Mathematics for Machine Learning

Ulrike von Luxburg

Winter 2026/27
Department of Computer Science, University of Tübingen

(Version as of October 29, 2026)

Table of contents

Part I

Linear Algebra 5

Lecture 1 (after Intro)

Groups and Fields 6
Complex Numbers 11
Vector space 20

Lecture 2

Basis and dimension 25
Linear Mappings 41

Lecture 3

Matrices and linear maps 53
Invertible maps and matrices 63
Change of basis 72

Lecture 4

Rank of a matrix 80
The determinant 84
Eigenvalues, eigenvectors, eigenspaces 100

Lecture 5

Characteristic polynomial 113
Diagonalization and Triangulization 120

Table of contents (2)

Lecture 6

Normed spaces . . . 136
All norms on $\mathbb{R}^n$ are equivalent (SKIPPED) . . . 146
Norms can be characterized by convex sets (SKIPPED) . . . 152
Scalar product . . . 164
Orthonormal basis and orthogonal projections . . . 173
Orthogonal matrices . . . 185
Symmetric matrices . . . 192
Positive definite matrices . . . 207
Variational characterization of eigenvalues . . . 218
Matrix norms . . . 229
Singular value decomposition . . . 235
Pseudo inverse . . . 254
Trace of a matrix . . . 261
Matrix series . . . 271
Implementing matrix operations . . . 277
Spaces of functions: Continuous, differentiable, integrable SKIPPED . . . 292
Operator norm SKIPPED . . . 310
Dual space SKIPPED . . . 316
Adjoint operator SKIPPED . . . 322

Table of contents (3)

Quotient space SKIPPED 325

Part I
Linear Algebra

Groups and Fields

Definition of a group

Definition 1 (Group)

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

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

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

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

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

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

Examples of groups

- $(\mathbb{R}^n, +)$, $(\mathbb{R}^+, \cdot)$ are groups.
- $(\mathbb{R}^-, \cdot)$ is not a group (why?)
- Set $S_n := \{\pi : \{1, \dots, n\} \to \{1, \dots, n\} \mid \pi \text{ is bijective}\}$
  Operation $\circ : S_n \times S_n \to S_n, \quad \pi_1 \circ \pi_2(i) = \pi_1(\pi_2(i))$
  Together, $(S_n, \circ)$ is a group (called the symmetric group with n elements).

Definition of a field

Definition 2 (Field)

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 identity element 1.

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

Examples of fields

- The real numbers with standard addition and multiplication: $(\mathbb{R}, +, \cdot)$
- The complex numbers (see also below): $(\mathbb{C}, +, \cdot)$
- For $n \in \mathbb{Z}$ consider the set $\mathbb{Z}_n := \{0, 1, \ldots, n-1\}$
  Define addition: $a +_n b := (a + b) \mod n$
  Define multiplication: $a \cdot_n b := (a \cdot b) \mod n$
  Then $(\mathbb{Z}_n, +_n, \cdot_n)$ is a field if and only if n is prime.

In this lecture, we will only deal with the fields $\mathbb{R}$ and $\mathbb{C}$.

Complex Numbers

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 use them to factorize polynomials.

Here are the very basics:

Quadratic equations over $\mathbb{R}$

Quadratic equation: For given parameters $a, b, c \in \mathbb{R}$, we 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, let's give it a name: $i := \sqrt{-1}$ (i for "imaginary number").

We can now define a new field of numbers, the complex numbers:

Definition 3 (Complex number)

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:

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

Imagine... (2)

We define addition and multiplciation in the natural way:

$$(a+bi)+(c+di) := (a+c)+(b+d)i$$
$$(a+bi)\cdot(c+di) := (ac-bd)+(ad+bc)i$$

With these operations, $\mathbb{C}$ 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 solutions $x_{1,2} \in \mathbb{R}$ "as usual".

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

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

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

This leads to solutions in $\mathbb{C}$.

Fundamental theorem of algebra

Theorem 4 (Fundamental theorem of algebra)

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). Assume that $a_n \neq 0$, so we have a polynomial of degree n.

Then, any such polynomial has exactly n roots $\gamma_1, \ldots, \gamma_n \in \mathbb{C}$ (not necessarily distinct) such that

$$a_0 + a_1 x + \cdots + a_n x^n = a_n (x - \gamma_1)(x - \gamma_2) \cdots (x - \gamma_n).$$

More notation

For a complex number $x = a + ib$ we define its "complex conjugate" by $\bar{x} := a - ib$.

Outlook

Dealing with complex numbers is a rich field in mathematics, but we will not touch it any futher.

Vector space

Definition of a vector space

Definition 5 (Vector space)

Let F be a field with operations $\oplus, \odot$ with identity elements 0 and 1. A vector space over the field F is a set V with two mappings

$+ : V \times V \to V$ ("vector addition")

$\cdot : F \times V \to V$ ("scalar multiplication") such that:

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

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

(V3) Associativity of scalar multiplication: $\forall a, b \in F$ and $\forall v \in V\colon\ (a \odot b) \cdot v = a \cdot (b \cdot v)$.

(V4) Distributive properties: $\forall a, b \in F$ and $\forall u, v \in V$:

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

(Careful, we have four different operations: $\oplus, \odot$ on F and $+, \cdot$ on V)

Vector space

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

A vector space is sometimes called a real vector space if $f = \mathbb{R}$ and a complex vector space if $F = \mathbb{C}$.

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 functions on a set $\mathcal{X}$) with operations

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

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

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

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

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

Subspaces

Definition 6 (Subspace)

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:

$$\forall \lambda, \mu \in F \; \forall u, v \in U : \; \lambda \cdot u + \mu \cdot v \in U.$$

Examples:

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

Basis and dimension

Linear combinations

Definition 7 (Linear combination)

Let V be a vector space over a field F, $u_1, \ldots, u_n \in V$ and $\lambda_1, \ldots, \lambda_n \in F$. Then $\sum_{i=1}^{n} \lambda_i u_i$ is called a linear combination.

The set of all linear combinations of $u_1, \ldots, u_n$ is called the span (linear hull) of $u_1, \ldots, u_n$:

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

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

Linear independence

Definition 8 (Linear independence)

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

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

Linear independence (2)

Examples:

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

Basis of a vector space

Definition 9 (Basis)

A subset B of a vector space V is called a (Hamel) basis if:

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

(B2) B is linearly independent

Examples

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

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0.5\\0.8\\0.4\end{pmatrix}, \begin{pmatrix}1.8\\0.3\\0.3\end{pmatrix}, \begin{pmatrix}-2.2\\-1.3\\3.5\end{pmatrix}$$

Basis of a vector space (2)

Does every vector space have a basis? Not obvious!

We proceed in two steps:

- We first consider vector spaces for which we assume that a finite basis exists.
- Later we briefly touch on the general case.

Finite-dimensional vector spaces

Definition 10 (finite-dimensional vector space)

A vector space is called finite-dimensional if it has a finite basis.

Can you come up with an example of a space that is not finite-dimensional?

(Note that we have not have defined the dimension of a space. So far we can only define the notion of "finite-dimensional".)

Reducing a set to a basis

ML Keywords: dictionary learning

Proposition 11 (Reduction of a set to a basis)

If $U = \{u_1, \ldots, u_n\}$ spans a vector space V, then the set U can be reduced to a basis of V.

Proof sketch

- If U is already linearly independent: done.
- If U is linear dependent:
  $\exists \lambda_1, \ldots, \lambda_n$ not all 0 such that $\sum_{i=1}^{n} \lambda_i u_i = 0$.
  Pick some k with $\lambda_k \neq 0$. Then:
  $$u_k = -\sum_{i \neq k} \frac{\lambda_i}{\lambda_k} u_i.$$
  Denote $\widetilde{U} := U \setminus \{u_k\}$. It is clear that $\operatorname{span}(\widetilde{U}) = \operatorname{span}(U)$.

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

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

Extending a set to a basis

Proposition 12 (Extension of a set to a basis)

Let $U = \{u_1, \dots, u_n\} \subset V$ be a set of linearly independent vectors and let V be a finite-dimensional vector space. Then U can be extended to a basis of V.

Proof sketch. Let $w_1, \dots, w_m$ be a basis of V (exists by definition of finite-dim). Consider the set $\{u_1, \dots, u_n, w_1, \dots, w_m\}$. Remove vectors "from the end" until the remaining vectors are linearly independent.

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

☺

Two finite bases have the same length

Corollary 13 (Length of basis of same vector space)

Let V be a finite-dimensional vector space. Then any two bases of V have the same length.

Proof.

- Let $B := \{b_1, \ldots, b_n\}$ and $C := \{c_1, \ldots, c_m\}$ be two bases, $n \leq m$. Consider the set $\{b_1, \ldots, b_n, c_1\}$. By construction it is linear dependent.
- By the same procedure as before, we can find a vector b_i such that $\{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n, c_1\}$ is linearly independent.
- We now apply this procedure iteratively: add vectors from C, remove vectors from B. Note that in each of the iterations, it is really possible to remove one of the b_i-vectors (because the c_i-vectors that we already added are linearly independent). We keep on iterating until we removed all the vectors b_i.

Two finite bases have the same length (2)

- At the end, this results in a set of n vectors $\{c_{i_1}, \ldots, c_{i_n}\} \subset \{c_1, \ldots, c_m\}$. By construction, this set is a basis (linearly independent, spans V) and has n elements. Adding one more vector would turn the set linearly dependent. So we found a subset $\{c_{i_1}, \ldots, c_{i_n}\} \subset C$ that is already a basis, and any any more vector will turn it linearly dependent. But because C is a basis as well, it needs to be the case that $\{c_{i_1}, \ldots, c_{i_n}\} = C$, thus $n = m$.

☺

Dimension of a vector space

Finally we are ready to define:

Definition 14 (Dimension of a vector space)

The length of a basis of a finite-dimensional vector space V is called the dimension of V.

Does every vector space have a basis?

What if a VS is not finite-dimensional?

One can prove that every VS has a basis, but this needs the axiom of choice / Zorn's lemma, which is deep maths beyond this lecture.

Just for those who are curious about it, the argument goes as follows:

Consider all linearly independent subsets of V, ordered by inclusion. Zorn's lemma states that if each ordered chain has an upper bound (a set of independent vecotrs that contains all subsets in the chain), then there exists a maximal set B. In our case, the upper bound of a chain can be constructed as the union of all subsets in the chain. It is still linearly independent, as linear independence is only defined for finite linear combinations. This set also spans V because otherwise we could extend it further, so it would not have been maximal. So B is a basis.

SKIPPED

Sum of vector spaces

Definition 15 (Sum of vector spaces)

Let U_1, U_2 be 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. In this case we denote the sum by $U_1 \oplus U_2$.

Complement of a subspace

Proposition 16 (Complement of a subspace)

Suppose V is finite-dimensional, and $U \subsetneq V$ is a subspace, $U \neq \{0\}$. Then there exists a subspace $W \subsetneq V$ such that $U \oplus W = V$.

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

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

Define

$$W = \mathsf{span}\{v_1, \ldots, v_m\}.$$

Does the job (one needs to prove it). ☺

Linear Mappings

Linear mapping

Definition 17 (Linear mapping)

Let U, V be a 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 by $\mathcal{L}(U, V)$. If $U = V$, then we write $\mathcal{L}(U)$.

Linear mapping (2)

Examples:

- $T : \mathcal{C}[a,b] \to \mathbb{R}, \quad f \mapsto \int_a^b f(x)\,dx$ (Integration)
- $D : \mathcal{C}^\infty[a,b] \to \mathcal{C}^\infty[a,b], \quad f \mapsto f'$ (Differentiation)

Linear mapping (3)

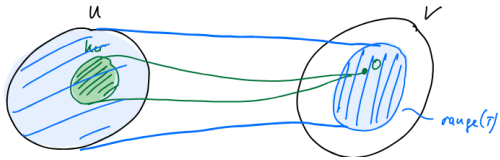
Definition 18 (Kernel and range)

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

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

The range or 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 19 (Properties of kernel and range)

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

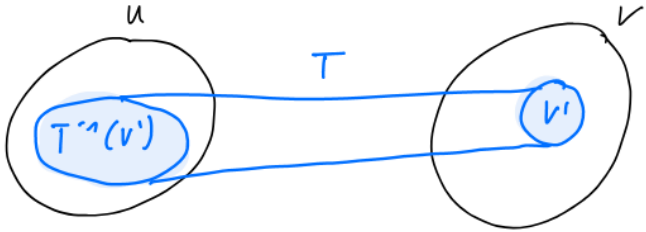
Proof sketch. Exercise. ☺

Pre-image

Definition 20 (Pre-image)

$T \in \mathcal{L}(U, V)$, $V' \subset V$, V' any set. The pre-image of V' is defined as

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



Pre-image (2)

Proposition 21 (Pre-image subspace)

If $V' \subset V$ is a subspace of V, then $T^{-1}(V')$ is a subspace of U.

Proof sketch. Exercise. ☺

Fundamental theorem for linear mappings

Theorem 22 (Fundamental theorem for linear mappings)

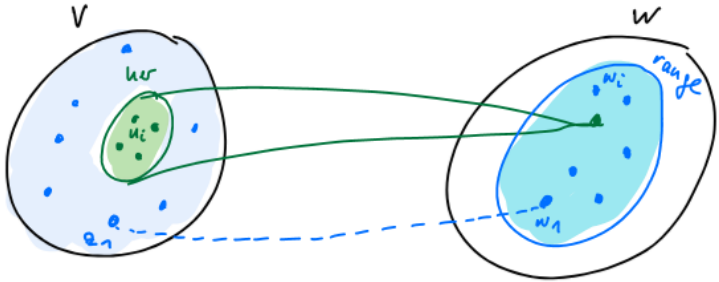
Let V be finite-dimensional, W any VS, $T \in \mathcal{L}(V, W)$.
Let $u_1, \ldots, u_n$ be a basis of $\ker(T) \subset V$.
Let $w_1, \ldots, w_m$ be a basis of range$(T) \subset W$.
Let $z_1 \in T^{-1}(w_1), \ldots, z_m \in T^{-1}(w_m)$.

Then $u_1, \ldots, u_n, z_1, \ldots, z_m \subset V$ form a basis of V.

In particular,

$$\dim(V) = \dim(\ker(T)) + \dim(\text{range}(T)).$$

Fundamental theorem for linear mappings (2)



Proof of the fundamental theorem for linear mappings

Step 1: $V \subset \operatorname{span}\{u_1, \ldots, u_n, z_1, \ldots, z_m\}$

Let $v \in V$, consider $Tv \in \operatorname{range}(T)$. By definition of the w_i, there exist $\lambda_1, \ldots, \lambda_m$ such that

$$
\begin{aligned}
Tv &= \lambda_1 w_1 + \ldots + \lambda_m w_m \\
&= \lambda_1 T(z_1) + \ldots + \lambda_m T(z_m) \\
&= T(\lambda_1 z_1 + \ldots + \lambda_m z_m) \\
\Rightarrow Tv - T(\lambda_1 z_1 + \ldots + \lambda_m z_m) &= 0 \\
&= T(\underbrace{v - (\lambda_1 z_1 + \ldots + \lambda_m z_m)}_{\in \ker(T)}) \\
\Rightarrow \exists \mu_1, \ldots, \mu_n \text{ s.t. } & v - (\lambda_1 z_1 + \ldots + \lambda_m z_m) = \mu_1 u_1 + \ldots + \mu_n u_n \\
\Rightarrow v &= \lambda_1 z_1 + \ldots + \lambda_m z_m + \mu_1 u_1 + \ldots + \mu_n u_n
\end{aligned}
$$

Proof of the fundamental theorem for linear mappings (2)

Step 2: $u_1, \ldots, u_n, z_1, \ldots, z_m$ are linearly independent

Assume that $\mu_1 u_1 + \ldots + \mu_n u_n + \lambda_1 z_1 + \ldots + \lambda_m z_m = 0$. (*)

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

we add something that is 0 (because in kernel)

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

exploit linearity

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

$$\Rightarrow \lambda_1 w_1 + \ldots + \lambda_m w_m = 0 \underset{w_1,\ldots,w_m \text{ basis}}{\Rightarrow} \lambda_1 = \ldots = \lambda_m = 0$$

$$\Rightarrow \mu_1 u_1 + \ldots + \mu_n u_n = 0 \text{ by (*)}$$

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

□

Finite-dim: Injective $\Leftrightarrow$ Surjective $\Leftrightarrow$ Bijective

Proposition 23 (Equivalent properties of finite-dim. linear mapping)

$T \in \mathcal{L}(V, V)$, V finite-dimensional. Then the following three statements are equivalent:

(i) T is injective.

(ii) T is surjective.

(iii) T is bijective.

Proof sketch. Direct consequence of fundamental theorem. (Can you see why? Exercise!)

☺ ⚠ Does not work in ∞-dim spaces!

Matrices and linear maps

Recap matrix notation

- Matrix $A \in \mathbb{R}^{m \times n}$ with entries a_{ij}:

$$A := \underbrace{\left.\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}\right\}}_{\text{n columns}} \text{ m rows}$$

- For matrix $A \in \mathbb{R}^{n \times n}$, vector $x \in \mathbb{R}^n$, the matrix-vector multiplication $Ax \in \mathbb{R}^n$ is defined component-wise by $(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j$.
- For two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$ the product $AB \in \mathbb{R}^{n \times k}$ is defined component-wise by $(AB)_{ij} = \sum_{s=1}^{m} a_{is} b_{sj}$.

But how exactly are these related to linear mappings?

Matrices represent linear maps

Consider V, W finite-dimensional, $v_1, \ldots, v_n$ be a basis of V, and $w_1, \ldots, w_m$ be a basis of W, and $T \in \mathcal{L}(V, W)$. We now want to derive a concise notation for this linear map, a matrix:

- If we know the results of T applied to the basis vectors v_j, then we can express $T(v)$ for arbitrary $v = \lambda_1 v_1 + \ldots + \lambda_n v_n \in V$ by

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

- For each basis vector v_j, we now express the image $T(v_j)$ in basis $w_1, \ldots, w_m$: There exist coefficients $a_{1j}, \ldots, a_{mj}$ s.t.

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

Matrices represent linear maps (2)

- We now store these coefficients in a matrix called $\mathcal{M}(T)$:

$$\mathcal{M}(T) := \underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix}}_{\text{n cols, one for each basis vector of V}} \Bigg\} \begin{array}{l} \text{m rows,} \\ \text{one for each} \\ \text{basis vector of } W \end{array}$$

- This matrix describes the linear mapping T w.r.t. the bases $v_1, \ldots, v_n$ of V and $w_1, \ldots, w_m$ of W.

Matrices represent linear maps (3)

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

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

The red term is the i-th entry of the product of matrix $\mathcal{M}(T)$ with vector λ.

Matrices represent linear maps (4)

Notation:

Let $T : V \to W$ be linear, let $\mathcal{B}$ be a basis of V, and $\mathcal{C}$ a basis of W. We denote by $\mathcal{M}(T, \mathcal{B}, \mathcal{C})$ the matrix corresponding to T w.r.t. bases $\mathcal{B}$ and $\mathcal{C}$.

If it is clear which bases we are using, we often drop the bases from the notation and simply write $\mathcal{M}(T)$.

Properties of matrices

Let V, W be vector spaces, and consider the bases fixed. Let $S, T \in \mathcal{L}(V, W)$. Then:

- $\mathcal{M}(S+T) = \mathcal{M}(S) + \mathcal{M}(T)$
- $\mathcal{M}(\lambda S) = \lambda \mathcal{M}(S)$
- If $T : U \to V$, $S : V \to W$ linear, then composition of linear mappings correspond to matrix multiplication:

$$\mathcal{M}(S \circ T) = \mathcal{M}(S) \cdot \mathcal{M}(T).$$

Notation for particular matrices

Identity matrix: $\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

Notation: I, I_n, Id, ...

Diagonal matrix: $\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

Upper triangular matrix: $\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

Lower triangular: ...

Matrix transpose

Definition 24 (Matrix transpose)

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

$$(A^t)_{ij} = A_{ji}.$$

Notation: differs, typical are A^t, $A^\intercal$, $A^\top$, A', ...

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

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

Linear mapping corresponding to the matrix transpose

The transpose matrix also has an interpretation as a transpose linear mapping, but it is beyond of our current knowledge to really understand it (I just put it here for completeness in case somebody wants to know the keywords):

- To each vector space V we can associate its "dual space" V^* that consists of all linear mappings from V to K.
- Given a linear map $T : V \to W$, the transpose is the linear map $T^* : W^* \to V^*$ such that $(T^* f)(v) = f(Tv)$.
- For real vector spaces, if $\mathcal{M}(T) = A$, then $\mathcal{M}(T^*) = A'$.
- For complex vector spaces, if $\mathcal{M}(T) = A$, then $\mathcal{M}(T^*) = A^*$.

SKIPPED

Invertible maps and matrices

Inverse of a linear map

Definition 25 (Inverse of a linear map)

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

$$S \circ T = \mathrm{Id}_V \quad \text{and} \quad T \circ S = \mathrm{Id}_W.$$

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

Remarks:

- Inverse maps are unique
- Not every linear map is invertible

Characterizing invertability

Proposition 26 (Characterizing invertability)

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

Proof.

"⇒":

- **Invertible ⇒ injective:**
  Suppose $T(u) = T(v)$. Then

$$u = T^{-1}(T(u)) = T^{-1}(T(v)) = v$$

  Thus T is injective.

Characterizing invertability (2)

- **Invertible $\Rightarrow$ surjective:**
  Consider $w \in W$. Then

$$w = T(T^{-1}(w)) \quad \Rightarrow \quad w \in \text{range of } T.$$

  Thus T is surjective.

Characterizing invertability (3)

"$\Leftarrow$":

Injective & surjective $\Rightarrow$ invertible.

- Let $w \in W$. There exists a unique $v \in V$ such that $T(v) = w$ (existence because T is surjective, uniqueness because T is injective). Define the mapping: $S(w) = v$. Then clearly, we have

$$T \circ S = \mathsf{Id}.$$

- Let $v \in V$. Then by associativity

$$T((S \circ T)(v)) = (T \circ S \circ T)(v) = (T \circ S)(T(v)) = \mathsf{Id} \circ T(v) = T(v).$$

Because T is injective, from the blue statement we can conclude:

$$(S \circ T)(v) = v$$

which implies that $S \circ T = \mathsf{Id}$.

Characterizing invertability (4)

- Linearity of S:
  (1) By linearity of T and because $T \circ S = Id$, we get

$$T(Sw_1 + Sw_2) = TSw_1 + TSw_2 = w_1 + w_2.$$

  From this statement, by injectivity of T, $Sw_1 + Sw_2$ is the unique element in V that T maps to $w_1 + w_2$.
  (2) The other way round, we know that S maps $w_1 + w_2$ to $S(w_1 + w_2)$.
  (3) Using the fact that S is the inverse of T, (1) and (2) imply that $Sw_1 + Sw_2$ and $S(w_1 + w_2)$ need to be the same element, thus

$$S(w_1 + w_2) = S(w_1) + S(w_2).$$

  A similar argument shows that we have $S(\lambda w) = \lambda S(w)$. Thus S is linear.

Inverse matrix

Definition 27 (Inverse matrix)

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

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

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

Inverting maps $\widehat{=}$ inverting matrices

Proposition 28 (Inverting maps $\widehat{=}$ inverting matrices)

The inverse matrix represents the inverse of the corresponding map, that is:
$T : V \to V$

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

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

Proof sketch. Exercise. ☺

Properties of inverse matrices

- The inverse matrix does not always exist!!!
- $(A^{-1})^{-1} = A, \quad (A \cdot B)^{-1} = B^{-1} \cdot A^{-1}.$
- A^t invertible $\Leftrightarrow$ A invertible
- $(A^t)^{-1} = (A^{-1})^t.$
- $A \in F^{n \times n}$ invertible $\Leftrightarrow$ $\mathrm{rank}(A) = n$ (see later)
- The set of all invertible matrices is called the 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 $Id : V \to V$. Assume we fix a basis of V (both in source and target space), then the corresponding matrix looks as follows:

$$M(Id, \mathcal{B}, \mathcal{B}) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}.$$

Representing the identity (2)

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 $\mathcal{M}(Id, \mathcal{A}, \mathcal{B})$ of the identity mapping look like?

Because $\mathcal{B}$ is a basis, we can write each of the vectors in $\mathcal{A}$ as linear combinations:

$$a_1 = t_{11}b_1 + t_{21}b_2 + \ldots + t_{n1}b_n$$
$$a_2 = \ldots$$

Now we form the corresponding matrix T:

$$T = \begin{pmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nn} \end{pmatrix}.$$

Representing the identity (3)

Sanity check: this matrix really represents the identity mapping:

- In the basis $\mathcal{A}$, the first basis vector a_1 has the representation $(1, 0, .., 0)'$ because $a_1 = 1 \cdot a_1 + 0 \cdot a_2 + \ldots + 0 \cdot a_n$
- Now observe that $T \cdot (1, 0, ..., 0)' = (t_{11}, t_{21}, ..., t_{n1})'$.
- This vector gives us Ta_1 expressed in basis $\mathcal{B}$ because $t_{11}b_1 + t_{21}b_2 + \ldots + t_{n1}b_n = a_1$ by definition of coefficients t_{ij}
- Hence, $Ta_1 = a_1$, so T is the identity mapping.

Change of basis is invertible

Proposition 29 (Change of basis is invertible)

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

Proof sketch. Exercise. ☺

Change of basis for an arbitrary mapping

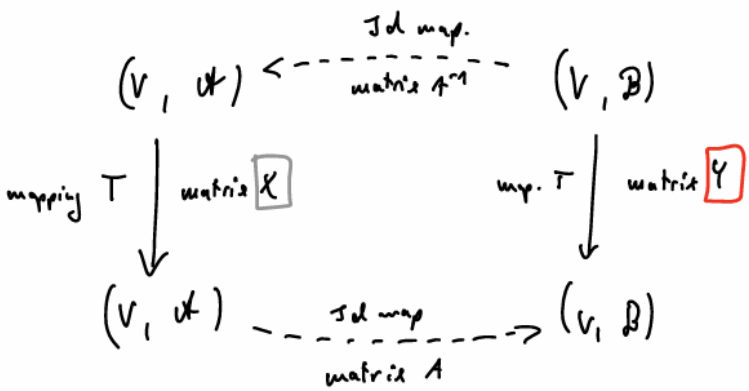
Proposition 30 (Chance of basis for an arbitrary mapping)

Let $\mathcal{A}, \mathcal{B}$ be two bases of V. Consider the transformation matrices

$$A = \mathcal{M}(\mathrm{Id}, \mathcal{A}, \mathcal{B}), \quad A^{-1} = \mathcal{M}(\mathrm{Id}, \mathcal{B}, \mathcal{A}).$$

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

Change of basis for an arbitrary mapping (2)



Change of basis for an arbitrary mapping (3)

In machine learning, we often use basis transforms, for example to represent a data matrix in a more convenient representation using PCA or SVD.

Rank of a matrix

Rank of a matrix

Definition 31 (Rank of a matrix)

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

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

The row rank is defined accordingly.

In machine learning, the rank describes how "complex" a matrix is. In low-rank methods, we try to only use matrices of a low rank (for example, to approximate matrices of high rank). This has advantages both computational and from the statistics / generalization point of view. Further keywords are recommender systems, compressed sensing,...

Rank of a matrix (2)

Proposition 32 (Row rank equals column rank)

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

Rank of a matrix (3)

Proposition 33

$T \in \mathcal{L}(V, W)$. Then the rank of the matrix that represents the map coincides with the dimension of the range of the linear mapping: $\text{rank}(\mathcal{M}(T)) = \dim(\text{range}(T))$.

The determinant

Motivation to study the determinant: geometry!

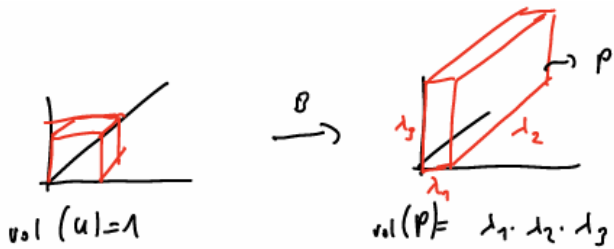
Consider the standard basis of $\mathbb{R}^3$: $e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

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

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

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

Motivation to study the determinant: geometry! (2)



We want to define a quantity (the determinant) that tells us how volumes change under **arbitrary** linear mappings.

Motivation to study the determinant: geometry! (3)

Which properties would such a quantity, let's call it d, have to satisfy?

- Assume that $T = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$ with respect to standard basis. Then we would like to get $d(T) = \lambda_1 \cdot \ldots \cdot \lambda_n$. (D1)
- If the linear mapping does not have full rank (the image 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 $d(T) = 0$. (D2)
- vol$(U) = 1$, so we would like that $d(Id) = 1$. (D3)

Motivation to study the determinant: geometry! (4)

Now let us relate this to the matrix T of the linear mapping. For simplicity, assume we work in the standard basis. The columns $t_1, \ldots, t_n$ of T are the images of the standard basis vectors. They describe the edges of the parallelotope obtained by applying T to the unit cube. With this intuition, let us rephrase the properties D1 - D3:

- (D1) Linearity in each column: If we stretch one edge by a factor $\lambda > 0$, the volume changes by the same factor. Similarly, adding two (non-negative) vectors in one column adds their signed volumes. More generally, we aim for linearity, the "volume" needs to allow for negative values, otherwise linearity is not possible; hence we use a "signed" volume, allowing d to be positive and negative.
- (D2) If two edges are identical, the parallelepiped is flat and has volume zero.
- (D3) The unit cube has volume 1, hence $d(Id) = 1$.

This is pretty much already the definition of the determinant, which we present now:

Definition of the determinant

Definition 34 (The determinant)

Consider a mapping $\det : F^{n\times n} \to F$. Then $\det$ is called a determinant if:

(D1) $\det$ is linear in each column of the matrix: 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

$$\det((a_1, \dots, a_i, \dots, a_n)) = \det((a_1, \dots, a_i', \dots, a_n)) + \det((a_1, \dots, a_i'', \dots, a_n))$$
$$\det((a_1, \dots, \lambda a_i, \dots, a_n)) = \lambda \det((a_1, \dots, a_i, \dots, a_n)).$$

(D2) If A has two identical columns, then $\det(A) = 0$.

(D3) $\det$ is normalized: $\det \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = 1$

Existence and uniqueness

Theorem 35 (Existence and uniqueness determinant)

The mapping $\det$ exists and is unique.

Proof sketch. Skipped. ☺

Properties of the determinant

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

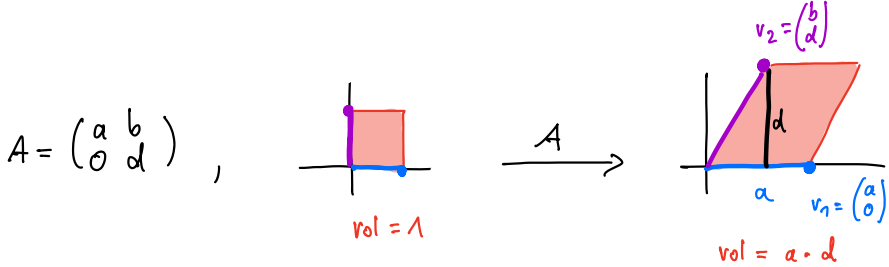
- The determinant of a linear mapping $T : V \to V$ does **not** depend on the basis.
- $\det(c \cdot A) = c^n \det(A)$ with n the dimension of the vector space.
- $\det(A \cdot B) = \det(A) \cdot \det(B)$.
- $\det(A^T) = \det(A)$.
- If A is invertible, then $\det(A^{-1}) = \frac{1}{\det(A)}$
- A invertible $\Leftrightarrow \det(A) \neq 0$.
- In general, $\det(A + B) \neq \det(A) + \det(B)$.

Properties of the determinant (2)

- Upper / lower triangular matrices:

$$A = \begin{pmatrix} \lambda_1 & * & \dots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} \implies \det(A) = \lambda_1 \cdots \lambda_n$$

Geometric intuition in case of 2 dimensions (and inductively for more dimensions):



Alternative definition of determinant

There exists a more straightforward definition of determinant. However, starting from this definition, proving the geometric properties is more cumbersome (and we haven't defined eigenvalues yet, so read this slide again towards the end of the linear algebra part):

Definition 36 (Alternative definition of determinant)

- 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.

Proof sketch. We skip the proof that this is the same object as "our" determinant. But in terms of intuition, I like this definition more than the previous one. ☺

Computing the determinant (in theory)

- For $n = 1$:

$$\det(a) = a$$

- For $n = 2$:

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

- For $n = 3$:

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

In general, there exists the formula of Laplace that expresses the determinant of an $n \times n$ matrix as a linear combination of determinant of many $(n-1) \times (n-1)$ submatrices. We will not need it in this lecture.

Computing the determinant (in practice)

In practice, determinants are typically computed by first bringing the matrix of interest in a particular form, such as an LU decomposition, Choleski decomposition, QR decomposition, etc. Depending on the type of the matrix, one or the other might be possible / simpler / more appropriate.

Example: Under certain conditions, a matrix admits an "LU-decomposition", in particular if the matrix is symmetric and positive definite (see later):

- Try to write matrix A as a product $A = L \cdot U$ where L is a lower triangular matrix and U upper triangular.
- The we can easily compute the determinant as product of the diagonal entries.
- If the matrix does not directly admit an LU decomposition, we can use a more general form, LU with pivoting, where we represent $PA = LU$ where P is a permutation matrix (which has $det(P) = \pm 1$).

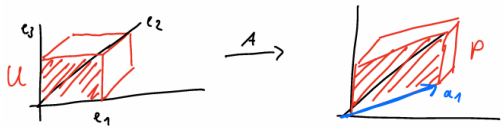
Geometric intuition again

Theorem 37 (Geometric intuition again)

Consider an $n \times n$ matrix A with columns $(a_1 \mid a_2 \mid \ldots \mid a_n) = A$.
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:

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

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



Applications to integrals

Proposition 38 (Change of variable formula)

Let $\Omega \subset \mathbb{R}^n$ be an open subset, $\sigma : \Omega \to \mathbb{R}^n$ a differentiable mapping, and $f : \sigma(\Omega) \to \mathbb{R}$. Then:

$$\int_{\sigma(\Omega)} f(y) \underbrace{dy}_{\text{volume element}} = \int_{\Omega} f(\sigma(x)) \underbrace{|\det(\sigma'(x))|\, dx}_{\text{volume element}}.$$

where σ' denotes the matrix of all partial derivatives $(\partial x_i / \partial x_j)_{ij}$ of σ.

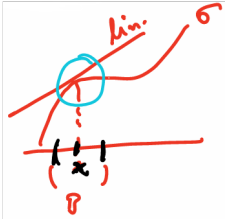
Remark: change of variables: $y = \sigma(x)$

Applications to integrals (2)

Intuition:

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

Then local volumes get stretched by the $|det(\sigma')|$.



Application to normal distribution

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

Normalization with $\det(\Sigma)^{1/2}$:

- Consider a random variable $Z \sim \mathcal{N}(0, Id)$ (standard multi-variate Gaussian)
- Apply a linear mapping to the random variable: $X = \mu + AZ$. Then $X \sim \mathcal{N}(\mu, \Sigma)$ where $\Sigma = AA'$.
- The linear mapping changes the volume by a factor of $|\det(A)|$, which coincides with $\sqrt{\det \Sigma}$
- To make sure that the density still integrates to 1, we need to recale it by this factor.

Eigenvalues, eigenvectors, eigenspaces

Eigenvalues

ML Keywords: PCA, SVD, Spectral Clustering, ...

Definition 39 (Eigenvalues)

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 $Tv = \lambda v$. A vector $v \neq 0$ with this property is called an eigenvector corresponding to eigenvalue λ. The set of all eigenvectors of λ, together with the zero vector, is called the eigenspace $E(\lambda, T)$.

Observe that the eigenspace can be written as $E(\lambda, T) = \ker(T - \lambda I)$ because

$$Tv = \lambda v \iff Tv - \lambda v = 0 \iff Tv - \lambda I v = 0 \iff (T - \lambda I)v = 0$$

Geometric intution

- Eigenvalue/eigenvector realizes a "stretching" $v \mapsto \lambda v$.
- Many mappings do not have eigenvectors, for example, a rotation.

Eigenvectors are not unique

If λ is an eigenvalue, it has many eigenvectors. For example, if v is an eigenvector, then also $a \cdot v$ $(a \in K)$ is an eigenvector:

$$T(a \cdot v) = a \cdot T(v) = a \cdot \lambda v = \lambda(a \cdot v)$$

Linear (in)dependence of eigenvectors?

- Eigenvectors corresponding to **different** eigenvalues are linearly independent.
  (Easy exercise: assume they are dependent and derive a contradiction)

- Eigenvectors that correspond to the same eigenvalue do not need to be independent. (Simple example: v eig. $\Rightarrow c \cdot v$ eig. (for $c \neq 0$), but v and $c \cdot v$ are not linearly independent)

  But they can be linearly independent. (Easy example: $A = I$, then every vector $v \neq 0$ is an eigenvector of eigenvalue 1.)

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

Tool: Polynomials of linear operators

In the following we will repeatedly use the following construction:

Let $p(x) = \sum_{i=0}^{n} a_i x^i$ be a polynomial. For a linear operator T, define

$$p(T) := \sum_{i=0}^{n} a_i T^i$$

Example:

$p : \mathbb{R} \to \mathbb{R}, p(x) = x^7 - 3x^5 + 4$, then

$p(T) : V \to V,\ p(T)(v) = (T^7 - 3T^5 + 4I)v = T^7 v - 3T^5 v + 4v$

Tool: Polynomials of linear operators (2)

Now let p, q be two polynomials. Then it is easy to see:

- If the polynomial p is a product of two polynomials, then the same holds true for the linear operators: $(p \cdot q)(T) = p(T)q(T)$
- In particular, if we consider a factorization of a polynomial $p(x)$, then the same factorization applies to the linear operator $p(T)$.

$$\begin{aligned} p(x) &:= a_0 + a_1 x + \cdots + a_n x^n = c \cdot (x - \lambda_1)(x - \lambda_2) \dots (x - \lambda_m) \\ \implies p(T) &= a_0 + a_1 T + \cdots + a_n T^n = c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \dots (T - \lambda_m I) \end{aligned}$$

- The order of factors does not matter: any two polynomials of the same operator commute: $p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$

Tool: Polynomials of linear operators (3)

Observe that because composition of injective maps in injective, we have:

$$p(T) \text{ injective and } q(T) \text{ injective} \implies (pq)(T) \text{ injective}$$

Or, the contrapositive statement:

$$(pq)(T) \text{ not injective} \implies p(T) \text{ or } q(T) \text{ not injective.}$$

We will repeatedly exploit this statement in the following.

Existence of eigenvalues

Theorem 40 (Existence of eigenvalues)

Every operator $T : V \to V$ on a finite-dimensional, **complex** vector space V has at least one eigenvalue.

Existence of eigenvalues (2)

Proof.

Step 1: Getting started

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

$$v, Tv, T^2v, \ldots, T^nv$$

has to be linearly dependent (it consists of $n+1$ vectors in an n-dim space). So we can find coefficients $a_0, a_1, \ldots, a_n$, not all zero, such that

$$p(T) := a_0v + a_1Tv + \cdots + a_nT^nv = 0.$$

Existence of eigenvalues (3)

Step 2: Factorization

Consider the polynomial on $\mathbb{C}$ with the same coefficients $a_0, a_1, \ldots, a_n$:

$$p(z) := a_0 + a_1 z + \cdots + a_n z^n.$$

On $\mathbb{C}$, we can factorize it into m linear factors (m could be smaller than n because a_n might be zero):

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

Hence we can also factorize $p(T)$:

$$p(T) = a_0 + a_1 T + \cdots + a_n T^n = c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \ldots (T - \lambda_m I)$$

Existence of eigenvalues (4)

Step 3: Putting it together

$$0 = a_0 v + a_1 T v + \cdots + a_n T^n v = c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \dots (T - \lambda_m I) \cdot v$$

$\Rightarrow v \in \ker(c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \dots (T - \lambda_m I))$

$\Rightarrow$ There must exist $i \in \{1, \dots, m\}$ such that $(T - \lambda_i I)$ not injective

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

Existence of eigenvalues (5)

⚠ The corresponding eigenvector is not necessarily v!

Construction of an eigenvector:
Apply the factors to v, starting from the right hand side of the polynomial: Consider $(T - \lambda_m I)v$, then $(T - \lambda_{m-1} I)(T - \lambda_m I)v$, etc. The product might be non-zero for a while, but eventually there has to be a first index j such that $(T - \lambda_j I)(T - \lambda_{j+1} I)....(T - \lambda_m)v = 0$ (because at latest when we have applied all factors, the result is 0). Then we found the eigenvector:

$$c \cdot (T - \lambda_1 I)(T - \lambda_2 I) \dots (T - \lambda_j I) \underbrace{(T - \lambda_{j+1} I) \dots (T - \lambda_m I) \cdot v}_{=: \text{ the eigenvector}} = 0$$

□

Characteristic polynomial

Motivation: an alternative characterization of eigenvalues

Observe the following identities: A is an $n \times n$-matrix

$$\begin{aligned} & \exists v \neq 0 : \ Av = \lambda v \\ \Leftrightarrow \ & \exists v \neq 0 : \ (A - \lambda I)v = 0 \\ \Leftrightarrow \ & \exists v \neq 0 : \ v \in \ker(A - \lambda I) \\ \Leftrightarrow \ & \operatorname{rank}(A - \lambda I) < n \\ \Leftrightarrow \ & \det(A - \lambda I) = 0 \end{aligned}$$

We now use this to build a tool to find eigenvalues, the characteristic polynomial.

Characteristic polynomial

Definition 41 (Characteristic polynomial)

The characteristic polynomial of an $n \times n$-matrix A is defined as

$$p_A : \mathbb{R} \to \mathbb{R},\ p_A(t) := \det(A - t \cdot I)$$

Characteristic polynomial (2)

Example:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$\det(A - tI) = \det\left(\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} - t \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = \det\begin{pmatrix} a_{11} - t & a_{12} \\ a_{21} & a_{22} - t \end{pmatrix}$$

$$= (a_{11} - t)(a_{22} - t) - a_{12}a_{21} = t^2 - t(a_{11} + a_{22}) + (a_{11}a_{22} - a_{12}a_{21})$$

Characteristic polynomial (3)

Observations

- The characteristic polynomial does not depend on the basis. To see this, consider a matrix A and basis transformation matrix U. Then the characteristic polynomial of UAU^{-1} is:

$$\det(UAU^{-1} - t \cdot I) = \det(UAU^{-1} - t \cdot UU^{-1}) = \det(U(A - t \cdot I)U^{-1})$$

$$= \det(U) \cdot \det(A - tI) \cdot \det(U^{-1}) = \det(A - tI)$$

- The roots of the characteristic polynomial correspond exactly to the eigenvalues of A because $Av = \lambda v \iff det(A - \lambda I) = 0$.
- Over $\mathbb{C}$, the characteristic polynomial always has n roots, so the matrix has n eigenvalues (not necessarily distinct).

Geometric and algebraic multiplicity

Definition 42 (Geometric and algebraic multiplicity)

For an operator A with eigenvalue λ, we define:

- Geometric multiplicity: the dimension of the corresponding eigenspace $E(\lambda, A)$.
- Algebraic multiplicity: the multiplicity of the root λ in the characteristic polynomial.

In general, the two notions **do not** coincide. Example:

The matrix $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ has the characteristic polynomial $p_A(t) = (1-t)^2$, thus eigenvalue 1 has algebraic multiplicity 2. To compute the eigenspace, we solve for $Av = v$, leading to eigenspace spanned by the vector $(1, 0)'$ only, so geometric multiplicity is 1.

Compute eigenvalues/vectors by hand

- Write down the characteristic polynomial, find the roots $\rightsquigarrow$ eigenvalues
- To compute the eigenvectors, then solve the linear systems that correspond to the eigenvalues that you found: $Ax = \lambda x$.

In practice ... see later (numerics)

Diagonalization and Triangulization

Diagonizable matrices

Definition 43 (Diagonizable matrices)

An operator $T \in \mathcal{L}(V)$ is diagonalizable if there exists a basis $\mathcal{B}$ of V such that the corresponding matrix is diagonal:

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

⚠ Not always the case, neither over $\mathbb{R}$ nor $\mathbb{C}$. We will later see some special cases where it always works (real symmetric matrices).

Diagonizable matrices (2)

Prime example for a matrix that is not diagonalizable over $\mathbb{C}$ is the matrix $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

- To find the eigenvalues we solve the characteristic polynomial

$$p_A(\lambda) = \det(A - \lambda I) = ... = (1 - \lambda)^2$$

  and find that the only solution / eigenvalue is $\lambda = 1$, with algebraic multiplicity 2. The eigenvalues are even in $\mathbb{R}$, we don't even need $\mathbb{C}$ to solve this.

- To find the eigenvectors we solve $Av = v$ and find that all vectors of the form $(x, 0)'$ are eigenvectors. We thus have a 1-dim eigenspace spanned by vector $(1, 0)'$ (the geometric multiplicity is 1). Hence A is not diagonalizable, neither over $\mathbb{R}$ nor over $\mathbb{C}$.

When is a matrix diagonizable?

Proposition 44 (When is a matrix diagonizable)

Let T be an operator on a finite-dimensional vector space V. Let $\lambda_1, \ldots, \lambda_m$ 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) $\dim V = \dim(\text{eigenspace}(\lambda_1)) + \cdots + \dim(\text{eigenspace}(\lambda_m))$.

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

(We skip the proof.)

Nice in theory. But hard to verify for a given matrix ...

Triangular matrices

Proposition 45 (Triangular matrices and invariant subspaces)

Let $T \in \mathcal{L}(V)$, $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ a basis. Then the following are equivalent:

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

(b) $\forall j = 1, \ldots, n : Tv_j \in \text{span}\{v_1, \ldots, v_j\}$

Triangular matrices (2)

Proof idea (illustrated on a small example): Consider an upper triangular 3×3 matrix.

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

$$Tv_2 = \begin{pmatrix} \lambda_1 & a_{12} & a_{13} \\ 0 & \lambda_2 & a_{23} \\ 0 & 0 & \lambda_3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a_{12} \\ \lambda_2 \\ 0 \end{pmatrix} = a_{12} \underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}_{v_1} + \lambda_2 \underbrace{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}_{v_2} \in \mathsf{span}(v_1, v_2).$$

$$Tv_3 \in \mathsf{span}(v_1, v_2, v3)$$

(Formal proof follows this argument to prove $(a) \implies (b)$; other direction easy.)

When are triangular matrices invertible?

Proposition 46 (Invertability of triangular matrices)

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

$$\begin{pmatrix} \lambda_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}.$$

When are triangular matrices invertible? (2)

Proof.

"$\Leftarrow$":

Let $v_1, \dots, v_n$ be 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 $0 \neq v_1 \in \text{range}(T)$.
- By triangular form, $\underbrace{Tv_2}_{\in \text{ range } T} = \underbrace{a_1 v_1}_{\in \text{ range } T} + \lambda_2 v_2$ for some scalar a_1.

  Thus $\lambda_2 v_2 \in \text{range}(T)$ and because $\lambda_2 \neq 0$, then $v_2 \in \text{range}(T)$. Similarly for $v_3, \dots, v_n$.

In this way, we see that $v_1, \dots, v_n$ are all non-zero and in range(T). Hence T is surjective, thus injective, and hence invertible.

When are triangular matrices invertible? (3)

"⇒":

Assume T is invertible.

- Clearly $\lambda_1 \neq 0$, otherwise $Tv_1 = 0$, contradicting invertibility.
- Suppose $\lambda_j = 0$. Then T maps span$(v_1, \ldots, v_j)$ into span$(v_1, \ldots, v_{j-1})$.
- Thus T is not injective on the subspace spanned by $v_1, \ldots, v_j$, so there exists v in this subspace such that $Tv = 0$.

But then T is not invertible. □

Entries of diagonal are eigenvalues

Proposition 47 (Entries of diagonal are eigenvalues)

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

Remark: Holds both over $\mathbb{R}$ and $\mathbb{C}$.

Entries of diagonal are eigenvalues (2)

Proof. Fix any $\lambda \in F$ and consider $T - \lambda I$, which is also upper triangular:

$$T = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}, \quad T - \lambda I = \begin{pmatrix} \lambda_1 - \lambda & * & \cdots & * \\ 0 & \lambda_2 - \lambda & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n - \lambda \end{pmatrix}.$$

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

$\iff$ by previous proposition, one of the diagonal entries $(\lambda - \lambda_i)$ of $T - \lambda I$ is zero

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

□

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

Proposition 48 (Triangularization over $\mathbb{C}$)

Let V be a complex finite-dimensional vector space, $T \in \mathcal{L}(V)$. Then there exists a basis $\mathcal{B}$ such that $\mathcal{M}(T, \mathcal{B}, \mathcal{B})$ has an upper triangular form.

⚠ Does not hold over $\mathbb{R}$!

Over $\mathbb{C}$ each matrix can be triangularized (2)

Proof by induction on n:

Base case: $n = 1$, clear.

Inductive hypothesis: statement holds for all dimensions up to $n - 1$.

Inductive step: from $n - 1$ to n:

- Already know that one complex eigenvalue exists: λ.
- Consider subspace $U := \text{range}(T - \lambda I)$
- Because $T - \lambda I$ is not surjective we have that $\dim(U) < n$
- The subspace U is invariant under T because for $u \in U$ we have

$$Tu = Tu - \lambda u + \lambda u = (T - \lambda I)u + \lambda u$$

  and both terms on the rhs live in U.
- Apply the induction hypothesis to the operator $T : U \to U$: exists basis $u_1, \ldots, u_m$ (in U) such that $T|_U$ has upper triangular form.
- Extend $u_1, \ldots, u_m$, say with vectors $w_1, \ldots, w_k$, to a basis of V.

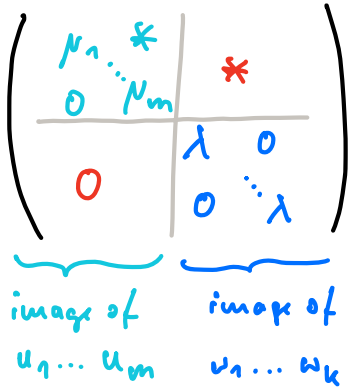
Over $\mathbb{C}$ each matrix can be triangularized (3)

- Now we need to show the triangular form still holds:

$$Tw_1 = Tw_1 - \lambda w_1 + \lambda w_1 = (T - \lambda I)w_1 + \lambda w_1 \in \text{span}(u_1, \ldots, u_m, w_1).$$
$$Tw_2 = Tw_2 - \lambda w_2 + \lambda w_2 = (T - \lambda I)w_2 + \lambda w_2 \in \text{span}(u_1, \ldots, u_m, w_2)$$
... etc ...

So in general we have $Tw_i \in \text{span}(u_1, \ldots, u_m, w_1, \ldots, w_i)$ and the matrix is triangular with the following structure:

Over $\mathbb{C}$ each matrix can be triangularized (4)



Thus T is upper triangular w.r.t. basis $\{u_1, \dots, u_m, w_1, \dots, w_k\}$.

☺

Over $\mathbb{C}$ each matrix can be triangularized (5)

Side remark:

- the dimension k of the blue part equals

$$k = dim(V) - dim(range(T - \lambda I)) = dim(ker(T - \lambda I)),$$

  which gives the geometric multiplicity of the eigenvalue λ.
- But the number of times λ appears on the diagonal could be larger that k, because some of the μ_i could also equal λ. The number of times λ appears on the diagonal is the algebraic multiplicity of the eigenvalue and is always larger or equal to the geometric one.
- If the algebraic multiplicity is larger than the geometric one, we cannot diagonalize the matrix: an eigenvalue appears more often on the diagonal than we have dimensions of the eigenspace. This cannot be fixed by moving into $\mathbb{C}$.

ML Keywords: loss functions, regularizers, sparsity

Normed spaces

Metric space

ML Keywords: inductive bias, k-nearest neigbor

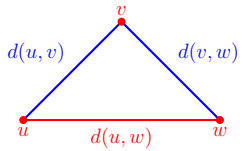
Definition 49 (Metric space)

Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called a metric if the following conditions hold for all $u, v, w \in X$:

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

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

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



Norm on a vector space

Definition 50 (Norm on a vector space)

Let V be a vector space. A norm on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for all $x, y \in V$, $\lambda \in \mathbb{F}$:

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

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

(N3) $\|x\| = 0 \impliedby x = 0$ (definiteness)

(N4) $\|x\| = 0 \implies x = 0$ (definiteness)

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

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

Definition 51 (Euclidean norm on $\mathbb{R}^d$)

The Euclidean norm on $\mathbb{R}^d$ is defined as:

$$\|x\| = \left(\sum_{i=1}^{d} x_i^2 \right)^{1/2}.$$

Intuition: The norm $\|x\|$ represents the "length of x" or the "distance between x and the origin".

Euclidean norm on $\mathbb{R}^d$ (2)

Easy to see: Every norm induces a metric through

$$d(x, y) := \|x - y\|.$$

However, not every metric arises from a norm (try to find a counterexample).

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

Definition 52 (p-Norms on $\mathbb{R}^d$)

Let $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$. Define the p-norm as:

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

- The case $p = 2$ coincides with the Euclidean norm.
- $\|x\|_\infty := \max |x_i|$ is also a norm.
- $\|x\|_0 :=$ number of non-zero coordinates of x is not a proper norm (see later).

Unit ball of a norm

Definition 53 (Unit ball of a norm)

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

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

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

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

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

Case $p \geq 1$ (convex balls):

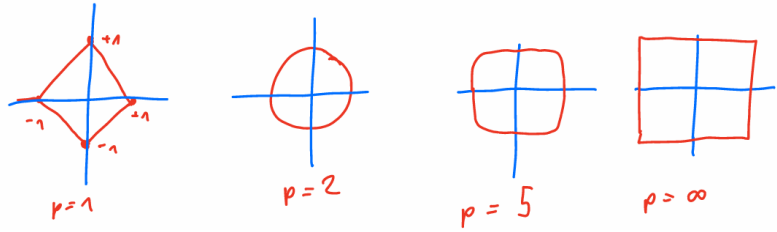
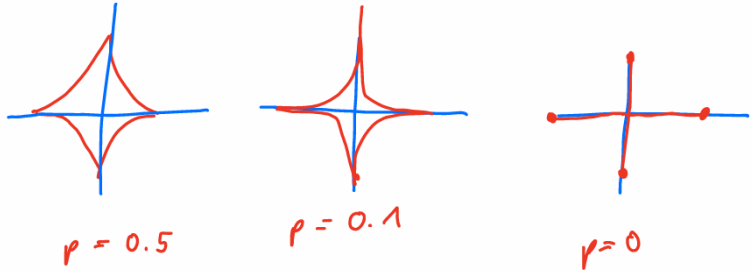


Illustration: Unit balls on $\mathbb{R}^2$ (2)

Case $p < 1$ (balls not convex):



When is a "p-norm" a norm?

Proposition 54 (When is a "p-norm" a norm?)

The p-norm $\|\cdot\|_p$ is a norm on $\mathbb{R}^d$ if and only if $p \geq 1$.

Proof sketch:

- Definiteness (N3) and (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 $p \geq 1$:

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

All norms on $\mathbb{R}^n$ are equivalent (SKIPPED)

SKIPPED

Equivalent norms

Definition 55 (Equivalent norms)

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

$$\forall x \in V: \quad \alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a.$$

SKIPPED

Equivalent norms (2)

Theorem 56 (Norm equivalence on $\mathbb{R}^n$)

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

SKIPPED

Equivalent norms (3)

Proof.

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

First inequality: $\exists c_1 > 0 : \forall x \, \|x\| \leq c_1 \|x\|_\infty$

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

$$\|x\| = \left\| \sum_i x_i e_i \right\| \leq \sum_i \|x_i e_i\| = \sum_i |x_i| \|e_i\|.$$

Since $|x_i| \leq \|x\|_\infty$, it follows:

$$\|x\| \leq \sum_i \|x\|_\infty \|e_i\| = \|x\|_\infty \sum_i \|e_i\| = \|x\|_\infty \cdot c_1,$$

where $c_1 := \sum_i \|e_i\|$.

SKIPPED

Equivalent norms (4)

Second inequality: $\exists c_2 > 0 : \forall x \, \|x\|_\infty \leq c_2 \|x\|$
Let $S := \{x \in \mathbb{R}^d \mid \|x\|_\infty = 1\}$ be the unit sphere with respect to $\|\cdot\|_\infty$. Consider the function:

$$f : S \to \mathbb{R}, \quad x \mapsto \|x\|.$$

- The mapping f is continuous with respect to $\|\cdot\|_\infty$, since:

$$|f(x) - f(y)| = |\|x\| - \|y\|| \leq \|x - y\| \leq c_1 \|x - y\|_\infty.$$

- The set S is closed and bounded. Thus, by the theorem of Heine-Borel, S is compact. Any continuous mapping on a compact set attains its minimum and maximum. Define:

$$\tilde{c_2} := \min\{f(x) \mid x \in S\}.$$

- Because $0 \notin S$ (sphere, not ball), we can conclude from definiteness that $\tilde{c_2} \neq 0$.

SKIPPED

Equivalent norms (5)

- Now compute: let $x \in \mathbb{R}^d$, $x \neq 0$, be arbitrary. Then $\frac{x}{\|x\|_\infty} \in S$, so by definition of $\tilde{c}_2$ and homogeneity of $\|\cdot\|$:

$$\tilde{c}_2 \leq \left\| \frac{x}{\|x\|_\infty} \right\| = \frac{\|x\|}{\|x\|_\infty}.$$

Thus:

$$\|x\|_\infty \leq \frac{1}{\tilde{c}_2} \|x\| \quad \text{with } c_2 := \frac{1}{\tilde{c}_2},$$

and for $x = 0$ the inequality holds trivially.

□

SKIPPED

Norms can be characterized by convex sets (SKIPPED)

SKIPPED

Convex set

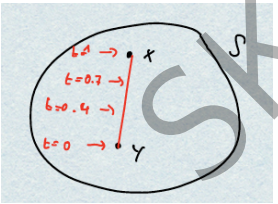
Definition 57 (Convex set)

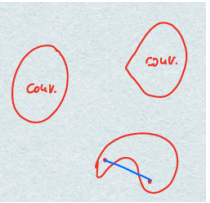
Consider a real vector space V, $S \subset V$.

S is called convex if $\forall t \in [0,1]$ and $\forall x, y \in S$,

$$tx + (1-t)y \in S$$

SKIPPED



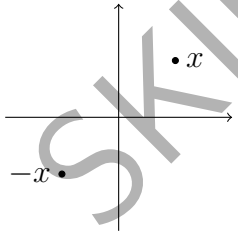


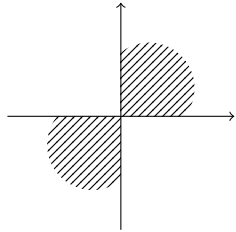
Symmetric set

Definition 58 (Symmetric set)

A set $C \subset V$ is called symmetric if

$$x \in C \Rightarrow -x \in C$$





SKIPPED

Convex sets induce norms

Theorem 59 (Convex sets induce norms)

(1) Let $C \subset \mathbb{R}^d$ be closed, convex, symmetric and have non-empty interior. Define

$$p(x) := \inf\left\{t > 0 \,\middle|\, \frac{x}{t} \in C\right\}$$
$$= \inf\{t > 0 \,|\, x \in t \cdot C\}$$

Then p is a semi-norm. If C is bounded, then p is a norm, and its unit ball coincides with C:

$$C = \left\{x \in \mathbb{R}^d \,\middle|\, p(x) \leq 1\right\}$$

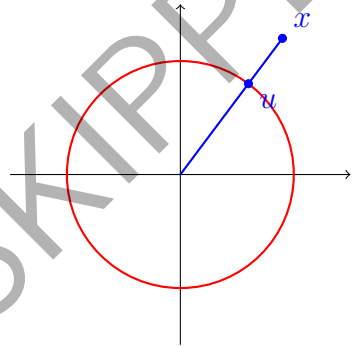
(2) For any norm $\|\cdot\|$ on $\mathbb{R}^d$, the set $C := \left\{x \in \mathbb{R}^d \,\middle|\, \|x\| \leq 1\right\}$ is bounded, symmetric, closed, convex, and has non-empty interior.

SKIPPED

Convex sets induce norms (2)

Illustration:

- Consider factor a by which I need to multiply x to end up on the unit sphere. Then define $\|x\| := 1/a$.
- Convex set = unit ball of norm 1



SKIPPED

Convex sets induce norms (3)

Proof.

Claim: $p(x)$ **is well defined**

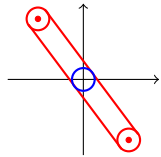
Want to prove: given $x \in \mathbb{R}^d$, the set

$$\left\{ t > 0 \,\middle|\, \frac{x}{t} \in C \right\} \neq \emptyset$$

We are going to prove: $\exists \varepsilon > 0$ such that

$$B_\varepsilon(0) := \left\{ e \in \mathbb{R}^d \,\middle|\, \|e\|_2 < \varepsilon \right\} \subset C$$

Intuition:



SKIPPED

Convex sets induce norms (4)

- By assumption, C has at least one interior point $v \in C^\circ \Rightarrow \exists \varepsilon > 0$ such that

$$B_\varepsilon(v) \subset C$$

- That is, $v + B_\varepsilon(0) = \{v + e \mid e \in B_\varepsilon(0)\} \subset C$
- By symmetry, applied to v itself: $v \in C \Rightarrow -v \in C$
- By convexity, for every $e \in B_\varepsilon(0)$ the midpoint of $v + e \in C$ and $-v \in C$ lies in C:

$$\frac{1}{2}(v + e) + \frac{1}{2}(-v) = \frac{e}{2} \in C$$

So $B_{\varepsilon/2}(0) \subset C$ (which proves the claim above, with $\varepsilon/2$ in place of ε), so the set $\left\{t > 0 \mid \frac{x}{t} \in C\right\}$ is non-empty.
The infimum

$$\inf\left\{t > 0 \,\middle|\, \frac{x}{t} \in C\right\}$$

exists because the set is a subset of $\mathbb{R}$ and bounded from below by 0.

SKIPPED

Convex sets induce norms (5)

Now we need to prove all axioms of a norm:

$p(0) = 0$

- We have seen: $0 \in C$
- $\forall t > 0 : \frac{0}{t} = 0 \in C$
- $\inf \left\{ t \,\middle|\, \frac{0}{t} \in C \right\} = 0 \Rightarrow p(0) = 0$

SKIPPED

Convex sets induce norms (6)

$p(\alpha x) = |\alpha| \cdot p(x)$

- For all $\alpha > 0$ we have:

$$\begin{aligned} p(\alpha x) &= \inf\left\{t > 0 \,\middle|\, \frac{\alpha x}{t} \in C\right\} \\ &= \inf\left\{\alpha s > 0 \,\middle|\, \frac{x}{s} \in C\right\} \\ &= \alpha \inf\left\{r > 0 \,\middle|\, \frac{x}{r} \in C\right\} \\ &= \alpha \cdot p(x) \end{aligned}$$

- By symmetry: $p(-x) = p(x)$, so the result also holds for $\alpha < 0$

$\rightarrow$ Combining these two statements gives **homogeneity**.

SKIPPED

Convex sets induce norms (7)

△-inequality Consider $x, y \in \mathbb{R}^d$, $s, t > 0$ such that $\frac{x}{s} \in C$, $\frac{y}{t} \in C$

Observe: $\frac{s}{s+t} + \frac{t}{s+t} = 1$

Thus, by convexity:

$$\frac{s}{s+t} \cdot \frac{x}{s} + \frac{t}{s+t} \cdot \frac{y}{t} = \frac{x+y}{s+t} \in C$$

$$\begin{aligned} p(x+y) &= \inf\left\{u > 0 \,\middle|\, \frac{x+y}{u} \in C\right\} \\ &\leq \inf\left\{s > 0 \,\middle|\, \frac{x}{s} \in C\right\} + \inf\left\{t > 0 \,\middle|\, \frac{y}{t} \in C\right\} \\ &= p(x) + p(y) \end{aligned}$$

SKIPPED

Convex sets induce norms (8)

We choose $s, t > 0$ such that

$$\frac{x}{s} \in C, \quad \frac{y}{t} \in C$$

Then, since $\frac{s}{s+t} + \frac{t}{s+t} = 1$, by convexity:

$$\frac{s}{s+t} \cdot \frac{x}{s} + \frac{t}{s+t} \cdot \frac{y}{t} = \frac{x+y}{s+t} \in C \Rightarrow p(x+y) \leq s+t$$

Now consider sequences $(s_i)_{i\in\mathbb{N}}$, $(t_i)_{i\in\mathbb{N}}$ with

$$\frac{x}{s_i} \in C, \quad s_i \to p(x), \qquad \frac{y}{t_i} \in C, \quad t_i \to p(y)$$

Then by the above argument:

$$p(x+y) \leq s_i + t_i \quad \Rightarrow \quad p(x+y) \leq p(x) + p(y)$$

SKIPPED

Convex sets induce norms (9)

Positive definiteness: $p(x) = 0 \Rightarrow x = 0$

$$p(x) = 0 \Leftrightarrow \inf\left\{t > 0 \,\middle|\, \frac{x}{t} \in C\right\} = 0$$

Then $\exists (t_k)_{k\in\mathbb{N}}$ such that $t_k \to 0$ and $\frac{x}{t_k} \in C \ \forall k$

Assume $x \neq 0$, then $\left(\frac{x}{t_k}\right)_{k\in\mathbb{N}}$ is unbounded.

↯ Contradiction, since C is bounded.

□

SKIPPED

Scalar product

Scalar product

Definition 60 (Scalar product)

Consider a 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 over $\mathbb{R}$),
  $\langle x, y \rangle = \overline{\langle y, x \rangle}$ (if over $\mathbb{C} \to$, need complex conjugate).
- Positive definiteness:
  (S4) $\langle x, x \rangle \geq 0$,
  (S5) $\langle x, x \rangle = 0 \iff x = 0$.

Examples

- Euclidean scalar product on $\mathbb{R}^n$: For $x = (x_1, ..., x_n)'$, $y = (y_1, ..., y_n)'$, define

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

- On $\mathbb{C}^n$:

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

- $C([a, b])$:

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt,$$

  is a scalar product.

Scalar product and angles

ML Keywords: cosine similarity

Proposition 61 (Scalar product and angles)

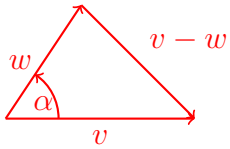
Consider the standard scalar product on $\mathbb{R}^n$. Then:

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

where α is the angle enclosed by v and w.

Scalar product and angles (2)

Proof sketch.



- In a general triangle, we have:

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\|v\|\|w\| \cos(\alpha).$$

- Using the scalar product, $\|v - w\|^2 = \langle v - w, v - w \rangle = \|v\|^2 + \|w\|^2 - 2\langle v, w \rangle$.

Thus:

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

Banach and Hilbert spaces

ML Keywords:
Reproducing Kernel Hilbert Space

Definition 62 (Banach and Hilbert spaces)

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

Scalar product implies norm

- Consider a vector space with a scalar product $\langle \cdot, \cdot \rangle$. Define $\| \cdot \| : V \to \mathbb{R}$:

$$\|x\| := \sqrt{\langle x, x \rangle}.$$

  Then $\| \cdot \|$ is a norm, we say it is the norm induced by $\langle \cdot, \cdot \rangle$.
- The other way around does not work in general: not every norm can be expressed by a scalar product

  Can you find a counterexample?

Norm implies metric

- Consider a vector space with norm $\|\cdot\|$. Define $d : V \times V \to \mathbb{R}$:

$$d(x, y) := \|x - y\|.$$

  Then d is a metric induced by the norm.
- The other way around does not work in general! (not every metric comes from a norm)

  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:

$$|\langle u, v\rangle| \leq \|u\|_2 \cdot \|v\|_2.$$

- Hölder inequality: Let $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then:

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

Orthonormal basis and orthogonal projections

Orthogonal vectors and sets

Definition 63 (Orthogonal vectors and sets)

Consider a pre-Hilbert space V.

- Two vectors $v_1, v_2 \in V$ are called orthogonal if $\langle v_1, v_2 \rangle = 0$.
  Notation: $v_1 \perp v_2$.
- Two sets $V_1, V_2 \subseteq V$ 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 $S \subseteq V$, we define its orthogonal complement $S^\perp$ as:

$$S^\perp := \{v \in V \mid v \perp s \text{ for all } s \in S\}.$$

Orthonormal vectors and sets

Definition 64 (Orthonormal vectors and sets)

- Two vectors v_1, v_2 are called orthonormal if they are orthogonal and additionally have norm 1:
$$\langle v_1, v_2 \rangle = 0, \quad \|v_1\| = 1, \|v_2\| = 1.$$
- A set of vectors $v_1, v_2, \ldots, v_n$ is called orthonormal if any two vectors are orthonormal.

Orthogonal/orthonormal basis

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

Proposition 65

In an orthonormal basis $u_1, \dots, u_n$, the representation of a vector $v = \sum a_i u_i$ is given as

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

Proof sketch. Exercise. ☺

⚠ We still do not know whether an orthonormal basis always exists.

Projection (general case)

Definition 66 (Projection general case)

An operator $A \in \mathcal{L}(V)$ is called a projection if $A^2 = A$.

Orthogonal Projection

Definition 67 (Orthogonal Projection)

Let U be a finite-dimensional 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 called the orthogonal projection of H on U.

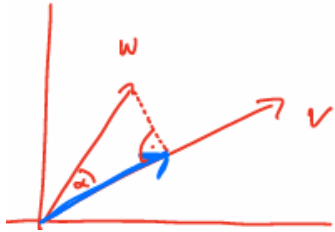
Orthogonal Projection (2)

Proof idea: Let $u_1, \ldots, u_k$ be an orthogonal basis of U (note that we still have to establish that this always exists, we do it on the next slides). Define $P_U : V \to U$ by

$$P_U(w) = \sum_{i=1}^{k} \frac{\langle w, u_i \rangle}{\|u_i\|^2} u_i.$$

- It is clear that P_U is linear.
- It is easy to see that $P_U(u) = u$ for all $u \in U$.
- For $w \in U^\perp$ it is easy to see that $P_U(w) = 0$.

Orthogonal Projection (3)



How to make a given basis orthogonal?

Proposition 68 (Orthonormal basis)

Let $v_1, \ldots, v_n$ be any basis of a finite-dimensional pre-Hilbert space. Then we can transform it into an orthonormal basis $u_1, \ldots, u_n$.

Proof.

We construct the **Gram-Schmidt orthogonalization**: A procedure that takes any basis $v_1, \ldots, v_n$ and transforms it into an orthonormal basis $u_1, \ldots, u_n$.

How to make a given basis orthogonal? (2)

Intuition: Iterative procedure. Successively remove the components in the directions of the previously constructed orthonormal vectors.

- **Step 1:** Start with
$$u_1 := \frac{v_1}{\|v_1\|}, \qquad U_1 := \mathsf{span}\{u_1\}.$$
- **Step k:** Assume that we have constructed orthonormal vectors $u_1, \dots, u_{k-1}$. Remove from v_k its components in the directions $u_1, \dots, u_{k-1}$:
$$\tilde{u}_k := v_k - \sum_{i=1}^{k-1} \langle v_k, u_i \rangle u_i.$$
Renormalize:
$$u_k = \frac{\tilde{u}_k}{\|\tilde{u}_k\|}.$$

How to make a given basis orthogonal? (3)

Then this constructs an orthogonal basis:

- The new vector is orthogonal to the old ones: For every $j < k$,

$$\langle \tilde{u}_k, u_j \rangle = \left\langle v_k - \sum_{i=1}^{k-1} \langle v_k, u_i \rangle u_i, u_j \right\rangle = \langle v_k, u_j \rangle - \langle v_k, u_j \rangle = 0.$$

- The new vector is non-zero: If $\tilde{u}_k = 0$, then we would have

$$v_k = \sum_{i=1}^{k-1} \langle v_k, u_i \rangle u_i \in \mathsf{span}\{v_1, \ldots, v_{k-1}\},$$

contradicting the linear independence of $v_1, \ldots, v_n$.

How to make a given basis orthogonal? (4)

Thus the procedure constructs an orthonormal basis.

In practice, it quickly results in numerical errors. For better numerical stability, see the QR factorization later.

□

Orthogonal matrices

Orthogonal matrices

Definition 69 (Orthogonal matrices)

Let $Q \in \mathbb{R}^{n \times n}$ be a matrix with orthonormal column vectors (with respect to the Euclidean scalar product). Then Q is called an orthogonal (!) matrix.

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

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 $\mathbb{R}^2$:**

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

Orthogonal matrices, examples (2)

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

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

Properties of orthogonal matrices

Let Q be orthogonal. Then:

- Q is always invertible, and $Q^{-1} = Q^T$.
- Columns are orthogonal $\iff$ rows are orthogonal.
- Q realizes an isometry:
$$\forall v \in V,\ \|Qv\| = \|v\|.$$
- Q preserves angles:
$$\langle Qv, Qu\rangle = \langle v, u\rangle \quad \forall u, v \in V.$$
- $|\det Q| = 1$.

The same properties hold for unitary matrices U, with $U^{-1} = \bar{U}^t$.

Proofs: Exercise.

⚠ Orthogonal rows and columns

Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}.$$

- Has full rank
- The columns are orthogonal, its rows are not orthogonal!
- The issue is that the norms of the rows / cols are not 1, and then we cannot conclude from orthogonal rows to orthogonal columns or vice versa. We can only make a conclusion if the norm of all rows / cols is 1 (as it is the case in an orthogonal matrix):

The statement "*columns orthogonal* $\iff$ *rows orthogonal*" does not hold for arbitrary matrices.

The statement "*columns orthonormal* $\iff$ *rows orthonormal*" holds for all square matrices.

Representation of isometric mappings

Let $S \in \mathcal{L}(V)$ for a real vector space V. Then equivalently:

(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:

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

where each of the little blocks

- either is a 1×1 matrix (λ a real number) being 1 or -1
- or is a 2×2 rotation matrix:

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

SKIPPED

ML Keywords: Covariance matrices, kernel matrices, similarity matrices, ...)

Symmetric matrices

Symmetric matrices

Definition 70 (Symmetric matrices)

- A matrix $A \in \mathbb{R}^{n \times n}$ is called symmetric if:

$$A = A^\top.$$

- A matrix $A \in \mathbb{C}^{n \times n}$ is called Hermitian if:

$$A = \overline{A}^\top.$$

Hermitian matrices have real-valued eigenvalues and orthogonal eigenvectors

Proposition 71

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

Hermitian matrices have real-valued eigenvalues and orthogonal eigenvectors (2)

Proof.

- λ eigenvalue of A with eigenvector x. Then:

$$\lambda\langle x, x\rangle = \langle \lambda x, x\rangle = \langle Ax, x\rangle = \langle x, Ax\rangle = \langle x, \lambda x\rangle = \overline{\lambda}\langle x, x\rangle$$

$$\implies \lambda = \overline{\lambda} \in \mathbb{R}.$$

- Let $(\lambda_1, x_1), (\lambda_2, x_2)$ be eigenpairs of A, $\lambda_1 \neq \lambda_2$. Then:

$$\lambda_1\langle x_1, x_2\rangle = \langle Ax_1, x_2\rangle = \langle x_1, Ax_2\rangle = \lambda_2\langle x_1, x_2\rangle$$

$$\implies 0 = \lambda_1\langle x_1, x_2\rangle - \lambda_2\langle x_1, x_2\rangle = (\lambda_1 - \lambda_2)\langle x_1, x_2\rangle.$$

$$\implies \text{either } \lambda_1 = \lambda_2 \text{ or if } \lambda_1 \neq \lambda_2, \text{ then } \langle x_1, x_2\rangle = 0.$$

$$\implies x_1 \perp x_2.$$

What can we conclude from here?

- Each matrix $A \in \mathbb{C}^{n \times n}$ has n eigenvalues in $\mathbb{C}$.
- For Hermitian matrices $A \in \mathbb{C}^{n \times n}$, all n eigenvalues now live in $\mathbb{R}$. The eigenvectors still might live in $\mathbb{C}^n$ (and not in $\mathbb{R}^n$).
- Symmetric matrices $A \in \mathbb{R}^{n \times n}$ are a special case of Hermitian (because $A = A^T = \overline{A^T}$), so it also has real-valued eigenvalues. Moreover, the eigenvectors can always be chosen in $\mathbb{R}^n$.

Self-adjoint operators

Definition 72 (Self-adjoint operators)

An operator $T \in \mathcal{L}(V)$ on a pre-Hilbert space V is called **self-adjoint** if:

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

Sometimes it is called a Hermitian operator (on $\mathbb{C}^n$) or symmetric operator (on $\mathbb{R}^n$).

Remark: Over $\mathbb{C}^n$, self-adjoint operators are represented by Hermitian matrices. Over $\mathbb{R}^n$, self-adjoint operators are represented by symmetric matrices.

Self-adjoint operator has real-valued eigenvalue and eigenvector

Proposition 73 (Self-adjoint operator has real-valued eigenvalue and eigenvector)

Let V be a finite-dimensional 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 the case of $\mathbb{R}$, the eigenvector also lies in $\mathbb{R}^n$.

⚠

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

Self-adjoint operator has real-valued eigenvalue and eigenvector (2)

Proof.

Consider the case $F = \mathbb{R}$.
Let $n := \dim V$. Choose $0 \neq v \in \mathbb{R}^n$, and consider:

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

These vectors have to be linearly dependent ($n+1$ vectors, dimension n).
Thus, there exist $a_0, \ldots, a_n \in \mathbb{R}$ (not all 0) such that:

$$a_0v + a_1Tv + \cdots + a_nT^nv = 0.$$

Self-adjoint operator has real-valued eigenvalue and eigenvector (3)

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

$$a_0+a_1x+a_2x^2+\cdots+a_nx^n = c(x^2+b_1x+c_1)\dots(x^2+b_Mx+c_M)(x-d_1)\dots(x-d_m),$$

where $c \neq 0$, $b_i, c_i, d_i \in \mathbb{R}$, and $M+m \geq 1$

Quadratic terms $x^2+b_ix+c_i$ cannot be decomposed into linear terms; in particular, they satisfy:

$$b_i^2 < 4c_i \quad \text{(otherwise, one could factorize them by the quadratic formula)}$$

Self-adjoint operator has real-valued eigenvalue and eigenvector (4)

Replace x by T:

$$0 = (a_0 I + a_1 T + \cdots + a_n T^n)v = \big(c\underbrace{(\dots)}_{\text{quadr.}}\underbrace{(\dots)}_{\text{linear}}\big)v.$$

Now we can prove: "Quadratic Operators" are invertible (See ★ below):
Thus, we multiply the equation with their inverses and obtain:

$$0 = (T - \lambda_1 I)\dots(T - \lambda_m I)v.$$

Because we had chosen $v \neq 0$, at least one of the terms $(T - \lambda_i I)$ has to exist $(m \geq 1)$. Then also, the product $(T - \lambda_1 I)\dots(T - \lambda_m I)$ is not injective, hence at least one of the factors is not injective. Thus, λ_i is an eigenvalue over $\mathbb{R}$. In particular, the eigenvector lives in $\mathbb{R}^n$ as well. □

Self-adjoint operator has real-valued eigenvalue and eigenvector (5)

(★) Invertibility of Quadratic Operators:
Proposition: Suppose T self-adjoint and $b, c \in \mathbb{R}$ satisfy $b^2 < 4c$. Then $T^2 + bT + cI$ is invertible

Intuition:

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 + \left(c - \frac{b^2}{4}\right) > 0, \text{ since } \left(c - \frac{b^2}{4}\right) > 0 \text{ by assumption.}$$

So $x^2 + bx + c$ is over invertible real numbers.

Self-adjoint operator has real-valued eigenvalue and eigenvector (6)

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

$$
\begin{aligned}
\langle (T^2 + bT + cI)v, v\rangle &= \langle T^2 v, v\rangle + b\langle Tv, v\rangle + c\langle v, v\rangle \\
&= \langle Tv, Tv\rangle + b\langle Tv, v\rangle + c\|v\|^2 \\
&\geq \|Tv\|^2 - |b|\|Tv\|\|v\| + c\|v\|^2 \quad \text{(Cauchy-Schwarz)} \\
&= \|Tv\| - \frac{|b|}{2}\|v\|)^2 + \big(c - \frac{b^2}{4}\big)\|v\|^2 > 0. \\
&\implies (T^2 + bT + cI)v \neq 0 \quad \forall v \neq 0, \\
&\implies T^2 + bT + cI \text{ is invertible.}
\end{aligned}
$$

□

Spectral Theorem for symmetric/Hermitian matrices

Theorem 74 (Spectral Theorem for symmetric/Hermitian matrices)

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 $D \in \mathbb{R}^{n\times n}$ such that:

$$\boxed{A = QDQ^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T.}$$

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

Spectral Theorem for symmetric/Hermitian matrices (2)

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

Base case: $n = 1$, clear.

Inductive step: $n - 1 \implies n$.

- A symmetric $\implies$ A has at least one eigenvector $u \in \mathbb{R}^n$ with $\lambda \in \mathbb{R}$ (previous theorem).
- Define $U := \text{span}\{u\}$. U is invariant under A, i.e., $\forall v \in U : Av \in U$.
- 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 dimension $n - 1$. □

Complex version of this theorem

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

Theorem 75 (Complex Spectral 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 such that:

$$A = UD\overline{U}^T.$$

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

Positive definite matrices

Positive definite matrices

Definition 76 (Positive definite matrices)

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

- Positive definite (pd) if $x^T A x > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
- Positive semi-definite (psd) if $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$.

Gram Matrices

Definition 77 (Gram matrices)

A matrix A is called a Gram matrix if there exists a set of vectors $v_1, \ldots, v_n$ such that:

$$a_{ij} = \langle v_i, v_j \rangle.$$

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 78 (Characterization of pd matrices over $\mathbb{C}$)

Let $A \in \mathbb{C}^{n\times n}$ be Hermitian. Then the following are equivalent:

(i) A is psd (pd).

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

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

$$\langle x, y \rangle_A := \overline{y}^T A x,$$

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 linearly independent (which are linearly independent)

Characterization of pd matrices over $\mathbb{C}$ (2)

- Over $\mathbb{C}$, positive definiteness implies self-adjointness.
- Over $\mathbb{R}$, this is **not** true: pd $\not\Rightarrow$ symmetric. To get a similar characterization of pd matrices over $\mathbb{R}$, we need to add a symmetry condition.

Characterization of pd matrices over $\mathbb{C}$ (3)

Example:

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

- Over $\mathbb{R}$, if $x \neq 0$, then:

$$x^T A x = x_1^2 + x_2^2 > 0 \qquad \text{on } \mathbb{R} \text{ if } x \neq 0.$$

  Thus, 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 $\mathbb{R}$

TODO

Roots of psd matrices

Theorem 79 (Roots of psd matrices)

Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive semi-definite (psd). Then there exists a matrix $B \in \mathbb{R}^{n \times n}$, also psd, such that:

$$A = B^2.$$

Sometimes, B is called the **square root** of A, and is denoted as:

$$B = A^{1/2}.$$

Roots of psd matrices (2)

Proof.

- By the Spectral Theorem, we can write:

$$A = UDU^T,$$

  where $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, with $\lambda_i \geq 0$ (since A is psd).

- Define:

$$\sqrt{D} := \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}).$$

Roots of psd matrices (3)

- Set:

$$B := U\sqrt{D}U^T.$$

- Then:

$$B^2 = (U\sqrt{D}U^T)(U\sqrt{D}U^T) = U(\sqrt{D})^2U^T = UDU^T = A.$$

□

Remark:

More generally, we can define $A^{1/k}$ for any $k \in \mathbb{N}$. Moreover, we can also define A^{-1} (in case A is psd and invertible) and A^k. For $p, q \in \mathbb{Z}$, we can define $A^{p/q}$.

Important psd matrices for ML

- Consider a data matrix $X \in \mathbb{R}^{n \times d}$, with n data points and d dimensions. We assume the data are centered, that is, each column of X has mean 0.
- Then, the matrix:
$$C := X^T X \in \mathbb{R}^{d \times d}$$
is called the covariance matrix (up to the factor $1/n$).
- 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 (psd).

(Prove it!)

Variational characterization of eigenvalues

Literature: Bhatia: Matrix Analysis

Rayleigh coefficient

Definition 80 (Rayleigh coefficient)

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. The function:

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

is called the Rayleigh coefficient.

Rayleigh coefficient $\to$ first eigenvalue

Proposition 81 (Rayleigh coefficient $\to$ first eigenvalue)

Let A be symmetric, and let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues and $v_1, \ldots, v_n$ the eigenvectors of A. Then:

- Smallest eigenvalue:

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

- Biggest eigenvalue:

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

Intuition for the proposition

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

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

Let y be a vector, also represented in this basis:

$$y = y_1 v_1 + y_2 v_2 + \cdots + y_n v_n, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Then:

$$y^T A y = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2.$$

Intuition for the proposition (2)

Among the vectors

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

the smallest result of $y^T Ay$ is given by the vector

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and the value would be λ_1. Therefore the minimum equals λ_1, and it is attained at v_1.

Formal proof (sketch)

Assume we start with the standard basis. Let $Q = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$ be the orthogonal matrix whose columns are the eigenvectors of A, so that

$$A = Q\Lambda Q^T, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

For $x \neq 0$ consider the transformed vector $y := Q^T x$. As Q is orthogonal, $\|y\| = \|x\|$ and $x = Qy$. Hence

$$R_A(x) = \frac{x^T A x}{x^T x} = \frac{(Qy)^T Q\Lambda Q^T (Qy)}{\|y\|^2} = \frac{y^T \Lambda y}{\|y\|^2} = \frac{\lambda_1 y_1^2 + \cdots + \lambda_n y_n^2}{\|y\|^2}.$$

Now we look at the minimum:

$$\min_{x \neq 0} R_A(x) = \min_{\|y\|=1} \left(\lambda_1 y_1^2 + \cdots + \lambda_n y_n^2\right) = \lambda_1,$$

Formal proof (sketch) (2)

since $\lambda_1 \leq \lambda_i$ for all i and $\sum_i y_i^2 = 1$. The minimum is attained for $y = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, that is for $x = Qy = v_1$, with value λ_1. □

Rayleigh coefficient $\rightarrow$ second eigenvalue

Proposition 82 (Rayleigh coefficient $\rightarrow$ second eigenvalue)

Consider a symmetric matrix A with eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Consider the optimization problem:

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

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

Proof intuition

Consider operator A restricted to the space

$$V_1^{\perp} := (\text{span}\{v_1\})^{\perp}.$$

We know that on this space, A is invariant and symmetric, so we can apply Rayleigh to this "smaller" space:

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

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

Min-max-Theorem / Courant-Fischer-Weyl

Theorem 83 (Min-max-Theorem / Courant-Fischer-Weyl)

Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. Then:

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

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

Proof intuition

- For the case $k = 1$: Case (2) is essentially what we have already proven:

$$\max_{\substack{U \text{ subspace} \\ \dim U = n-k+1}} \min_{x \in U \setminus \{0\}} R_A(x) = \min_{x \in V \setminus \{0\}} R_A(x) = \lambda_1.$$

  Here, $\max_{\substack{U \text{ subspace} \\ \dim U = n-k+1}}$ can be dropped, and the result corresponds to the previous result.
- Case (1) follows similar principles.
- Case $k = 2$: Similar to the previous argument.
- General case: Induction.

ML Keywords: Low rank approximation, Perturbation analysis

Matrix norms

Motivation

We 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 $A \in \mathbb{R}^{m \times n}$, we define the following norms:

- Maximum norm (entry-wise maximum):

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

- Frobenius norm:

$$\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{\operatorname{tr}(A^T A)} = \sqrt{\sum \sigma_i^2}$$

where σ_i are the singular values of A.

Definitions of matrix norms (2)

- Spectral norm (operator norm):

$$\|A\|_2 = \sigma_{\max}(A)$$

where $\sigma_{\max}(A)$ is the largest singular value.
Alternatively, the spectral norm can also be defined as:

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

where $\|x\|$ denotes the Euclidean norm, on $\mathbb{R}^n$ in the denominator and on $\mathbb{R}^m$ in the numerator.

Definitions of matrix norms (3)

- Induced (operator) norms:

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

In particular

$\|A\|_1 = \max_j \sum_i |a_{ij}|$ (max. absolute column sum), $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ (max

The spectral norm $\|A\|_2$ above is the case $p = 2$.

- Nuclear norm:

$$\|A\|_* = \mathrm{tr}(\sqrt{A^T A})$$

Note: Many more matrix norms exist beyond those listed here.

Simple inequalities

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

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2$$

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

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

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

Singular value decomposition

ML motivation: recommender systems

- Netflix ratings: huge matrix of user-movie-ratings!

Figure: The dot represents 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 the matrix.

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) (2)

Proposition 84 (Singular Value Decomposition (SVD))

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

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

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":

$$\Sigma = \underbrace{\begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}}_{\text{case } r=n} \quad \text{or} \quad \underbrace{\begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_r & 0 & \dots & 0 \end{pmatrix}}_{\text{case } r=m}.$$

Exactly r of the diagonal values $\sigma_1, \sigma_2, \dots$ are non-zero. In the general case $r < m$ and $r < n$, so Σ has both zero rows and zero columns.

Illustration

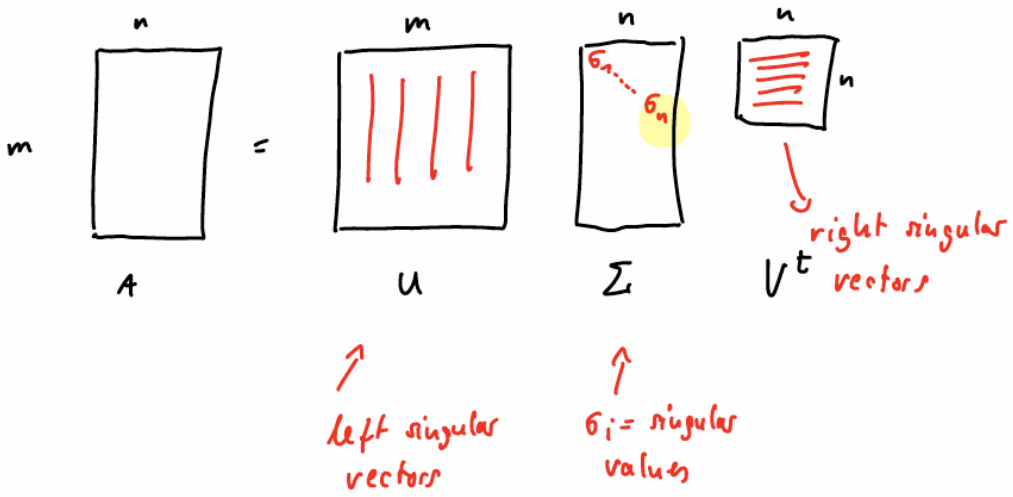


Illustration (2)

Matrix	Dimension	Meaning
U	$m \times m$	Left singular vectors
Σ	$m \times n$	Singular values (σ_i)
V^T	$n \times n$	Right singular vectors

Remark: It has $\min(m, n)$ singular values.

Proof sketch

Given $A \in \mathbb{R}^{m \times n}$, we consider

$$B := A^\top A \in \mathbb{R}^{n \times n}.$$

Observe:

- B is symmetric:
$$(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A.$$
- B is positive semi-definite:
$$x^\top B x = \langle x, Bx \rangle = \langle x, A^\top A x \rangle = \langle Ax, Ax \rangle = \|Ax\|^2 \geq 0.$$

So there exists an orthonormal basis of eigenvectors $v_1, \ldots, v_n$ of B with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ (ordered decreasingly). Exactly $r := \operatorname{rank}(A)$ of them are non-zero.

Proof sketch (2)

Define:

- $\sigma_i := \sqrt{\lambda_i}$, so $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$.
- For $i = 1, \ldots, r$ only (for $i > r$ we would divide by 0) set

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

  These are orthonormal, because $\langle Av_i, Av_j \rangle = v_i^\top A^\top A v_j = \lambda_j \langle v_i, v_j \rangle$. Extend $u_1, \ldots, u_r$ to an orthonormal basis $u_1, \ldots, u_m$ of $\mathbb{R}^m$.
- $U = \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix} \in \mathbb{R}^{m \times m}$, $\quad V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \in \mathbb{R}^{n \times n}$, both with the vectors as columns.
- $\Sigma \in \mathbb{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i$ for $i \leq r$, and all other entries 0.

Proof sketch (3)

Now we need to show that with these definitions, we have

$$A = U\Sigma V^\top.$$

Proof sketch (4)

Verification:

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

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

- Now multiply with $V^\top$:
  - Rows of $V^\top$ 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 term with $i = j$ will result in a factor of 1.

So we will be left with matrix A.

Conclusion: 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:

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

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

Proof.

Exercise. □

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

- SVD always exists, no matter how A looks like! It can be rectangular, does not need to be symmetric, etc.
- 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) (2)

If $A \in \mathbb{R}^{n \times n}$ is **symmetric**, then the SVD is "nearly the same" as the eigenvalue decomposition:
If (λ_i, v_i) are the eigenvalues/vectors of A, then:
$(|\lambda_i|, v_i)$ are the singular values/vectors of A.

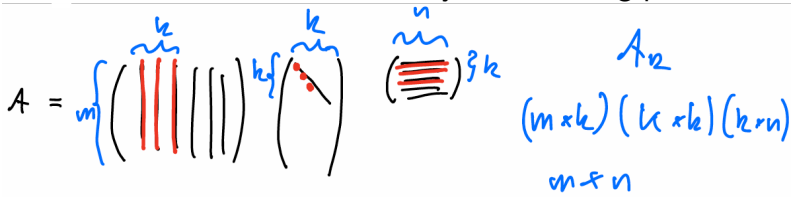
In particular, left- and right-singular vectors are the same (up to signs).

Relationship between SVD (A) and eig (AA^T) (AA^T is symmetric!)

- For general (not necessarily square) matrices A:
  - Left-singular vectors of A are the eigenvectors of AA^T.
  - Right-singular vectors of A are the eigenvectors of A^TA.
- $\lambda \neq 0$ is an eigenvalue of $AA^T \iff \sqrt{\lambda}$ is a singular value of A.

Rank-k-approximation

Given the SVD $A = U\Sigma V^T$, where the entries $\sigma_1, \sigma_2, \ldots$ are sorted in decreasing order, let $k \leq r$. We define a new matrix A_k by the following procedure:



To build A_k:

- Take the first k columns of U
- the first k entries of Σ
- and the first k rows of V^T.

Rank-k-approximation (2)

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 85 (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 rank-k 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:

$$\|A - A_k\| = \begin{cases} \sigma_{k+1} & \text{in case of } \|\cdot\|_2, \\ \sqrt{\sum_{i=k+1}^{\min(m,n)} \sigma_i^2} & \text{in case of } \|\cdot\|_F. \end{cases}$$

Proof sketch. Skipped. ☺

SVD and matrix norms

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

- $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2}$,
- $\|A\|_2 = \sigma_1$.

Proof sketch. Skipped. ☺

Relation to machine learning

Computational reasons for a rank-k-approximation:
The running time of many ML algorithms scales heavily in the (implicit) dimension. Often they can be implemented efficiently if matrices are (sparse or) low rank.

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

- Typical assumption: Data lies on a **low-dimensional** manifold ($\Rightarrow$ locally low rank).
- (Data is sparse.)

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

Pseudo inverse

Pseudo-inverse

Definition 86 (Pseudo-inverse)

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:

$$
\begin{aligned}
&(1) \quad AA^{\#}A = A \quad (\text{"nearly inverse"}) \\
&(2) \quad A^{\#}AA^{\#} = A^{\#} \quad (\text{"nearly inverse"}) \\
&(3) \quad (AA^{\#})^T = AA^{\#} \quad (\text{symmetry}) \\
&(4) \quad (A^{\#}A)^T = A^{\#}A \quad (\text{symmetry})
\end{aligned}
$$

Remark: If A were invertible, $A^{\#} = A^{-1}$.

Pseudo-inverse (2)

Intuition:

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

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

- This cannot be inverted, as inversion would reconstruct the original point.
- But we can "invent" a reconstruction, for example:

$$R : \mathbb{R}^2 \to \mathbb{R}^3, \quad R \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix}.$$

- Now we have:

$$ARA = A, \quad AA^{\#}A = A, \text{ with } R = A^{\#}$$

Intuition: How could we define a pseudo-inverse?

- Consider a diagonal matrix:

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}, \text{ with } \lambda_1, \ldots, \lambda_k \neq 0.$$

- Then set $A^{\#} \in \mathbb{R}^{n \times m}$ to be the "diagonal" matrix with $1/\lambda_1, \ldots, 1/\lambda_k$ on the diagonal and 0 everywhere else. Note that its shape is transposed with respect to A.
- This does the job for diagonal matrices.

Intuition: How could we define a pseudo-inverse? (2)

- How do we generalize to all matrices? $\Rightarrow$ Use SVD!

Moore-Penrose pseudo-inverse

Proposition 87 (Moore-Penrose pseudo-inverse)

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

$$A^{\#} := V\Sigma^{\#}U^T$$

with $\Sigma^{\#} \in \mathbb{R}^{n \times m}$ defined as:

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

Proof sketch. Easy, just do it. ☺

Moore-Penrose pseudo-inverse (2)

Proposition 88

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

Proof sketch. Easy, just do it. ☺

Trace of a matrix

Trace

Definition 89 (Trace)

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

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

Properties

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

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

- $\mathrm{tr}(A \cdot B) = \mathrm{tr}(B \cdot A)$.
  ⚠ $\mathrm{tr}(A \cdot B) \neq \mathrm{tr}(A) \cdot \mathrm{tr}(B)$.
- The trace does **not** depend on the basis. Let $T \in \mathcal{L}(V)$, and let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two bases of V. Then:

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

Properties (2)

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

$$\tilde{A} = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \text{ w.r.t. some basis} \{v_1, \ldots, v_n\}$$

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

Properties (3)

- With the characteristic polynomial written in monic form, $p_A(t) := \det(tI - A)$, the trace equals the negative of the coefficient in front of t^{n-1}:

$$p_A(t) = \det(tI - A) = t^n + a_{n-1}t^{n-1} + \ldots, \qquad \operatorname{tr}(A) = -a_{n-1}.$$

(Careful: elsewhere we also write $\det(A - tI) = (-1)^n p_A(t)$, which is monic only for even n.)

Trace vs. determinant

$$\mathrm{tr}(A) = \textbf{sum} \text{ of eigenvalues}$$

$$\det(A) = \textbf{product} \text{ of eigenvalues}$$

Riddle

Consider a real-valued matrix $A \in \mathbb{R}^{n \times n}$. Over $\mathbb{C}$, we can always find eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and bring the matrix into triangular form $\tilde{A}$. Then:

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

and

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

but the two are identical because trace is ind. of the 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:

Riddle (2)

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 $2\cos\theta$.
- The characteristic polynomial of A is:

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

which expands to:

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

Riddle (3)

- The roots of $p(t)$ are:

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

These are complex eigenvalues:

$$\lambda_{1,2} = \cos\theta \pm i\sin\theta.$$

- The diagonal representation of A is:

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

The trace is:

$$\operatorname{tr}(A) = \lambda_1 + \lambda_2 = (\cos\theta + i\sin\theta) + (\cos\theta - i\sin\theta) = 2\cos\theta \in \mathbb{R}.$$

What is going on?

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

Reason: 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:

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

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

Matrix series

Spectral radius

Definition 90 (Spectral radius)

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 considers all complex eigenvalues of A.)

Spectral radius (2)

Proposition 91

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

(In Frobenius norm: $\|A^k - 0\|_F \to 0$.)

Spectral radius (3)

Proof sketch.

- "$\Rightarrow$": Let v be the eigenvector of the eigenvalue λ that defines $\rho(A)$. Then:

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

  This implies $\lim_{k\to\infty} \lambda^k = 0$, and hence $|\lambda| < 1$.

- "$\Leftarrow$": Proof for symmetric matrices:
  Write $A = UDU^T$, then $A^k = UD^kU^T$.
  Since $\|A^k\|_F = \|D^k\|_F$, and $D^k \to 0$, it follows that $\|A^k\|_F \to 0$.

For general matrices: Use Jordan normal form (proof skipped). ☺

Neumann series

Proposition 92 (Neumann series)

- The series $\sum_{k=0}^{\infty} A^k$ converges if and only if $\rho(A) < 1$.
- If $\rho(A) < 1$, then $(I - A)$ is invertible and:

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

Over $\mathbb{R}$: $\frac{1}{1-a} = \sum_{k=0}^{\infty} a^k$ for $|a| < 1$.

Proof intuition:

For symmetric matrices, apply matrix powers to diagonal matrix D. For general matrices, proof requires 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:

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

Implementing matrix operations

Literature:
Golub/van Loan: Matrix computations

Triangular matrices are great

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

- Solving a linear 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 the bottom:

$$x_3 = b_3/a_{33},$$

  then plug into the second-to-last row:

$$x_2 = (b_2 - a_{23}x_3)/a_{22},$$

  and so on.

Triangular matrices are great (2)

- 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 systems: theory

Solving linear systems is one of the most basic tasks and appears as a subset of more complex algorithms.

$$Ax = b$$

Proposition 93 (Linear systems)

Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

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

(ii) The system $Ax = b$ has at least one solution if and only if $b \in \mathrm{Range}(A)$.

(iii) If w is a solution of $Ax = b$, then the full set of solutions is given by:

$$w + \ker(A) := \{w + v \mid v \in \ker(A)\}.$$

Gaussian 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 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, clever tricks such as selecting the best row for pivoting (reordering, rearranging, etc.) can be done.

Computational complexity in general case: $\mathcal{O}(n^3)$.

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:

$$\boxed{A = L \cdot U},$$

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

LU decomposition exists under certain conditions. In particular, for symmetric positive-definite matrices, it always exists.

Computational complexity: $\mathcal{O}(n^3)$.

LU decomposition to solve a linear system

Once we have an LU decomposition ($\mathcal{O}(n^3)$), the solution to the problem $Ax = b$ can be found by solving two linear systems (trivial due to triangular form):

$$\begin{aligned} &(1) \quad Ly = b \qquad (\mathcal{O}(n^2)) \\ &(2) \quad Ux = y \qquad (\mathcal{O}(n^2)) \end{aligned}$$

$$\Rightarrow Ax = LUx = Ly = b$$

LU decomposition to compute the inverse

To compute the inverse of a matrix A, we need to find a matrix X that satisfies $AX = I$. This corresponds to solving n linear systems:

$$Ax_i = e_i,$$

where e_i are the standard basis vectors. Once we have the LU decomposition of A, we can simply solve these systems.

Complexity complexity: $\mathcal{O}(n^3)$

Cholesky factorization

For the special case of symmetric, positive-definite matrices, we can simplify the LU decomposition:

$$\boxed{A = LL^T},$$

where L is a lower triangular matrix. This is numerically very stable, uses less memory than LU, and is faster than LU (still $O(n^3)$, but with better constants).

QR decomposition

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

$$\boxed{A = QR},$$

where Q is an orthogonal matrix, and R is upper triangular. Several algorithms exist, with different advantages and disadvantages.

Complexity complexity: $\mathcal{O}(n^3)$

QR decomposition to find an orthogonal basis

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

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

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

Iterative procedure:

- Set $A^{(0)} = A$.
- For $k = 1, 2, \ldots$:
  - Compute the QR factorization of $A^{(k-1)}$:
  $$A^{(k-1)} = Q^{(k)} R^{(k)}.$$
  - Recombine in reversed order:
  $$A^{(k)} := R^{(k)} Q^{(k)}.$$

Note:

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

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

Largest eigenvalue and eigenvector

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

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Power method (vanilla version):

- Start with a random vector x_0.
- Iteratively compute:

$$v_{k+1} = \frac{Av_k}{\|Av_k\|}.$$

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

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

Conjugate gradient method for linear systems of sparse symmetric positive-definite matrices

Problem:
Many of the algorithms we have seen (LU, QR, etc.) cannot fully exploit sparsity of a matrix.

Bad: in ML, many matrices are very sparse.

Solution: Use iterative methods that are based on matrix-vector multiplications. (example: power method)

Iterative methods for sparse matrices

- We want to solve $Ax = b$.
- Consider the minimization problem:

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

  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).

Spaces of functions: Continuous, differentiable, integrable
SKIPPED

SKIPPED

Space of continuous functions

Definition 94 (Space of continuous functions)

Let T be a metric space,

$$C^b(T) := \{f \colon T \to \mathbb{R} \mid f \text{ is continuous and bounded}\}$$

$$\text{i.e.,} \quad \exists c \in \mathbb{R} \text{ such that } \forall t \in T \colon |f(t)| < c$$

As a norm on $C^b(T)$ we use

$$\|f\|_\infty := \sup_{t \in T} |f(t)|$$

SKIPPED

Space of continuous functions (2)

Proposition 95

The space $C^b(T)$ with the norm $\|\cdot\|_\infty$ is a Banach space, i.e., 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

SKIPPED

Space of differentiable functions

Definition 96 (Space of differentiable functions)

Let $[a, b] \subset \mathbb{R}$, then

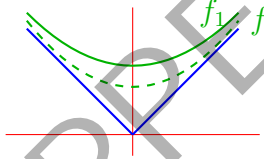
$$C^1([a,b]) := \{f \colon [a,b] \to \mathbb{R} \mid f \text{ is continuously differentiable}\}$$

SKIPPED

Space of differentiable functions (2)

Which norm?

- Consider $\|\cdot\|_\infty$. With this norm, C^1 is **not** complete!



$\rightarrow$ limit function, not differentiable!

- $$\|f\|_{C^1} := \sup_{t\in[a,b]} \max\{|f(t)|,\ |f'(t)|\}$$

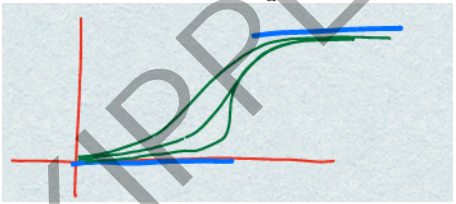
$$\|f\| := \|f\|_\infty + \|f'\|_\infty$$

$C^1([a,b])$ with any of these two norms is a Banach space.

SKIPPED

Spaces of integrable functions?

- Consider $C^b([a,b])$ with the norm $\|f\|_1 := \int_a^b |f(t)|\,dt$

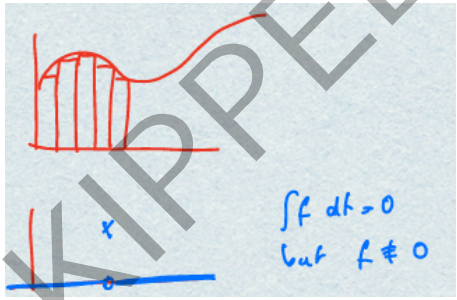


We can see: $\|\cdot\|_1$ is a norm, but the space is not complete.

SKIPPED

Spaces of integrable functions? (2)

- Consider $\mathcal{R}([a,b])$ of all Riemann-integrable functions on $[a,b] \subset \mathbb{R}$, together with $\|\cdot\|_1$.



However, on $\mathcal{R}([a,b])$, $\|\cdot\|_1$ is **not** a norm: it is not true that

$$\|f\| = 0 \quad \Rightarrow \quad f = 0$$

SKIPPED

The space $\mathcal{L}^p$

Definition 97 (The space $\mathcal{L}^p$)

For $1 \leq p < \infty$, we define

$$\mathcal{L}^p([a,b]) := \left\{ f : [a,b] \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ measurable wrt Lebesgue measure and} \\ \int |f(t)|^p \, dt < \infty \end{array} \right\}$$

As a norm we use

$$\|f\|_p := \left(\int |f(t)|^p \, dt \right)^{1/p}$$

SKIPPED

The space $\mathcal{L}^p$ (2)

Definition 98 (The space $\mathcal{L}^\infty$)

The case $p = \infty$ is defined separately:

$$\mathcal{L}^\infty([a,b]) := \{f : [a,b] \to \mathbb{R} \mid f \text{ measurable}, \ \exists C : |f(t)| \leq C \text{ for almost all } t\}$$

with the norm

$$\|f\|_\infty := \operatorname{ess\,sup}_{t\in[a,b]} |f(t)| = \inf\{C \geq 0 \mid |f(t)| \leq C \text{ for almost all } t\}.$$

SKIPPED

The space $\mathcal{L}^p$ (3)

Proposition 99

$\|f\|_p$ is a semi-norm on $\mathcal{L}^p$.

Proof.

- Vector space: $\mathcal{L}^p$ is a vector space.
- Semi-norm: observe: $\|f\|_p = 0 \Rightarrow f = 0$ almost everywhere (a.e.)
  So we do **not** have that $\|f\|_p = 0 \Rightarrow f = 0$.

□

⚠ $\|f\|_p$ 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.

SKIPPED

The space $\mathcal{L}^p$ (4)

Proposition 100

$\mathcal{L}^p$ is complete under $\|\cdot\|_p$.

Proof.

SKIPPED

If $(f_i)_{i\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{L}^p$, then want to prove that $\lim_{i\to\infty} f_i \in \mathcal{L}^p$.
This is equivalent to proving the following:

Let $(f_i)_i$ be a sequence such that

$$a := \sum_{i=1}^{\infty} \|f_i\|_p < \infty \qquad (\bigstar)$$

Then there exists $f \in \mathcal{L}^p$ such that $\sum_{i=1}^{n} f_i \to f$ in $\|\cdot\|_p$ as $n \to \infty$.

The space $\mathcal{L}^p$ (5)

Define

$$\hat{g} := \sum_{i=1}^{\infty} |f_i|$$

Note: This might not yet be a well-defined function from $[a, b] \to \mathbb{R}$, might be ∞ at certain points.

Let

$$\hat{g}_n := \sum_{i=1}^{n} |f_i| \in \mathcal{L}^p$$

By Minkowski:

$$\|\hat{g}_n\|_p \overset{\text{def}}{=} \left\| \sum_{i=1}^{n} |f_i| \right\|_p \overset{\text{Mink.}}{\leq} \sum_{i=1}^{n} \|f_i\|_p \overset{\text{ass.}(\bigstar)}{<} a$$

$$\Rightarrow \hat{g}_n \to \hat{g} \text{ monotonically}$$

SKIPPED

The space $\mathcal{L}^p$ (6)

By theorem of monotonic convergence, g is measurable and we have

$$\lim_{n\to\infty} \int \hat{g}_n^p \, d\lambda = \int \lim_{n\to\infty} \hat{g}_n^p \, d\lambda \overset{\text{def}}{=} \int \hat{g}^p \, d\lambda \leq a^p$$

$\Rightarrow \hat{g} < \infty$ a.e., that is, there exists a set N of measure 0 s.t. on $[a,b] \setminus N$, $\hat{g}$ is finite

Now we can define

$$g(t) := \begin{cases} \hat{g}(t) & t \in [a,b] \setminus N \\ 0 & t \in N \end{cases} \qquad \in \mathcal{L}^p$$

SKIPPED

The space $\mathcal{L}^p$ (7)

From this, it now follows that

$$f(t) := \sum_{i=1}^{\infty} f_i(t), \quad t \notin N$$

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,

$\sum_{n=1}^{\infty} f_n$ converges to f in $\|\cdot\|_p$ because of the Theorem of dominated convergence.

SKIPPED

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.

SKIPPED

From $\mathcal{L}^p$ to L^p (2)

The formal construction goes as follows:

Define

$$N := \ker(\|\cdot\|_p) := \{f \in \mathcal{L}^p \mid \|f\|_p = 0\}$$

It is a subspace of $\mathcal{L}^p$.

Now consider the quotient space of $\mathcal{L}^p$ w.r.t. this subspace:

$$L^p([a,b]) := \mathcal{L}^p([a,b])/N$$

Define a norm on L^p by

$$\| \underbrace{[f]}_{\text{new}} \|_p := \| \underbrace{f}_{\text{old}} \|_p$$

SKIPPED

From $\mathcal{L}^p$ to L^p (3)

This norm is well-defined: if $f, \tilde{f} \in [f]$, then $\|f\|_p = \|\tilde{f}\|_p$.

This "norm" is a norm, because

$$\|[f]\|_p = 0 \quad \Rightarrow \quad [f] = [0]$$

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.

SKIPPED

From $\mathcal{L}^p$ to L^p (4)

- 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 SKIPPED

SKIPPED

Continuous = bounded

Theorem 101 (Continuous = bounded)

Let X, Y be normed spaces and $T : X \to Y$ be linear. Then the following statements are equivalent:

(i) T is continuous.
$$\forall x \in X\ \forall \varepsilon > 0\ \exists \delta > 0\ \forall y \in X : \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon$$

(ii) T is continuous at 0.

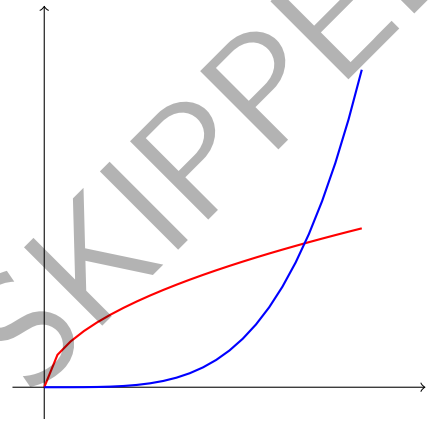
(iii) T is bounded:
$$\exists M > 0\ \forall x \in X : \|Tx\| \leq M \cdot \|x\|$$

(iv) T is uniformly continuous.
$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall x \in X\ \forall y \in X : \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon$$

SKIPPED

Continuous = bounded (2)

- uniformly continuous
- not uniformly continuous



SKIPPED

Operator norm

Definition 102 (Operator norm)

Let X, Y be normed spaces, and let $T : X \to Y$ be linear and continuous.

$$\|T\| := \sup_{x \in X, x \neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\| \leq 1}} \|Tx\| = \sup_{\substack{x \in X \\ \|x\| = 1}} \|Tx\|$$

is called the operator norm of T.

Observe: coincides with the matrix norm $\|\cdot\|_2$ as we had defined it earlier.

SKIPPED

Examples

- **Evaluation operator:** $T : \mathcal{C}([0,1]) \to \mathbb{R}, \quad Tf = f(0).$
  As norm consider $\|\cdot\|_\infty$ on $\mathcal{C}([0,1])$, $|\cdot|$ on $\mathbb{R}$. Then:

$$\|T\| = 1, \qquad \sup_{f \in \mathcal{C}([0,1])} \frac{|Tf|}{\|f\|_\infty} = \sup \frac{|f(0)|}{\|f\|_\infty} = \cdots \text{ exercise } \cdots = 1$$

- **Integral operator:** $T : \mathcal{C}([0,1]) \to \mathbb{R}, \quad Tf = \int_0^1 f(t)\,dt$

  With the same norm as above, T is continuous and

$$\|T\| = 1$$

SKIPPED

Examples (2)

- **Differential operator:** $D : \mathcal{C}^1([0,1]) \to \mathcal{C}([0,1]), \quad f \mapsto f'$
  - 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 SKIPPED

SKIPPED

Dual space

Definition 103 (Dual space)

Let V be a vector space, $T : V \to F$ is called a **functional** (linear).

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 vector space, then the space of all linear, continuous functionals from V to F is called the (topological) dual space V' of V.

SKIPPED

Dual space (2)

Remark: If V is finite dimensional, 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

$$\|T\| := \sup_{v \in V, v \neq 0} \frac{|Tv|}{\|v\|}$$

SKIPPED

Proposition 104 (Dual space is normed space)

V' is a vector space, and the operator norm is indeed a norm on V'.

Dual space (3)

Proposition 105 (Dual space is Banach space)

If V is a normed vector space (but not necessarily complete), then V' with the operator norm is a Banach space.

SKIPPED

Examples

- $K \subset \mathbb{R}$ compact set, $\mathcal{C}(K)$ space of continuous functions with $\|\cdot\|_\infty$.
  Then $\mathcal{C}(K)'$ is equivalent to the space $\mathcal{M}(K)$, 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 $L^q(S)$.

SKIPPED

Riesz representation theorem

Theorem 106 (Riesz representation theorem)

Let H be a Hilbert space, H' its dual. Then the mapping

$$\Phi : H \to H', \quad y \mapsto \langle \cdot, y \rangle$$

is bijective, isometric, and satisfies $\Phi(\lambda x) = \overline{\lambda}\,\Phi(x)$.

Stated differently: For any mapping $x' \in H'$, there exists a unique $y \in H$ such that

$$x'(x) = \langle x, y \rangle.$$

SKIPPED

Adjoint operator SKIPPED

SKIPPED

Adjoint operator

Definition 107 (Adjoint operator)

Let $T : H_1 \to H_2$ be linear and continuous, H_1, H_2 Hilbert spaces. Then there exists an operator $T^* : H_2 \to H_1$ such that

$$\langle Tx, y \rangle_{H_2} = \langle x, T^* y \rangle_{H_1}$$

for all $x \in H_1, y \in H_2$. T^* is called the adjoint of T.

SKIPPED

Remark: The existence of this operator is a consequence of the Riesz representation theorem.

Adjoint operator (2)

Definition 108 (Self-adjoint Operator)

An operator $T : H_1 \to H_1$ is called self-adjoint if

$$\langle Tx, y \rangle = \langle x, Ty \rangle.$$

SKIPPED

Quotient space SKIPPED

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

SKIPPED

Equivalence relation

Definition 109 (Equivalence relation)

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

SKIPPED

Equivalence relation (2)

Example: V vector space, $W \subset V$ subspace.
$v \sim u \Leftrightarrow v - u \in W$

Example: Consider the space $L^1(\mathbb{R})$ of all functions $f : \mathbb{R} \to \mathbb{R}$ that are Lebesgue integrable. Define:

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

SKIPPED

Equivalence class

Definition 110 (Equivalence class)

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

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

Proposition 111

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

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

SKIPPED

Quotient spaces

Constructing quotient spaces:

V vector space, $W \subset V$ subspace, equivalence relation:

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

Denote the equivalence classes as $[v]$.

Observe: the equivalence classes have the form

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

SKIPPED

Quotient spaces (2)

Definition 112 (Quotient "space")

Define the quotient "space" as

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

If $[v], [u] \in V/W$, define:

$$[v] + [u] := [v + u]$$

$$\lambda[v] := [\lambda v]$$

SKIPPED

Quotient spaces (3)

These operations are **well-defined:**

- Suppose $v' \sim v$ (i.e. $v' \in [v]$, $[v] = [v']$) and $u' \sim u$. Is $[v] + [u] = [v'] + [u']$ well-defined?
- $v \sim v' \Leftrightarrow \exists w \in W : v - v' = w$
  $u \sim u' \Leftrightarrow \exists \tilde{w} \in W : u - u' = \tilde{w}$
- Then:

$$[v] + [u] = [v + u], \quad [v'] + [u'] = [v' + u']$$

Need to check: $(v + u) \sim (v' + u')$?

$$(v + u) - (v' + u') = (v - v') + (u - u') \in W \quad \Rightarrow \quad v + u \sim v' + u'$$

- Similarly, for scalar mult.

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

SKIPPED

Quotient spaces (4)

Proposition 113

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

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

SKIPPED