{C}$ for the space of all such numbers:

$$\mathbb{C} = \{a + ib \mid a, b \in \mathbb{R}\}.$$

__Observe__: $\mathbb{C}$ is a field.

Quadratic equations over $\mathbb{C}$

Consider the quadratic equation again: $ax^2+bx+c=0$.

Observe that it now always has a solution in $\mathbb{C}$:

$$x_{1/2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

Case 1: $b^2-4ac \geq 0$. Then $x_{1,2} \in \mathbb{R}$ "as usual".

Case 2: $b^2-4ac < 0$. Then can write

$$\sqrt{\underbrace{b^2-4ac}_{<0}} = \sqrt{-1\,\underbrace{(4ac-b^2)}_{>0}} = \underbrace{\sqrt{-1}}_{i} \cdot \underbrace{\sqrt{4ac-b^2}}_{>0}$$

$$= i \cdot r \quad \text{where } r = \sqrt{4ac-b^2} \in \mathbb{R}.$$

Fundamental theorem of algebra

Theorem :

Consider a polynomial $a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$.

(where the a_i can be real or complex numbers).

If $a_n \neq 0$, this is a polynomial of degree n.

Any such polynomial has exactly roots $r_1, \ldots, r_n \in \mathbb{C}$ (not necessarily distinct) such that

$$a_0 + a_1 x + \ldots + a_n x^n = a_n (x - r_1)(x - r_2) \ldots (x - r_n)$$

Outlook

Dealing with complex numbers is a rich field in mathematics,

but we will not touch it ...

Vector space

Definition of a vector space

ML keyword: data representation

<u>Def</u> Let F be a field with id. elements 0 and 1.
A vector space over the field F is a ~~the~~ set V with a mapping $+ : V \times V \to V$ ("vector addition") and a mapping
$\cdot : F \times V \to V$ ("scalar multiplication") such that:

(V1) $(V, +)$ is a commutative group.

(V2) Multiplicative identity: $\forall v \in V : \ 1 \cdot v = v$

(V3) Distributive properties: $\forall a, b \in F \quad \forall u, v \in V$

$$a \cdot (u + v) = a \cdot u + a \cdot v$$
$$(a + b)\, u = a \cdot u + b \cdot u$$

Elements of V are called vectors, elements of F are called scalars.

Examples of vector spaces

- $\mathbb{R}^n$ with the standard operations.

- Function spaces:

  - $\mathbb{R}^{\mathcal{X}} := \{ f : \mathcal{X} \to \mathbb{R} \}$ the space of all real valued fcts on a set $\mathcal{X}$. Define:

    $+ : \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}, \quad (f+g)(x) := f(x) + g(x)$

    $\cdot : \mathbb{R} \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}, \quad (\lambda \cdot f)(x) := \lambda \cdot (f(x))$

    Then $(\mathbb{R}^{\mathcal{X}}, +, \cdot)$ is a real vector space.

  - $\mathcal{C}(\mathcal{X}) := \{ f : \mathcal{X} \to \mathbb{R} \mid f \text{ is continuous} \}$

  - $\mathcal{C}^r([a,b]) = \{ f : [a,b] \to \mathbb{R} \mid f \text{ is } r \text{ times cont. differentiable} \}$

Subspaces

Def Let V be a vector space, $U \subset V$ non-empty set. We call U a subspace of V if it is closed under linear combinations: $\forall \lambda, \mu \in F \ \forall u, v \in U: \ \lambda u + \mu \cdot v \in U$

Examples:

- $\mathcal{C}(X)$ is a subspace of $\mathbb{R}^X$.
- The set S of symmetric matrices of size $n \times n$ is a subspace of $\mathbb{R}^{n \times n}$.

Basis and dimension

Linear combinations

<u>Def</u> V vector space over F, $u_1, \dots, u_n \in V$, $\lambda_1, \dots, \lambda_n \in F$. Then $\sum_{i=1}^{n} \lambda_i u_i$ is called a **linear combination**. The set of all lin. comb. of $u_1, \dots, u_n$ is called the **span** (linear hull) of $u_1, \dots, u_n$.

Notation:

$$\operatorname{span}(u_1, \dots, u_n) := \left\{ \sum_{i=1}^{n} \lambda_i u_i \;\middle|\; \lambda_i \in F \right\}.$$

The set $U := \{u_1, \dots, u_n\}$ is the **generator of** span (U).

Linear independence

Def A set of vectors $v_1, \dots, v_n$ is called **linearly independent** if the following holds:

$$\sum_{i=1}^{n} \lambda_i v_i = 0 \Rightarrow \lambda_1 = \dots = \lambda_n = 0 .$$

Examples:

- The vectors $\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}2\\1\\0\end{pmatrix}, \begin{pmatrix}3\\4\\1\end{pmatrix} \in \mathbb{R}^3$ are lin. indep.
- The functions $\sin(x)$ and $\cos(x) \in \mathbb{R}^{\mathbb{R}}$ are lin. ind.
- Any set of $d+1$ vectors in $\mathbb{R}^d$ is lin. dependent.
- Any set that contains the 0-vector is not lin. indep.

Basis of a vector space

Def A subset B of a vectorspace V is called a (Hamel) basis if

(B1) $\operatorname{Span}(B) = V$

(B2) B is lin. independent.

Example

- The canonical basis of $\mathbb{R}^3$ is $\begin{pmatrix}1\\3\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}$

- Another basis of $\mathbb{R}^3$ is given by $\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}$ or $\begin{pmatrix}0.5\\0.8\\0.4\end{pmatrix}, \begin{pmatrix}1.8\\0.3\\0.3\end{pmatrix}, \begin{pmatrix}-2.2\\-1.3\\3.5\end{pmatrix}$

Reducing a set to a basis

<u>Proposition</u>: If $\mathcal{U} = \{u_1, \dots, u_n\}$ spans a VS V, then the set $\mathcal{U}$ can be reduced to a basis of V.

ML keyword: dictionary learning

Proof sketch

- If U is already lin. independent: done.
- If U is lin. dependent:

$\Rightarrow \exists\, \lambda_1, \dots, \lambda_n$ not all 0 such that $\sum \lambda_i u_i = 0$.

Pick some k with $\lambda_k \neq 0$. Then:

$$u_k = \sum_{i \neq k} \frac{\lambda_i}{\lambda_k} u_k$$

Denote $\tilde{U} := U \setminus \{u_k\}$. It is clear that $\operatorname{span}(\tilde{U}) = \operatorname{span}(U)$

If $\tilde{U}$ is not lin. independent, we repeat this process until the remaining set is lin. independent.

Note that this will eventually be the case, at latest if the set only consists of one vector. ∎

Finite-dim vector space

<u>Def</u> A VS is called finite-dim if it has a finite basis.

We need to do a bit more work to define the dim of a VS.

Can you come up with an example of an infinite space?

Extending a set to a basis

Prop Let $U = \{u_1, \dots, u_n\} \subset V$ be a set of lin. ind. vectors and let V be a finite-dim VS. Then U can be extended to a basis of V.

Proof (Sketch) Let $w_1, \dots, w_m$ be a basis of V. Consider the set

$\{u_1, \dots, u_n, w_1, \dots, w_m\}$. Remove vectors "from the end" until the remaining vectors are lin. independent.

- remaining set spans V
- remaining set is linearly ind. by construction
- remaining set contains U.

$\blacksquare$

Two finite bases have the same length

Corollary Let V be a finite-dim VS. Then any two bases of V have the same length.

Proof sketch: Let $\{\underline{b_1,..,b_n}\} =: B$ and $\{\underline{c_1,\dots,c_m}\} =: C$ be two bases, $n \leq m$.

Consider the set $\{\underline{b_1,\dots,b_n},\underline{c_1}\}$. By construction is lin. dependent. So by the same procedure as before, we can find a vector b_i such that $\{\underline{b_1,..,b_{i-1}},\underline{b_{i+1}\dots,b_n},\underline{c_1}\}$ is ind. Keep on applying this procedure: add vectors from C, remove vectors from B. At the end, this results in a set of n vectors $\{c_{i_1},\dots,c_{i_n}\} \subset \{c_1,\dots,c_m\}$. By construction, they are lin. ind and span V. If now we had $m > n$, then the set $\{c_1,\dots,c_m\}$ cannot be lin. ind. any more.

Dimension of a vector space

<u>Def</u> The length of a basis of a finite-dim VS is called the <u>dimension</u> of V.

Linear Mappings

Linear mapping

<u>Def</u> Let U, V VS over F. A mapping $f: U \to V$ is called linear if $\forall u_1, u_2 \in U$, $\forall \lambda \in F$

$$f(u_1 + u_2) = f(u_1) + f(u_2)$$

$$f(\lambda u_1) = \lambda f(u_1)$$

The set of all linear mappings from $U \to V$ is denoted $\mathcal{L}(U, V)$. If $U = V$, then we write $\mathcal{L}(U)$.

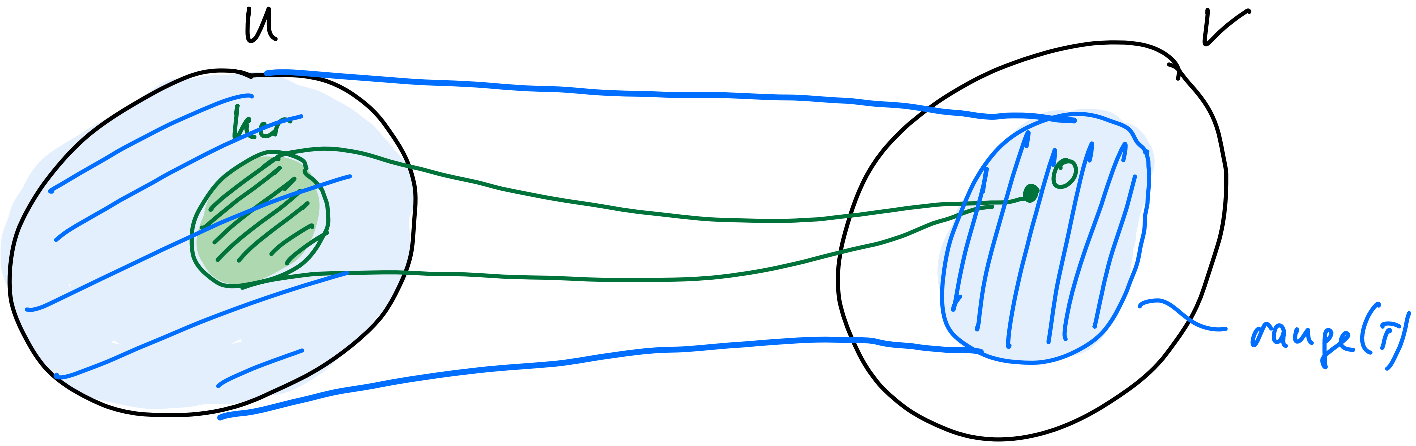
<u>Examples</u>:
- $T: \mathcal{C}[a,b] \to \mathbb{R}$, $f \mapsto \int_a^b f(x)\,dx$ (integration)
- $D: \mathcal{C}^\infty[a,b] \to \mathcal{C}^\infty[a,b]$, $f \mapsto f'$ (Differentiation)

Def $T \in \mathcal{L}(U,V)$. Then **kernel** of T (**null space**) is defined as

$$\ker(T) := \operatorname{null}(T) := \{u \in U \mid Tu = 0\} \subset U$$

The **range of T** (**image of T**) is defined as

$$\operatorname{range}(T) := \operatorname{Im}(T) := \{Tu \mid u \in U\} \subset V$$



Properties of kernel and range

Proposition:

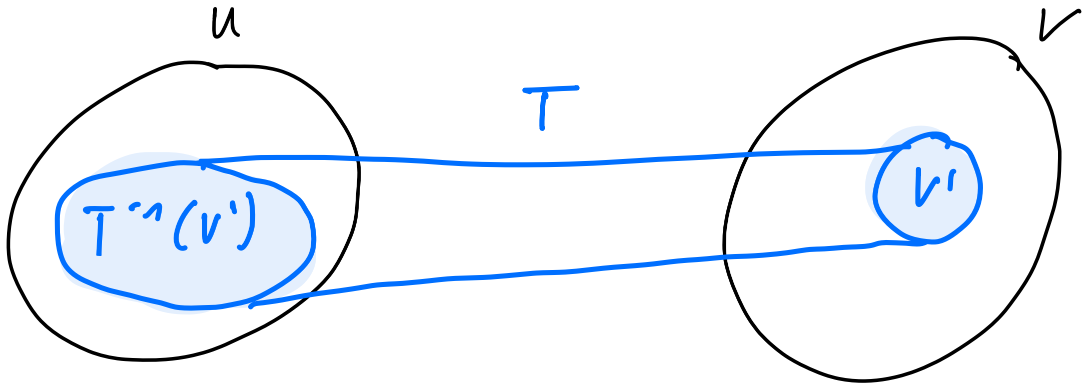
- ker (T) and range (T) are subspaces.
- T is injective $\Leftrightarrow$ $\ker T = \{0\}$
- T is surjective $\Leftrightarrow$ range $T = V$

Proof: Exercise.

Pre-image

<u>Def</u> : $V' \subset V$, V' any set. The <u>pre-image</u> of V' is defined as

$$T^{-1}(V') := \{ u \in U \mid Tu \in V' \}.$$



<u>Prop</u> : If $V' \subset V$ is a subspace of V, then $T^{-1}(V')$ is a subspace of U.

<u>Proof</u> : exercise!

Fundamental theorem for linear mappings

<u>Theorem</u>: Let V be finite-dim, W any VS, $T \in \mathcal{L}(V, W)$.

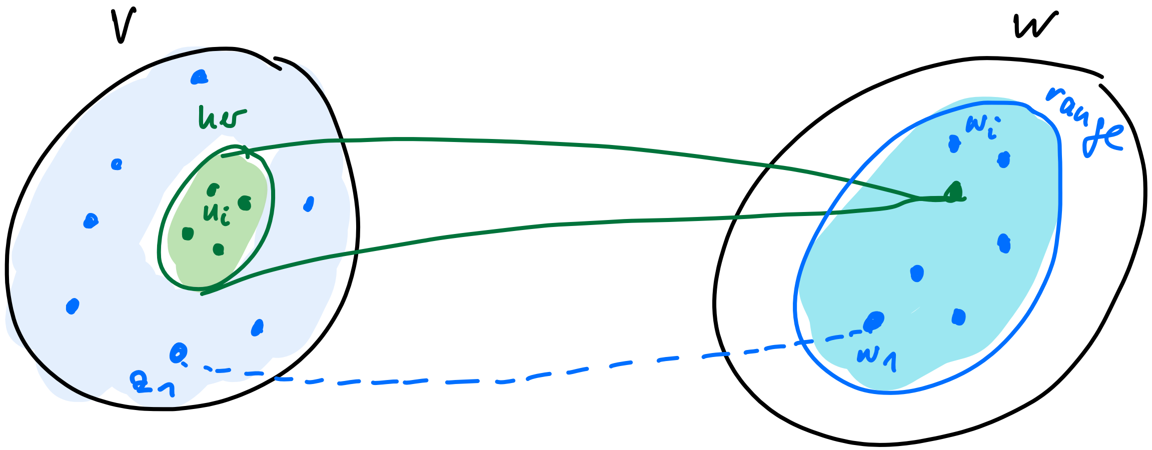
Let $u_1, \dots, u_n$ be a basis of $\ker(T) \subset V$

Let $w_1, \dots, w_m$ be a basis of $\operatorname{range}(T) \subset W$.

Let $z_1 \in T^{-1}(w_1), \dots, z_m \in T^{-1}(w_m)$. Then

Then $u_1, \dots, u_n, z_1, \dots, z_m \subset V$ forms a basis of V.

In particular, $\dim(V) = \dim(\ker(T)) + \dim(\operatorname{range}(T))$.



Proof of the theorem

$\underline{\text{Step 1}}$: $V \subset \text{span}\{u_1, \dots, u_n, z_1, \dots, z_m\}$

Let $v \in V$, consider $Tv \in \text{range}(T)$.

$\Rightarrow \exists\, \lambda_1, \dots, \lambda_m$ s.t.

$$\underline{Tv} = \lambda_1 w_1 + \dots + \lambda_m w_m$$

$$= \lambda_1 T(z_1) + \dots + \lambda_m T(z_m)$$

$$= \underline{T(\lambda_1 z_1 + \dots + \lambda_m z_m)}$$

$$\Rightarrow \quad Tv - T(\lambda_1 z_1 + \dots + \lambda_m z_m) = 0$$

$$= T(\underbrace{v - (\lambda_1 z_1 + \dots + \lambda_m z_m)}_{\in \ker(T)})$$

$$\Rightarrow \quad \exists\, \mu_1, \dots, \mu_n \text{ s.t. } \quad v - (\lambda_1 z_1 + \dots + \lambda_m z_m) = \mu_1 u_1 + \dots + \mu_n u_n$$

$$\Rightarrow \quad v = \lambda_1 z_1 + \dots + \lambda_m z_m + \mu_1 u_1 + \dots + \mu_n u_n$$

$$\boxed{\text{Step 2}: u_1,\dots,u_n, z_1,\dots,z_m \text{ are lin. indep.}}$$

Assume that

$$\mu_1 u_1 + \dots + \mu_n u_n + \lambda_1 z_1 + \dots + \lambda_m z_m = 0 \quad \circledast$$

$$\lambda_1 w_1 + \dots + \lambda_m w_m = \lambda_1 T(z_1) + \dots + \lambda_m T(z_m)$$

we add something that is 0 (because in kernel)

$$= \lambda_1 T(z_1) + \dots + \lambda_m T(z_m) + \underbrace{\mu_1 T(u_1) + \dots + \mu_n T(u_n)}_{=0}$$

exploit linearity

$$= T(\underbrace{\lambda_1 z_1 + \dots + \lambda_m z_m + \mu_1 u_1 + \dots + \mu_n u_n}_{=0 \text{ by } \circledast}) = 0$$

$$\Rightarrow \lambda_1 w_1 + \dots + \lambda_m w_m = 0 \quad \xRightarrow[\text{basis}]{w_1,\dots,w_m} \quad \lambda_1 = \dots = \lambda_m = 0$$

$$\Rightarrow \mu_1 u_1 + \dots + \mu_n u_n = 0 \quad \text{by } \circledast$$

$$\Rightarrow \mu_1 = \dots = \mu_n = 0 \quad \text{because } u_1,\dots,u_n \text{ basis.}$$

∎

Injective $\Leftrightarrow$ surjective $\Leftrightarrow$ bijective

<u>Prop</u> $T \in \mathcal{L}(V,V)$, V finite-dim. Then the following three statements are equivalent:

(i) T injective.

(ii) T surjective.

(iii) T bijective.

<u>Proof:</u> Direct consequence of theorem. (can you see why? Exercise!)

⚠ Does not hold in ∞-dim spaces!

Matrices and linear maps

Matrices

A matrix is the following object:

$$A = \begin{matrix} m \\ \text{rows} \end{matrix} \left\{ \overbrace{\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ & & \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}}^{n \text{ col.}} \right. = (a_{ij})_{\substack{i=1\dots m \\ j=1\dots n}}$$

Matrices represent linear maps

Consider $T \in \mathcal{L}(V, W)$, V, W finite-dim,

let $v_1, \dots, v_n$ be a basis of V

$w_1, \dots, w_m$ basis of W

- If we know the results of T applied to the basis vectors v_i, then we can express $T(v)$ for arbitrary v:

$v = \lambda_1 v_1 + \dots + \lambda_n v_n$ arbitrary vector

$$T(v) = T(\lambda_1 v_1 + \dots + \lambda_n v_n)$$

$$= \lambda_1 T(v_1) + \dots + \lambda_n T(v_n)$$

- For basis vector v_j, we can express the image $T(v_j)$ in basis $w_1, \dots, w_m$:

  there exist coefficients $a_{1j}, \dots, a_{mj}$ s.t.

$$\boxed{T(v_j) = a_{1j} w_1 + \dots + a_{mj} w_m}$$

- We now stack these coefficients in a matrix called $M(T)$:

$$\begin{array}{c} m \text{ rows,} \\ \text{one for each} \\ \text{basis vector} \\ \text{of } W \end{array} \quad M := \underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix}}_{\substack{\text{col } j \\ \\ n \text{ cols, one for each} \\ \text{basis vector of } V}} = \begin{array}{l} \text{matrix of mapping } T \\ \text{with respect to} \\ \text{the bases} \\ v_1, \dots, v_n \text{ of } V \\ w_1 \dots w_m \text{ of } W. \end{array}$$

- The result of $T(v)$ can now be expressed by a matrix-vector multiplication:

$$T(v) = \sum_{j=1}^{n} \lambda_j \, T(v_j)$$

$$= \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{m} a_{ij} \, w_i$$

$$= \sum_{i=1}^{m} \underbrace{\left(\sum_{j=1}^{n} a_{ij} \lambda_j \right)}_{\left(M \cdot \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \right)_i} w_i$$

i-th entry of product of matrix M with vector λ

Notation for matrices of linear maps

<u>Notation</u> Let $T : V \to W$ be linear, let $\mathcal{B}$ a basis of V, $\mathcal{C}$ basis of W. We denote by

$$M(T, \mathcal{B}, \mathcal{C})$$

the matrix corresponding to T wrt bases $\mathcal{B}$ and $\mathcal{C}$.

Properties of matrices

V, W vector spaces, consider the basis fixed. Let $S, T \in \mathcal{L}(V, W)$.

Then:

- $M(S+T) = M(S) + M(T)$
- $M(\lambda S) = \lambda M(S)$
- $T: U \to V$, $S: V \to W$ linear, then
  $M(S \circ T) = M(S) \cdot M(T)$

Matrix transpose

Def Given a matrix $A = (a_{ij})_{ij} \in F^{m \times n}$, the **transpose matrix** is given as

$$(A^t)_{kj} = A_{jk}$$

Notation: A^t, A'

If $F = \mathbb{C}$, then the **conjugate transpose** matrix is defined as

$$(A^*)_{ij} = \overline{a}_{ji}$$

Sum of vector spaces

Def Assume that we have U_1, U_2 subspaces of V. The sum of the two spaces is defined as

$$U_1 + U_2 := \{ u_1 + u_2 \mid u_1 \in U_1, u_2 \in U_2 \}$$

The sum is called a direct sum, if each element in the sum can be written in exactly one way.

Notation: $U_1 \oplus U_2$

Complement of a subspace

Prop Suppose V is finite-dim, and $U \subset V$ is a subspace. Then there exists a subspace $W \subset V$ such that $U \oplus W = V$.

Proof (sketch) Let the set $\{u_1, \dots, u_k\}$ basis of of U. Extend it to a basis of V, say the resulting set is

$$\{\underbrace{u_1, \dots, u_k}_{\rightsquigarrow U}, \underbrace{v_1, \dots, v_m}_{\rightsquigarrow W}\}.$$

Define

$$W = \operatorname{span}\{v_1, \dots, v_m\}.$$

$\square$

Invertible maps and matrices

Inverse of a linear map

Def $T \in \mathcal{L}(V,W)$ is called invertible if there exists a linear map $S \in \mathcal{L}(W,V)$ such that

$$S \circ T = Id_V \quad \text{and} \quad T \circ S = Id_W$$

The map S is called the inverse of T, denoted by T^{-1}.

Rem
- Inverse maps are unique.
- not every lin. map is invertible

Characterizing invertability

Prop A linear map is invertible iff it is injective and surjective.

Proof "⇒" Invertible ⇒ injective:

suppose $T(u) = T(v)$. Then $\underline{u} = T^{-1}(T(u))$

$= T^{-1}(Tv) = \underline{v}$ ⇒ injective

Invertible ⇒ surjective:

Consider $w \in W$. Then

$w = T(T^{-1}(w))$ ⇒ $w \in$ range of T

⇒ surjective.

"$\Leftarrow$" inj & surj $\Rightarrow$ invertible.

- Let $w \in W$. There exists unique $v \in V$ s.th. $T(v) = w$. (because T surj.)

  Define the mapping: $S(w) = v$. Clearly here $\underline{T \circ S = \mathrm{Id}}$.

- Let $v \in V$. Then $T(\underline{(S \circ T)(v)}) =$

  $$= \underbrace{(T \circ S)}_{\mathrm{Id}}(Tv) = \mathrm{Id} \circ Tv = T\underline{v}.$$ Because T is injective, then

  $$(S \circ T)v = v \quad \Rightarrow \quad \underline{S \circ T = \mathrm{Id}}$$

- Linearity: $T(Sw_1 + Sw_2) = \underbrace{TS}_{\mathrm{Id}}w_1 + \underbrace{TS}_{\mathrm{Id}}w_2 = w_1 + w_2$

$\overset{\text{inj.}}{\Rightarrow}$ $Sw_1 + Sw_2$ is the unique el. in V that T maps to $w_1 + w_2$

$\Rightarrow$ By def of S, we thus get $S(w_1 + w_2) = Sw_1 + Sw_2$

Similarly for scalar multiplication.

$\blacksquare$

Inverse matrix

<u>Def</u> A square matrix $A \in F^{n \times n}$ is invertible if there exists a square matrix $B \in F^{n \times n}$ such that

$$A \cdot B = B \cdot A = Id = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

The matrix B is called the inverse matrix, and is denoted by A^{-1}.

Inverting maps ≙ inverting matrices

Prop The inverse matrix represents the inverse of the corr. lin. maps, that is: $T : V \to V$

$$\underset{\text{matrix of}}{M}\,\big(\underbrace{T^{-1}}_{\text{inverse map}}\big) = \underbrace{\big(M(\underbrace{T}_{\text{(of the original matrix)}})\big)}_{\text{inverse matrix}}^{-1}$$

In particular, a matrix is invertible iff the corr. map is invertible.

Proof: Exercise.

Properties of inverse matrices

- The inverse matrix does <u>not</u> always exist.
- $(A^{-1})^{-1} = A$, $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$
- A^t invertible $\Leftrightarrow$ A invertible,
  $(A^t)^{-1} = (A^{-1})^t$
- $A \in F^{n \times n}$ invertible $\Leftrightarrow$ rank $(A) = n$
- The set of all invertible matrices is called general linear group:

$$GL(n, F) = \{ A \in F^{n \times n} \mid A \text{ invertible} \}$$

Change of basis

Representing the identity

Consider the identity mapping $J : V \to V, x \mapsto x$.
Assume we fix a basis of V (both in source and target space), then the corr. matrix looks as follows:

$$M(J, \mathcal{B}, \mathcal{B}) = \begin{pmatrix} 1 & \cdots & 0 \\ 0 & & 1 \end{pmatrix}$$

Now consider $\mathcal{A} = \{a_1, \ldots, a_n\}$ and $\mathcal{B} = \{b_1, \ldots, b_n\}$ both bases on V. How does the matrix of the id. mapping

$$J : (V, \mathcal{A}) \to (V, \mathcal{B}) \quad \text{look like?}$$

Because $\mathfrak{B}$ is basis, we can write each of the vectors in $\mathfrak{A}$ as lin. comb. of vectors in $\mathfrak{B}$:

$$a_1 = \boxed{t_{11} b_1 + t_{21} b_2 + \dots + t_{n1} b_n}$$

$$a_2 = \dots$$

Now we form the corr. matrix T

$$T = \left(\begin{matrix} \boxed{\begin{matrix} t_{11} \\ \vdots \\ t_{n1} \end{matrix}} & \begin{matrix} \cdots \\ \\ \cdots \end{matrix} & \begin{matrix} t_{1n} \\ \vdots \\ b_{nn} \end{matrix} \end{matrix}\right)$$

This matrix represents the identity mapping:

- In the basis $\mathcal{A}$, the first basis vector a_1 has the representation $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

$a_1 = 1 \cdot a_1 + 0 \cdot a_2 + 0 \cdot a_3 \cdots + 0 \cdot a_n$

- $T \cdot \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix}$ this vector gives us Ta_1 expressed in basis $\mathcal{B}$

by def of coefficients t_{ij}

- $t_{11} b_1 + \dots + t_{n1} b_n = a_1$

- $T a_1 = a_1$.

Change of basis is invertible

Prop Let $\mathcal{A}, \mathcal{B}$ be two bases of V. Then the matrices $M(\mathrm{Id}, \mathcal{A}, \mathcal{B})$ and $M(\mathrm{Id}, \mathcal{B}, \mathcal{A})$ are invertible and each is the inverse of each other.

Proof Exercise / skipped.

Change of basis for an arbitrary mapping

Prop Let $\mathcal{A}, \mathcal{B}$ be two bases of V. Consider the transformation matrix $A = M(\mathrm{Id}, \underline{\mathcal{A}}, \underline{\mathcal{B}})$, and $A^{-1} = M(\mathrm{Id}, \underline{\mathcal{B}}, \underline{\mathcal{A}})$.

Let $T: V \to V$ linear, and $\boxed{X} = M(T, \underline{\mathcal{A}}, \underline{\mathcal{A}})$. Then $\boxed{Y := A \cdot X \cdot A^{-1}}$ represents T in basis $\mathcal{B}$, that is

$Y = M(T, \underline{\mathcal{B}}, \underline{\mathcal{B}})$.

$$\begin{tikzcd}
(V, \mathcal{A}) \arrow[d, "\text{mapping } T"', "\text{matrix } \boxed{X}"] & & (V, \mathcal{B}) \arrow[ll, dashed, "\text{Id map.}"', "\text{matrix } A^{-1}"] \arrow[d, "\text{map. } T"', "\text{matrix } \boxed{Y}"] \\
(V, \mathcal{A}) \arrow[rr, dashed, "\text{Id map}", "\text{matrix } A"'] & & (V, \mathcal{B})
\end{tikzcd}$$

Rank of a matrix

Rank of a matrix

ML keywords:
low rank methods:
Rank describes how "complex"
a matrix is.
Recommender
systems, compressed sensing, ...

Def $A \in F^{m \times n}$. The column rank of A is

$$\dim(\operatorname{span}(\text{column vectors of } A))$$

The row rank is defined accordingly.

Prop For a matrix, the row and column rank always coincide. We now call it the rank of the matrix.

Prop $T \in \mathcal{L}(V, W)$. Then $\operatorname{rank}(M(T)) = \dim(\operatorname{range}(T))$.

Proofs: skipped

The determinant

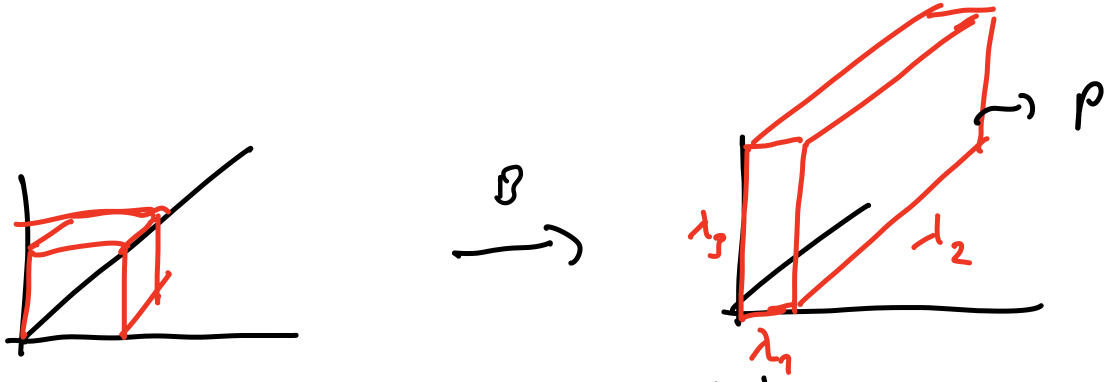
Motivation to study the determinant: geometry!

Consider the standard basis of $\mathbb{R}^3$: $e_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$, $e_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$, $e_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$.

Consider a linear mapping that just streches these vectors: $T = \begin{pmatrix}\lambda_1 & & 0\\ & \lambda_2 & \\ 0 & & \lambda_3\end{pmatrix}$

Let U be the unit cube and $P = T(U)$ its mapping.

The volume of P is then $\lambda_1 \cdot \lambda_2 \cdot \lambda_3$



$\text{vol}(U) = 1$ $\qquad$ $\text{vol}(P) = \lambda_1 \cdot \lambda_2 \cdot \lambda_3$

Want to define a quantity "det" that tells us how volumes change under <u>arbitrary</u> linear mappings.

Which properties would such a mapping d have to satisfy?

- $\mathrm{vol}(U) = 1$, so we would like that $d(I) = 1$. $\rightsquigarrow \boxed{D3}$
- $T = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ with respect to standard basis $\Rightarrow d(T) = \lambda_1 \cdot \ldots \cdot \lambda_n$ $\rightsquigarrow \boxed{D2}$
- If the linear mapping does not have full rank (the image $\mathrm{vol}(U)$ is not "full-dimensional", but for example the cube in $\mathbb{R}^3$ is just mapped to a "plane" in 2 dim), then volume is 0, hence we would like $d(U) = 0$. $\rightsquigarrow \boxed{D2}$

Now let's look at the formal definition:

Definition of the determinant

<u>Def</u> Consider a linear mapping $d: F^{n \times n} \to F$. Then d is called a <u>determinant</u> if:

(D1) d is linear in each column of the matrix:

$$\begin{pmatrix} | & | & \\ a_1 & a_2 & \cdots \\ | & | & \end{pmatrix}$$

Let A be a matrix with columns $a_1, \dots, a_n$.

Consider column a_i, assume $a_i = a_i' + a_i''$ for some $a_i', a_i'' \in F^{n \times 1}$. Then it holds that

- $\det((a_1, \dots, \underline{a_i}, \dots, a_n)) =$

$$\det((a_1, \dots, \underline{a_i'}, \dots, a_n)) + \det((a_1, \dots, \underline{a_i''}, \dots, a_n))$$

- $\det((a_1, \dots, \underline{\lambda a_i}, \dots, a_n)) = \lambda \cdot \det((a_1, \dots, \underline{a_i}, \dots, a_n))$

(D2) d is alternating: if A has two identical columns, then $\det A = 0$.

(D3) d is normed: $\det \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = 1$.

Existence and uniqueness

Theorem : The mapping α exists and is unique.

Proof : Skipped

Properties of the determinant

Based on (D1), (D2), (D3) we can now prove many important properties of the determinant:

- The determinant of an linear mapping does not depend on the basis.
- $\det(c \cdot A) = c^n \det(A)$
- $\det(A \cdot B) = (\det A) \cdot (\det B)$
- $\det(A^t) = \det(A)$
- $\det(A^{-1}) = 1/\det(A)$ (if A is invertible)

→ cont.

- A invertible $\iff$ $\det(A) \neq 0$

- $\det(A+B) \neq \det(A) + \det(B)$

- If A is upper triangular, that is

$$A = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

then $\det A = \lambda_1 \cdot \ldots \cdot \lambda_n$.

Same for lower triangular matrices.

Computing the determinant (in theory)

Special cases:

$n=1$ $\quad \det(a) = a$

$n=2$ $\quad \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}\cdot a_{22} - a_{12}\cdot a_{21}$

$n=3$ $\quad \det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a\cdot\det\begin{pmatrix} e & f \\ h & i \end{pmatrix} - b\cdot\det\begin{pmatrix} d & f \\ g & i \end{pmatrix} + c\cdot\det\begin{pmatrix} d & e \\ g & h \end{pmatrix}$

In general, there exists the formula of Laplace that expresses the determinant of an $n\times n$ matrix as a lin. comb. of det. of many $(n-1)\times(n-1)$-submatrices.

Alternative definition of determinant

There exists a more straight-forward definition of det. However, starting from this definition, proving the geometric properties is more cumbersome.

Def (alternative):

- If V is a vector space over $\mathbb{C}$ and $T: V \to V$, then $\det(T)$ is the product of all eigenvalues (repeated according to multiplicity).
- If V is a vector space over $\mathbb{R}$ and $T: V \to V$, then $\det(T)$ is the product of its eigenvalues over $\mathbb{C}$, repeated according to multiplicity.

We skip the proof that this is the same object as "our" determinant.

Computing the determinant (in practice)

- LU decomposition:

  Any matrix A can be written as a product

$$A = L \cdot U$$

where L is a lower triangular matrix and U upper triangular:

$$L = \begin{pmatrix} l_{11} & & 0 \\ & l_{22} & \\ * & & \ddots \; l_{nn} \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & & * \\ & \ddots & \\ 0 & & u_{nn} \end{pmatrix}$$

- $\det(A) = \det(L \cdot U) = \det(L) \cdot \det(U) =$

$$= \left(\prod_{i=1}^{n} l_{ii} \right) \cdot \left(\prod_{i=1}^{n} u_{ii} \right)$$

Geometric intuition again

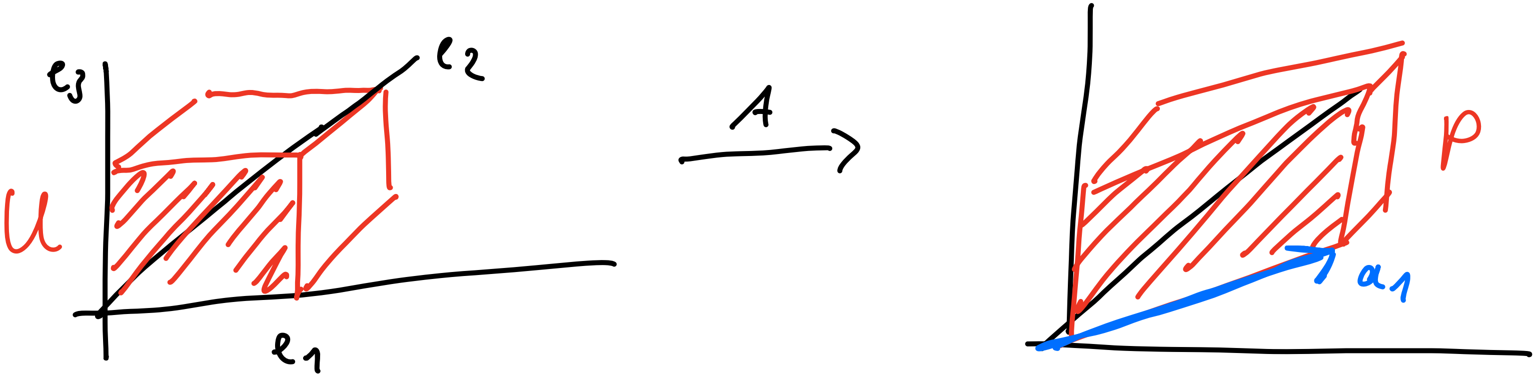
Theorem: Consider an $n \times n$ matrix A with columns $(a_1 | a_2 | \cdots | a_n) = A$.

Consider the unit cube $U = \{c_1 e_1 + \ldots + c_n e_n \mid 0 \leq c_i \leq 1\}$ and its image P under the mapping A:

$$U \longmapsto P := \{ c_1 a_1 + \ldots + c_n a_n \mid 0 \leq c_i \leq 1 \}$$ parallelotope.

Then $\det(A)$ gives us the (signed) volume of P

Proof (for general matrices A!): skipped.



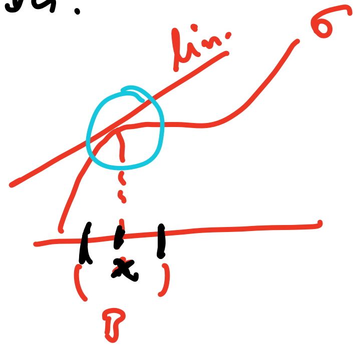
Applications to integrals

Proposition: $\Omega \subset \mathbb{R}^n$ open subset, $\sigma : \Omega \to \mathbb{R}^n$ differentiable, $f : \sigma(\Omega) \to \mathbb{R}$. Then:

$$\int_{\sigma(\Omega)} f(y)\, \underbrace{dy}_{\substack{\text{volume}\\ \text{element}}} = \int_{\Omega} f(\sigma(x))\, \underbrace{\left|\det\left(\overset{\text{derivative, linear}}{\sigma'}(x)\right)\right| dx}_{\text{volume element}}$$

Intuition: σ differentiable, that is we can locally (on a small ball B around x) approximate σ by a linear function.

And volumes thus get stretched by the factor given by the determinant.



Another occurrence of the det in ML

density of the multivariate Gaussian:

$$p(x) = \frac{1}{(2\pi)^{d/2}} \quad \frac{1}{\boxed{\det(\Sigma)^{1/2}}} \quad \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma (x-\mu)\right)$$

normalization term related to the volume, so that density integrates to 1 in the end

Eigenvalues, eigenvectors,
eigenspaces

Eigenvalues

ML keywords: PCA, SVD, spectral clustering, ...

Def Let $T: V \to V$. A scalar $\lambda \in F$ is called an **eigenvalue** if there exists a $v \in V$, $v \neq 0$, such that $Tv = \lambda \cdot v$. A vector $v \neq 0$ with this property is called an **eigenvector** corresponding to eigenvalue λ. The set of all eigenvectors of λ is called the **eigenspace** $E(\lambda, T) = \ker(T - \lambda I)$.

$$\left\{\begin{array}{l} Tv = \lambda v \\ Tv - \lambda v = 0 \\ Tv - \lambda I v = 0 \\ (T - \lambda I) v = 0 \end{array}\right.$$

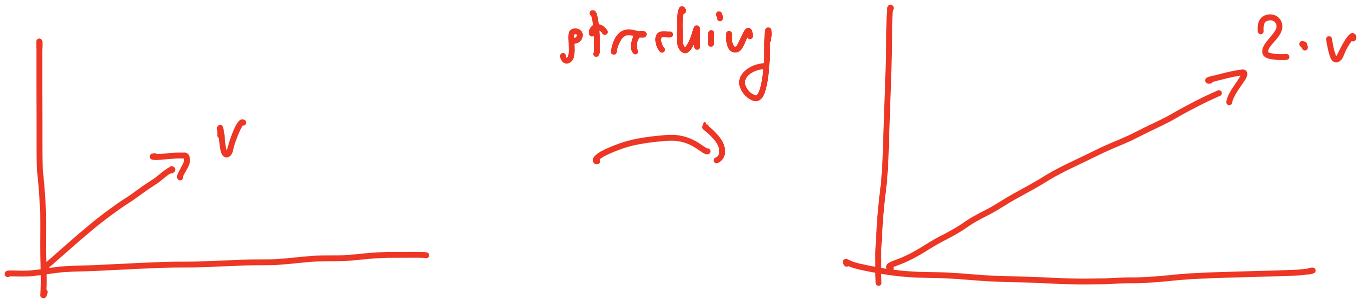
Geometric intuition

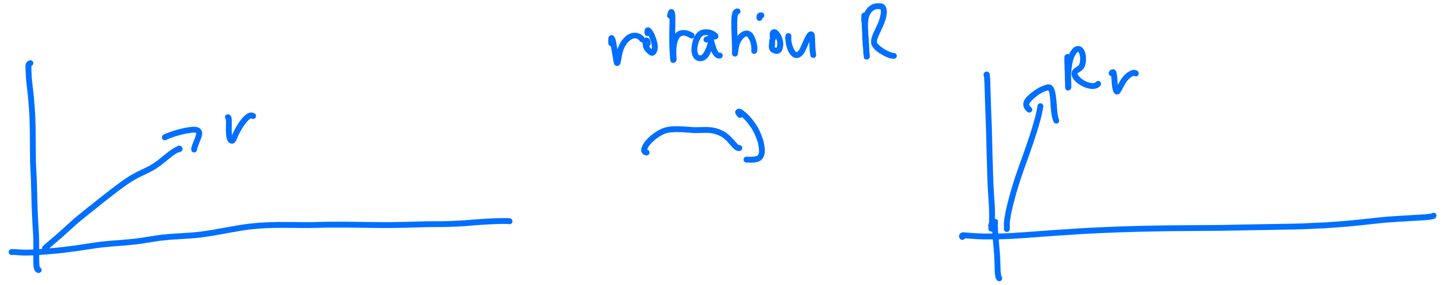
- Eigenvalue/eigenvector realizes a "streching"

$$v \longmapsto \lambda v$$

- Many mappings do not have eigenvectors

for example, a rotation over $\mathbb{R}^2$.





Eigenvectors are not unique

- If λ is an eigenvalue, it has many eigenvectors!

  For example, if v is eigenvector, then also $a \cdot v$ $(a \in K)$ is an eigenvector!

$$T(\underbrace{a \cdot v}_{}) = a \cdot T(v) = a \cdot \lambda \cdot v = \lambda (\underbrace{a \cdot v}_{})$$

Linear (in)dependence of eigenvectors?

- Eigenvectors corr. to <u>distinct</u> eigenvalues are linearly independent. (easy exercise; assume they are dependent and drive a contradiction)

- Eigenvectors that corr. to the same eigenvalue do not need to be independent

  Simple example: v eig. $\Rightarrow$ $c\cdot v$ eig. but v and $c\cdot v$ are not lin. ind.

  They can be lin. independent:

  Easy example: $A = I$, then every vector v is an eigenvector of eigenvalue 1.

- The eigenspace $E(\lambda, T)$ is always a lin. subspace of V.

Theorem: Every operator $T: V \to V$ on a finite-dim, complex (!) vs V has at least one eigenvalue.

Proof

Step 1: Getting started

Let $n = \dim V$. Choose a vector $v \in V$, $v \neq 0$. Then the set

$$v, \; Tv, \; T^2 v, \; \ldots, \; T^n v$$

has to be linearly dependent (it consists of $n+1$ vectors in an n-dim space). Find coefficients $a_0, a_1, \ldots, a_n$ such that

$$p(T) := a_0 v + a_1 Tv + \ldots + a_n T^n v = 0 .$$

Now we want to show that we can factorize this "polynomial of operators":

$$a_0 + a_1 T + a_2 T^2 + \ldots + a_n T^n$$

$$\overset{!}{=} c(T - \lambda_1 I)(T - \lambda_2 I) \ldots (T - \lambda_m I) .$$

Step 2: factorization

Excursion to complex-valued polynomials:

Consider a polynomial on $\mathbb{C}$ with these coefficients:

$$p(z) := a_0 + a_1 \cdot z + \dots + a_n z^n$$

n and m can be different, eg if $a_n = 0$.

Over $\mathbb{C}$, we can factorize it:

$$p(z) = c \cdot (z - \lambda_1)(z - \lambda_2) \dots (z - \lambda_m)$$

Not difficult to see that we can then also factorize $p(T)$:

* below

$$p(T) = a_0 + a_1 T + \dots + a_n T^n$$

$$= c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \dots (T - \lambda_m I) \quad (*)$$

Step 3: putting it together

Hence, $0 = a_0 v + a_1 T v + \dots + a_n T^n v =$

$$= c(T-\lambda_1 I)(T-\lambda_2 I) \dots (T-\lambda_m I)\cdot v$$

$\Rightarrow v \in \ker$ (big operator $\circledast$)

$\Rightarrow$ there must exist $i \in \{0, \dots, m\}$ such that

$(T-\lambda_i I)$ not injective

$\Rightarrow \lambda_i$ is an eigenvalue of T !

∎

⚠ The corresponding eigenvector is not necessarily v!

Reason:

$$c\,(T-\lambda_1 I)\dots \overbrace{(T-\lambda_i I)}^{\text{not injective}} \underbrace{(T-\lambda_{i+1})\dots(T-\lambda_n I)\,v}_{=:\,w} = 0$$

w is the eigenvector!

(we started with <u>any</u> vector v)

∎

⊛ Polynomials of operators

Let $p(x) = \sum_{i=0}^{n} a_i x^i$ be a polynomial.

For a linear operator T, define $p(T) := \sum_{i=0}^{n} a_i T^i$

Then the following properties ensure that "factorization" works:

Let p, q be two polynomials. Then:

- If the polynomial factorizes, so does the related operator:

$$(p \cdot q)(T) = p(T) \cdot q(T)$$

- Order of factors does not matter:

$$p(T) \cdot q(T) = q(T) \cdot p(T).$$

Moreover, if $p(T) \cdot q(T) v = 0 \Rightarrow$

Diagonalization and Triangularization

Diagonalizable matrices

<u>Def</u> An operator $T \in \mathcal{L}(V)$ is **diagonalizable** if there exists a basis $\mathcal{B}$ of V such that the corr. matrix is diagonal:

$$M(T,\mathcal{B},\mathcal{B}) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

⚠ Not always the case, neither over $\mathbb{R}$ nor over $\mathbb{C}$!

Will later see some special cases where it always works (symmetric matrices).

When is a matrix diagonalizable?

Proposition: Let T be an operator on a finite-dim$^{=n}$ vector space V. Let $\lambda_1, \dots, \lambda_m$ denote the distinct eigenvalues of T. Then the following statements are equivalent:

(a) T is diagonalizable.

(b) V has a basis consisting of eigenvectors of T

(c) $\dim V = \dim(\text{eigenspace}(\lambda_1)) + \dots + \dim(\text{eigenspace}(\lambda_m))$.

Remark: Later will see that this is also equivalent to saying:

(d) V has n eigenvalues if counted with multiplicity, and for all of them the algebraic and geometric multiplicities are the same.

Triangular matrices

A matrix is called upper triangular, if it has the form

$$\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Geometric intuition

Prop $T \in \mathcal{L}(V)$, $\mathcal{B} = \{v_1, v_2 \dots, v_n\}$ a basis.

Then equivalent:

(a) $M(T, \mathcal{B})$ is upper triangular.

(b) $Tv_j \in \operatorname{span}\{v_1, \dots, v_j\}$ $\forall j = 1, \dots, n$

Proof idea:

$$Tv_1 = \begin{pmatrix} \lambda_1 & a_{12} & a_{13} \\ & \lambda_2 & a_{23} \\ 0 & & \lambda_3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ 0 \\ 0 \end{pmatrix} = \lambda_1 \cdot v_1$$

$$Tv_2 = \begin{pmatrix} \lambda_1 & a_{12} & a_{13} \\ & \lambda_2 & a_{23} \\ 0 & & \lambda_3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a_{12} \\ \lambda_2 \\ 0 \end{pmatrix} = a_{12} \overbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}^{v_1} + \lambda_2 \overbrace{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}^{v_2}$$

$$\in \operatorname{span}(v_1, v_2)$$

When are triangular matrices invertible?

Proposition: Let $T: V \to V$ have an upper-triangular matrix. Then T is invertible if and only if all entries $\lambda_1, \dots, \lambda_n$ in the diagonal are non-zero.

$$\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Proof

Proof "$\Leftarrow$" : Let $v_1, \dots, v_n$ the basis for which T is upper-triangular.

Assume all $\lambda_i \neq 0$.

- By triangular form, $Tv_1 = \lambda_1 v_1 \overset{\text{ass.}}{\neq} 0$, thus $\overset{\neq 0}{v_1} \in \text{range}(T)$.
- By triangular form, $\underbrace{Tv_2}_{\in \text{range } T} = \underbrace{a_1 v_1}_{\in \text{range } T} + \underbrace{\lambda_2 v_2}_{\text{also has to be in range } T}$ for some scalar a.

  Thus $\lambda_2 v_2 \in \text{range}(T)$, and because $\lambda_2 \neq 0$, then also $v_2 \in \text{range}(T)$. etc

- In this way, we can see that $v_1, \dots, v_n$ are all in range (T).
- Thus T surjective, thus injective, thus invertible.

Proof "$\Rightarrow$" Assume T invertible.

- Clearly $t_1 \neq 0$ (otherwise $Tv_1 = 0$ thus not invertible).
- Suppose $t_j = 0$. Then T maps $\operatorname{span}(v_1, \dots, v_j)$ into $\operatorname{span}(v_1, \dots, v_{j-1})$
- Thus T is not injective on the subspace spanned by $v_1, \dots, v_j$, so there exists v in this subspace s.t. $Tv = 0$.

But then T is not invertible.

∎

Entries of diagonal are eigenvalues

Prop Suppose $T \in \mathcal{L}(V)$, V any finite-dim VS, T has an upper triangular form. Then the entries on the diagonal are precisely the eigenvalues of T.

holds both over $\mathbb{R}$ and $\mathbb{C}$

Proof

Fix any $\lambda \in F$ and consider $T - \lambda I$:

$$T = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \qquad T - \lambda I = \begin{pmatrix} \lambda_1 - \lambda & & * \\ & \ddots & \\ 0 & & \lambda_n - \lambda \end{pmatrix}$$

λ eigenvalue of T $\iff$ $T - \lambda I$ not invertible

$\iff$ One of the diagonal entries of $T - \lambda I$ is zero

previous proposition

$\iff$ $\lambda = \lambda_i$ for some i.

∎

Algebraic multiplicity of eigenvalues

<u>Definition:</u> The number of times each eigenvalue occurs on the diagonal is called the algebraic multiplicity of the eigenvalue.

<u>Remark:</u> More often, the algebraic mult. is defined using the characteristic polynomial, which we skipped. The two definitions are equivalent.

<u>Definition:</u> The geometric multiplicity of an eigenvalue is the dimension of the corresponding eigenspace.

⚠ In general, the two do not agree!

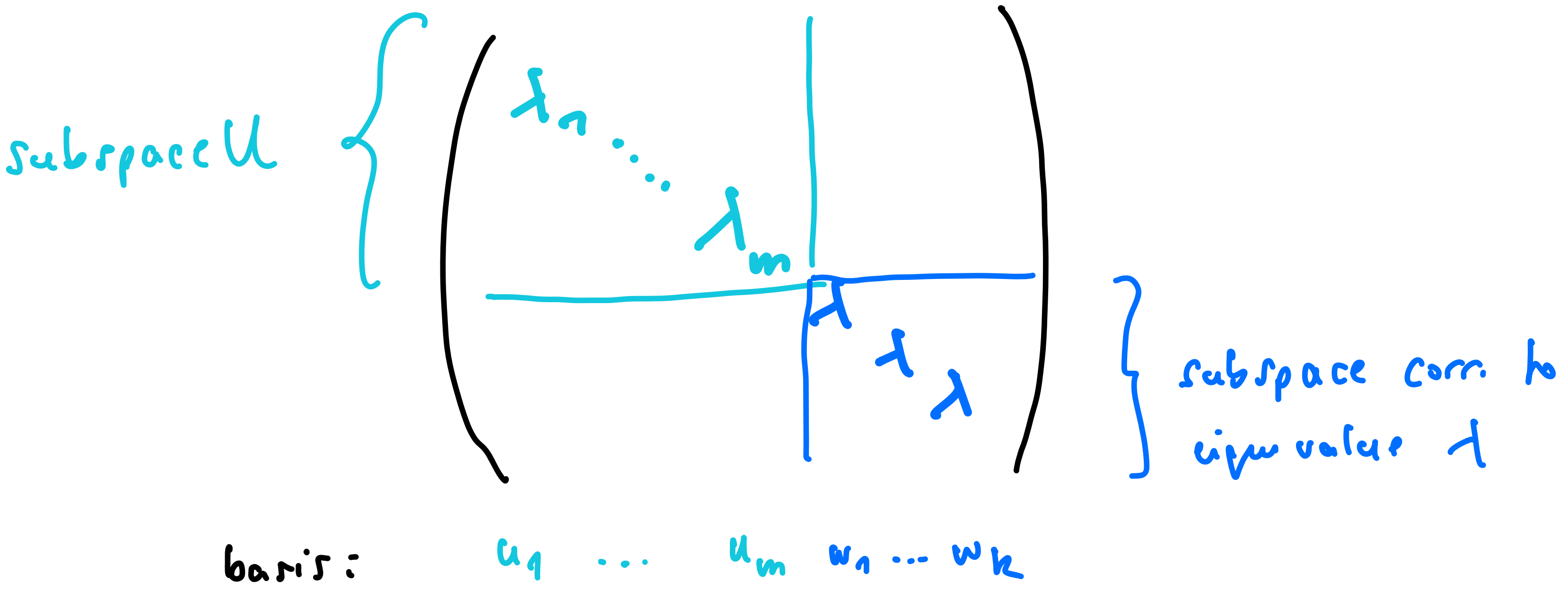
Over $\mathbb{C}$ each matrix can be triangularized

Prop V complex finite-dim VS, $T \in \mathcal{L}(V)$. Then $M(T)$ has an upper triangular form for some basis.

⚠ Does not hold over $\mathbb{R}$!

Proof idea as an image:

- split off the part of the space that belongs to eigenvectors for λ
- On remainder U apply induction hypothesis
- then add any basis for eig(λ) and show that it does not destroy upper-triang. form.


subspace U
λ_1
λ_m
λ
λ
λ
subspace corr. to eigenvalue λ
basis: $u_1 \dots u_m \; w_1 \dots w_k$

Proof: induction on n

- know already that one complex eigenvalue exists: λ
- Consider subspace $U := \text{range}(T - \lambda I)$ "complement of eig (λ)"
- Easy to see that U is invariant under T, that is $TU \subset U$ and that $\dim(\text{range}(U)) < n$.
- We can now apply the induction hypothesis to the operator $T: U \to U$: exists basis $\underline{u_1, \ldots, u_m}$ $(m < n)$ such that T_U has upper triangular form.
- Now extend $\underline{u_1, \ldots, u_m}$ by $\underline{w_1, \ldots, w_k}$ to a basis of V again.
- $Tw_1 = Tw_1 - \lambda w_1 + \lambda w_1 = (T - \lambda I) w_1 + \lambda w_1 \in \text{span}\{\underline{u_1, \ldots, u_m}, \underline{w_1}\}$

  Similarly: $Tw_i \in \text{span}\{\underline{u_1, \ldots, u_m}, \underline{w_1, \ldots, w_i}\}$.
- Thus T is upper-triangular wrt basis $\{\underline{u_1, \ldots, u_m}, \underline{w_1, \ldots, w_k}\}$. ∎

Normed spaces

ML keywords : loss functions, regularizers, sparsity,

Metric space

<u>Definition</u>: Let X be a set. A function $d: X \times X \to \mathbb{R}$ is called a metric if the following conditions hold:

$\forall u, v, w \in X$

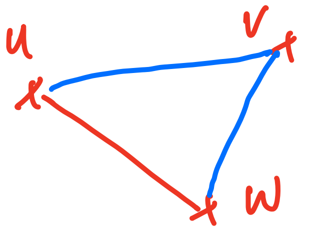
(1) $d(x,y) > 0$ if $x \neq y$ and

$d(x,x) = 0$

(2) $d(x,y) = d(y,x)$ (symmetry)

(3) $d(u,v) + d(v,w) \geq d(u,w)$

ML keywords:
inductive bias
k-nearest neighbor



Norm on a vector space

Def Let V be a vector space. A **norm** on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that $\forall x, y \in V$, $\lambda \in F$ the conditions are true:

(N1) $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$ (homogeneous)

(N2) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

$$\left.\begin{array}{ll} (N3) & \|x\| = 0 \Leftarrow x = 0 \\ (N4) & \|x\| = 0 \Rightarrow x = 0 \end{array}\right\} \text{(definiteness)}$$

$\|\cdot\|$ is a **semi-norm** if (N1) - (N3) are satisfied.

Euclidean norm on $\mathbb{R}^d$

Euclidean norm on $\mathbb{R}^d$: $\|x\| = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}$

Intuition norm (x) = " length of x "

= distance $(x, 0)$

Every norm induces a metric : $d(x,y) := \|x-y\|$

But not vice versa (try to find a counter-example!)

p - Norms on $\mathbb{R}^d$

Def Consider $V = \mathbb{R}^d$. Define $\|\cdot\|_p : \mathbb{R}^d \to \mathbb{R}$,

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \quad \text{for } 0 < p < \infty$$

not always a norm

The case $p=2$ coincides with the Euclidean norm.

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$

Furthermore we define:

$$\|x\|_\infty := \max |x_i| \qquad \text{(is a norm)}$$

$$\|x\|_0 := \text{number of non-zero coordinates}$$

$$= \sum_{i=1}^{d} \mathbb{1}_{\{x_i \neq 0\}}$$

is not a proper norm, see later.

Unit ball of a norm

Def The unit ball of a norm is the set of points such that norm ≤ 1:

$$B_p := \{ x \in \mathbb{R}^d \mid \|x\|_p \leq 1 \}$$

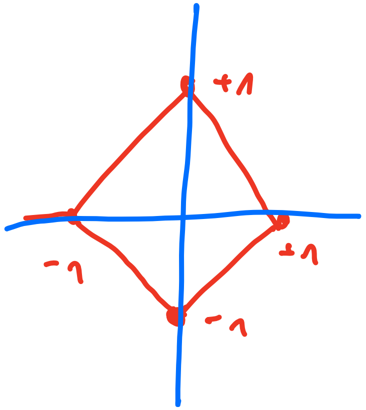
The unit sphere is the set of points such that norm $= 1$:

$$S_p := \{ x \in \mathbb{R}^d \mid \|x\|_p = 1 \}$$

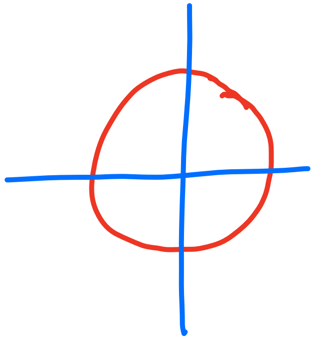
Illustration: unit balls on $\mathbb{R}^2$

$p \geq 1$:

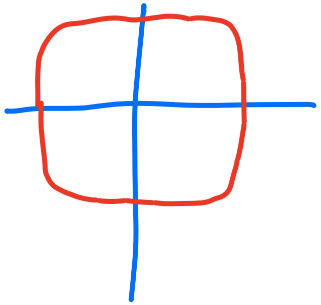
(convex balls)



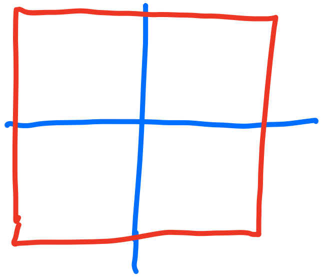
$p = 1$



$p = 2$



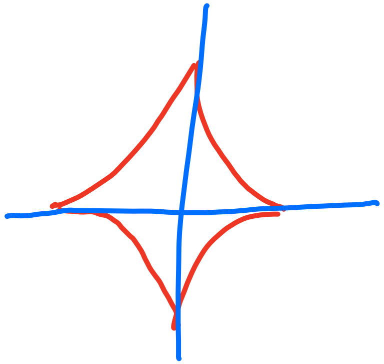
$p = 5$



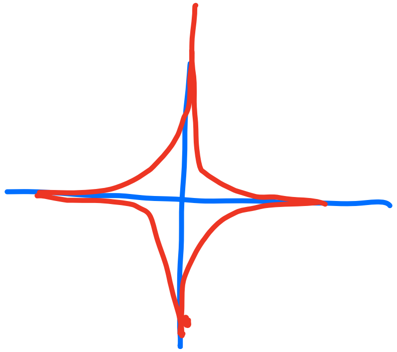
$p = \infty$

$p < 1$:

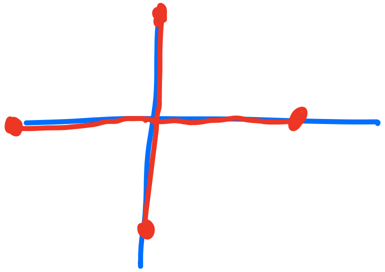
(balls not convex)



$p = 0.5$



$p = 0.1$



$p = 0$

When is a "p-norm" a norm?

Prop $\|\cdot\|_p$ is a norm on $\mathbb{R}^d$ iff $p \geq 1$.

Proof sketch

- Definiteness (N3, N4) hold for any $p > 0$
- Homogeneity holds for any $p > 0$
- Triangle inequality: this is the critical point, it only holds for $p \geq 1$. This is due to the famous Minkowski inequality, which holds iff $p \geq 1$:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p}$$

$\blacksquare$

All norms on $\mathbb{R}^n$

are equivalent

Equivalent norms (definition)

Definition: Let V be a vector space and $\|\cdot\|_a$ and $\|\cdot\|_b$ two norms on V. Then the two norms are called (topologically) equivalent if there exist constants $\alpha, \beta > 0$ such that

$$\forall x \in V: \quad \alpha \|x\|_a \leq \|x\|_b \leq \beta \cdot \|x\|_a \qquad \circledast$$

Theorem All norms on $\mathbb{R}^n$ are (topologically) equivalent.

Proof: W.l.o.g. we prove that if $\|\cdot\|$ is any norm on $\mathbb{R}^d$, then it is equivalent to $\|\cdot\|_\infty$.

→

$$\boxed{\underline{\text{First inequality}}: \quad \exists\, c_1 > 0 : \quad \forall x \quad \|x\| \leq c_1 \|x\|_\infty}$$

Let $x = \sum x_i e_i$ the representation of x in the standard basis of $\mathbb{R}^d$.

$$\begin{aligned}
\|x\| &= \Big\| \sum_{i=1}^{d} x_i e_i \Big\| \\
&\leq \sum_i \| x_i e_i \| \\
&= \sum_i \underbrace{|x_i|}_{\leq \|x\|_\infty} \|e_i\| \\
&\leq \sum_i \|x\|_\infty \cdot \|e_i\| \\
&= \|x\|_\infty \cdot \underbrace{\sum_i \|e_i\|}_{=:\, c_1}
\end{aligned}$$

Second inequality: $\exists c_2 > 0 \ \forall x \ \|x\|_\infty \le c_2 \cdot \|x\|$

Let $S := \{x \in \mathbb{R}^d \mid \|x\|_\infty = 1\}$ be the unit sphere wrt $\|\cdot\|_\infty$

Consider $f : S \to \mathbb{R}$, $x \longmapsto \|x\|$.

- The mapping f is continuous wrt $\|\cdot\|_\infty$:

  (this follows directly from the fact that

$$\begin{aligned} |f(x) - f(y)| &= \big|\, \|x\| - \|y\| \,\big| \\ &\le \|x-y\| \le c_1 \cdot \|x-y\|_\infty \end{aligned}$$

).

- The S is closed and bounded, thus by Theorem of Heine-Borel, S compact. Any continuous mapping on a compact set takes its min and max.

$$\tilde{c}_2 := \min \{f(x) \mid x \in S\}$$

- Because $0 \notin S$ (sphere, not ball), we can conclude from the definiteness that $\tilde{c}_2 \neq 0$.

- Now compute: For $x \in S$ we have

$$\underbrace{\tilde{c}_2}_{\text{choice of } \tilde{c}_2} \leq \|x\| = \left\| \frac{x}{1} \right\| = \left\| \frac{x}{\|x\|_\infty} \right\| = \frac{\|x\|}{\|x\|_\infty}$$

$$\Rightarrow \quad \|x\|_\infty \leq \underbrace{\frac{1}{\tilde{c}_2}}_{=: c_2} \|x\|$$

∎

Scalar product

Scalar product, definition

<u>Def</u> Consider vector space V. A mapping $\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$ is called a scalar product if

linearity
(S1) $\langle x_1 + x_2 , y \rangle = \langle x_1 , y \rangle + \langle x_2 , y \rangle$
(S2) $\langle \lambda x , y \rangle = \lambda \langle x , y \rangle$

symmetry
(S3) $\langle x , y \rangle = \langle y , x \rangle$ (if on $\mathbb{R}$)
$\langle x , y \rangle = \overline{\langle y , x \rangle}$ (if on $\mathbb{C}$)

complex conjugate

positive definite
(S4) $\langle x , x \rangle \geq 0$
(S5) $\langle x , x \rangle = 0 \iff x = 0$

Examples

- Euclidean scalar product on $\mathbb{R}^n$: $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

- On $\mathbb{C}^n$, $\langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y_i}$

- $\mathcal{C}([a,b])$: $\langle f, g \rangle = \int_a^b f(t) g(t) \, dt$

is a scalar product (but space would not be complete)

Scalar product and angles

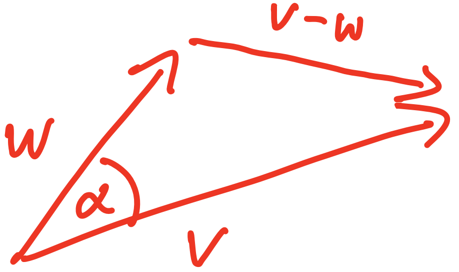
Prop Consider the standard scalar product on $\mathbb{R}^n$. Then

$$\langle v, w \rangle = \|v\| \, \|w\| \cos(\alpha)$$

where α is the angle enclosed by v and w.

Proof

- In a general triangle we have
  $$\|v-w\|^2 = \|v\|^2 + \|w\|^2 - 2\|v\| \, \|w\| \cos\alpha$$



- $\|v-w\|^2 = \langle v-w, v-w \rangle = \|v\|^2 + \|w\|^2 - 2\langle v, w\rangle$

$$\Rightarrow \langle v, w \rangle = \|v\| \, \|w\| \cos(\alpha).$$

∎

ML keyword: cosine similarity

Banach and Hilbert spaces

Def A vector space with a norm is called a **normed space**. If a normed space is complete (each Cauchy sequence converges), then V is called a **Banach space**. A VS with a scalar product is called a **pre-Hilbert-space**. If it is additionally complete, then it is called **Hilbert space**.

ML keyword: RKHS: reproducing kernel Hilbert space

Relationship between norm and scalar product

Scalar product $\Rightarrow$ norm

$\not\Leftarrow$

Consider a VS with a scalar product $\langle \cdot, \cdot \rangle$. Define $\|\cdot\| : V \to \mathbb{R}$ as $\boxed{\|x\| := \sqrt{\langle x, x \rangle}}$. Then $\|\cdot\|$ is a norm on V, the norm induced by $\langle \cdot, \cdot \rangle$.

The other way round does not work in general!

(Can you find a counterexample?)

Relationship between norm and metric

$$\text{norm} \Rightarrow \text{metric}$$

$$\not\Leftarrow$$

Consider a VS V with norm $\|\cdot\|$. Then

$$d : V \times V \to \mathbb{R} \;, \quad d(x,y) := \|x-y\|$$

is a metric on V, the metric induced by the norm.

The other direction does not work in general.

(Can you find a counterexample?)

Important inequalities

Consider $u, v \in \mathbb{R}^d$, and denote by $\|\cdot\|_p$ the p-norm.

Cauchy-Schwarz inequality:

$$|\langle u, v\rangle| \le \|u\|_2 \|v\|_2$$

Hölder inequality: Let $p, q \ge 1$ with $1/p + 1/q = 1$. Then:

$$|\langle u, v\rangle| \le \sum_{i=1}^{d} |u_i v_i| \le \|u\|_p \|v\|_q$$

Orthonormal basis and orthogonal projection

Orthogonal vectors and sets

<u>Def</u> Consider a pre-Hilbert space V.

- Two vectors $v_1, v_2 \in V$ are called <u>orthogonal</u> if $\langle v_1, v_2 \rangle = 0$.

  Notation: $v_1 \perp v_2$

- Two sets $V_1, V_2 \subset V$ are called orthogonal if

  $\forall v_1 \in V_1 \ \forall v_2 \in V_2 : \ \langle v_1, v_2 \rangle = 0$

- For a set $S \subset V$ we define its orthogonal complement $S^\perp$ as follows:

  $$S^\perp := \{ v \in V \mid v \perp s \ \ \forall s \in S \}$$

Orthonormal vectors and sets

Def:

- Two vectors v_1, v_2 are called orthonormal if they are orthogonal and additionally the two vectors have norm 1:

  - $\langle v_1, v_2 \rangle = 0$
  - $\|v_1\| = 1, \quad \|v_2\| = 1$

- A set of vectors $v_1, v_2, \dots, v_n$ is called orthonormal if any two vectors are orthonormal.

Orthogonal / orthonormal basis

We are particularly interested in orthogonal / orthonormal bases of a space:

<u>Proposition</u>: In an orthonormal basis $u_1, \dots, u_n$, the representation of a vector v is given as

$v = \sum a_i u_i$

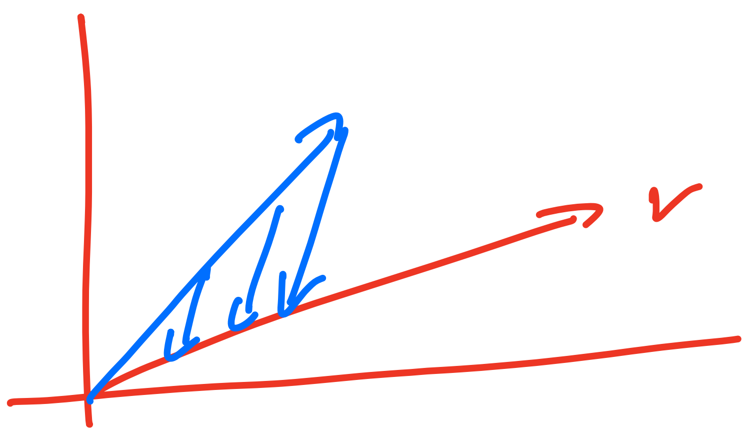
$$v = \sum_{i=1}^{n} \langle v, u_i \rangle u_i$$

<u>Proof</u> exercise

⚠ We still don't know whether an orthonormal basis always exists...

Projection (general case)

<u>Def</u> $A \in \mathcal{L}(V)$ is called a projection if $A^2 = A$.



blue vector gets projected on red vector (not orthogonal)

Orthogonal projection

<u>Theorem & Def</u> : Let U be a finite-dim subspace of a pre-Hilbert space H. Then there exists a linear projection $P_U : H \to U$ with $\ker(P_U) = U^\perp$. P_U is then called the <u>orthogonal projection</u> of H on U.

existence: see next theorem...

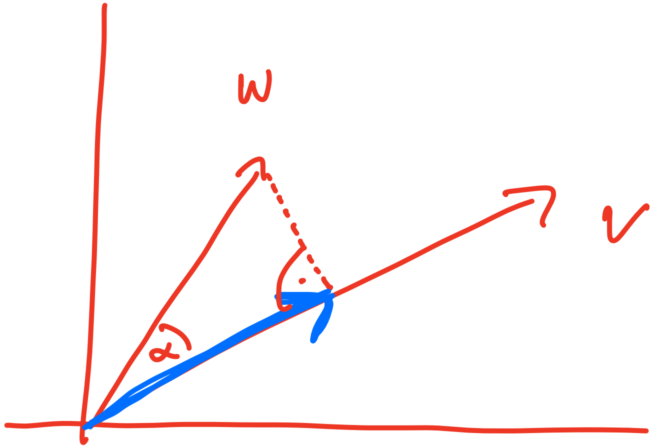
Proof idea: Let $u_1, \dots, u_k$ be an orthogonal basis of U.

Define $p_U : V \to U$ by $p_U(w) = \sum_{i=1}^{k} \frac{\langle w, u_i \rangle}{\| u_i \|} u_i$

Obviously:

- p_U linear
- $w \in U^\perp$

  $\Rightarrow p_U(w) = 0.$



How to make a given basis orthogonal?

Proposition: Let $v_1, \dots, v_n$ be any basis of a pre-Hilbert space. Then we can transform it into an orthonormal basis $u_1 \dots, v_n$.

Proof:

Gram-Schmidt orthogonalization

Is a procedure that takes any basis $v_1, \dots, v_n$ of a finite-dim VS and transforms it into another basis $u_1, \dots, u_n$ that is orthogonal:

Intuition: iterative procedure

<u>Step 1</u>: $u_1 := \frac{v_1}{\|v_1\|}$, $\quad U_1 := \text{span}\{u_1\}$

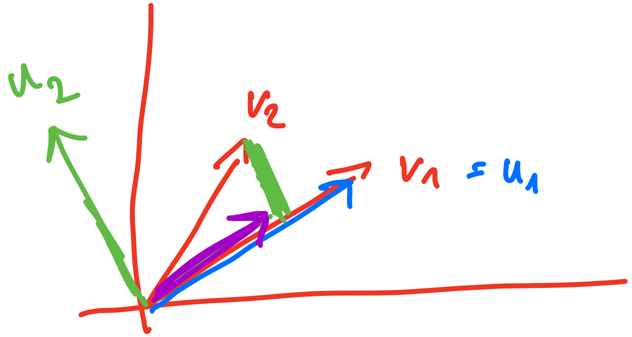
<u>Step k</u>: Assume that we already identified $u_1, \dots, u_{k-1}$.

- Project v_k on U_{k-1}, and keep "the rest":

$$\underline{\tilde{u}_k} := \underline{v_k} - \underline{P_{U_{k-1}}(v_k)}$$

- Renormalize:

$$u_k = \tilde{u}_k / \|\tilde{u}_k\|$$



Works in theory (would need to prove that, skipped)

In practice, it quickly results in large numerical errors.

→ QR factorization, see later.

Orthogonal matrices

Orthogonal matrices, definition

<u>Def</u> Let $Q \in \mathbb{R}^{n \times n}$ be a matrix with orthonormal (!) column vectors (wrt Euclidean scalar product). Then Q is called an <u>orthogonal (!) matrix.</u>

If $Q \in \mathbb{C}^{n \times n}$ and the columns are orthonormal (wrt the standard scalar product on $\mathbb{C}^n$), then it is called a <u>unitary matrix</u>.

Orthogonal matrices, examples

- Identity: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, Reflection: $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

- Permutation of coordinates: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

- Rotation in $\mathbb{R}^2$: $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$

- Rotation in $\mathbb{R}^3$:
  - Rotation about one of the axes:

$$R_{\theta,1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$

  - General rotation: can be written as a product of "elementary" rotations

Properties of orthonormal matrices

Let Q be orthogonal. Then:

- columns are orthogonal $\Leftrightarrow$ rows are orthogonal

  * see also slide below

- Q is always invertible, and $Q^{-1} = Q^t$

- Q realizes an <u>isometry</u>: $\forall v \in V: \|Qv\| = \|v\|$

- Q preserves angles: $\langle Qu, Qv \rangle = \langle u, v \rangle \quad \forall u,v \in V$

- $|\det Q| = 1$

The respective properties also hold for unitary matrices U:

- $U^{-1} = \overline{U}^t$

<u>**Proofs**</u>: assignment!

⚠ Orthogonal rows and colums

Consider the projection matrix $A = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix}$. The colums are obviously not orthogonal. The rows formally satisfy that $\langle \binom{0}{0}, \binom{2}{1} \rangle = 0$. The property that "rows orthogonal <=> cols orthogonal" does not hold here. But note that A is *not* an orthogonal matrix because the latter requires all rows / cols to have norm 1 (in particular, also full rank).

The statement "rows orthogonal <=> colums orthogonal" clearly does not hold for arbitrary matrices.

SKIPPED

Representation of isometric mappings

<u>Theorem</u> Let $S \in \mathcal{L}(V)$ for a real VS V. Then equivalent:

(a) S is an isometry: $\|Sv\| = \|v\| \quad \forall v \in V$

(b) there exists an orthonormal basis of V such that the matrix of S has the following form:

$$M = \begin{pmatrix} \square & & & & & & \\ & \square & & & & 0 & \\ & & \square & & & & \\ & & & \square & & & \\ & & & & \square & & \\ & 0 & & & & \ddots & \\ & & & & & & \square \end{pmatrix}$$

where each of the little blocks

- either is a 1×1 matrix ($\hat{=}$ a real number) being 1 or -1
- or is a 2×2 rotation matrix: $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$

<u>Proof</u> skipped.

Symmetric matrices

In ML, symmetric matrices are all over the place:
covariance matrix, kernel matrix, similarity matrix, ...

Symmetric matrices

Def A matrix $A \in \mathbb{R}^{n \times n}$ is called symmetric if $A = A^t$.

A matrix $A \in \mathbb{C}^{n \times n}$ is called hermitean if $A = \bar{A}^t$.

Hermitean matrices have real-valued eigenvalues and orthogonal eigenvectors

Prop Let $A \in \mathbb{C}^{n \times n}$ be hermitean. Then all eigenvalues of A are real-valued. Eigenvectors that correspond to different eigenvalues are orthogonal.

Proof • λ eig. value of A with eigenvector x. Then

$$\underline{\lambda \langle x, x \rangle} = \langle \lambda x, x \rangle = \langle Ax, x \rangle \overset{\text{hermitean}}{=}$$

$$= \langle x, Ax \rangle = \langle x, \lambda x \rangle \underset{\text{property of scalar prod. on } \mathbb{C}}{=} \underline{\bar{\lambda} \langle x, x \rangle}$$

$$\Rightarrow \lambda = \bar{\lambda} \in \mathbb{R}$$

- (λ_1, x_1), (λ_2, x_2) eigs of A, $\lambda_1 \neq \lambda_2$. Then:

$\underline{\lambda_1 \langle x_1, x_2 \rangle}$ = ...as above ... = $\underline{\lambda_2 \langle x_1, x_2 \rangle}$

$$\Rightarrow \quad 0 = \lambda_1 \langle x_1, x_2 \rangle - \lambda_2 \langle x_1, x_2 \rangle =$$

$$= (\lambda_1 - \lambda_2) \langle x_1, x_2 \rangle$$

$\Rightarrow$ either $\lambda_1 = \lambda_2$

or if $\lambda_1 \neq \lambda_2$, then $\langle x_1, x_2 \rangle = 0$

$\Rightarrow x_1 \perp x_2$. ∎

What can we conclude from here ?

- Each matrix $\in \mathbb{C}^{n\times n}$ has n eigenvalues $\in \mathbb{C}$
- For hermitian matrices $\in \mathbb{C}^{n\times n}$ all n eigenvalues now live in $\mathbb{R}$
- The eigenvectors still might live in $\mathbb{C}^{n\times n}$ (and not in $\mathbb{R}^{n\times n}$).

Symmetric matrices $\in \mathbb{R}^{n\times n}$ are a special case of hermitian, (because $A \underset{sym}{=} A^t \underset{\mathbb{R}}{=} \bar{A}^t$), so it also has n real-valued eigenvalues. But the eigenvectors might still live in $\mathbb{C}^n$.

Self-adjoint operators

<u>Def</u> An operator $T \in \mathcal{L}(V)$ on a pre-Hilbert space V is called <u>self-adjoint</u> if

$$\langle Tv, w \rangle = \langle v, Tw \rangle .$$

Sometimes it is called a Hermitean operator (on $\mathbb{C}^n$)

symmetric operator (on $\mathbb{R}^n$).

<u>Remark</u> Over $\mathbb{C}^n$, self-adjoint operators are represented by hermitean matrices. On $\mathbb{R}^n$, self-adjoint op. are represented by symmetric matrices.

Self-adjoint operator has real-valued eigenvalue and eigenvector

Prop: V vector space over $\mathbb{C}$ or $\mathbb{R}$, $T \in \mathcal{L}(V)$, $T \neq 0$, self-adjoint.
Then T has at least one eigenvalue and it is real-valued.
In particular in case of $\mathbb{R}$, also the eigenvector lives in $\mathbb{R}^n$.

⚠ Already know: over $\mathbb{C}$, every matrix has an eigenvalue. But it could be a complex number. Now: if T is self-adjoint, have a real eigenvalue.

Consider case $F = \mathbb{R}$.

Proof $n := \dim V$. Chose $\overset{\in \mathbb{R}^n}{v} \neq 0$, and consider

$$v, Tv, T^2v, \dots, T^n v. \in \mathbb{R}^n$$

These vectors have to be lin. dependent ($n+1$ vectors, $\dim = n$)

There exist $a_0, \dots, a_n \in \mathbb{R}$ (not all 0) such that

$$a_0 v + a_1 Tv + \dots + a_n T^n v = 0.$$

Consider the polynomial with these coefficients and decompose it over $\mathbb{R}$ into linear and quadratic factors:

$$a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n =$$

decompose over $\mathbb{R}$!

$c \neq 0,\ b_i, c_i, \lambda_i \in \mathbb{R}$

$M + m \geq 1$

$$= \underset{\neq 0}{c}\,\underbrace{(x^2+b_1x+c_1)\dots(x^2+b_Mx+c_M)}_{\text{quadratic terms that cannot be decomposed into linear terms;}} \cdot \underbrace{(x-\lambda_1)\cdot\ldots\cdot(x-\lambda_m)}_{\text{linear terms}}$$

in particular they satisfy

$b_i^2 < 4c_i$ (otherwise one could factorize them by the quadratic formula)

Replace the x by T:

$$0 = (a_0 + a_1 T + \dots + a_n T^n) v = \Big(c \underbrace{(\dots \quad)}_{\text{quadr.}} \underbrace{(\dots\dots)}_{\text{lin. terms}} \Big) \cdot v$$

Now can prove: the "quadratic operators" are invertible. (see (*) below).

So we multiply the equation with their inverses and obtain

$$0 = (T - \lambda_1 I) \cdot \dots (T - \lambda_n I) v$$

Because we had chosen $v \neq 0$, at least one of the terms $(T - \lambda_i I)$ has to exist (i.e. $m \geq 1$)

But then also the product $(T - \lambda_1 I) \cdot \dots \cdot (T - \lambda_m I)$ is not injective, thus at least one of the factors j not injective, thus λ_j eigenvalue. over $\mathbb{R}$. In particular, the eigenvector lives in $\mathbb{R}^n$ as well.

∎

(*) <u>Proposition</u>: Suppose T self-adjoint and $b, c \in \mathbb{R}$ satisfy $b^2 < 4c$. Then $T^2 + bT + cI$ is invertible

<u>Proof</u> Intuition:

$$x^2 + bx + c = \underset{\geq 0}{\left(x + \frac{b}{2}\right)^2} + \underbrace{\left(c - \frac{b^2}{4}\right)}_{>0 \text{ by ass.}} > 0$$

So $x^2 + bx + c$ is an invertible real number.

Now do it: Consider any $v \neq 0$.

$$\langle (T^2 + bT + cI)v, v \rangle =$$

$$= \underbrace{\langle T^2 v, v \rangle}_{\substack{= \langle Tv, Tv \rangle \\ = \|Tv\|^2}} + b \underbrace{\langle Tv, v \rangle}_{\leq \|Tv\| \|v\| \text{ Cauchy-Schwartz}} + c\|v\|^2$$

$$\geq \|Tv\|^2 - |b| \, \|Tv\| \, \|v\| + c\|v\|^2$$

$$= \underbrace{\left(\|Tv\| - \frac{|b|\,\|v\|}{2} \right)^2}_{\geq 0} + \underbrace{\left(c - \frac{b^2}{4} \right)}_{> 0 \text{ by ass.}} \|v\|^2 > 0$$

$$\Rightarrow \quad (T^2 + bT + cI)v \neq 0 \text{ for all } v \neq 0$$

$$\Rightarrow \quad T^2 + bT + cI \text{ invertible.} \qquad \blacksquare$$

Spectral theorem for symmetric / hermitian matrices

<u>Theorem</u> : A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is orthogonally diagonalizable: there exists an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ s.t.

$$A = Q D Q^t = \sum_{i=1}^{n} \lambda_i q_i q_i^t .$$

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \quad Q = \begin{pmatrix} | & | & \\ q_1 & q_2 & \cdots \\ | & | & \end{pmatrix}$$

Proof (sketch) By induction on $n := \dim V$

Base case $n = 1$: clear

Inductive step $n-1 \rightsquigarrow n$

previous theorem $\lambda \in \mathbb{R}$

- A symmetric $\Rightarrow$ A has atleast one eigenvector $u \in \mathbb{R}^n$
- $U := \operatorname{span}\{u\}$. U is invariant under A. $\forall v \in U: Av \in U$
- Consider $U^\perp$ and the restriction of A to $U^\perp$.

  On $U^\perp$, A is again a symmetric operator

  and $\dim(U^\perp) = n-1$.
- Apply the induction hypothesis on this space of dim $n-1$.

  Does the job! ∎

Complex version of this theorem

(often not as relevant to ML as the real version above)

Theorem A hermitian matrix $A \in \mathbb{C}^{n \times n}$ is unitarily diagonalizable: there exists a unitary matrix U and a diagonal matrix D s.th.

$$A = U D \bar{U}^t$$

In particular, the entries of D are real-valued.

Positive definite matrices

Positive definite matrices

Def A matrix $A \in \mathbb{R}^{n \times n}$ is called

semi-definite (psd)

positive definite (pd) if $\forall x \in \mathbb{R}^n, x \neq 0$:

$$x^t A x \underset{\geq}{>} 0.$$

Def A matrix A is called a Gram matrix if there exists a set of vectors $v_1, \dots, v_n$ s.th.

$$a_{ij} = \langle x_i, x_j \rangle.$$

Observe: On $\mathbb{C}^{n \times n}$, Gram matrices are hermitian

On $\mathbb{R}^{n \times n}$, Gram matrices are symmetric.

Characterization of pd matrices over $\mathbb{C}$

Theorem: $A \in \mathbb{C}^{n \times n}$ hermitean. Then equivalent:

(i) A is psd (pd)

(ii) All eigenvalues of A are ≥ 0 (>0)

(iii) The mapping $\langle \cdot , \cdot \rangle_A : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ with

$$\langle x, y \rangle_A := \bar{y}^t A x$$

satisfies all properties of a scalar product

except one: if $\langle x, x \rangle_A = 0$ this does not imply $x = 0$.

(iv) A is a Gram matrix of n vectors $a_{ij} = \langle x_i , x_j \rangle$

which are not necessarily lin. independent

which are lin. independent

⚠ Observe that implicitly this theorem implies that over $\mathbb{C}$ we have pd $\Rightarrow$ self-adjoint.

Over $\mathbb{R}$, this is *not* true: pd $\not\Rightarrow$ symmetric

Thus to get a similar characterization of pd matrices over $\mathbb{R}$, we need to add a symmetry condition

Example: $A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

on $\mathbb{R}$, if $x \neq 0$

$$x^t A x = x_1^2 + x_2^2 > 0$$

- So A is pd but not symmetric.
- Over $\mathbb{C}$, the same matrix is not pd because $x_1^2 + x_2^2$ can be negative!

Characterization of symmetric, pd matrices over $\mathbb{R}$

To Do

Add

Roots of psd matrices

Theorem: Let $A \in \mathbb{R}^{n\times n}$ be symmetric, psd. Then there exists a matrix $B \in \mathbb{R}^{n\times n}$, B psd such that $A = B^2$.

Sometimes B is called the square root of A, sometimes denoted as $B = (A)^{1/2}$.

Proof

- Spectral theorem $\Rightarrow$ $A = U D U^t$, $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

- psd $\Rightarrow$ eigenvalues $\lambda_i \geq 0$

- Define $\sqrt{D} := \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{pmatrix}$ and set

$B := U \sqrt{D} U^t$. Does the job. $\blacksquare$

Remark: more generally, one can define $A^{1/k}$ for any $k \in \mathbb{N}$. And because we can also define A^{-1} (in case pd) and A^k, one can define $A^{p/q}$ for $p, q \in \mathbb{Z}$.

Important psd matrices for ML

- Consider a data matrix $X \in \mathbb{R}^{n \times d}$: n data pts, d dims

- Then the matrix $C := X^t X \in \mathbb{R}^{d \times d}$ is called the covariance matrix.

- The matrix $K := X X^t \in \mathbb{R}^{n \times n}$ is called a kernel matrix (for the linear kernel), and in this special case it is also the Gram matrix.

- All these matrices are symmetric and positive semi-definite

(prove it!)

Variational characterization of eigenvalues

ML keyword: spectral clustering

Literature: Bhatia: Matrix Analysis.

Rayleigh coefficient

<u>Def</u> Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix.

$$R_A : \mathbb{R}^n \setminus \{0\} \to \mathbb{R} , \quad x \longmapsto \frac{x^t A x}{x^t x} = \frac{x^t A x}{\|x\|^2}$$

is called the Rayleigh coefficient.

Rayleigh coeff. ⇝ first eigenvalue

<u>Prop</u> Let A be symmetric, let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ be the eigenvalues and $v_1, \dots, v_n$ the eigenvectors of A.

Then:

$$\min_{x \in \mathbb{R}^n} R_A(x) = \min_{\|x\|=1} x^t A x = \lambda_1 \quad , \text{ attained at } x = v_1$$

$$\max_{x \in \mathbb{R}^n} R_A(x) = \max_{\|x\|=1} x^t A x = \lambda_n \quad , \text{ attained at } v_n.$$

Intuition for the proposition

Assume A is expressed in terms of the basis $v_1, \dots, v_n$ by

$$A = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Let y be a vector, also represented in this basis: $y = y_1 v_1 + y_2 v_2 + \dots + y_n v_n$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

⊛ $y^t A y = \underline{\lambda_1} y_1^2 + \dots + \lambda_n y_n^2$

Among the vectors $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$

$$\begin{pmatrix} 0.5 \\ 0.5 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

the smallest result of $y^t A y$ would be given by the vector $\underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{= v_1}$, and the value would be λ_1

Formal proof (sketch)

Assume we start with the standard basis.

Let $Q = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}$ be the basis transformation that brings A in diagonal form: Q orthogonal such that

$$A = Q^t \Lambda Q \quad \text{with } \Lambda \text{ diagonal.}$$

For a vector $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ in the original basis, we now consider the transformed vector $y := Q^t x$ and compute its Rayleigh coefficient:

$$R_A(y) = \frac{\overbrace{(Q^t x)^t}^{y^t} \; \overbrace{(Q^t \Lambda Q)}^{A} \; \overbrace{(Q^t x)}^{y}}{\underbrace{(Q^t x)^t}_{y^t} \; \underbrace{(Q^t x)}_{y}}$$

$$(Q^t x)^t = x^t Q$$

$$= \frac{x^t \cancel{Q \, Q^t} \Lambda \cancel{Q \, Q^t} x}{x^t \cancel{Q \, Q^t} x}$$

$$= \frac{x^t \Lambda x}{x^t x} = \frac{\lambda_1 x_1^2 + \dots + \lambda_n x_n^2}{\|x\|}$$

Now we look at the minimum of $R_A(y)$:

$$\min_{\|y\|=1} R_A(y) = \min_{\|x\|=1} \lambda_1 x_1^2 + \dots + \lambda_n x_n^2$$

This min. is attained for $x = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$, that is

$y = Q^t x = v_1$, with value $R(y) = \lambda_1$. ∎

Rayleigh coeff ⇝ second eigenvalue

<u>Prop</u> Consider a symmetric matrix A with eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Consider the optimization problem

$$\min_{\substack{\|x\|=1 \\ x \perp v_1}} R(x)$$

This problem is solved by $x = v_2$, $R(v_2) = \lambda_2$.

Proof intuition

Consider operator A restricted to the space $V_1^{\perp} := \left(\text{span}\{v_1\}\right)^{\perp}$. We know that on this space, A is invariant and symmetric, so we can apply Rayleigh to this "smaller" space.

$$V_1^{\perp} = \text{span}\{v_2, \dots, v_n\}$$

If we apply Rayleigh to $V_1^{\perp}$, then we get the solutions λ_2, v_2.

Min-max-theorem / Courant-Fischer-Weyl

<u>Theorem</u> $A \in \mathbb{R}^{n \times n}$ symmetric; eigenvalues $\lambda_1 \leq \dots \leq \lambda_n$. Then:

$$\lambda_k = \min_{\substack{U \text{ subspace} \\ \dim U = k}} \max_{x \in U \setminus \{0\}} R_A(x) \qquad (1)$$

$$= \max_{\substack{U \text{ subspace,} \\ \dim U = n-k+1}} \min_{x \in U \setminus \{0\}} R_{A}(x) \qquad (2)$$

Proof intuition

- For case $k=1$, case (2) is pretty much what we have proved already:

$$\max_{\underbrace{\substack{U \text{ subspace} \\ \dim U = \underbrace{n-k+1}_{n}}}_{\text{can drop it}}} \ \min_{x \in U \setminus \{0\}} R_A(x) =$$

$$= \min_{x \in V \setminus \{0\}} R_A(x) \overset{\text{previous result}}{=} \lambda_1 .$$

Case (1) follows similar principles.

- case $k=2$ .... similar to the previous statement.
- General case: induction.

Matrix norms

ML keywords:

- low rank approximation
- Perturbation analysis

Motivation

Want to quantify the "similarity" of various matrices.

So we define "norms" on the space of all matrices.

⚠ Not all of them are proper norms.

Definitions of matrix norms

Given a matrix $A \in \mathbb{R}^{m \times n}$. Define the following norms:

$$\|A\|_{max} = \|A\|_{\infty} = \max_{ij} |a_{ij}|$$

trace, see later

$$\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{tr\,(A^t A)}$$

Frobenius norm

$$= \sqrt{\sum \sigma_i^2}$$ where σ_i are the singular values of A.

see next slide

$\|A\|_2 = \sigma_{max}(A)$ where σ_{max} is the largest singular value

$$= \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$ Euclidean norm on vectors in $\mathbb{R}^m$

"Operator norm", "spectral norm"

trace see later

$\|A\|_* = \mathrm{tr}\left(\sqrt{A^t A}\right)$ nuclear norm

Many more matrix norms exist ...

Simple inequalities

Let $A \in \mathbb{R}^{m \times n}$. Then:

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

$$\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\,\|A\|_\infty$$

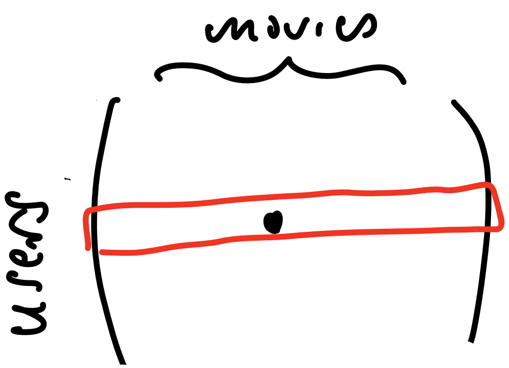
$$\|A\|_2 \le \sqrt{\|A\|_1 \cdot \|A\|_\infty}$$

Many, many more... see eg the book on Matrix Analysis by Bhatia.

Singular value decomposition

ML motivation: recommender systems

- Netflix ratings: huge matrix!



rating of user i for movie j

- ratings of a particular user are "not random" but have structure
- "compress" the matrix into something much smaller that also better represents the "structure" of this matrix.

Singular value decomposition

<u>Proposition</u> Consider $A \in \mathbb{R}^{m \times n}$ of rank r. Then we can write A in the form

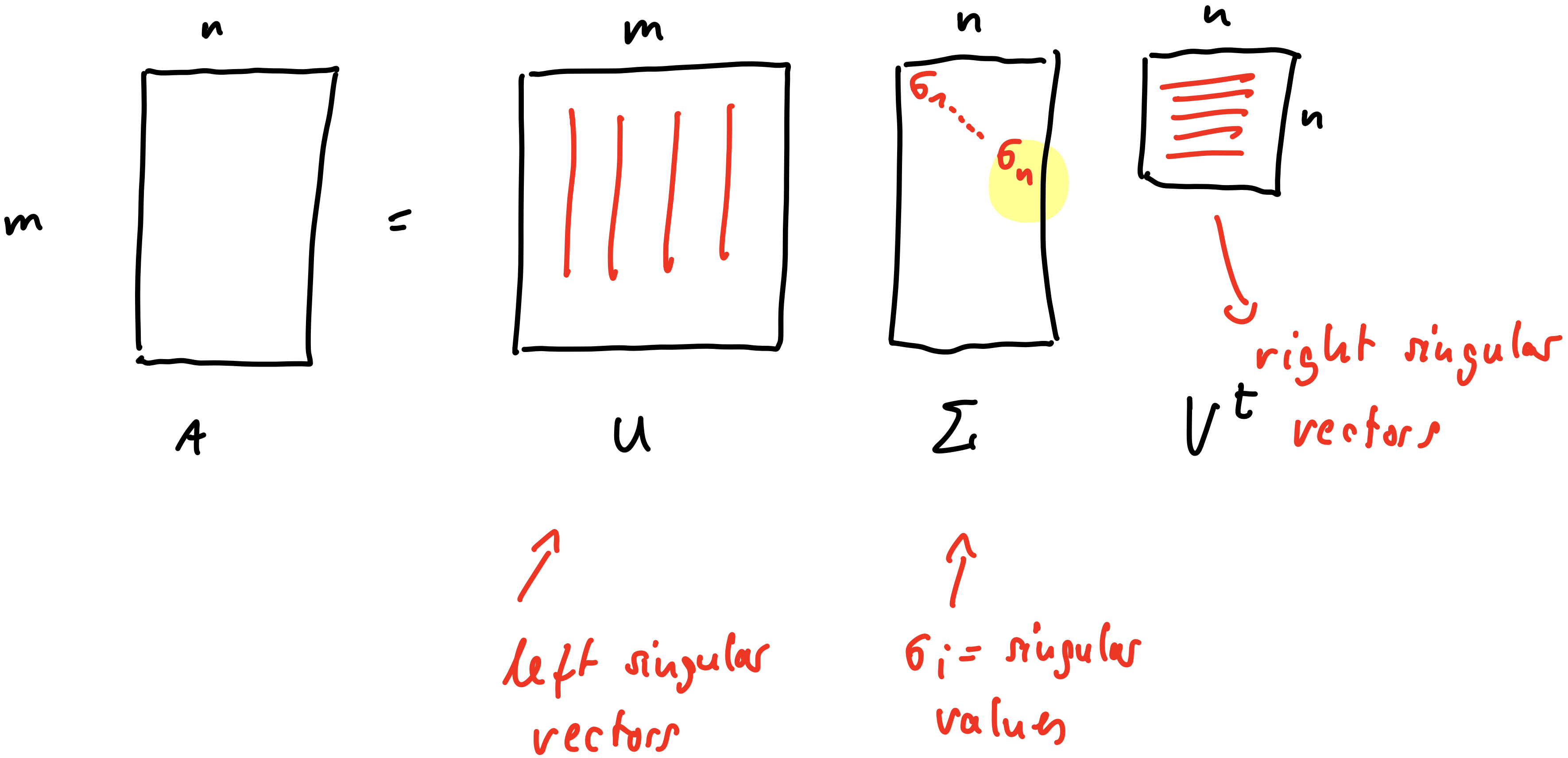
$$\boxed{A = U \cdot \Sigma \cdot V^t}$$

where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ is "diagonal".

$$\begin{array}{cc} & n \\ m & \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & O & \end{pmatrix} \end{array} \qquad \begin{array}{cc} & n \\ m & \begin{pmatrix} \sigma_1 & & 0 & \\ & \ddots & & O \\ 0 & & \sigma_m & \end{pmatrix} \end{array}$$

Exactly r of the diagonal values $\sigma_1, \sigma_2, \ldots$ are non-zero.

Illustration



Proof sketch

Given $A \in \mathbb{R}^{m\times n}$, we consider $\underline{B} := \underbrace{A^t}_{n\times m}\underbrace{A}_{m\times n} \in \mathbb{R}^{n\times n}$.

observe:

- B is symmetric:

$$(A^tA)^t = A^t(A^t)^t = A^tA$$

- B is positive semi-definite:

$$\langle (A^t)^t x, Ax \rangle$$

$$x^tBx = \langle x, Bx \rangle = \langle x, A^tAx \rangle$$

$$= \langle Ax, Ax \rangle$$

$$= \|Ax\|^2 \geq 0$$

$$\langle x, Cy \rangle$$
$$= \langle C^t x, y \rangle$$

So there exists an orthonormal basis of eigenvectors $\underline{v_1, \dots, v_n}$ with eigenvalues $\underline{\lambda_1, \dots, \lambda_n} \geq 0$.

Define:

- $\underline{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \in \mathbb{R}^{m \times n}$ where $\sigma_i = \sqrt{\lambda_i}$

- $\underline{U} = \begin{pmatrix} | \\ u_i \\ | \end{pmatrix}$ matrix with columns $u_i := \dfrac{A v_i}{\sigma_i}$

- $\underline{V} = \begin{pmatrix} | \\ v_i \\ | \end{pmatrix}$ matrix with v_i as columns

Now we need to show that with these definitions we have $A = U \cdot \Sigma \cdot V^t$:

- Columns of $U \cdot \Sigma$ are given as

$$\sigma_i u_i = \sigma_i \frac{Av_i}{\sigma_i} = Av_i$$

- Now multiply with V^t:
  - rows of V^t are the v_i,
  - exploit that if $i \neq j$ then $v_i \perp v_j$ and $\|v_i\| = 1$.
  - The terms consisting of i, j with $i \neq j$ cancel, the terms with $i = j$ will result in a factor of 1.

So we will be left with matrix A.

Finally, it is easy to see that U, V are orthogonal matrices, and that the number of non-zero entries in Σ coincides with the rank.

Basic properties of SVD

- The rank of a matrix coincides with the number of non-zero singular values

- If the matrix A has rank r, then

$$\ker(A) = \operatorname{span}\{v_{r+1}, \dots, v_n\}$$

$$\operatorname{range}(A) = \operatorname{span}\{u_1, \dots, u_r\}$$

Proof: Exercise

Key differences between SVD (A) and eig (A)

- SVD always exists, no matter how A looks like! can be rectangular, does not need to be symmetric, ...

- U, V are orthonormal! (not true for eigenvectors in general).

- singular values are always real and non-negative.

Key differences between SVD (A) and eig (A)

If $A \in \mathbb{R}^{n \times n}$ is <u>symmetric</u>, then the SVD is "nearly the same" as the eigenvalue decomposition: (λ_i, v_i) are the eigenvalues/vectors of A, then

$(|\lambda_i|, v_i)$ are the singular values / vectors of A.

In particular, left- and right singular vectors are the same (up to signs).

Relationship SVD(A) and eig$(\underbrace{AA^t}_{\text{symmetric!}})$

- For general (not nec. square) matrices A:

Left-singular vectors of A are the eigenvectors of AA^t.

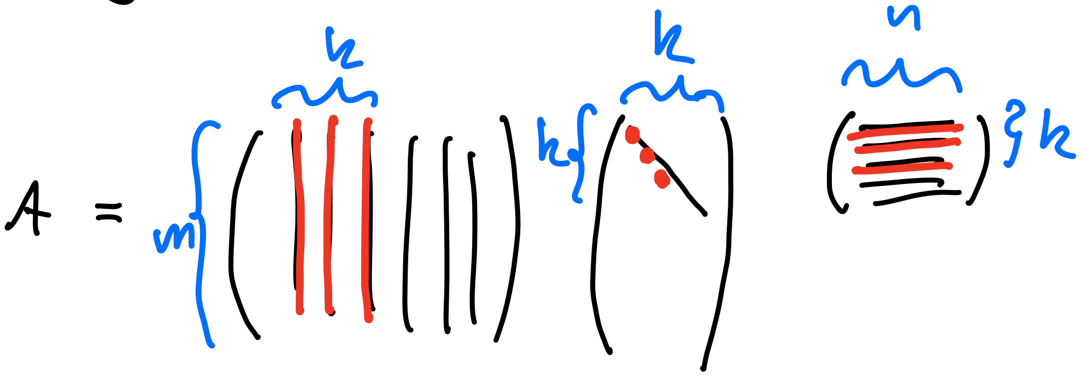
Right- A^tA.

- $\lambda \neq 0$ is an eigenvalue of AA^t $\Longleftrightarrow$

$\sqrt{|\lambda|} \neq 0$ is singular value of A

Rank-k-approximation

Given SVD $A = U \Sigma V^t$, entries $\sigma_1, \sigma_2, \dots$ sorted in decreasing order, $k \in \mathbb{N}$. Now we are going to define a new matrix A_k by the following procedure:



A_k

$(m \times k)(k \times k)(k \times n)$

$m \times n$

- take first k cols of U,
  first k entries of Σ
  first k rows of V^t

$\Bigg\} \; \boxed{A_k}$

More formally:

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^t$$

Observe: $\operatorname{rank}(A_k) = k$.

Best rank-k approximation

Theorem (Eckart-Young-Mirsky)

Let $\|\cdot\|$ be either the Frobenius-norm $\|\cdot\|_F$ or the two-norm $\|\cdot\|_2$.

Consider a matrix $A \in \mathbb{R}^{m\times n}$, A_k the low-rank-matrix constructed above, and $B \in \mathbb{R}^{m\times n}$ any other rank-k matrix.

Then $\|A - A_k\| \leq \|A - B\|$.

In particular,
$$\|A - A_k\| = \begin{cases} \sigma_{k+1} & \text{in case of } \|\cdot\|_2 \\ \left(\sum\limits_{i=k+1}^{\min(m,n)} \sigma_i^2\right)^{1/2} & \text{in case } \|\cdot\|_F. \end{cases}$$

Proof: skipped.

SVD and matrix norms

Consider $A \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$.
($p = \min\{n, m\}$).

Then:

- $\| A \|_F = \sigma_1^2 + \ldots + \sigma_p^2$
- $\| A \|_2 = \sigma_1$

Proof: skipped

Relation to machine learning

Computational reasons for a rank-k-approximation:

The running time of many ML algorithms scales heavily in the (implicit) dim. Often they can be implemented efficiently if matrices are (sparse or) low rank.

Statistical reason: ML only works if the data "is simple".

Typical assumptions:

- lives on a low-dim manifold (~ locally low rank)

(• is sparse)

Replacing a matrix with its SVD is supposed to get rid of noise.

Pseudo-inverse

ML keyword:
linear regression!

Pseudo-inverse

Definition For $A \in \mathbb{R}^{m \times n}$, a pseudo-inverse of A is defined as the matrix $A^{\#} \in \mathbb{R}^{n \times m}$ which satisfies the following conditions:

If A would be invertible $A A^{-1} = \mathrm{Id} \Rightarrow A A^{-1} A = A$

$$\left.\begin{array}{ll} (1) & \underbrace{A\,\overbrace{A^{\#}}^{\neq \mathrm{Id} \text{ in general}}}_{\neq \mathrm{Jd} \text{ in general}} A = A \\ (2) & A^{\#} A A^{\#} = A^{\#} \end{array}\right\} \text{"nearly inverse"}$$

$$\left.\begin{array}{ll} (3) & (A A^{\#})^{t} = A A^{\#} \\ (4) & (A^{\#} A)^{t} = A^{\#} A \end{array}\right\} \text{symmetry}$$

Intuition:

- Consider a projection from $\mathbb{R}^3 \to \mathbb{R}^2$,

$$A\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} = \begin{pmatrix}x_1\\x_2\end{pmatrix}$$

- Cannot invert, obviously (inverting would mean to reconstruct the original point).

- But I could "invent" a reconstruction, for example: $R: \mathbb{R}^2 \to \mathbb{R}^3$,

$$R\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}x_1\\x_2\\17\end{pmatrix}$$

- Now we have:

$$\begin{aligned} A\,\underset{\shortparallel}{R}\,A &= A\\ A\,A^{\#}A &= A\end{aligned}$$

Intuition: how could we define a pseudo-inverse?

Consider a diagonal matrix $A = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$

with $\lambda_1, \dots, \lambda_k \neq 0$.

Then set $A^{\#} := \begin{pmatrix} 1/\lambda_1 & & & & & \\ & \ddots & & & & \\ & & 1/\lambda_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$

Does the job.

How to do it for general matrices? $\rightsquigarrow$ SVD!

Moore-Penrose pseudoinverse

Proposition: Let $A \in \mathbb{R}^{m \times n}$, $A = U \Sigma V^t$ its SVD. Then the following matrix is a pseudo-inverse:

$$\boxed{A^\# := V \Sigma^\# U^t} \quad \text{with} \quad \Sigma^\# \in \mathbb{R}^{m \times n}$$

$$\Sigma^\#_{ii} = \begin{cases} 1/\Sigma_{ii} & \text{if } \Sigma_{ii} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof: easy, just do it.

__Proposition:__ If A is invertible, then $A^{-1} = A^{\#}$.

__Proof__ : easy, just do it.

Trace of a matrix

Trace

Def The trace of a square matrix $A \in F^{n \times n}$ is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{n} a_{ii} .$$

Properties

- $\mathrm{tr} : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a linear operator

  In particular, $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.

- $\mathrm{tr}(A\cdot B) = \mathrm{tr}(B\cdot A)$

  ⚠ $\mathrm{tr}(A\cdot B) \neq \mathrm{tr}(A)\cdot\mathrm{tr}(B)$

- trace does not depend on the basis :

  Let $T \in \mathcal{L}(V)$, and $\mathcal{B}_1$ and $\mathcal{B}$ two bases of V. Then:

  $\mathrm{tr}(M(T,\mathcal{B}_1)) = \mathrm{tr}(M(T,\mathcal{B}_2))$.

- The trace of an operator equals the sum of its complex eigenvalues, summed according to multiplicity:

$$\tilde{A} = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$ wrt some basis $v_1, \dots, v_n$

$$\Rightarrow \operatorname{tr}(\tilde{A}) = \sum_{i=1}^{n} \lambda_i$$

- trace equals the negative of the coefficient in front of t^{n-1} in the char. polynomial

$$p_A(t) = t^n + \boxed{a_{n-1}}\, t^{n-1} + \dots +$$

Trace vs determinant

tr (A) = sum of eigenvalues

det (A) = product of eigenvalues

Riddle

Consider a real-valued matrix $A \in \mathbb{R}^{n \times n}$.

Over $\mathbb{C}$, we can always find eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ and bring the matrix in triangular form, $\tilde{A}$. Then:

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii} \in \mathbb{R}$$

$$\operatorname{tr}(\tilde{A}) = \sum \underbrace{\lambda_i}_{\in \mathbb{C}} \in \mathbb{C}$$

but the two are identical because trace is independent of basis, hence $\operatorname{tr}(\tilde{A}) \in \mathbb{R}$! ?!?

So even over $\mathbb{C}$, the trace always is a real number. Seems confusing...

Let's look at an example:

Example: Consider a rotation matrix

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

- A does not have any real eigenvalues.
- The trace is given as $2 \cdot \cos\theta$.
- The char. poly. of A is

$$p(t) = \det(A - tI) = \det\begin{pmatrix} (\cos\theta) - t & -\sin\theta \\ \sin\theta & (\cos\theta) - t \end{pmatrix}$$

$$= (\cos\theta - t)^2 + \sin^2\theta$$

$$= t^2 - 2\cos\theta \cdot t + \underbrace{\cos^2\theta + \sin^2\theta}_{=1}$$

$$= t^2 - (2\cos\theta) \cdot t + 1$$

- The roots of the char. pol.

$$\lambda_{1/2} = \frac{2\cos\theta \pm \sqrt{\boxed{(2\cos\theta)^2 - 4}}}{2}$$

$$4\cos^2\theta - 4 = 4(\cos^2\theta - 1) = 4(-\sin\theta)$$

eigenvalues $\in \mathbb{C}$ $\longrightarrow$ $= \cos\theta \pm i \cdot \sin\theta$

- The matrix has a diagonal representation

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

$$\operatorname{tr}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \cos\theta + i\sin\theta + \cos\theta - i\sin\theta$$

sum in $\mathbb{R}$ $\longrightarrow$ $= 2\cos\theta$

What is going on?

For any matrix $A \in \mathbb{C}^{n \times n}$, if $\lambda \in \mathbb{C}$ is an eigenvalue, then also $\bar{\lambda}$ is an eigenvalue.

Reason: the complex eigenvalues have their roots in solving quadratic equations, as in the last example:

$$\lambda_{1,2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

- If $b^2-4ac > 0$, eigenvalues $\in \mathbb{R}$
- If $b^2-4ac < 0$, then eigenvalues are

$$\lambda_{1,2} = \frac{-b \pm i\sqrt{4ac-b^2}}{2a} = \frac{-b}{2a} \pm i\sqrt{4ac-b^2}$$

so they are complex conjugates, and $\lambda_1 + \lambda_2 \in \mathbb{R}$.

Matrix series

Spectral radius

Def The spectral radius of a matrix $A \in \mathbb{R}^{n \times n}$ or $A \in \mathbb{C}^{n \times n}$ is defined as

$$\rho(A) := \max \{ |\lambda| \, ; \, \lambda \in \mathbb{C} \text{ eigenvalue of } A \}$$

(Note that even for real matrices, this definition looks at all complex eigs of A).

Proposition $\lim_{k\to\infty} A^k = 0 \iff \rho(A) < 1$

in Frobenius norm: $\|A^k - 0\|_F \to 0$

Proof "$\Rightarrow$" Let v the eigenvector of the eig λ that defines $\rho(A)$.

$$0 \overset{\text{ass.}}{=} \lim_{k\to\infty} A^k v \overset{\text{eig}}{=} \lim_{k\to\infty} \lambda^k v = \left(\lim_{k\to\infty} \lambda^k\right) v$$

This implies $\lim \lambda^k = 0$, thus $|\lambda| < 1$.

"$\Leftarrow$" Proof for symmetric matrices:

$A = U D U^t$, then $A^k = U D^k U^t$.

$$\| A \|_F = \| D \|_F \to 0.$$

For general matrices, need to exploit another normal form, the Jordan normal form. Skipped.

Neumann series

Prop • The series $\sum_{i=0}^{\infty} A^k$ converges if and only if $\rho(A) < 1$.

• If $\rho(A)<1$, then $(I-A)$ is invertible and

$$(I-A)^{-1} = \sum_{i=0}^{\infty} A^k$$

over $\mathbb{R}$:

$$\frac{1}{1-a} = \sum_{k=0}^{\infty} a^k$$

Proof intuition: for symmetric matrices easy, can apply matrix power to diagonal matrix D as above.

In general a bit more work, skipped.

Matrix exponential

For any matrix A the series

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!} = I + A + \frac{A^2}{2} + \dots$$

converges. It is called the matrix exponential.

The matrix $\exp(A)$ is always invertible and

$$\left(\exp(A)\right)^{-1} = \exp(-A)$$

Implementing matrix operations

Literature: Golub / van Loan : Matrix computations

Triangular matrices are great

If we have a matrix in triangular form, then many standard quantities can be easily computed:

- Solving a liner system : $Ax = b$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

  Starting from bottom: $x_3 = b_3 / a_{33}$,

  then plug in the second-low row , $x_2 = (b_2 - a_{23} x_3) / a_{22}$

  ...

- eigenvalues are the entries on the diagonal
- det (A) is the product of the diagonal entries

So many numerical algorithms are based on triangular matrices.

Linear system: theory

Solving linear systems is about the most basic task and occurs as a substep of more complex algorithms all over the place.

$$Ax = b$$

<u>Prop</u> $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

(i) The set of solutions of $Ax = 0$ is given by $\ker(A)$ and is a subspace of $\mathbb{R}^m$.

(ii) The system $Ax = b$ has at least one solution iff $b \in \text{Range}(A)$

(iii) If w is a solution of $Ax = b$, then the full set of solutions is given by $w + \ker(A) := \{ w + v \mid v \in \ker(A) \}$

Gauss algorithm to solve linear systems

Intuition:

- Take the first equation and use it to eliminate the first variable in all other equations.
- Then use the second equation to eliminate the second variable from all the remaining equations.
- and so on.
- At the end, we are left with an upper triangular matrix
- We then solve the system starting from the bottom...

Of course one can do clever tricks such as selecting the best row for replacing others (pivoting, rearranging, ...).

Computational complexity in general case: $O(n^3)$

LU decomposition

Idea: One can see that the Gaussian elimination algorithm implicitly does something more general.

Given a square matrix A, it decomposes A into a product

$$A = L \cdot U$$

where L is a lower triangular matrix and U an upper triangular matrix.

LU decomposition exists under certain conditions. In particular for symmetric pd matrices it always exists.

Computational complexity: $O(n^3)$

LU decomposition to solve a linear system

$O(n^3)$ Once we have a LU decomposition, the solution to the problem $Ax = b$ can then be found by solving two linear systems (trivial due to triang. form):

$$\left.\begin{array}{lll} O(n^2) & (1) & Ly = b \\ O(n^2) & (2) & Ux = y \end{array}\right\} \Rightarrow Ax = L\underbrace{Ux}_{y} \overset{(2)}{=} Ly \overset{(1)}{=} b.$$

LU decomposition to compute the inverse

To compute the inverse of a matrix, we have to find a matrix X that solves $A \cdot X = I$. This corresponds to solving n linear systems: $Ax_i = e_i$

Once we have the LU-decomposition of A, we can simply solve them.

$\mathcal{O}(n^3)$

Cholesky factorization

For the special case of symmetric, psd matrices, one can simplify the LU decomposition:

$$A = L L^t$$

Is numerically very stable, uses less memory than LU, and is a bit faster than LU (still $\mathcal{O}(n^3)$, but with better constants)

QR decomposition

Any matrix $A \in \mathbb{R}^{n \times n}$ can be decomposed as

$$A = QR$$

where Q is an orthogonal matrix and R is upper triangular.

Several algorithms exist, with different advantages/disadvantages.

Computational complexity: $O(n^3)$

QR decomposition to find orthogonal basis

In particular, if A has full rank (i.e., its columns form a basis of $\mathbb{R}^n$), then Q contains an orthonormal basis of $\mathbb{R}^n$.

More generally, the first k columns of Q form an orthonormal basis of the subspace spanned by the first columns of A.

QR decomposition to compute *all* eigenvalues of *dense* matrices A

Iterative procedure:

- $A^{(0)} =: A$
- For $k = 1, 2, \ldots$
  - Compute QR factorization of $A^{(k-1)}$:

    $$A^{(k-1)} = Q^{(k)} R^{(k)}$$

  - Recombine in reversed order:

    $$A^{(k)} := R^{(k)} Q^{(k)}$$

Note: $A_{k+1} = R_k Q_k = (Q_k^{-1} Q_k) R_k Q_k = Q_k^{-1} (Q_k R_k) Q_k = Q_k^{-1} A_k Q_k$

Under certain assumptions, A^k converges to a triangular matrix or even to a diagonal matrix (eg if A is symmetric).

Largest eigenvalue and eigenvector

Let $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \dots, \lambda_n$ such that

$|\lambda_1| > |\lambda_2| \dots > |\lambda_n|$.

Power method (vanilla version)

- start with a random vector x_0
- iteratively compute $$v_{k+1} = \frac{A \cdot v_k}{\| A v_k \|}$$

- If $|\lambda_1| > |\lambda_2|$, then v_k converges to the eigenvector v_1. The speed of convergence depends on the spectral gap $|\lambda_2| / |\lambda_1|$.

- Problematic if eigenspace has dim > 1, or if it is unknown whether A has an eigenvalue in the first place.

Iterative methods for sparse matrices

Many of the algorithms we have seen (LU, QR, ...) cannot really exploit sparsity of a matrix. Bad, in ML many matrices are very sparse.

Alternatively, one uses iterative methods that are based on matrix-vector-multiplications (example: power method)

Conjugate gradient method for linear systems of sparse symmetric pd matrices

- want to solve $Ax = b$
- Consider the minimization problem $\min \Phi(x)$ with

$$\Phi(x) = \frac{1}{2} x^t A x - x^t b.$$

Minimum is achieved by setting $x := A^{-1} b$.

- So we can find the solution x to our system $Ax = b$ by minimizing Φ.
- The gradient is $\nabla \Phi(x) = Ax - b$ and can be computed just with a matrix-vector product (good for sparsity). ...
- Now apply optimization methods (conjugate gradient descent).

Skipped material

Not treated this year, but the videos still exist if you are interested.

Skipped

Quotient spaces

Def Consider a set S. A subset $R \subset S \times S$ is called an equivalence relation on S if $\forall\, x, y, z \in S$:

(E1) $(x,x) \in R$ (reflexivity)

(E2) $(x,y) \in R \Rightarrow (y,x) \in R$ (symmetry)

(E3) $(x,y) \in R,\ (y,z) \in R \Rightarrow (x,z) \in R$ (transitivity)

Notation: $(x,y) \in R \Leftrightarrow x \sim y$

Example V VS, $W \subset V$ subspace.

$$v \sim u \;:\Leftrightarrow\; v - u \in W$$

Example: Consider the space $\mathcal{L}(\mathbb{R})$ of all functions $f : \mathbb{R} \to \mathbb{R}$ that are Lebesgue integrable. Define

$$f \sim g \;:\Leftrightarrow\; f = g \text{ almost everywhere}$$

Def The equivalence class of an element $a \in S$ under equivalence relation $\sim$ is defined as

$$[a] := \{ b \in S \mid b \sim a \}$$

Prop Two equivalence classes $[a]$ and $[b]$ are either identical or disjoint.

Consequence : An equ. relation on S results in a disjoint partition of equivalence classes.

Constructing quotient spaces:

V VS, $W \subset V$ subspace, equivalence relation

$$v \sim u \Leftrightarrow v - u \in W$$

Denote the equivalence classes as $[v]$.

Observe: The equ. classes have the form

$$[v] = v + W = \{ u \in S \mid \exists w \in W : u = v + w \}$$

$$= \{ v + w \mid w \in W \} \subset V$$

Define the quotient "space" as

$$V/W := \{ [v] \mid v \in V \}$$

$$[v], [u] \in V/W$$

$$[v] + [u] :\Leftrightarrow [v+u]$$

$$\lambda [v] :\Leftrightarrow [\lambda v]$$

These operations are <u>well-defined</u>:

- suppose $v' \sim v$ $\quad$ (i.e. $v' \in [v]$, $[v] = [v']$)

  $u' \sim u$

$$[v] + [u] \overset{?}{=} [v'] + [u']$$

$$v \sim v' \iff \exists\, w \in W \quad v - v' = w$$

$$u \sim u' \iff \exists\, \tilde{w} \in W : \; u - u' = \tilde{w}$$

$$\left.\begin{array}{l} [v] + [u] = [v+u] \\ [v'] + [u'] = [v'+u'] \end{array}\right\} ? \quad (v+u) \sim (v'+u')$$

$$(v+u) - (v'+u')$$

$$= \underbrace{(v-v')}_{w} + \underbrace{(u-u')}_{\tilde{w}} \in W$$

- similarly, for scalar mult.

$(V/W \,,\, + \,,\, \cdot\,)$ is a vector space: exercise.

Prop : Consider $g : V \to V/W$, $v \mapsto [v]$. Then:

- g is linear
- $\ker(g) = W$
- $\text{range}(g) = V/W$
- If V has finite dim, then $\dim V/W = \dim V - \dim W$.

SKIPPED

Characteristic polynomial

Motivation:

A $n \times n$ - matrix
$v \neq 0$

$$Av = \lambda v$$

$$\Leftrightarrow (A - \lambda I) v = 0$$

$$\Leftrightarrow v \in \ker (A - \lambda I)$$

$$\Leftrightarrow \operatorname{rank} (A - \lambda I) < n$$

$$\Leftrightarrow \det (A - \lambda I) = 0$$

Def The characteristic polynomial of an $n \times n$-matrix A is defined as

$$p_A(t) := \det(A - t \cdot I)$$

Example: $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$

$$\det(A - t \cdot I) = \det\left(\begin{pmatrix} a_{11} & -a_{12} \\ a_{21} & a_{22} \end{pmatrix} - t \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$

$$= \det \begin{pmatrix} a_{11} - t & a_{12} \\ a_{21} & a_{22} - t \end{pmatrix}$$

$$= (a_{11} - t)(a_{22} - t) - a_{12} \cdot a_{21}$$

$$= t^2 + t(-a_{11} - a_{22}) - a_{12} \cdot a_{21} + a_{11} \cdot a_{22}$$

Observations

- $p_A(t)$ is a polynomial with degree n
- Char. pol. does <u>not</u> depend on the basis:

<u>Proof</u> Consider A, basis transformation matrix U.

Want to look at char. pol. of UAU^{-1}.

$$\det\left(UAU^{-1} - t\cdot \underbrace{I}\right)$$

$$= \det\left(UAU^{-1} - t\cdot \overbrace{U\cdot U^{-1}}\right)$$

$$= \det\left(U\,(A - t\cdot I)\,U^{-1}\right)$$

$$= \det(U)\cdot\det(A - t\cdot I)\cdot\det(U^{-1})$$

$$= \det(A - tI)$$

- The roots of the characteristic poly. correspond exactly to the eigenvalues of A.

- Over $\mathbb{C}$, the char. poly. always has n roots, so the matrix has "n eigenvalues" (not nec. distinct).

- A is invertible $\Leftrightarrow$ 0 is not an eigenvalue.

  If 0 is an eigenvalue, e.v v with

  $$Av = 0 \cdot v = 0$$

  $\Leftrightarrow$ ker (A) non-trivial $\Leftrightarrow$ A not invertible

- Let $A \in \mathcal{L}(V)$, λ eig. of A. Then λ^k is an eig. of A^k.

- Let A be invertible; λ eig of A. Then $1/\lambda$ is an eig. of A^{-1}.

Def For an operator A with eigenvalue λ, we define its geometric multiplicity as the dimension of the corr. eigenspace $E(\lambda, A)$.

The algebraic multiplicity is the multiplicity of the root λ in the char. poly.

In general, the two notions do not coincide.

Computing eigs in theory

- Write down the char. pol., find the roots.

  ⇝ eigenvalues

- To compute the eigenvectors, solve the lin. system

  $A x = \lambda x$

In practice ... see later (numerics)

SKIPPED

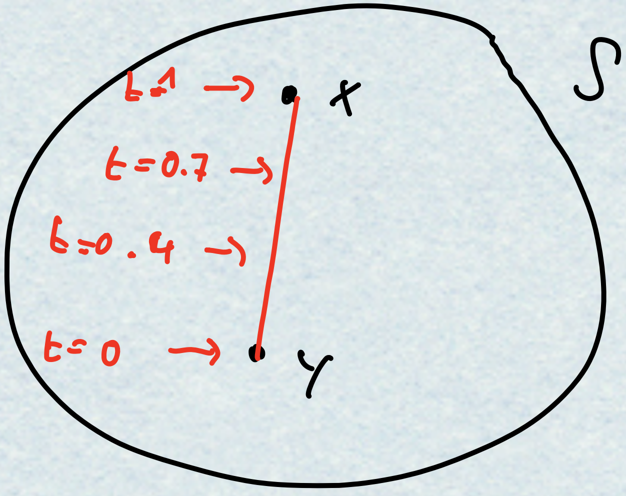
Norms can be characterized by convex sets

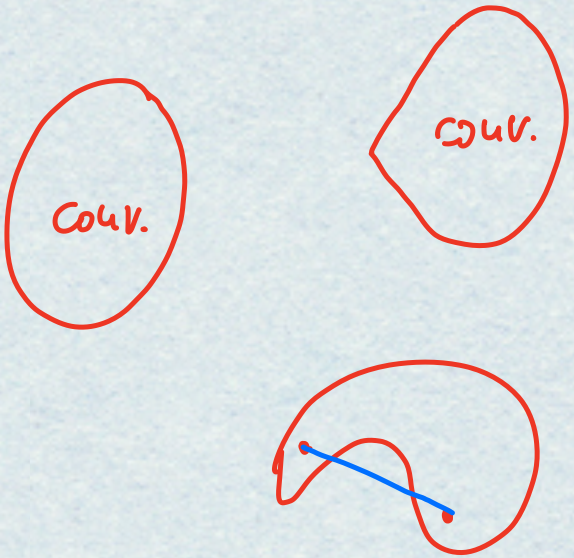
Convex set

<u>Def</u> Consider a real VS V, $S \subset V$. S is called **<u>convex</u>** if $\forall\, t$, $0 \le t \le 1$ and $\forall\, x, y \in S$,

$$t \cdot x + (1-t)\, y \in S \leftarrow$$

<u>Intuition</u>

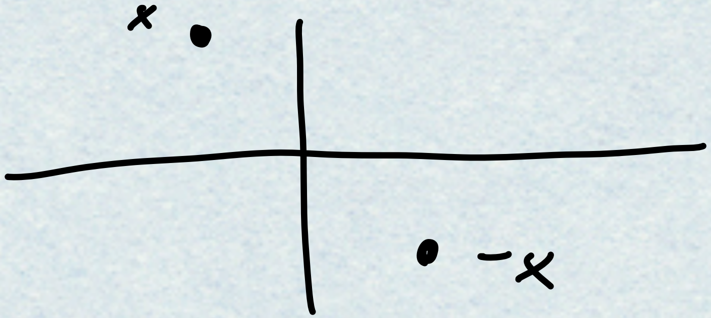


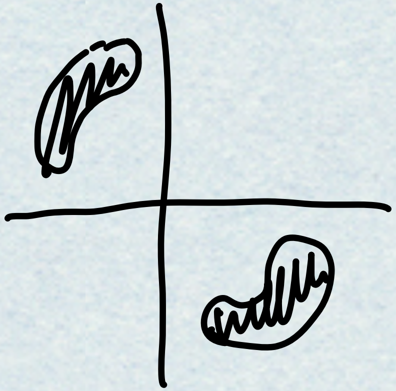


Symmetric set

<u>Def</u> A set $C \subset V$ is called symmetric if

$$x \in C \Rightarrow -x \in C$$





Convex sets induce norms

Theorem:

(1) Let $C \subset \mathbb{R}^d$ closed, convex, symmetric and has non-empty interior. Define

$$p(x) := \inf \left\{ t > 0 \;\middle|\; \frac{x}{t} \in C \right\}.$$

Then

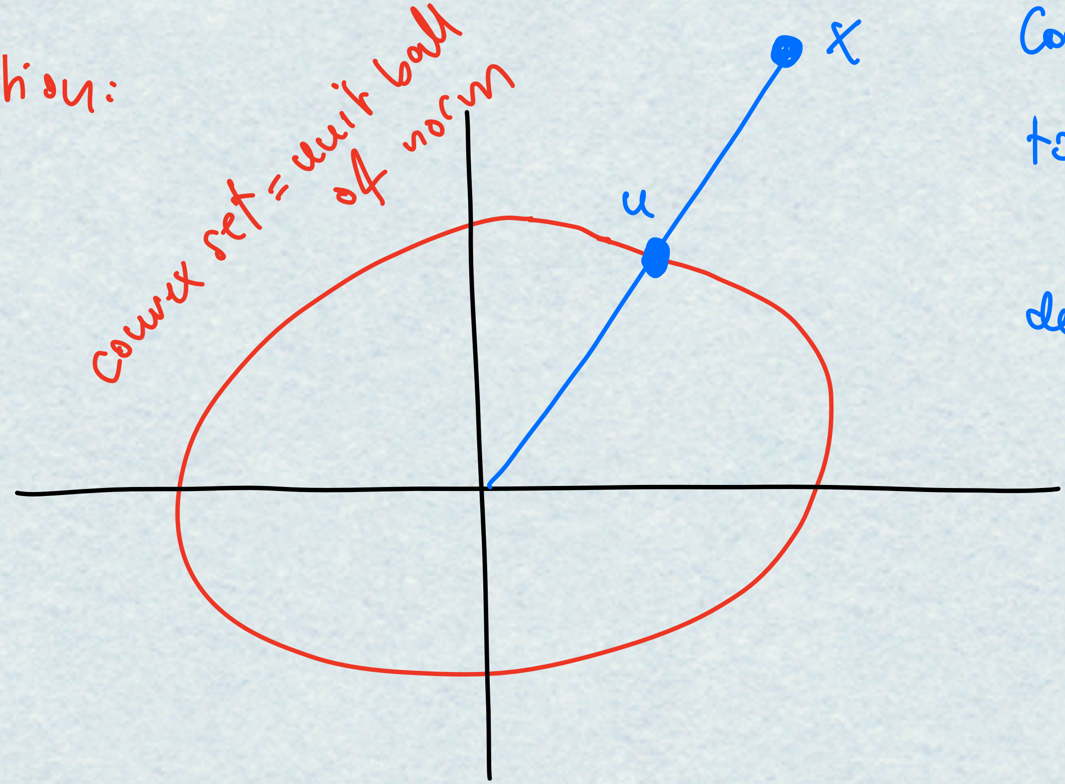
$$= \inf \{ t > 0 \mid x \in t \cdot C \},$$ this might be more intuitive.

p is a semi-norm. If C is bounded, then p is a norm, and its unit ball coincides with C:

$$C = \{ x \in \mathbb{R}^d \mid p(x) \leq 1 \}$$

(2) For any norm $\|\cdot\|$ on $\mathbb{R}^d$, the set $C := \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ is bounded, symmetric, closed, convex, and has non-empty interior.

Illustration: part (1)



Consider factor a by which I need to multiply x to end up on the unit sphere. Then define $\|x\| := 1/a$.

Proof :

The next couple of pages treat the proof of this theorem.

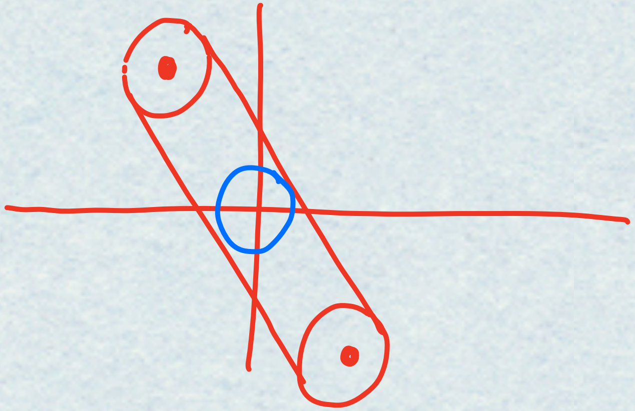
Proof p(x) is well defined

Want to prove: given $x \in \mathbb{R}^d$, the set $\{t > 0 \mid \frac{x}{t} \in C\} \neq \emptyset$

We are going to prove: $\exists \, \varepsilon > 0$ such that

$$B_\varepsilon(0) = \{e \in \mathbb{R}^n \mid \|e\|_2 < \varepsilon\} \subset C.$$

Intuition:



- By ass, C has at least one interior point

  $v \in C^\circ \Rightarrow \exists\, \varepsilon$ such that

  $B_\varepsilon(v) \subset C$

  $\shortparallel$

  $v + B_\varepsilon(0) = \{ v + e \mid e \in B_\varepsilon(0) \}$

- By symmetry, $v + e \in C \Rightarrow -(v+e) \in C$
- By convexity, $\frac{1}{2}(v+e) + \frac{1}{2}(-v-e) = e \in C$

So $B_\varepsilon(0) \subset C$, so the set $\{ t > 0 \mid \frac{x}{t} \in C \}$

is non-empty.

The infimum of $\inf \overbrace{\{ t > 0 \mid \frac{x}{t} \in C \}}^{S_x}$ exists

because $S_x \subset \mathbb{R}$, 0 is a lower bound.

Now we need to prove all axioms of a norm:

$$\boxed{p(0) = 0}$$

- Have seen: $0 \in C$
- $\forall t > 0$: $\frac{0}{t} = 0 \in C$
- $\inf \{ t \mid \frac{0}{t} \in C \} = 0$.

$\Rightarrow p(0) = 0$

$$\boxed{p(\alpha x) = |\alpha| \, p(x)}$$

- For all $\alpha > 0$ we have

$$p(\alpha \cdot x) = \inf\left\{ t > 0 \;\middle|\; \frac{\alpha x}{t} \in C \right\} = \leftarrow s := \frac{t}{\alpha}$$

$$= \inf\left\{ \alpha \cdot s > 0 \;\middle|\; \frac{x}{s} \in C \right\}$$

$$= \alpha \cdot \underbrace{\inf\left\{ s > 0 \;\middle|\; \frac{x}{s} \in C \right\}}_{p(x)}$$

$$\Rightarrow p(\alpha x) = \alpha \cdot p(x)$$

- By symmetry we also get

$$\begin{aligned} p(-x) &= \inf\left\{ t>0 \;\middle|\; \frac{-x}{t} \in C \right\} = \qquad -\frac{x}{t} \in C \Rightarrow \frac{x}{t} \in C \\ &= \inf\left\{ t>0 \;\middle|\; \frac{x}{t} \in C \right\} = p(x) \end{aligned}$$

- Combining the two statements gives homogeneity.

$\boxed{\triangle\text{-inequality}}$ Consider $x, y \in \mathbb{R}^d$, $s, t > 0$ such that

$$\frac{x}{s} \in C, \quad \frac{y}{t} \in C.$$

Observe: $\frac{s}{s+t} + \frac{t}{s+t} = 1$. Thus, by convexity,

$$\underbrace{\frac{s}{s+t}} \cdot \underbrace{\frac{x}{s}}_{\in C} + \underbrace{\frac{t}{s+t}} \cdot \underbrace{\frac{y}{t}}_{C} \in C \qquad \circledast$$

two scalars that sum up to 1

Want to prove:

$$p(x+y) = \inf\left\{u > 0 \;\middle|\; \frac{x+y}{u} \in C\right\}$$

$$\leq \underbrace{\inf\left\{s > 0 \;\middle|\; \frac{x}{s} \in C\right\}}_{p(x)} + \underbrace{\inf\left\{t > 0 \;\middle|\; \frac{y}{t} \in C\right\}}_{p(y)}$$

$$\frac{s}{s+t}\,\frac{x}{s} + \frac{t}{s+t}\,\frac{y}{t} \in C$$

$$\frac{x+y}{\underset{= u_0}{\boxed{s+t}}} \in C$$

$$p(x+y) = \inf\left\{ u > 0 \;\middle|\; \frac{x+y}{u} \in C \right\} \leq u_0$$

$$= \underbrace{s}_{p(x)} + \underbrace{t}_{p(y)}$$

s was chosen such that $\quad \frac{x}{s} \in C$

t $\quad \frac{y}{t} \in C$

Consider a sequence $(s_i)_{i \in \mathbb{N}}$ such that

$$\frac{x}{s_i} \in C \quad \text{and} \quad s_i \to p(x)$$

Similarly $(t_i)_{i \in \mathbb{N}}$ such that $\frac{y}{t_i} \in C$ and $t_i \to p(y)$.

By the argument above, we know that

$$\forall i : \quad p(x+y) \leq \underbrace{s_i}_{\searrow \, p(x)} + \underbrace{t_i}_{\searrow \, p(y)}$$

$$\Rightarrow \quad p(x+y) \leq p(x) + p(y).$$

$$\boxed{p(x) = 0 \Rightarrow x = 0}$$

$$p(x) = 0 \iff \inf\left\{ t \geq 0 \;\middle|\; \frac{x}{t} \in C \right\} = 0$$

$\Rightarrow$ There exists a sequence $(t_k)_{k \in \mathbb{N}}$ such that

$t_k \to 0$ and $\frac{x}{t_k} \in C$ $\forall k$.

Now assume that $x \neq 0$. Then the sequence

$\left(\frac{x}{t_k}\right)_{k \in \mathbb{N}}$ is unbounded. ↯ Contradiction because C is bounded.

∎

SKIPPED

Spaces of functions:

(continuous, differentiable, integrable)

$\mathcal{C}^b$ $\mathcal{C}^1$ L_p

Space of continuous functions

Def Let T be a metric space,

$$\mathcal{C}^b(T) := \{ f : T \to \mathbb{R} \mid f \text{ is continuous and bounded} \}$$

$\exists c \in \mathbb{R} :$
$\forall t \in T : |f(t)| < c$

As norm on $\mathcal{C}^b(T)$ we now use

$$\|f\|_\infty := \sup_{t \in T} |f(t)|$$

Prop Then the space $\mathcal{C}^b(T)$ with norm $\|\cdot\|_\infty$ is a Banach space: a complete, normed vector space.

Proof outline:

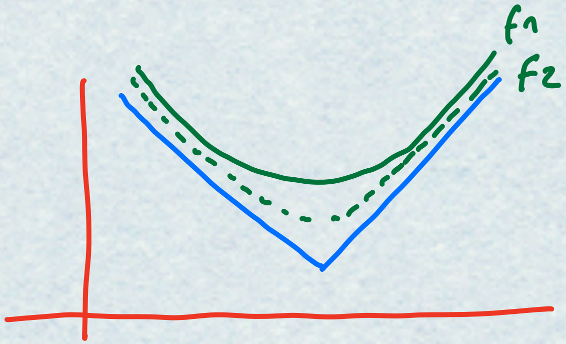
- need to check vector space axioms
- norm axioms
- completeness: follows from the fact that $\|\cdot\|_\infty$ induces uniform convergence

Space of differentiable functions

Let $[a,b] \subset \mathbb{R}$, $\mathcal{C}^1([a,b]) = \{ f : [a,b] \to \mathbb{R} \mid f$ is cont. differentiable$\}$

Which norm?

- Consider $\|\cdot\|_\infty$. With this norm, $\mathcal{C}^1$ is not complete!



limit function, not differentiable!

- Consider $\|f\| := \sup\limits_{t \in [a,b]} \max\{ |f(t)|, |f'(t)| \}$

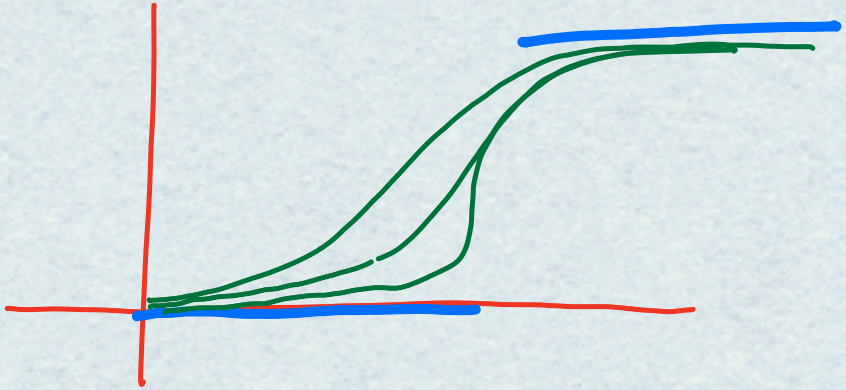
$$|||f||| := \|f\|_\infty + \|f'\|_\infty$$

$\mathcal{C}^1([a,b])$ with any of these two norms is a Banach space.

Spaces of integrable functions?

- Consider $\mathcal{C}^b([a,b])$ with the norm $\|f\|_1 := \int_a^b |f(t)|\,dt$

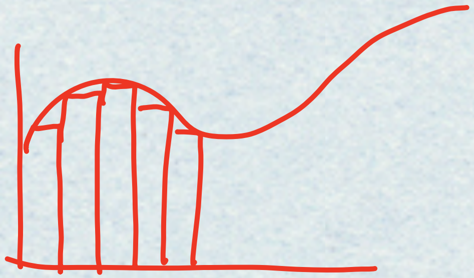
Can see: $\|\cdot\|_1$ is a norm, but the space is not complete.

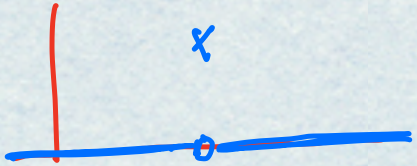


- Consider $R([a,b])$ of all Riemann-integrable functions on $[a,b] \subset \mathbb{R}$, together with $\|\cdot\|_1$.

However, on $R([a,b])$, $\|\cdot\|_1$ is not a norm: it is not true that

$$\|f\| = 0 \Rightarrow f = 0$$





$\int f\,dt = 0$ but $f \not\equiv 0$

The space $\mathcal{L}_p$

<u>Def</u> For $1 \leq p < \infty$, we define

$\mathcal{L}_p([a,b]) := \{ f : [a,b] \to \mathbb{R} \mid f$ measurable wrt the Lebesgue measure and $\int |f(t)|^p \, dt < \infty \}$

$\|f\|_p := \left(\int |f|^p \, d\lambda \right)^{1/p}$

Proposition 1 : $\|f\|_p$ is a semi-norm on $\mathcal{L}_p$.

Proof :

- Vector space (clear)
- semi-norm : observe: $\|f\|_p = 0 \Rightarrow f = 0$ almost everywhere

so we do *not* have that $\|f\|_p = 0 \Rightarrow f = 0$. a.e. $\square$

⚠ $\|f\|_p$ is not a norm! For example, the function

$$f(x) = \begin{cases} 1 & \text{if } x = 17 \\ 0 & \text{otherwise} \end{cases}$$

has integral 0, but is not the 0-function.

Proposition 2 $\mathcal{L}_p$ is complete under $\|\cdot\|_p$.

Proof: If $(f_i)_{i \in \mathbb{N}}$ is a Cauchy-sequence in $\mathcal{L}_p$, then want to prove that $\lim\limits_{i \to \infty} f_i \in \mathcal{L}_p$.

This is equivalent to proving the following:

Let $(f_i)_i$ be a sequence such that

$$a := \sum_{i=1}^{\infty} \|f_i\|_p < \infty \qquad \circledast$$

Then there exists $f \in \mathcal{L}_p$ such that $f_i \to f$ (in $\|\cdot\|_p$).

Define

$$\hat{g} := \sum_{i=1}^{\infty} |f_i|$$

Note: this might not yet be a well-defined fct from $[a,b]$ to $\mathbb{R}$, might be ∞ at certain points.

$$\hat{g}_n := \sum_{i=1}^{n} |f_i| \in \mathcal{L}_p$$

By Minkowski,

$$\| \hat{g}_n \|_p \overset{def}{=} \Big\| \sum_{i=1}^{n} |f_i| \Big\|_p \overset{Mink.}{\leq} \sum_{i=1}^{n} \| f_i \|_p \overset{ass. *}{<} a$$

$\hat{g}_n \to g$ monotonously

By theorem of monotonic convergence, g is measurable

and we have

$$\lim_{n\to\infty} \int \hat{g}_n^{\,p} \, d\lambda = \int \lim \hat{g}_n^{\,p} \, d\lambda$$

$$\overset{\text{by def}}{=} \int \hat{g}^p \, d\lambda$$

$$\leq a^p .$$

$\Rightarrow$ $\hat{g} < \infty$ a.e., that is there exists a set N of measure 0 such that on $[a,b] \setminus N$, $\hat{g}$ is finite.

Now we can define

$$g(t) = \begin{cases} \hat{g}(t) & t \in [a,b] \setminus N \\ 0 & t \in N \end{cases} \qquad \in \mathcal{L}_p$$

From this it now follows that $f(t) := \sum_{i=1}^{\infty} f_i(t)$, $t \notin N$

exists. For $t \in N$, we set $f(t) = 0$.

Now f is measurable, and in $\mathcal{L}_p$

because $\int |f|^p \, d\lambda \leq \int \hat{g}^p \, d\lambda < \infty$

Finally, $\sum^{\infty} f_n$ converges to f in $\|\cdot\|_p$

because of the theorem of dominated convergence. $\blacksquare$

From $\mathcal{L}_p$ to L_p

We constructed a space $\mathcal{L}_p$ with the Lebesgue integral as a semi-norm. This means,

given $f \in \mathcal{L}_p$, we can change the p values of f in a set of measure 0, resulting in $\tilde{f}$, but the norm "does not see a difference":

$$\| f - \tilde{f} \| = 0$$

To fix this, we want to consider functions to be "equivalent" if they only differ by a set of measure 0.

The formal construction goes as follows:

Define $N := \ker(\| \cdot \|_p) := \{ f \in \mathcal{L}_p \mid \|f\|_p = 0 \}$

is a subspace of $\mathcal{L}_p$.

Now consider the quotient space of $\mathcal{L}_p$ wrt this subspace:

$$L_p([a,b]) := \mathcal{L}_p([a,b]) / N$$

Define a norm on L_p by $\| \underbrace{[f]}_{\text{new}} \|_p = \underbrace{\|f\|}_{\text{old}}{}_p$

This norm is well-defined: if $f, \tilde{f} \in [f]$,

then $\|f\|_p = \|\tilde{f}\|_p$.

This "norm" is a norm, because

$$\|[f]\|_p = 0 \Rightarrow [f] = [0].$$

Conclusion: L_p with $\|\cdot\|_p$ is a Banach space!

For simplicity, in future we write $\|f\|_p$ for $\|[f]\|_p$.

Elements ("functions") in L_p are equivalence classes $[f]$ consisting of all functions that coincide a.e.

⚠

- It does not make sense to evaluate $f(0)$ because $\{0\}$ has Lebesgue measure 0.
- quite annoying for machine learning, where we always want to evaluate functions on input points.
- often use alternative spaces instead, for example reproducing kernel Hilbert spaces.

SKIPPED

Operator norm

Continuous = bounded

Theorem X, Y normed spaces, $T: X \to Y$ linear. Then the following statements are equivalent:

(i) T is continuous. → $\left\{ \begin{array}{l} \forall x \in X \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall y \in X: \\ \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon \end{array} \right.$

(ii) T is continuous at 0.

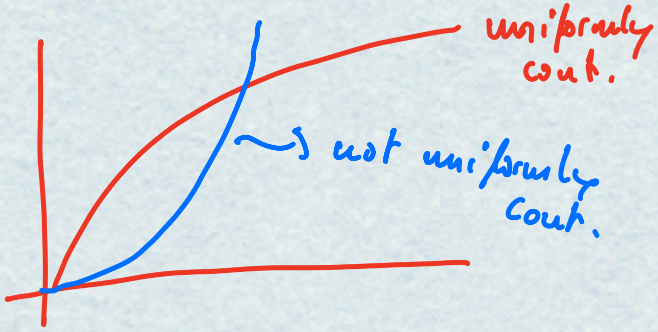
(iii) T is bounded:

$$\exists M > 0 \ \forall x \in X: \|Tx\| \leq M \cdot \|x\|$$

(iv) T is uniformly continuous.

$\forall \varepsilon > 0 \ \exists \delta > 0 \ \forall x \in X \ \forall y \in X:$

$\|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon$



Definition

Def X, Y normed spaces, $T: X \to Y$ linear and continuous.

$$\|T\| := \sup_{x \in X} \frac{\|Tx\|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\| \leq 1}} \|Tx\| = \sup_{\substack{x \in X \\ \|x\| = 1}} \|Tx\|$$

is called the **operator norm** of T.

Observe: coincides with the matrix norm $\|\cdot\|_2$ as we had defined it earlier.

Examples

- <u></u>**Evaluation operator**: $T : \mathcal{C}[0,1] \longrightarrow \mathbb{R}$, $Tf = f(0)$.

  As norms consider $\|\cdot\|_\infty$ on $\mathcal{C}[0,1]$, $|\cdot|$ on $\mathbb{R}$. Then

  $\|T\| = 1$.

$$\sup_{f \in \mathcal{C}[0,1]} \frac{|Tf|}{\|f\|_\infty} = \sup \frac{|f(0)|}{\|f\|_\infty} = \underline{\text{exercise}} = 1$$

- **Integral operator**: $T : \mathcal{C}[0,1] \to \mathbb{R}$, $Tf = \int_0^1 f(t)\,dt$

  With the same norms as above, T is cont. and has

  $\|T\| = 1$.

Examples

- Differential operator: $D : \mathcal{C}^1[0,1] \to \mathcal{C}[0,1]$, $f \mapsto f'$.
  - Consider $\|\cdot\|_\infty$ on $\mathcal{C}^1$ and $\mathcal{C}$. Then D is linear, but not continuous!
  - Consider $|||f||| := \|f\|_\infty + \|f'\|_\infty$ on $\mathcal{C}^1$. With this norm, D is continuous and bounded.

SKIPPED

Dual space

Dual space

Definition V VS, $T : V \to F$ linear is called a functional.

Given a vector space V, the algebraic dual space V^* consists of all linear functionals on V:

$$V^* := \mathcal{L}(V, F).$$

If V is a normed VS, then the space of all linear, continuous functionals from V to F is called the (topological) dual space V' of V.

Remark If V is finite dim, then $V^* = V'$ because then linear mappings are always continuous. In general, this is not true.

We endow the dual space with the operator norm

$$\|T\| := \sup_{x \in X} \frac{\|Tx\|}{\|x\|} .$$

Proposition: V' is a vector space, and the operator norm is indeed a norm on V'.

Prop: If V is a normed VS (but not necessarily complete), then V' with the operator norm is a Banach space.

Examples

- $K \subset \mathbb{R}$ compact set, $\mathcal{C}(K)$ space of cont. fcts with $\|\cdot\|_\infty$. Then $(\mathcal{C}(K))'$ is equivalent to the space $M(K)$, the space of all (Radon) measures over K.

- $S \subset \mathbb{R}$ measurable set, $1 \leq p < \infty$, q such that $\frac{1}{p} + \frac{1}{q} = 1$. Then: the dual of $L_p(S)$ is given as $L^q(S)$.

Riesz representation theorem

Theorem: H Hilbert space, H' its dual. Then the mapping $\Phi : H \to H'$, $y \mapsto \langle \cdot , y \rangle$ is bijective, isometric, and satisfies $\Phi(\lambda x) = \bar{\lambda}\, \Phi(y)$.

Stated differently: for any mapping $x' \in H'$ there exists a unique $y \in H$ such that $x'(x) = \langle x, y \rangle$.

Adjoint operator

Definition

Def Let $T \in \mathcal{L}(H_1, H_2)$, H_1, H_2 Hilbert spaces. Then there exists an operator $T^*: H_2 \to H_1$ such that

$$< Tx, y >_{H_2} = < x, T^* y >_{H_1} .$$

for all $x \in H_1$, $y \in H_2$. T^* is called the adjoint of T.

Remark The existence of this operator is a consequence of the Riesz representation theorem.

Def An operator $T: H_1 \to H_1$ is called self-adjoint if $< Tx, y > = < x, Ty >$