Mathematics for Machine Learning

Ulrike von Luxburg

Winter 2024/25
Department of Computer Science, University of Tübingen

(Version as of January 6, 2026)

Table of contents

Part I

Table of contents (2)

Table of contents (3)

Table of contents (4)

Table of contents (5)

Table of contents (6)

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

Examples of groups

- $(\mathbb{R}^n, +)$, $(\mathbb{R}^+, \cdot)$ are groups.
- $(\mathbb{R}^-, \cdot)$ is not a group.
- $S_n := \{\pi : \{1, \dots, n\} \to \{1, \dots, n\} \mid \pi \text{ is bijective}\}$
  $\circ : S_n \times S_n \to S_n, \quad \pi_1 \circ \pi_2(i) = \pi_1(\pi_2(i))$
  $(S_n, \circ)$ is a group.

Definition of a field

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

- $(\mathbb{R}, +, \cdot)$
- $(\mathbb{C}, +, \cdot)$
- $n \in \mathbb{Z}$, consider $\mathbb{Z}_n := \{0, 1, \ldots, n-1\}$
  $a +_n b := (a + b) \mod n$
  $a \cdot_n b := (a \cdot b) \mod n$
  Then $(\mathbb{Z}_n, +_n, \cdot_n)$ is a field if and only if n is prime.

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

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:

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

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

Quadratic equations over $\mathbb{C}$

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

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

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

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

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

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

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

Fundamental theorem of algebra

Theorem 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)

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

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

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

Outlook

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

Vector space

Definition of a vector space

ML Keywords: data representation

Definition 5 (Vector space)

Let F be a field with identity elements 0 and 1. A vector space over the field F is a set V with a mapping

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

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

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

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

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

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

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

Examples of vector spaces

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

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

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

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

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

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

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

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 $\mathbb{R}^{n \times n}$.

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 $u_1, \ldots, u_n$.

Notation:

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

The set $U := \{u_1, \ldots, u_n\}$ is 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:

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

Examples:

- The vectors $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix} \in \mathbb{R}^3$ are 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

Example

- 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

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

Reducing a set to a basis

ML Keywords: dictionary learning

Proposition 10 (Reduction of a set to a basis)

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

Proof sketch

Proof.

- If U is already linearly independent: done.
- If U is linear dependent:
  $\Rightarrow \quad \exists \lambda_1, \ldots, \lambda_n$ 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 consist of one vector.

Finite-dimensional vector spaces

Definition 11 (finite-dim. vector space)

A VS is called finite-dimensional, if it has a finite basis.

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

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

Extending a set to a basis

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-dim. VS. Then U can be extended to a basis of V.

Proof.

Let $w_1, \dots, w_m$ be a basis of V. Consider the set $\{u_1, \dots, u_n, w_1, \dots, w_m\}$. Remove vectors "from the end" until the remaining vectors are linearly independent.

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

Two finite bases have the same length

Corollary 13 (Length of basis of same VS)

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

Proof.

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

Dimension of a vector space

Definition 14 (Dimension of a vector space)

The length of a basis of a finite-dim. VS V is called the dimension of V.

Linear Mappings

Linear mapping

Definition 15 (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 $U \to V$ is denoted by $\mathcal{L}(U, V)$.

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

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 (2)

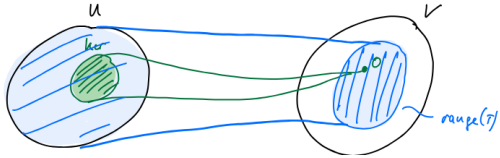
Definition 16 (Kernel and range)

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

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

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

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



Properties of kernel and range

Proposition 17 (Properties of kernel and range)

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

Proof.

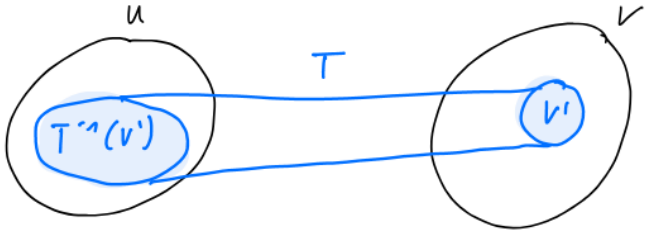
Exercise. □

Pre-image

Definition 18 (Pre-image)

$V' \subset V$, 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 19 (Pre-image subspace)

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

Proof.

Exercise. □

Fundamental theorem for linear mappings

Theorem 20 (Fundamental theorem for linear mappings)

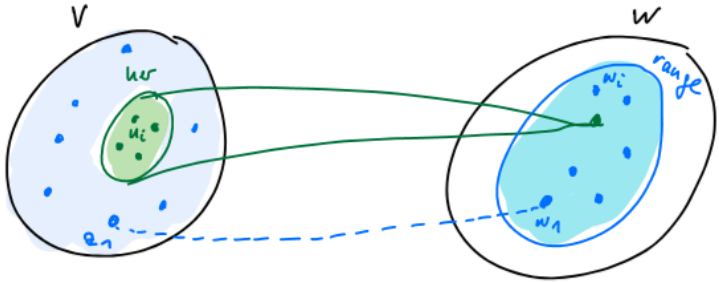
Let V be finite-dimensional, W any VS, $T \in \mathcal{L}(V, W)$.
Let $u_1, \ldots, u_n$ be a basis of $\ker(T) \subset V$.
Let $w_1, \ldots, w_m$ be a basis of range$(T) \subset W$.
Let $z_1 \in T^{-1}(w_1), \ldots, z_m \in T^{-1}(w_n)$.

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 $Tv \in \operatorname{range}(T)$.

$$\Rightarrow \exists \lambda_1, \ldots, \lambda_m \text{ such that}$$

$$\begin{aligned} Tv &= \lambda_1 w_1 + \ldots + \lambda_m w_m \\ &= \lambda_1 T(z_1) + \ldots + \lambda_m T(z_m) \\ &= T(\lambda_1 z_1 + \ldots + \lambda_m z_m) \end{aligned}$$

$$\begin{aligned} \Rightarrow Tv - T(\lambda_1 z_1 + \ldots + \lambda_m z_m) &= 0 \\ = T(\underbrace{v - (\lambda_1 z_1 + \ldots + \lambda_m z_m)}_{\in \ker(T)}) & \end{aligned}$$

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

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

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

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

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

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

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

□

Injective $\Leftrightarrow$ Surjective $\Leftrightarrow$ Bijective

Proposition 21 (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.

Direct consequence of theorem. (Can you see why? Exercise!) □

⚠ Does not work in ∞-dim spaces!

Matrices and linear maps

Matrices

A matrix is the following object:

$$A := \underbrace{\left.\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}\right\}}_{\text{n columns}} \text{ m rows } = (a_{ij})_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$$

Matrices represent linear maps

Consider $T \in \mathcal{L}(V, W)$, V, W finite-dimensional.
Let $v_1, \ldots, v_n$ be a basis of V.
Let $w_1, \ldots, w_m$ be a basis of W.

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

$$v = \lambda_1 v_1 + \ldots + \lambda_n v_n \quad \text{(arbitrary vector)}$$

$$\begin{aligned} T(v) &= T(\lambda_1 v_1 + \ldots + \lambda_n v_n) \\ &= \lambda_1 T(v_1) + \ldots + \lambda_n T(v_n). \end{aligned}$$

Matrices represent linear maps (2)

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

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

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

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

$\Rightarrow$ matrix of 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:

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

$\Rightarrow$ i-th entry of the product of matrix M with vector λ.

Notation for matrices of linear maps

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

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

the matrix corresponding to T w.r.t. bases $\mathcal{B}$ and $\mathcal{C}$.

Properties of matrices

Let V, W be vector spaces, and consider the basis 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

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

Matrix transpose

Definition 22 (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: A^t, $A^\intercal$.

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

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

Sum of vector spaces

Definition 23 (Sum of vector spaces)

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

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

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

Notation: $U_1 \oplus U_2$

Complement of a subspace

Proposition 24 (Complement of a subspace)

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

Proof.

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

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

Define

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

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}.

Remark:

- 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

$$u = T^{-1}(T(u)) = T^{-1}(T(v)) = v \Rightarrow \text{injective}$$

Characterizing invertability (2)

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

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

$$\Rightarrow \text{surjective}$$

Characterizing invertability (3)

"$\Leftarrow$":

Injective & surjective $\Rightarrow$ invertible.

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

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

- Let $v \in V$. Then

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

  Because T is injective, we conclude:

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

Characterizing invertability (4)

- Linearity of S

$$T(S(w_1 + w_2)) = \underline{TS}w_1 + \underline{TS}w_2 = w_1 + w_2.$$

$\overset{\text{inj.}}{\Rightarrow}$ $Sw_1 + Sw_2$ is the unique element in V mapping to $w_1 + w_2$.

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

Similarly for scalar multiplication.

□

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

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

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:
$T : V \to V$

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

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

Proof.

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 $\Leftrightarrow$ $\mathrm{rank}(A) = n$.
- 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

Assume we fix a basis of V (both in source and target space), then the corresponding matrix looks as follows:

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

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

$$\mathcal{J} : (V, A) \to (V, \mathcal{B})$$

look like?

Representing the identity (2)

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

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

This matrix represents the identity mapping

- In the basis A, the first basis vector a_1 has the representation $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $a_1 = 1 \cdot a_1 + 0 \cdot a_2 + \ldots + 0 \cdot a_n$
- $T \cdot \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix}$ This vector gives us Ta_1 expressed in basis $\mathcal{B}$.
- $t_{11}b_1 + t_{21}b_2 + \ldots + t_{n1}b_n = a_1$ by def. of coefficients t_{ij}
- $Ta_1 = a_1$.

Change of basis is invertible

Proposition 29 (Change of basis is invertible)

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

Proof.

Exercise. □

Chance of basis for an arbitrary mapping

Proposition 30 (Chance of basis for an arbitrary mapping)

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

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

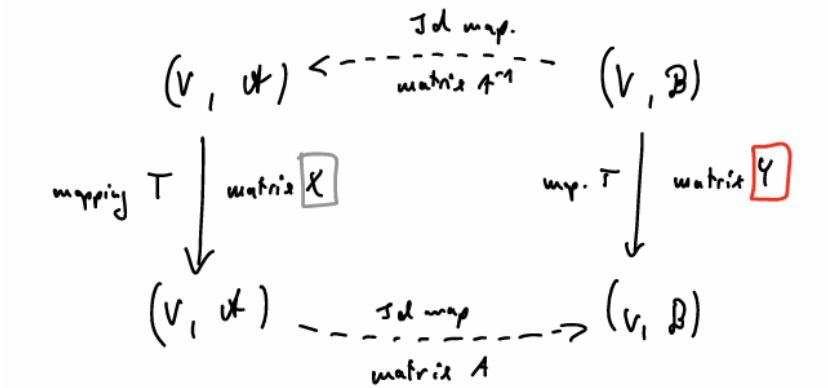
Let $T : V \to V$ be linear, and $\boxed{X} := M(T, \underline{A}, \underline{A})$. Then

$$\boxed{Y := A \cdot X \cdot A^{-1}}$$

represents T in basis $\mathcal{B}$, that is

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

Chance of basis for an arbitrary mapping (2)



Rank of a matrix

Rank of a matrix

ML Keywords: low rank methods, recommender systems, compressed sensing, ...

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.

Rank describes how "complex" a matrix is.

Rank of a matrix (2)

Proposition 32 (Row rank equals column rank)

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

Proposition 33

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

The determinant

Motivation to study the determinant: geometry!

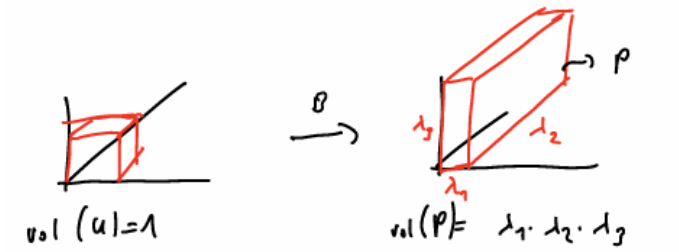
Consider the standard basis of $\mathbb{R}^3$: $e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

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 mapping.

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, "det", that tells us how volumes change under arbitrary linear mappings.

Motivation to study the determinant: geometry! (3)

Which properties would such a mapping d have to satisfy?

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

Now let's look at the formal definition:

Definition of the determinant

Definition 34 (The determinant)

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

(D1) d is linear in each column of the matrix:
Let A be a matrix with columns $a_1, \ldots, a_n$.
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

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

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

(D3) d 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 d exists and is unique.

Proof.

Skipped. □

Properties of the determinant

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

- The determinant as linear mapping does not depend on the basis.
- $\det(c \cdot A) = c^n \det(A)$, with n the dimension of A.
- $\det(A \cdot B) = \det(A) \cdot \det(B)$.
- $\det(A^T) = \det(A)$.
- $\det(A^{-1}) = \frac{1}{\det(A)}$ (if A is invertible).
- A invertible $\Leftrightarrow \det(A) \neq 0$.
- $\det(A + B) \neq \det(A) + \det(B)$.

Properties of the determinant (2)

- If A is upper triangular, that is

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

then $\det(A) = \lambda_1 \cdots \lambda_n$. Same for lower triangular matrices.

Computing the determinant (in theory)

Special cases:

- 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 $n = 3$:

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

Computing the determinant (in theory) (2)

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.

Alternative definition of determinant

There exists a more straightforward definition of determinant. However, starting from this definition, proving the geometric properties is more cumbersome.

Definition 36 (Alternative)

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

Proof.

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

Computing the determinant (in practice)

- LU decomposition:
  Any matrix A can be written as a product

$$A = L \cdot U$$

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

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

- Determinant computation:

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

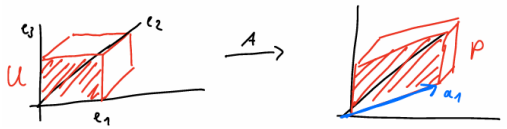
Geometric intuition again

Theorem 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 (signed) volume of P.



Applications to integrals

Proposition 38 (Applications to integrals)

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

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

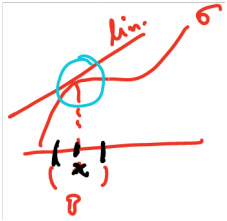
Where σ' is a linear derivative.

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.

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



Another occurence of the det in ML

Density of the multivariate Gaussian:

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

Normalization with $\det(\Sigma)^{1/2}$ relates to the volume, so that density integrates to 1 in the end.

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 $Tv = \lambda v$.

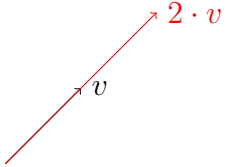
A vector $v \neq 0$ with this property is called an eigenvector corresponding to eigenvalue λ.

The set of all eigenvectors of λ is called the eigenspace:

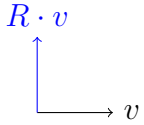
$$E(\lambda, T) \underbrace{=}_{\begin{cases} Tv = \lambda v \\ Tv - \lambda v = 0 \\ Tv - \lambda I v = 0 \\ (T - \lambda I)v = 0 \end{cases}} \ker(T - \lambda I).$$

Geometric intution

- Eigenvalue/eigenvector realizes a "stretching" $v \mapsto \lambda v$.



- Many mappings do not have eigenvectors, for example, a rotation over $\mathbb{R}^2$.



Eigenvectors are not unique

If λ is an eigenvalue, it has many eigenvectors! For example, if v is 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 distinct 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., but v and $c \cdot v$ are not linearly independent

Linear (in)dependence of eigenvectors? (2)

- They can be linearly independent:
  Easy example: $A = I$, then every vector v is an eigenvector of eigenvalue 1.

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

Existence of eigenvalues

Theorem 40 (Existence of eigenvalues)

Every operator $T : V \to V$ on a finite-dim., 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). Find coefficients $a_0, a_1, \ldots, a_n$ such that

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

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

$$a_0 + a_1T + a_2T^2 + \cdots + a_nT^n \overset{!}{=} c(T - \lambda_1 I)(T - \lambda_2 I)\ldots(T - \lambda_m I).$$

Existence of eigenvalues (3)

Step 2: Factorization (Excursion to complex-volumed polynomials)

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

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

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

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

It is not difficult to see that we can then also factorize $p(T)$:

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

Existence of eigenvalues (4)

Step 3: Putting it together

Hence,

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

$\Rightarrow v \in \ker(\bigstar)$

$\Rightarrow$ There must exist $i \in \{0, \dots, m\}$ such that $(T - \lambda_i I)$ not injective

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

□

⚠ The corresponding eigenvector is not necessarily v!

Reason:

$$c \cdot (T - \lambda_1 I) \dots (T - \lambda_{i-1} I) \underbrace{(T - \lambda_i I)}_{\text{not injective}} \underbrace{(T - \lambda_{i+1} I) \dots (T - \lambda_m I) \cdot v}_{:= w,\ w \text{ is the eigenvector!}} = 0$$

□

Existence of eigenvalues (5)

(★) Polynomids of operators

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

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

Then the following properties ensure that "factorization" works. Let p, q be two polynomials. Then:

- If the polynomial factorizes, so does the related operator:
  $(p \cdot q)(T) = p(T) \cdot q(T)$
- Order of factors does not matter:
  $p(T) \cdot q(T) = q(T) \cdot p(T)$

Hence, if $p(T) \cdot q(T)$ not injective $\implies$ $p(T)$ not injective or $q(T)$ not injective.

Diagonalization and Triangulization

Diagonizable matrices

Definition 41 (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 (symmetric matrices).

When is a matrix diagonizable?

Proposition 42 (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) $\dim V = \dim(\text{eigenspace}(\lambda_1)) + \cdots + \dim(\text{eigenspace}(\lambda_m))$.

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

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

Triangular matrices

Definition 43 (Upper Triangular matrix)

A matrix is called upper triangular if it has the form

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

Geometric intuition

Proposition 44 (Geometric intuition)

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) $Tv_j \in \text{span}\{v_1, \ldots, v_j\} \quad \forall j = 1, \ldots, n.$

Geometric intuition (2)

Proof idea:

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

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

When are triangular matrices invertible?

Proposition 45 (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 $v_1, \ldots, v_n$ 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, $\underbrace{Tv_2}_{\in \text{ range } T} = \underbrace{a_1 v_1}_{\in \text{ range } T} + \underbrace{\lambda_2 v_2}_{\text{also in range } T}$ for some scalar a.

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

In this way, we see that $v_1, \ldots, v_n$ are all in $\text{range}(T)$. 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 46 (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 $T - \lambda I$:

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

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

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

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

□

Algebraic multiplicity of eigenvalues

Definition 47 (Algebraic multiplicity)

The number of times each eigenvalue occurs on the diagonal is called the algebraic multiplicity of the eigenvalue.

Remark: More often, the algebraic multiplicity is defined using the characteristic polynomial, which we skipped. The two definitions are equivalent.

Algebraic multiplicity of eigenvalues (2)

Definition 48 (Geometric multiplicity)

The geometric multiplicity of an eigenvalue is the dimension of the corresponding eigenspace.

⚠ In general, the two do not agree!

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

Proposition 49 (Triangularization over $\mathbb{C}$)

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

⚠ Does not hold over $\mathbb{R}$!

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

Proof idea as an image:

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

subspace U

$$\begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_m & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \lambda & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \lambda \end{bmatrix}$$

subspace corr. to eigenvalue λ

Basis: $u_1, \ldots, u_m, w_1, \ldots, w_k$

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

Proof (induction on n):

- Already know that one complex eigenvalue exists: λ.
- Consider subspace $U := \text{range}(T - \lambda I)$, the "complement of eig(λ)".
- Easy to see that U is invariant under T, that is, $T(U) \subseteq U$, and that $\dim(\text{range}(U)) < n$.
- 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.

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

- Extend $u_1, \ldots, u_m$ by $w_1, \ldots, w_k$ to a basis of V again.
- For w_1:

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

- Similarly, $Tw_i \in \text{span}(u_1, \ldots, u_m, w_1, \ldots, w_i)$.

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

ML Keywords: loss functions, regularizers, sparsity

Normed spaces

Metric space

ML Keywords: inductive bias, k-nearest neigbor

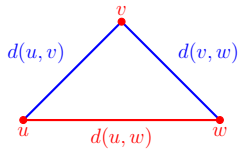
Definition 50 (Metric space)

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 51 (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 52 (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". Every norm induces a metric:

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

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

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

Definition 53 (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 54 (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 $\mathbb{R}^2$

$p \geq 1$ (convex balls):

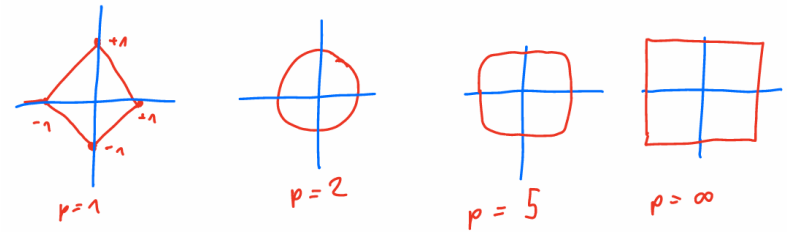
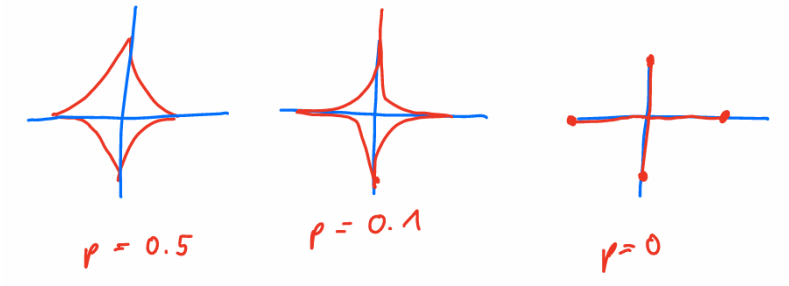


Illustration: Unit balls on $\mathbb{R}^2$ (2)

$p < 1$ (balls not convex):



When is a "p-norm" a norm?

Proposition 55 (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 non-negativity (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

Equivalent norms

Definition 56 (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:

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

Equivalent norms (2)

Theorem 57 (Norm equivalence on $\mathbb{R}^n$)

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

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

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 $\tilde{c_2} \neq 0$.

Equivalent norms (5)

- Now compute: For $x \in S$,

$$\tilde{c_2} \leq \|x\| = \left\|\frac{x}{1}\right\| = \left\|\frac{x}{\|x\|_\infty}\right\| = \frac{\|x\|}{\|x\|_\infty}.$$

Thus:

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

□

Scalar product

Scalar product

Definition 58 (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 $\mathbb{C} \to$ 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 $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$,

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

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

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

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

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

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

Scalar product and angles

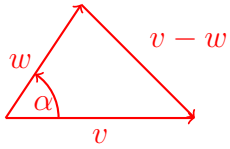
ML Keywords: cosine similarity

Proposition 59 (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.

- 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 60 (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.

Relationship between norm and scalar product

scalar product $\Longrightarrow$ norm

$\not\Longleftarrow$

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

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

  Then $\| \cdot \|$ is a norm induced by $\langle \cdot, \cdot \rangle$.
- The other way around does not work in general! (not every norm comes from a scalar product)
  Can you find a counterexample?

Relationship between norm and metric

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

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

  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 61 (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:

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

Orthonormal vectors and sets

Definition 62 (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 63

In an orthonormal basis $u_1, \ldots, u_n$, the representation of a vector v is given as:

$$v = \sum_{i=1}^{n} \langle v, u_i \rangle u_i \qquad \text{, with } v = \sum a_i u_i$$

Proof.

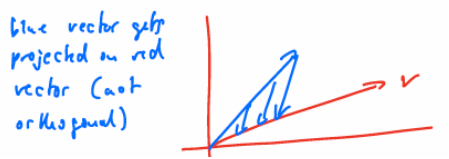
Exercise. □

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

Projection (general case)

Definition 64 (Projection general case)

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



Orthogonal Projection

Definition 65 (Orthogonal Projection)

Let U be a finite-dimensional subspace of a pre-Hilbert space H. Then there exists a linear projection

$$P_U : H \to U$$

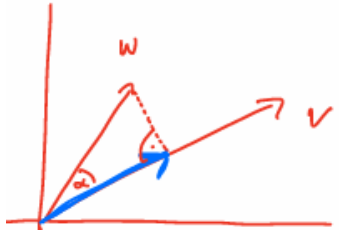
with $\ker(P_U) = U^\perp$. P_U is called the orthogonal projection of H on U.

Orthogonal Projection (2)

Proof idea: Let $u_1, \ldots, u_k$ be an orthogonal basis of U. Define $P_U : V \to U$ by

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

- It is clear that P_U is linear.
- For $w \in U^\perp \implies P_U(w) = 0$.



How to make a given basis orthogonal?

Proposition 66

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

Proof.

Gram-Schmidt orthogonalization

Is a procedure that takes any basis $v_1, \ldots, v_n$ of a finite-dimensional vector space and transforms it into another basis $u_1, \ldots, u_n$ that is orthogonal.

How to make a given basis orthogonal? (2)

Intuition: iterative procedure

- **Step 1:** Start with $u_1 := \frac{v_1}{\|v_1\|}$, $U_1 := \text{span}\{u_1\}$.
- **Step k:** Assume that we have identified $u_1, \ldots, u_{k-1}$. Project v_k onto U_{k-1} and keep the "rest":

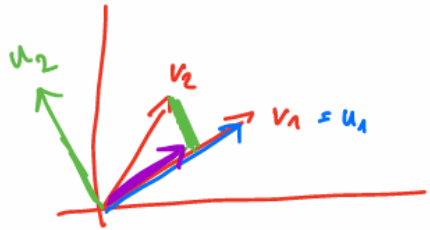
$$\tilde{u}_k := v_k - P_{U_{k-1}}(v_k).$$

Renormalize:

$$u_k = \frac{\tilde{u}_k}{\|\tilde{u}_k\|}.$$

This works in theory (proof omitted here). In practice, it quickly results in numerical errors. For better numerical stability, see the QR factorization later. □

How to make a given basis orthogonal? (3)



Orthogonal matrices

Orthogonal matrices

Definition 67 (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):**

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

General rotations can be written as a product of elementary rotations.

Properties of orthogonal matrices

Let Q be orthogonal. Then:

- Columns are orthogonal $\iff$ rows are orthogonal.
- Q is always invertible, and $Q^{-1} = Q^T$.
- Q acts as a geometry:
$$\forall v \in V,\ \|Qv\| = \|v\|.$$
- 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: Assignment!

⚠ Orthogonal rows and columns

Consider the projection matrix:

$$A = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix}.$$

- The columns are obviously not orthogonal.
- The rows formally satisfy $\langle (0,0), (2,1) \rangle = 0$, but this does not imply A is orthogonal.
- Orthogonal matrices require all rows/columns to have norm 1 (and full rank).

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

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

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

Symmetric matrices

Symmetric matrices

Definition 68 (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 69

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. But the eigenvectors might still live in $\mathbb{C}^n$.

Self-adjoint operators

Definition 70 (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 71 (Self-adjoint operator has real-valued eigenvalue and eigenvector)

Let V be a 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:

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = c(x^2 + b_1 x + c_1) \dots (x^2 + b_M x + c_M)(x - d_1) \dots (x - d_m),$$

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

Quadratic terms $x^2 + b_i x + c_i$ 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 72 (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 73 (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 74 (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 75 (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 76 (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 dependent)

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 77 (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.
- Then, the matrix:
$$C := X^T X \in \mathbb{R}^{d \times d}$$
is called the covariance matrix.
- The matrix:
$$K := X X^T \in \mathbb{R}^{n \times n}$$
is called a kernel matrix (for the linear kernel), and in this special case, it is also the Gram matrix.
- All these matrices are symmetric and positive semi-definite (psd).

(Prove it!)

Variational characterization of eigenvalues

Literature: Bhatia: Matrix Analysis

Rayleigh coefficient

Definition 78 (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 79 (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, $\lambda_1 = v_1$.

Formal proof (sketch)

Assume we start with the standard basis. Let $Q = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$ be the basis transformation that brings A in diagonal form; Q is orthogonal, such that:

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

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

$$R_A(y) = \frac{(Q^T x)^T A (Q^T x)}{(Q^T x)^T (Q^T x)} = \frac{x^T Q \Lambda Q^T x}{x^T x}.$$

Simplifying:

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

Formal proof (sketch) (2)

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

$$\min R_A(y) = \min_{\|y\|=1} \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2.$$

This minimum is attained for $x = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, that is $y = Q^T x = v_1$, with value

$R_A(y) = \lambda_1$. □

Rayleigh coefficient $\to$ second eigenvalue

Proposition 80 (Rayleigh coefficient $\to$ 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 81 (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):

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

- 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 for vectors in $\mathbb{R}^m$.

- 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 Ohaifa.

Singular value decomposition

ML motivation: recommender systems

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

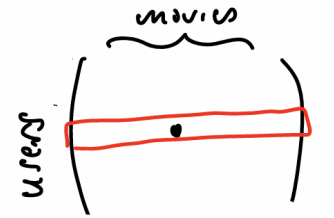


Figure: The dot represents rating of user i for movie j.

- 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)

Proposition 82 (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":

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_r & 0 & \dots & 0 \end{pmatrix}.$$

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

Illustration

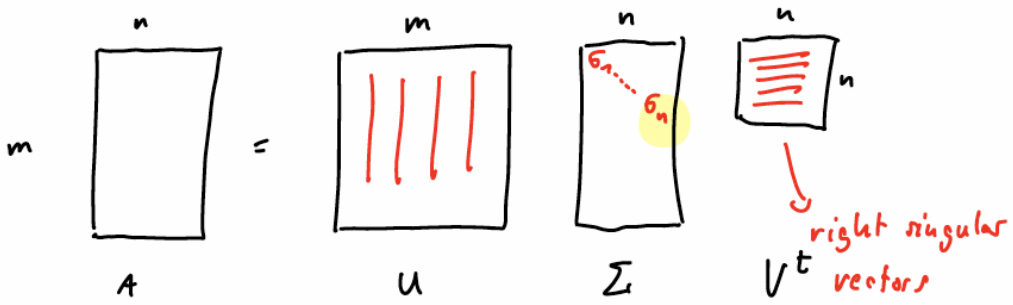


Illustration (2)

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 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 $v_1, \ldots, v_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n \geq 0$.

Proof sketch (2)

Define:

- $\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \sigma_i = \sqrt{\lambda_i}.$
- $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$, matrix with columns $u_i := \frac{Av_i}{\sigma_i}$.
- $V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$, matrix with v_i as columns.

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

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

Proof sketch (3)

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.

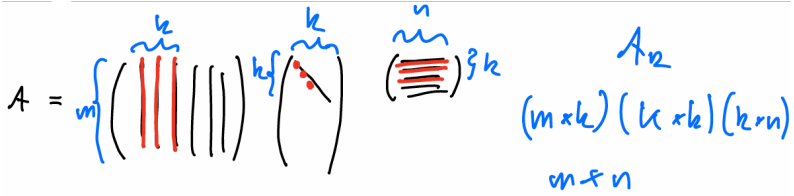
In particular, left- and right-singular vectors are the same (up to signs).

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 83 (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.

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.

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 84 (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:

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

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

- Then set:

$$A^{\#} := \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\lambda_k & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

- This does the job for diagonal matrices.
- How do we generalize to all matrices? $\Rightarrow$ Use SVD!

Moore-Penrose pseudo-inverse

Proposition 85 (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:

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

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.

Easy, just do it. □

Moore-Penrose pseudo-inverse (2)

Proposition 86

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

Proof.

Easy, just do it. □

Trace of a matrix

Trace

Definition 87 (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:

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

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

- The trace equals the negative of the coefficient in front of t^{n-1} in the characteristic polynomial:

$$p_A(t) = t^n + a_{n-1} t^{n-1} + \ldots$$

Trace vs. determinant

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

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

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 matrix $A \in \mathbb{C}^{n\times n}$, 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 88 (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 89

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

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

Spectral radius (3)

Proof.

- "$\Rightarrow$": 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 $\|A\|_F = \|D\|_F$, 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 90 (Neumann series)

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

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

Over $\mathbb{R}$: $\frac{1}{1-a} = \sum_{k=0}^{\infty} a^k$ 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 91 (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 form a basis of $\mathbb{R}^n$), then Q contains an orthonormal basis of $\mathbb{R}^n$.

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

Skipped material SKIPPED

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

Equivalence relation

Definition 92 (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: $(x, y) \in R \;\Leftrightarrow\; x \sim y$

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:

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

Equivalence class

Definition 93 (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 94

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.

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

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

Quotient spaces (2)

Definition 95 (Quotient "space")

Define the quotient "space" as

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

If $[v], [u] \in V/W$, define:

$$[v] + [u] := [v + u]$$

$$\lambda[v] := [\lambda v]$$

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.

Quotient spaces (4)

Proposition 96

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 $\dim(V/W) = \dim V - \dim W$

Characteristic polynomial SKIPPED

Motivation

Motivation: A is an $n \times n$-matrix, $v \neq 0$

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

Characteristic polynomial

Definition 97 (Characteristic polynomial)

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

$$p_A(t) := \det(A - t \cdot I)$$

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

- $p_A(t)$ is a polynomial with degree n.
- The characteristic polynomial does not depend on the basis.

Proof.

Consider A, basis transformation matrix U. Want to look at char. poly. of UAU^{-1}:

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

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

□

Characteristic polynomial (4)

- The roots of the characteristic polynomial correspond exactly to the eigenvalues of A.
- Over $\mathbb{C}$, the characteristic polynomial always has n roots, so the matrix has n eigenvalues (not necessarily distinct).
- A is invertible $\Leftrightarrow$ 0 is not an eigenvalue.
  Reason: If 0 is an eigenvalue, then $Av = 0 \cdot v = 0 \Rightarrow \ker(A)$ non-trivial $\Leftrightarrow$ A not invertible.
- Let $A \in \mathcal{L}(V)$, λ eigenvalue of A, then λ^k is an eigenvalue of A^k.
- Let A be invertible, λ eigenvalue of A, then $1/\lambda$ is an eigenvalue of A^{-1}.

Geometric and algebraic multiplicity

Definition 98 (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.

Compute eigs in theory

- Write down the characteristic polynomial, find the roots $\rightarrow$ eigenvalues.
- To compute the eigenvectors, solve the linear system $Ax = \lambda x$.

In practice ... see later (numerics)

Norms can be characterized by convex sets SKIPPED

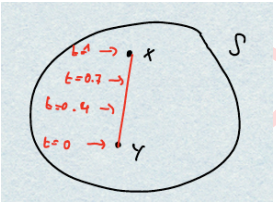
Convex set

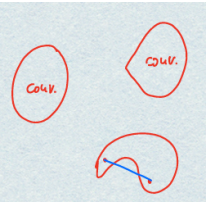
Definition 99 (Convex set)

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

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

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



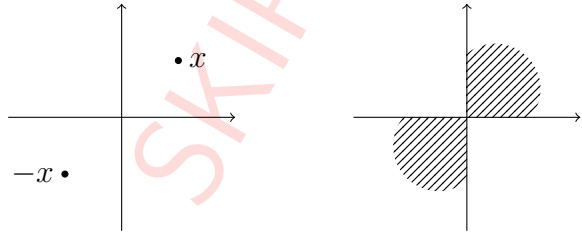


Symmetric set

Definition 100 (Symmetric set)

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

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



Convex sets induce norms

Theorem 101 (Convex sets induce norms)

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

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

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

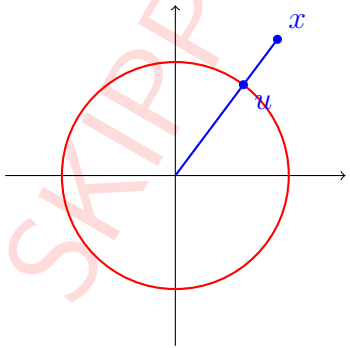
$$C = \left\{x \in \mathbb{R}^d \,\middle|\, p(x) \leq 1\right\}$$

(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.

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



Convex sets induce norms (3)

Proof.

Claim: $p(x)$ is well defined

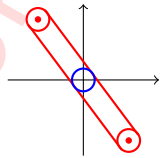
Want to prove: given $x \in \mathbb{R}^d$, the set

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

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

$$B_\varepsilon(0) := \{ e \in \mathbb{R}^n \mid \|e\|_2 < \varepsilon \} \subset C$$

Intuition:



Convex sets induce norms (4)

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

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

- $v + B_\varepsilon(0) = \{v + e \mid e \in B_\varepsilon(0)\}$
- By symmetry: $v + e \in C \Rightarrow -(v + e) \in C$
- By convexity:

$$\frac{1}{2}(v + e) + \frac{1}{2}(-v - e) = 0 \in C \Rightarrow e \in C$$

So $B_\varepsilon(0) \subset C$, 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.

Convex sets induce norms (5)

Now we need to prove all axioms of a norm:

$p(0) = 0$

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

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.

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:

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

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

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:

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

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.

□

Spaces of functions: Continuous, differentiable, integrable
SKIPPED

Space of continuous functions

Definition 102 (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

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

Space of continuous functions (2)

Proposition 103

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

Space of differentiable functions

Definition 104 (Space of differentiable functions)

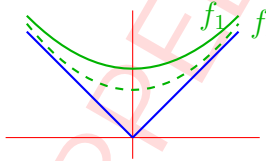
Let $[a, b] \subset \mathbb{R}$, then

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

Space of differentiable functions (2)

Which norm?

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



$\rightarrow$ limit function, not differentiable!

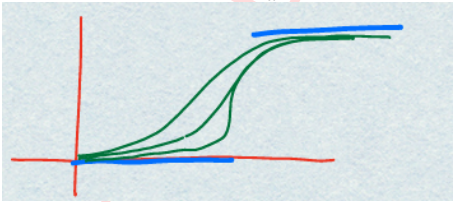
- $$\|f\|_1 := \sup_{t\in[a,b]} \max\{|f(t)|,\ |f'(t)|\}$$

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

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

Spaces of integrable functions?

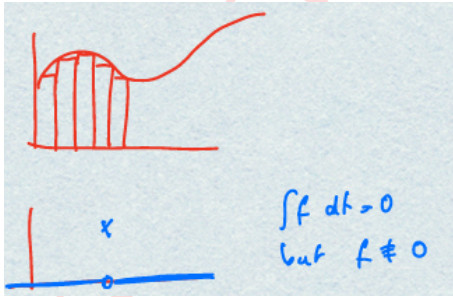
- Consider $C^b([a,b])$ with the norm $\|f\|_1 := \int_a^b |f(t)|\,dt$



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

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

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

The space $\mathcal{L}^p$

Definition 105 (The space $\mathcal{L}^p$)

For $1 \leq p \leq \infty$, 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

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

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

Proposition 106

$\|f\|_p$ 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.

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

Proposition 107

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

Proof.

If $(f_i)_{i\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{L}^p$, then want to prove that $\lim_{i\to\infty} f_i \in \mathcal{L}^p$.
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 $f_i \to f$ in $\|\cdot\|_p$.

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

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:

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

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

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

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

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

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

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.

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.

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

$$\| \underbrace{[f]}_{\text{new}} \|_p := \| \underbrace{f}_{\text{old}} \|_p$$

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.

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.

Operator norm SKIPPED

Continuous = bounded

Theorem 108 (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.
$$\forall x \in X\ \forall \varepsilon > 0\ \exists \delta > 0\ \forall y \in X : \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon$$

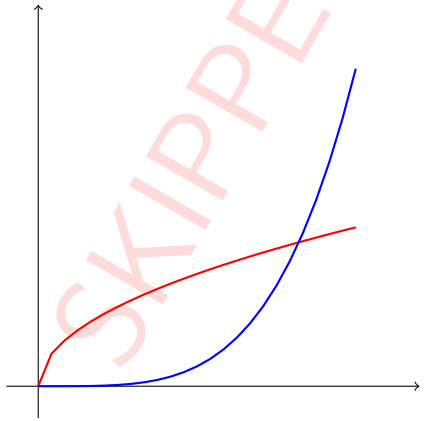
(ii) T is continuous at 0.

(iii) T is bounded:
$$\exists M > 0\ \forall x \in X : \|Tx\| \leq M \cdot \|x\|$$

(iv) T is uniformly continuous.
$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall x \in X\ \forall y \in X : \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \varepsilon$$

Continuous = bounded (2)

- uniformly continuous
- not uniformly continuous



Operator norm

Definition 109 (Operator norm)

Let X, Y be normed spaces, and let $T : X \to Y$ be linear and continuous.

$$\|T\| := \sup_{x \in X} \frac{\|Tx\|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\| \leq 1}} \|Tx\| = \sup_{\substack{x \in X \\ \|x\| = 1}} \|Tx\|$$

is called the operator norm of T.

Observe: coincides with the matrix norm $\|\cdot\|_2$ as we had defined it earlier.

Examples

- 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

$$\|T\| = 1$$

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.

Dual space SKIPPED

Dual space

Definition 110 (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.

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

$$\|T\| := \sup_{x \in X} \frac{|Tx|}{\|x\|}$$

Proposition 111 (Dual space is normed space)

V' is a vector space, and the operator norm is indeed a norm on V'.

Dual space (3)

Proposition 112 (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.

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 $L^q(S)$.

Riesz representation theorem

Theorem 113 (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

$$x'(x) = \langle x, y \rangle.$$

Adjoint operator SKIPPED

Adjoint operator

Definition 114 (Adjoint operator)

Let $T \in \mathcal{L}(H_1, H_2)$, H_1, H_2 Hilbert spaces. Then there exists an operator $T^* : H_2 \to H_1$ such that

$$\langle Tx, y \rangle_{H_2} = \langle x, T^* y \rangle_{H_1}$$

for all $x \in H_1, y \in H_2$. T^* is called the adjoint of T.

Remark: The existence of this operator is a consequence of the Riesz representation theorem.

Adjoint operator (2)

Definition 115 (Self-adjoint Operator)

An operator $T : H_1 \to H_1$ is called self-adjoint if

$$\langle Tx, y \rangle = \langle x, Ty \rangle.$$

Part II
Calculus

ML Motivation

ML is all about optimizing functions to fit the training data, and we typically use gradients to do this. So we need to know everything about differential calculus in $\mathbb{R}^d$.

To be able to define all of this, we first need to look at sequences and convergence.

And if you want to be a Bayesian, you need to integrate all the time...

ML Keywords:
Convergence of a learning algorithm!

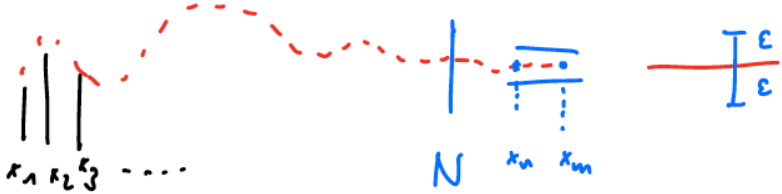
Sequences and convergence

Cauchy sequence

Definition 116 (Cauchy sequence)

A sequence $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^d$ is called a Cauchy sequence if

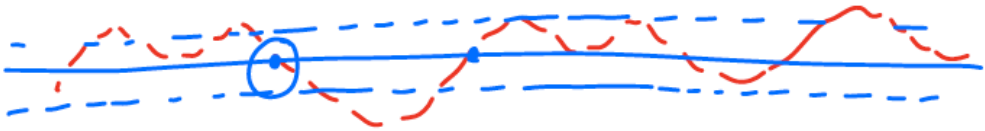
$$\forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n, m > N : |x_n - x_m| < \varepsilon$$



Accumulation point

A point $x \in \mathbb{R}$ is called an accumulation point of the sequence $(x_n)_n$ if

$$\forall \varepsilon > 0 \, \forall N \in \mathbb{N} \, \exists n > N : |x_n - x| < \varepsilon$$



⚠ In $\mathbb{R}^d$, we replace the absolute value with a norm: $\|x_n - x\|$.

Convergence

A sequence $(x_n)_n$ converges to $x \in \mathbb{R}^d$ if

$$\forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n > N : |x_n - x| < \varepsilon$$

Notation:

$$\lim_{n \to \infty} x_n = x \quad , \quad x_n \xrightarrow{n \to \infty} x$$

First observations

- A sequence can have many accumulation points (or none at all).
- Even if the sequence has just one accumulation point, it is not necessarily a Cauchy sequence.
- If $(x_n)_n$ converges to x, then x is the only accumulation point and the sequence is Cauchy.

Examples

- $x_n = \frac{1}{n}$ on $]0, 1] = \{x \in \mathbb{R} \mid 0 < x \leq 1\}$.
  $(x_n)_n$ is Cauchy, but does not converge within $]0, 1]$.
  It does converge to 0 on $[0, 1]$.

- Consider the sequence $x_n = \begin{cases} \frac{1}{n}, & \text{if } n \text{ is even} \\ n, & \text{if } n \text{ is odd} \end{cases}$

  It has an accumulation point but is not Cauchy.

Maximum and upper bound

Assume we are on $\mathbb{R}$ (or more general, on a space that has a total ordering). Let $U \subset \mathbb{R}$ be a subset.

- $x \in \mathbb{R}$ is called a maximum element of U if $x \in U$ and $\forall u \in U : u \leq x$.
- x is called an upper bound of U if $\forall u \in U : u \leq x$.
  ⚠ x does not have to be in U!
- x is called supremum of U if it is the smallest upper bound.

Analogously, minimum, lower bound, infimum.

Examples

- 1 is the maximum of $[0, 1]$. It is also the supremum of $[0, 1]$.
- $]0, 1[$ does not have a maximum element.
- 5 is an upper bound of $]0, 1[$.
- 1 is also an upper bound of $]0, 1[$.
- 1 is the supremum of $]0, 1[$.

Bounded sequence

A sequence $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ is called bounded if there exist $a, b \in \mathbb{R}$ such that $x_n \in [a, b]$ for all $n \in \mathbb{N}$.

Theorem 117 (Heine-Borel)

Any bounded sequence in $\mathbb{R}$ has at least one accumulation point.

Limsup and Liminf

For a sequence $(x_n)_n \subset \mathbb{R}$ we define:

$$\liminf_{n\to\infty} x_n := \lim_{n\to\infty} \left(\inf_{m\geq n} x_m \right)$$

$$\limsup_{n\to\infty} x_n := \lim_{n\to\infty} \left(\sup_{m\geq n} x_m \right)$$

Observations

- For a bounded sequence $(x_n)_n$, the $\limsup$ is the largest accumulation point of $(x_n)_n$, and the $\liminf$ is the smallest.

- The $\liminf$ is the largest $y \in \mathbb{R}$ such that

$$\forall \varepsilon > 0 \, \exists N \, \forall n > N : x_n > y - \varepsilon.$$

Basic concepts in topology

Open and closed sets

Definition 118 (Epsilon-ball)

Let (X, d) be a metric space, and denote for $x \in X, \varepsilon > 0$

$$B_\varepsilon(x) = \{y \in X \mid d(x, y) \leq \varepsilon\}.$$

Open and closed sets (2)

Definition 119 (Open and closed sets)

A subset $U \subset X$ of a metric space is called closed if all Cauchy sequences converge and have their limit point in U.

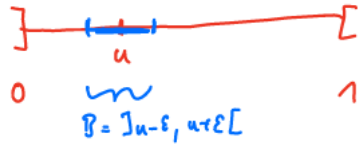
A set $U \subset X$ is called open if

$$\forall u \in U \, \exists \varepsilon > 0 : B_\varepsilon(u) \subset U.$$

⚠ Topologies, open, and closed sets are also defined if a metric does not exist...

Examples

- The set $[0, 1]$ is closed.
- The set $]0, 1[$ is open:



- A set U can be <u>neither</u> open nor closed: $[0, 1[$.

Open vs. closed

Proposition 120 (Open vs. closed)

- Complements of open sets are closed.
- Complements of closed sets are open.

Interior, closure

Definition 121 (Interior, closure)

A point $u \in U$ is an interior point of U if there exists a $\varepsilon > 0$ such that $B_\varepsilon(u) \subset U$.

$U = [0, 1]$, then $x \in]0, 1[$ are interior points.

The (topological) closure of a set U is defined as the set of points that can be approximated by Cauchy sequences in U:

$$w \in \overline{U} \Leftrightarrow \forall \varepsilon > 0 \, \exists z \in U : d(w, z) < \varepsilon.$$

Notation: $\overline{U}$ is the closure of U.

The (topological) interior of a set U is defined as the set of interior points of U.
Notation: U°

Boundary

The (topological) boundary of a set U is defined as the set

$$\overline{U} \setminus U^\circ.$$

Note: Literature is not always consistent here... sometimes we also read $U \setminus U^\circ$ instead of $\overline{U} \setminus U^\circ$.

$$X = [0,1[, \quad \overline{X} = [0,1], \quad X^\circ =]0,1[$$
$$\Rightarrow \text{boundary}_1(X) = \overline{X} \setminus X^\circ = \{0,1\}.$$

$$(\text{boundary}_2(X) = X \setminus X^\circ = \{0\})$$

Boundary (2)

Definition 122 (Bounded set)

A set $U \subset X$ is bounded if there exists $D > 0$ such that

$$\forall u, v \in U\,,\; d(u, v) < D.$$

Dense sets

Definition 123 (Dense sets)

A set U is dense in X if we can approximate every $x \in X$ by a sequence in U.
Formally,

$$\forall x \in X \, \forall \varepsilon > 0 \, B_\varepsilon(x) \cap U \neq \emptyset.$$

Examples:

- $\mathbb{Q}$ is a dense subset of $\mathbb{R}$.
- Let $C^1[0,1]$ be the set of functions $f : [0,1]$ that are differentiable, and $C[0,1]$ the continuous functions. Then $C^1[0,1]$ is dense in $C[0,1]$ with respect to the $\|\cdot\|_\infty$-norm.

ML keyword: Can we approximate the underlying target fact f by the facts f_n that can be constructed by our learning algorithm?

Continuity

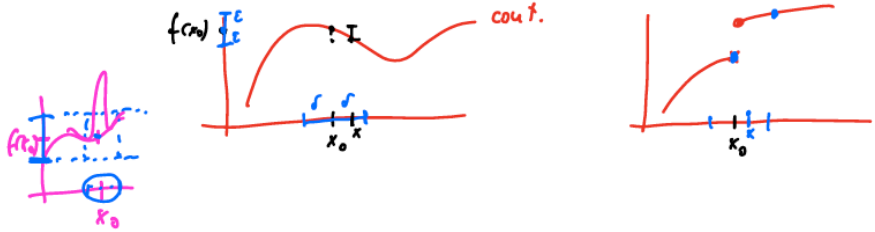
Continuous function

Definition 124 (Continuous function)

A function $f : X \to Y$ between two metric spaces (X, d), (Y, d) is called continuous at $x_0 \in X$ if

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x \in X : d(x, x_0) < \delta \Rightarrow d(f(x), f(x_0)) < \varepsilon.$$

Continuous function (2)



A function $f : X \to Y$ is called continuous if it is continuous for every $x_0 \in X$.

$$\forall x_0 \in X \, \forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x \in X : d(x, x_0) < \delta \Rightarrow d(f(x), f(x_0)) < \varepsilon.$$

Alternative definitions

- $f : X \to Y$ is continuous at x_0 if for every sequence $(x_n)_n \subset X$ we have

$$x_n \to x_0 \quad \Rightarrow f(x_n) \to f(x_0).$$

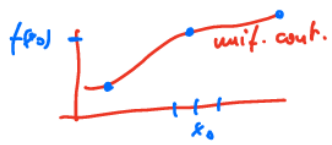
- A function f between two metric spaces (X, d), (Y, d) is continuous if and only if pre-images of open sets are open:

$$B \subset Y \underbrace{\text{ open }}_{\text{in } Y} \quad \Rightarrow \quad \underbrace{f^{-1}(B)}_{:=\{x \in X \mid f(x) \in B\}} \quad := \{x \in X \mid f(x) \in B\} \text{ is open in } X.$$

Uniformly continuous

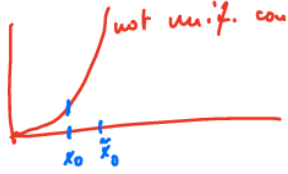
A function $f : X \to Y$ is called uniformly continuous if

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x_0 \in X \forall x \in X : d(x, x_0) < \delta \Rightarrow d(f(x), f(x_0)) < \varepsilon.$$



Given ε, I can choose δ that works for all x_0

Intuition: bounded derivative



Cannot choose δ to be the same for all x_0

Intuition: unbounded derivative

Examples

- $f :]0, 1] \to \mathbb{R}$, $f(x) = \frac{1}{x}$ is continuous but not uniformly continuous.
- $f : [-c, c] \to \mathbb{R}$ (for some constant c), $f(x) = x^2$:
  - is continuous and uniformly continuous.
  - The same function would not be uniformly continuous if it were defined on all of $\mathbb{R}$.

Proposition 125 (Bounded implies uniformly continuous)

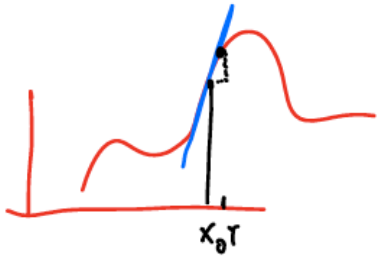
Let $f : [a, b] \to \mathbb{R}$ be continuous. Then it is already uniformly continuous.

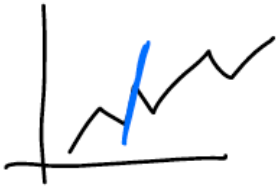
Lipschitz continuous

A function $f : X \to Y$ is called Lipschitz continuous with Lipschitz constant L if

$$\forall x, y \in X : d(f(x), f(y)) \leq L \cdot d(x, y).$$

Intuition: "bounded derivative"





Lipschitz continuity $\Longrightarrow$ uniform continuity

Proposition 126 (Lipschitz continuous $\Longrightarrow$ uniform continuous)

If f is Lipschitz continuous $\Rightarrow$ f is uniformly continuous.

Proof.

Easy. □

⚠ The other way round is not true:

$$f(x) : [0, \infty[\to \mathbb{R},\ x \mapsto \sqrt{x}$$

is uniformly continuous but not Lipschitz continuous.

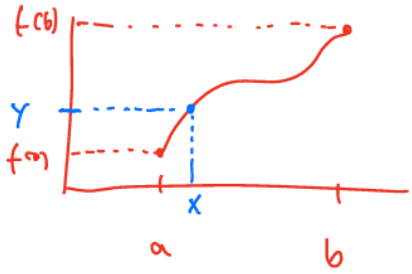
Intermediate value theorem

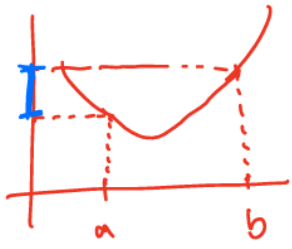
Theorem 127 (Intermediate value theorem)

If $f : [a, b] \to \mathbb{R}$ is continuous, then f attains all values between $f(a)$ and $f(b)$:

$$\forall y \in [f(a), f(b)] \, \exists x \in [a, b] : f(x) = y.$$

Intermediate value theorem (2)





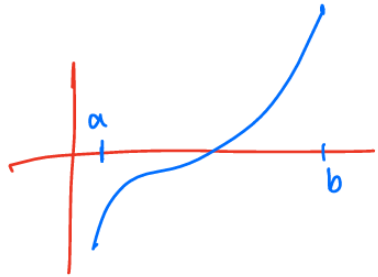
Application: finding zeros

If you want to find x with $f(x) = 0$:

- Find a with $f(a) < 0$,
- and b with $f(b) > 0$.

Then there must exist $x \in [a, b]$ with $f(x) = 0$.

Bisection search ...



Inverse function

Let $f : A \to B$ be a function, denote by $f(A) \subset B$ the range of f. A mapping $g : f(A) \to A$ is called the inverse of f, notation f^{-1}, if

$$g \circ f = \mathrm{id} \quad \text{and} \quad f \circ g = \mathrm{id}.$$

⚠ Not every function has an inverse. Example: $f(x) = x^2$.

⚠ Sometimes we also use the notation f^{-1} to denote the pre-image (which does not need to be unique).

Invertible function

Proposition 128 (Invertible function)

Let $D \subset \mathbb{R}$, $f : D \to \mathbb{R}$ be continuous and strictly monotone

$$(a < b \Rightarrow f(a) < f(b))$$

Then f is invertible, and the inverse is continuous as well.

- Invertibility follows from monotonicity.

- Continuity of the inverse follows directly from continuity of f.

Sequence of functions

Pointwise convergence

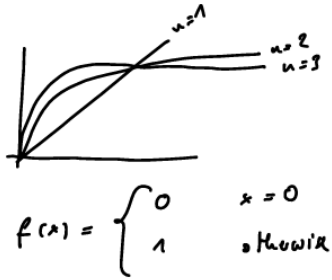
Definition 129 (Pointwise convergence)

Consider functions $f_n : D \to \mathbb{R}, \quad D \subset \mathbb{R}$.
We say that the sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to $f : D \to \mathbb{R}$ if

$$\forall x \in D : \ f_n(x) \to f(x).$$

Example

$$f_n, f : [0,1] \to \mathbb{R},\ f_n(x) = x^{1/n}.$$



⚠ This example also shows: $f_n \to f$ pointwise, and all f_n are continuous, but this does not imply that f is continuous.

Uniform convergence

Definition 130 (Uniform convergence)

A sequence $(f_n)_n$ of functions *converges uniformly* to f if

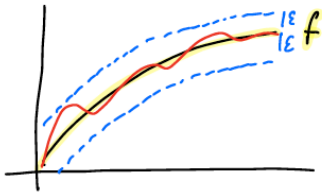
$$\forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n > N \, \underline{\forall x \in D} : |f_n(x) - f(x)| < \varepsilon.$$

Pointwise:

$$\forall x \in D \, \forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n > N : |f_n(x) - f(x)| < \varepsilon.$$

Intuition

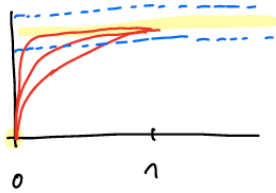
uniform convergence:
given ε, there exists N such that all f_n with $n > N$ are contained in ε-tube.



f_n

uniform

Close to 0 there will always be points x such that the red function are not yet in ε-tube.
$\Rightarrow$ Not uniformly conv.



not uniform,
only pointwise

Alternative definition

The sequence $(f_n)_n$ converges uniformly to f if and only if

$$\|f_n - f\|_\infty \to 0.$$

Additional notes:

- $C(D) := \{f : D \to \mathbb{R} \mid f \text{ is continuous}\}$ is a vector space.
- $\|\cdot\|_\infty$ is a norm on $C(D)$, defined as

$$\|f\|_\infty := \sup_{x \in D} |f(x)|.$$

Remark

$$\text{uniform convergence} \Rightarrow \text{pointwise convergence}$$
$$\not\Leftarrow$$

Uniform convergence preserves continuity

Theorem 131 (Uniform convergence preserves continuity)

Let $f_n, f : D \to \mathbb{R}$, $D \subset \mathbb{R}$. Assume all f_n are continuous, and $f_n \to f$ uniformly. Then f is continuous.

Proof

Consider $x_0 \in D$. Suppose some $\varepsilon > 0$ is given.
Observe that for every $n \in \mathbb{N}$:

$$(\bigstar) \qquad |f(x_0) - f(x)| \leq |f(x_0) - f_n(x_0)| + |f_n(x_0) - f_n(x)| + |f_n(x) - f(x)|.$$

Step 1:

Uniform convergence $\Rightarrow \exists n \in \mathbb{N}$ such that for all $x, x_0 \in D$:

$$|f_n(x) - f(x)| < \frac{\varepsilon}{3}$$
$$|f_n(x_0) - f(x_0)| < \frac{\varepsilon}{3}.$$

Proof (2)

Step 2:

Now consider the function f_n. By assumption f_n is continuous, so there exists $\delta > 0$ such that

$$|x - x_0| < \delta \Rightarrow |f_n(x) - f_n(x_0)| < \frac{\varepsilon}{3}.$$

Combining both steps: For given $\varepsilon > 0$, there exists $\delta > 0$ such that for all x

$$|x - x_0| < \delta \overset{(\bigstar)}{\Rightarrow} |f(x) - f(x_0)| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Thus, f is continuous at x_0. □

Derivatives (1-dim)

Derivative

Definition 132 (Derivative)

Let $U \subset \mathbb{R}$ be an interval, $f : U \to \mathbb{R}$. The function f is called differentiable at $a \in U$ if

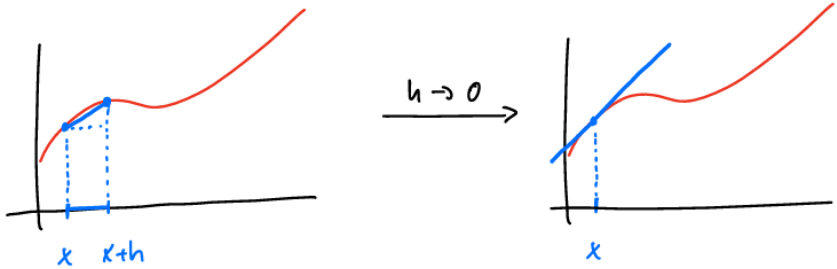
$$f'(a) := \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

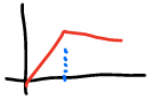
exists.

We often write

$$f'(a) = \frac{df}{dx}(a).$$

Illlustration





Differentiable functions

Definition 133 (Differentiable functions)

The function f is called differentiable if it is differentiable for all $a \in U$.

It is continuously differentiable if it is differentiable and the function $f' : U \to \mathbb{R}$, $a \mapsto f'(a)$ is continuous.

For $D \subset \mathbb{R}$, we denote

$$C^1(D) := \{f : D \to \mathbb{R} \mid f \text{ is continuously differentiable}\}.$$

Higher-order derivatives

We can repeat the process of taking derivatives:

$$f' = \frac{df}{dx}, \quad f'' = \frac{df'}{dx}.$$

Notation: $f^{(n)}$ denotes the n-th derivative (if it exists).

$C^n(D) := \{f : D \to \mathbb{R} \mid f \text{ is } n \text{ times continuously differentiable}\}$.

Differentiable implies continuous

Theorem 134 (Differentiable implies continuous)

Let f be differentiable at a. Then there exists a constant c_a such that, on a small ball around a, we have

$$|f(x) - f(a)| \leq c_a \cdot |x - a|.$$

In particular, f is continuous at a.

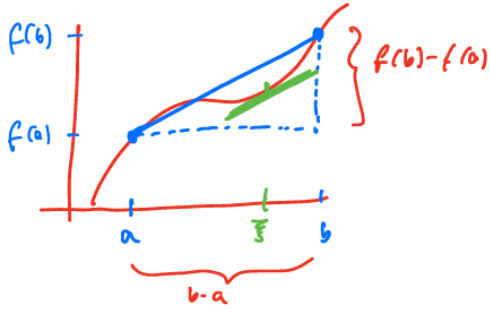
Intermediate value theorem for derivatives

Theorem 135 (Intermediate value theorem for derivatives)

If $f \in C^1([a,b])$, then there exists $\xi \in [a,b]$ such that

$$\frac{f(b)-f(a)}{b-a} = f'(\xi).$$

Intermediate value theorem for derivatives (2)



Exchanging limes and derivative

Theorem 136 (Exchanging limes and derivative)

Let $f_n : [a, b] \to \mathbb{R}$, $f_n \in C^1([a, b])$. If the limit

$$f(x) := \lim_{n\to\infty} f_n(x)$$

exists for all $x \in [a, b]$, and the derivatives f_n' converge uniformly, then f is continuously differentiable, and we have

$$(f')(x) = \left(\lim_{n\to\infty} f_n\right)'(x) \overset{!}{=} \lim_{n\to\infty} (f_n')(x).$$

Exchanging limes and derivative (2)

Explanation:

- First take the limit of f_n, we obtain f, and then compute its derivative.
- Alternatively, first compute all f_n', then take the limit of these derivatives.

⚠ Uniform convergence is crucial, otherwise, the result would be wrong!

Riemann-integral (1-dim) SKIPPED

2024 skipped, but assumed that you know this already!

Construction of the Riemann integral

Consider a function $f : [a, b] \to \mathbb{R}$. Assume that f is bounded:

$$\exists l, u \in \mathbb{R}\, \forall x \in [a, b] : \; l \leq f(x) \leq u.$$

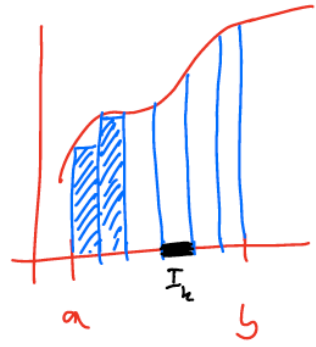
Consider $x_0, x_1, \ldots, x_n$ with

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

Construction of the Riemann integral (2)

These points introduce a partition of $[a, b]$ into n intervals:

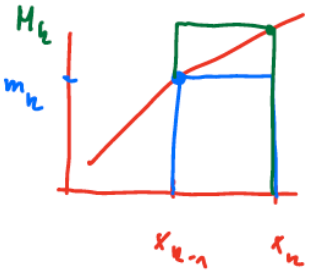
$$I_k := [x_{k-1}, x_k].$$



Construction of the Riemann integral (3)

Define

$$m_k := \inf(f(I_k))$$
$$M_k := \sup(f(I_k))$$



(exists because f is bounded).

Construction of the Riemann integral (4)

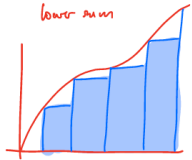
Define the lower sum :

$$s\left(f, \{x_0, x_1, \ldots, x_n\}\right) = \sum_{k=1}^{n} |I_k| \cdot m_k$$

and upper sum

$$S\left(f, \{x_0, x_1, \ldots, x_n\}\right) = \sum_{k=1}^{n} |I_k| \cdot M_k$$

where $|I_k| = x_k - x_{k-1}$ is the length of I_k.



Construction of the Riemann integral (5)

Now define:

$$J_* := \sup_{\text{partitions}} (s(f, \text{partition}))$$
$$J^* := \inf_{\text{partitions}} (S(f, \text{partition})).$$

We call f Riemann-integrable if $J_* = J^*$. Then we denote:

$$J_* = J^* = \int_a^b f(t)\, dt.$$

Monotone resp. continuous functions are Riemann-integrable

Theorem 137 (Monotone resp. continuous functions are Riemann-integrable)

- $f : [a, b] \to \mathbb{R}$ monotone $\Rightarrow$ integrable (i.e., $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$).
- $f : [a, b] \to \mathbb{R}$ continuous $\Rightarrow$ integrable (even true if f is continuous everywhere except at finitely many points).

Many functions are not Riemann-integrable

Example:

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{otherwise} \end{cases} = \mathbb{1}_{\mathbb{Q}}.$$

- For any interval $I_k = [x_{k-1}, x_k]$, we have:

$$M_k = 1, \quad m_k = 0.$$

- Thus:

$$J_* = (b-a) \cdot 0, \quad J^* = (b-a) \cdot 1.$$

Since $J_* < J^*$, the function is not Riemann-integrable.

Further shortcomings of the Riemann integral

- One cannot prove theorems about "integral" with "$\lim$":

$$\lim_{n\to\infty} \int f_n dt \stackrel{?}{=} \int \lim_{n\to\infty} f_n \, dt.$$

- Hard to extend to "other spaces."

$\rightarrow$Solution: Lebesgue integral!

Fundamental Theorem of Calculus

Fundamental Theorem of Calculus

Theorem 138 (Theorem I)

Let $f : [a, b] \to \mathbb{R}$ be (Riemann-) integrable and continuous at $\xi \in [a, b]$. Let $c \in [a, b]$ be a constant. Define the function

$$F(x) := \int_c^x f(t)\, dt.$$

Then F is differentiable at ξ and $F'(\xi) = f(\xi)$.

If $f \in C([a, b])$, then $F \in C^1([a, b])$ and

$$F'(x) = f(x) \quad \forall x \in [a, b].$$

Fundamental Theorem of Calculus (2)

Theorem 139 (Theorem II)

Let $F : [a, b] \to \mathbb{R}$ be continuously differentiable. Then

$$\int_a^b F'(t)\,dt = F(b) - F(a).$$

Algebraic version of the Fundamental Theorem (informal)

Informal, algebraic version:

The integral operator $I : C([a,b]) \to C_c^1([a,b])$

$$C_c^1([a,b]) := \{f \in C^1([a,b]) \mid f(c) = 0\}$$

is an isomorphism (linear, bijective), and its inverse is the differential operator.

Proof Part I

Proof I: Need to prove that F is differentiable at ξ.

Consider

$$
\begin{aligned}
A(h) &:= \frac{F(\xi+h) - F(\xi)}{h} \\
&= \frac{1}{h}\left(\int_c^{\xi+h} f(t)\,dt - \int_c^{\xi} f(t)\,dt\right) \\
&= \frac{1}{h}\int_{\xi}^{\xi+h} f(t)\,dt \qquad \text{Want to prove: converges to } f(\xi) \text{ as } h \to 0
\end{aligned}
$$

Proof Part I (2)

Want to prove:

$$A(h) - f(\xi) \xrightarrow{!} 0$$

$$= \frac{1}{h} \int_{\xi}^{\xi+h} f(t)\,dt - \underbrace{f(\xi)}_{B} \qquad B = \frac{1}{n} \int_{\xi}^{\xi+h} B dt$$

$$= \frac{1}{h} \int_{\xi}^{\xi+h} f(t)\,dt - \int_{\xi}^{\xi+h} f(\xi)\,dt$$

$$= \frac{1}{h} \int_{\xi}^{\xi+h} f(t) - f(\xi)\,dt$$

Intuition: $f(t) - f(x)$ is small due to continuity of f at ξ

Proof Part I (3)

Formally: Given $\varepsilon > 0$ we can find $h > 0$ such that

$$|f(t) - f(\xi)| < \varepsilon \quad \forall t \in [\xi, \xi + h]$$

Then:

$$\frac{1}{h}\int_{\xi}^{\xi+h} |f(t) - f(\xi)|\, dt \leq \frac{1}{h}\int_{\xi}^{\xi+h} \varepsilon\, dt = \frac{1}{h} \cdot \varepsilon \cdot \int_{\xi}^{\xi+h} 1\, dt = \frac{1}{h} \cdot \varepsilon \cdot h = \varepsilon.$$

□

Proof Part II

Proof II:

Know that F' continuous. Thus, by Theorem I the function

$$G(x) := \int_a^x F'(t)\,dt \quad \text{is differentiable and}$$

(i) $G(a) = 0$ (by def. of G)

(ii) $G'(x) = F'(x)$ on $[a, b]$ (by Theorem I)

Consider

$$H(x) := F(x) - G(x)$$

Proof Part II (2)

By (ii) we know that $H'(x) = F'(x) - G'(x) = 0$ for all x.
Hence it is a constant function.

We know that

$$H(a) = F(a) - G(a) = F(a), \text{ thus}$$

(iii) $H(x) \equiv F(a)$ ($\equiv$ means: "constant")

Proof Part II (3)

Consider $x = b$:

$$F(a) \overset{(iii)}{=} H(b) \overset{\text{def}}{=} F(b) - G(b) \overset{\text{def}}{=}$$

$$= F(b) - \int_a^b F'(t)\,dt$$

$$\Rightarrow \int_a^b F'(t)\,dt = F(b) - F(a)$$

□

Power series

Power series

Definition 140 (Power series)

A series of the form

$$p(x) := \sum_{n=0}^{\infty} a_n x^n$$

is called a power series.

A power series $p(x) = \sum_{n=0}^{\infty} a_n x^n$ converges if the sequence of partial sums

$$p_N(x) := \sum_{n=0}^{N} a_n x^n$$

converges in the usual sense as $N \to \infty$.

Radius of convergence

Theorem 141 (Radius of convergence)

For every power series $p(x) = \sum_{n=0}^{\infty} a_n x^n$, there exists a constant r, $0 \leq r \leq \infty$, called the radius of convergence, such that

- The series converges for all x with $|x| < r$.
- If $|x| > r$, the series diverges.

⚠ No general statement is possible for $|x| = r$.

Computing the radius of convergence

The radius of convergence only depends on the a_n and can be computed by various formulas:

- $r = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|$ (if the limit exists),
- $r = \frac{1}{L}$ where $L = \limsup_{n\to\infty} (|a_n|)^{1/n}$ (if the $\limsup$ exists).

A first example

Consider the power series

$$p(x) = \sum_{n=0}^{\infty} n^c x^n \quad \text{for some constant } c.$$

Radius of convergence:

$$r = \lim \left| \frac{a_n}{a_{n+1}} \right| = \lim \frac{n^c}{(n+1)^c} = \lim \left(\frac{n}{n+1} \right)^c = 1.$$

The result is independent of c.

A first example (2)

Case $c = -1$: The series $\sum \frac{1}{n} x^n$ has a radius of convergence $r = 1$.

- For $|x| > 1$, it diverges.
- For $|x| < 1$, it converges.
- For $|x| = 1$, we analyze more closely:
  - For $x = +1$, the series diverges because

$$\sum \frac{1}{n} \cdot 1^n = \sum \frac{1}{n} \to \infty.$$

  - For $x = -1$, it converges:

$$\sum \frac{1}{n}(-1)^n = -\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots\right) = -\log(2).$$

A first example (3)

Case $c = 0$: The series $\sum n^c x^n = \sum x^n$ diverges for $|x| = r$.
Radius of convergence remains $r = 1$.

- For $|x| < 1$, the series converges.
- For $|x| > 1$, the series diverges.
- For $|x| = 1$:
  - For $x = +1$,
$$\sum x^n = \sum 1 = N \to \infty \quad \text{(diverges)}.$$
  - For $x = -1$,
$$\sum x^n = \sum (-1)^n = -1 + 1 - 1 + 1 - 1 + \ldots \quad \text{(does not converge)}.$$

More examples

- Exponential series:

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \text{has } r = \infty,$$

$$\text{because} \quad \left| \frac{a_n}{a_{n+1}} \right| = \frac{\frac{1}{n!}}{\frac{1}{(n+1)!}} = n + 1 \to \infty.$$

- Factorial series:

$$\sum n! x^n \quad \text{has } r = 0,$$

$$\text{because} \quad \left| \frac{a_n}{a_{n+1}} \right| = \frac{n!}{(n+1)!} = \frac{1}{n+1} \to 0.$$

From power series to Taylor series, intuition

Observation: Given a power series $\boxed{f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n}$.

Let's compute its derivative:

$$\begin{aligned} f'(x) &= \left(a_0 + a_1(x-a) + a_2(x-a)^2 + a_3(x-a)^3 + \dots\right)' \\ &= \quad a_1 \quad + 2a_2(x-a) + 3a_3(x-a)^2 + \dots, \end{aligned}$$

or equivalently,

$$f'(x) = \sum_{n=1}^{\infty} \underline{n \cdot a_n \cdot (x-a)^{n-1}}.$$

From power series to Taylor series, intuition (2)

Similarly,

$$f''(x) = \dots,$$

$$f^{(k)}(x) = \sum_{n=k}^{\infty} a_n \cdot n \cdot (n-1) \cdot \dots \cdot (n-k+1) \cdot (x-a)^{n-k}.$$

In particular, for $x = a$, we have:

$$f^{(k)}(a) = a_k \cdot k!,$$

or, stated otherwise:

$$\boxed{a_k = \frac{f^{(k)}(a)}{k!}}.$$

From power series to Taylor series, formally

Theorem 142 (From power series to Taylor series)

Let $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$ with $r > 0$. Then, for x with $|x-a| < r$, we have:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Proof.

Intuition: Start with a power series that converges. Then, we have the recursion formula derived above that expresses the coefficients a_n in terms of the derivatives. □

Does this construction also work "the other way round"?

Question:

Does it work the other way round?

That is, given any function (possibly with nice assumptions), can we simply build the series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \qquad (\stackrel{?}{=} f(x))?$$

and "hope" that it converges to the function?

Taylor series

Taylor series

Theorem 143 (Taylor series)

Let $J \subseteq \mathbb{R}$ be an open interval, $f : J \to \mathbb{R}$, and suppose $f \in \mathcal{C}^{n+1}(J)$. For $a, x \in J$, define:

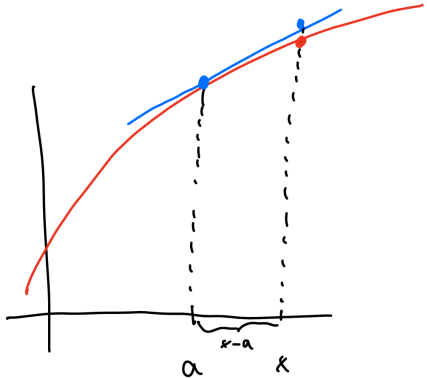
$$T_n(x,a) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \quad \text{(Taylor series up to degree } n\text{)},$$

$$R_n(x,a) := \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\,dt \quad \text{(Remainder term)}.$$

Then

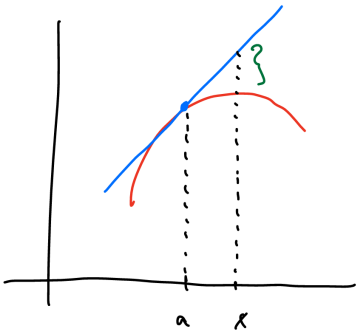
$$f(x) = T_n(x,a) + R_n(x,a).$$

Intuition about Taylor series



line is a good approx. of f around a

$$f(x) \approx f(a) + f'(a)(x-a)$$



line is not such a good approx., also need to look at curvature

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2$$

Proof

The proof follows from the Fundamental Theorem of Calculus, by induction on n.
Base case $n = 0$:

We need to prove:

$$f(x) = f(a) + \int_a^x f'(t)\,dt \quad \text{(which follows directly from the Fundamental Theorem).}$$

Inductive step $n \to n + 1$:

- Consider: $F(t) = \frac{(x-t)^{n+1}}{(n+1)!} f^{(n+1)}(t)$.
- Take its derivative
- Integrate, and exploit Fundamental Theorem □

Taylor with Lagrange remainder

Theorem 144 (Taylor with Lagrange remainder)

If $f \in \mathcal{C}^{n+1}(J)$, $a, x \in J$, then there exists some $\xi \in J$ such that:

$$R_n(x, a) = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(\xi).$$

Proof

Let $J = [a, b]$.

- Consider two functions $F, G \in \mathcal{C}^{n+1}([a, b])$. Assume that

$$F(a) = G(a) = 0, \text{ and } G' \neq 0 \text{ on } [a, b]. \qquad (\bigstar)$$

Now:

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F(b)}{G(b)} \overset{\text{interm. value thm}}{=} \frac{F'(\xi)}{G'(\xi)} \quad \text{for some } \xi \in [a, b].$$

Assume that F' and G' also satisfy $(\bigstar)$. We can iterate...

$$\frac{F(b)}{G(b)} = \frac{F^{(n+1)}(\xi)}{G^{(n+1)}(\xi)} \quad \text{for some } \xi \in [a, b]. \qquad (\bigstar)$$

Proof (2)

- Now choose $F(x) := f(x) - T_n(x, a) = R_n(x, a)$ and $G(x) := (x - a)^{n+1}$.
- For all k in $0 \leq k \leq n$, we have by construction that:

$$f^{(k)}(a) = T_n^{(k)}(a), \quad \text{so in particular } f^{(k+1)}(a) = 0, \quad \text{and } G^{(k)}(a) = 0.$$

- For $n + 1$, we now have:

$$F^{(n+1)}(x) = f^{(n+1)}(x), \quad G^{(n+1)}(x) = (n + 1)!.$$

By $(\bigstar)$, we obtain:

$$F(b) = R_n(x, a) = G(b) \cdot \frac{F^{(n+1)}(\xi)}{G^{(n+1)}(\xi)},$$

$$= \frac{(x - a)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi).$$

□

Taylor convergence

Theorem 145 (Taylor convergence)

Suppose $f \in \mathcal{C}^\infty(J)$, $x, a \in J$. Define:

$$T(x) := \lim_{n\to\infty} T_n(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Then we have $f(x) = T(x)$ if $R_n(x, a) \to 0$ as $n \to \infty$.

For example, this is the case if there exist constants $\alpha, C > 0$ such that:

$$|f^{(n)}(t)| \leq \alpha \cdot C^n \quad \forall t \in J,\, n \in \mathbb{N}.$$

This follows directly from the Lagrange remainder.

Examples

- Exponential series:

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

power series with $r = \infty$, $\exp$ always coincides with its Taylor series.

- Logarithm:

$$f(x) = \log(1+x), \quad \text{Taylor series about } a = 0.$$

Convergence radius of Taylor series is $r = 1$. For x outside of $]-1, 1[$, the Taylor series does not make sense at all.

Examples (2)

- Special example:

$$f(x) = \begin{cases} \exp(-1/x^2) & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Has the property that $\forall n \in \mathbb{N} : f^{(n)}(\underline{0}) = 0$.

Consider the Taylor series derived about $a = 0$.

All terms will be 0, so $T_n(x) = 0$, $r = \infty$. But of course $f \neq 0$, so we get:

$$T_n(x) \neq f(x) \quad \forall x \neq 0.$$

Taylor series converges everywhere, but not to the function f!

Lebesgue measure on R^n SKIPPED

Goal

We want to construct a measure on $\mathbb{R}^n$. We want that rectangles of the form

$$[a_1, b_1[\times[a_2, b_2[\times \cdots \times [a_n, b_n[$$

have the "natural volume" given by $\prod_{i=1}^{n}(b_i - a_i)$

$$\Rightarrow \text{vol}(R) \overset{!}{=} c_1 \cdot c_2$$

And we want "nice" mathematical properties.

Earlier approaches

First approaches (Jordan, Riemann) attempted the following:

"Outer approximation":

$$A \subset \bigcup_{i=1}^{\infty} \text{rect}_i$$

Earlier approaches (2)

"Inner approximation":

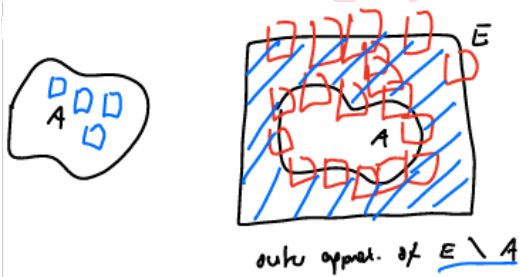
$$\bigcup_{i=1}^{\infty} \text{rectangles}_i \subset A$$

A would be called "measurable" if outer and inner approximation "converge".

Problem: Too many sets turn out to be not measurable (e.g. $\mathbb{Q}$)

Now: generalization of this approach

- Allow for countable coverings
- Replace inner approximation by an outer approximation of the complement:



$$\mu(E) = \underline{\mu(E \setminus A)} + \mu(A)$$

$$\mu(A) = \mu(E) - \underline{\mu(E \setminus A)}$$

- Need σ-algebra as underlying structure.

Outer Lebesgue measure

Set the "natural volume" of rectangles:

$$R = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$$

$$|R| := \prod_{i=1}^{n} (b_i - a_i)$$

Outer Lebesgue measure (2)

Definition of outer Lebesgue measure:
Let $A \subset \mathbb{R}^n$ be arbitrary. We define

$$\lambda(A) := \inf \left\{ \sum_{i=1}^{\infty} |R_i| \;\middle|\; A \subset \bigcup_{i=1}^{\infty} R_i,\ R_i \text{ rectangle} \right\}$$

We cover A by a countable union of rectangles, then take $\inf$.

Observe: $\lambda(A) \in \mathbb{R} \cup \{\infty\}$.

We want to make λ into a measure. Problem: if we use $\mathcal{P}(\mathbb{R}^n)$ as σ-algebra, we run into contradictions.

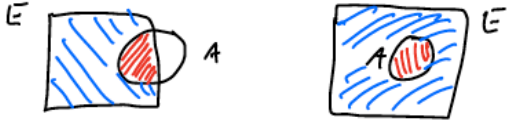
We need to restrict ourselves to a smaller σ-algebra...

Measurable set

Definition 146 (Measurable set)

We say that a set $A \subset \mathbb{R}^n$ is measurable if for all $E \subset \mathbb{R}^n$

$$\lambda(E) = \underline{\lambda(E \cap A)} + \underline{\lambda(E \setminus A)}$$



Denote by $\mathcal{L}$ all measurable subsets of $\mathbb{R}^n$.

Outer measure as measure...

Theorem 147 (Outer measure as measure)

The set $\mathcal{L}$ forms a σ-algebra on $\mathbb{R}^n$. The outer measure λ (defined above) is in fact a measure on

$$(\mathbb{R}^n, \mathcal{L}).$$

On rectangles it coincides with the "natural volume".

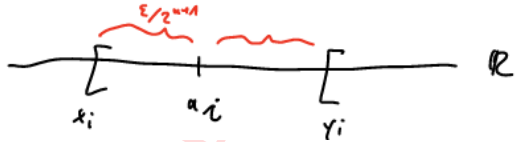
Examples:

- $\lambda(\{x\}) = 0$
- $\lambda(\mathbb{R}) = \infty$
- $A \subset \mathbb{R}$ countable $\Rightarrow \lambda(A) = 0$. In particular, $\mathbb{Q}$ is measurable and has $\lambda(\mathbb{Q}) = 0$.

Proof sketch

For $\varepsilon > 0$, define for all $a_i \in A$ the interval $[x_i, y_i[$ such that

$$x_i = a_i - \frac{\varepsilon}{2^{i+1}}, \quad y_i = a_i + \frac{\varepsilon}{2^{i+1}}$$



$$A \subset \bigcup_{i=1}^{\infty} [x_i, y_i[$$

$$\Rightarrow \lambda(A) \leq \sum_{i=1}^{\infty} \lambda([x_i, y_i[) = \sum_{i=1}^{\infty} \frac{\varepsilon}{2^{i+1}} = \varepsilon$$

Taking the $\inf$. over all coverings shows that $\lambda(A) = 0$.

Comparing $\mathcal{L}$ (σ-algebra of Lebesgue measurable sets) with the Borel-σ-algebra $\mathcal{B}$

(1) $\mathcal{B} \subset \mathcal{L}$:

- Open intervals are measurable, thus in $\mathcal{L}$.
- Any open set $A \subset \mathbb{R}^n$ can be written as a countable union of open intervals:

$$A \subset \bigcup_{i=1}^{\infty} I_i, \quad I_i \text{ open intervals}$$

(2) For every Lebesgue-measurable set L, there exists a set $B \in \mathcal{B}$ and $N \in \mathcal{L}$ with $\lambda(N) = 0$ such that

$$L = B \cup N$$

Summary: $\mathcal{L} \approx \mathcal{B}$ (up to sets of measure 0).

A non-measurable set SKIPPED

Construction (quite abstract!)

Consider $[0, 1[$. Define an equivalence relation on $[0, 1[$ as follows:

$$x \sim y \Leftrightarrow x - y \in \mathbb{Q}$$

$\frac{\pi}{4}$, $\frac{\pi}{4} + \frac{1}{2}$, $\frac{\pi}{4} + 1$, $\frac{\pi}{4} + \frac{789}{800}$ would be equivalent

Construction (quite abstract!) (2)

Consider the equivalence classes

$$\frac{\pi}{4} + \mathbb{Q} = \left\{ \frac{\pi}{4} + q \mid q \in \mathbb{Q} \right\}$$

$$\frac{5}{3} + \mathbb{Q}$$

$$\frac{\sqrt{2}}{2} + \mathbb{Q}$$

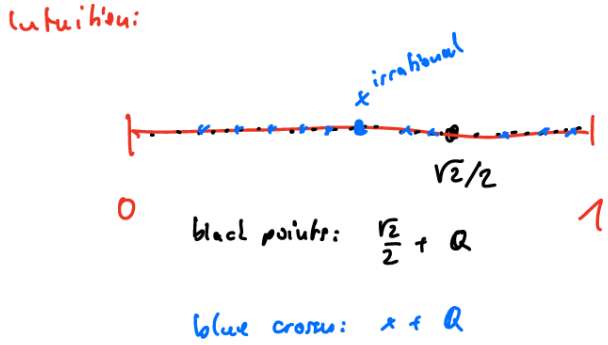
$$\vdots$$

We pick a representative of each of the classes, and denote by $\mathcal{N}$ the set of all such representations.

$\mathcal{N}$ is not Lebesgue-measurable!

Proposition 148 ($\mathcal{N}$ is not Lebesgue-measurable)

$\mathcal{N}$ is not Lebesgue-measurable.

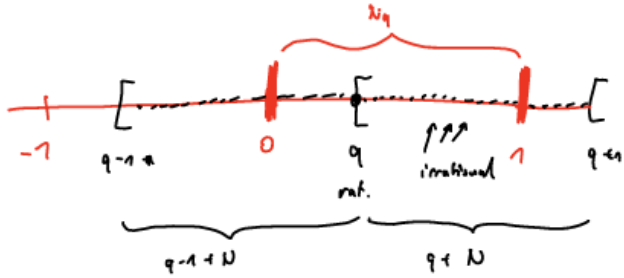


Proof

Proof by contradiction:

Assume $\mathcal{N}$ is measurable.
We now construct the following sets: For $q \in [0,1[\cap \mathbb{Q}$

$$N_q := ((q + \mathcal{N}) \cup (q - 1 + \mathcal{N})) \cap [0,1[$$



Proof (2)

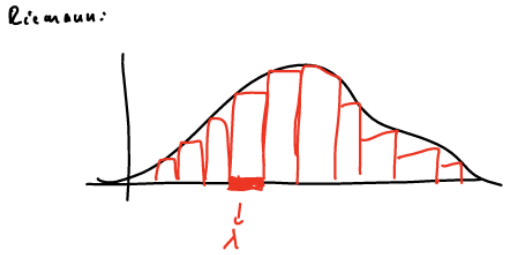
- If $\mathcal{N}$ is measurable, then $q + \mathcal{N}$ is measurable $\forall q \in [0,1[$, and $\lambda(N_q) = \lambda(\mathcal{N})$
- $[0,1[= \bigcup_{q \in [0,1] \cap \mathbb{Q}} N_q$
- $N_q \cap N_p \neq \emptyset \quad \Rightarrow \quad N_p = N_q$
  $\Rightarrow \bigcup N_q$ is disjoint
- σ-additivity:

$$\underbrace{\lambda([0,1[)}_{1} = \lambda\left(\bigcup_q N_q\right) \overset{\downarrow}{=} \sum_{q \in [0,1] \cap \mathbb{Q}} \underbrace{\lambda(N_q)}_{=\lambda(\mathcal{N})}$$

- Could be that $\lambda(N_q) = 0$. But then $\sum_q \lambda(N_q) = 0$ ↯
- Could be that $\lambda(N_q) > 0$. But then $\sum_q \lambda(N_q) = \infty$ ↯

The Lebesgue-integral on R^n SKIPPED

Intuition: Partition Y instead of X



Measurable functions

Definition 149 (Measurable functions)

A function $f\colon (X, \mathcal{F}) \to (Y, \mathcal{G})$ between two measurable spaces is called measurable if pre-images of measurable sets are measurable:

$$\forall G \in \mathcal{G}: \quad f^{-1}(G) \in \mathcal{F}$$

$$f^{-1}(G) := \{x \in X \mid f(x) \in G\}$$

Lebesgue-integral for simple functions

Definition 150 (Lebesgue-integral for simple functions)

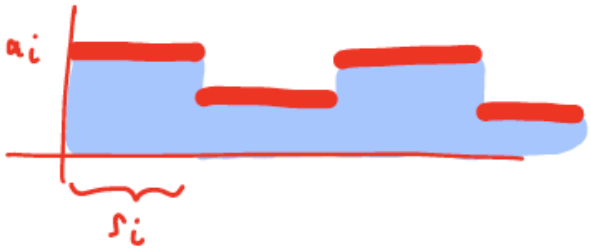
A function $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ is called a simple function if there exist measurable sets $S_i \subset \mathbb{R}^n$, $a_i \in \mathbb{R}$ such that

$$\Phi = \sum_{i=1}^{n} a_i \cdot \mathbb{1}_{\{S_i\}} \quad \text{where } S_i = \Phi^{-1}(a_i)$$

For such a simple function we can define its Lebesgue integral as

$$\int \Phi \, d\lambda := \sum_{i=1}^{n} a_i \cdot \lambda(S_i)$$

Lebesgue-integral for simple functions (2)



Lebesgue integral for non-negative functions

For a non-negative function $f^+ : \mathbb{R}^n \to [0, \infty[$ we define its Lebesgue integral

$$\int f^+ \, d\lambda := \sup \left\{ \int \Phi \, d\lambda \;\middle|\; \Phi \leq f_1, \ \Phi \text{ simple} \right\} \quad \text{(might be } \infty\text{)}$$

Φ approximates f by simple functions. Note: the sets S_i can be complicated sets, not just intervals!

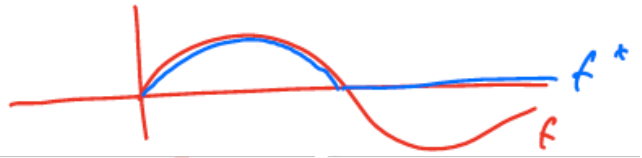
Lebesgue integral for general functions

- For a general function $f\colon \mathbb{R}^n \to \mathbb{R}$ we split the function into positive and negative parts:

$$f = f^+ - f^-$$

where

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



- Note: f^+ and f^- are measurable if f is measurable.

Lebesgue integral for general functions (2)

- If both f^+ and f^- satisfy

$$\int f^+ d\lambda < \infty, \quad \int f^- d\lambda < \infty,$$

then we call f integrable and define

$$\int f \, d\lambda := \int f^+ \, d\lambda - \int f^- \, d\lambda$$

Much more powerful notion than Riemann integral.

Example:

$$\int \mathbb{1}_{\mathbb{Q}} \, d\lambda = 1 \cdot \lambda(\mathbb{Q}) = 0$$

Monotone convergence

Theorem 151 (Monotone convergence)

Consider a sequence of functions $f_k \colon \mathbb{R}^n \to [0, \infty[$ that is pointwise non-decreasing:

$$\forall x \in \mathbb{R}^n : \quad f_{k+1}(x) \geq f_k(x)$$

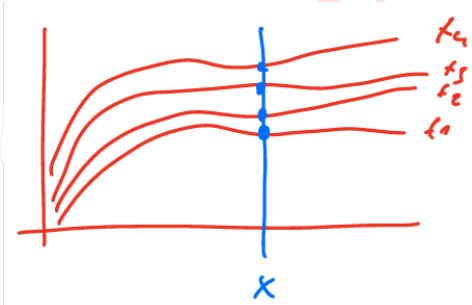
Assume that all f_k are measurable, and that the pointwise limit exists:

$$\forall x : \quad \lim_{k \to \infty} f_k(x) = f(x)$$

Then:

$$\int f(x)\,dx = \int \lim_{k \to \infty} f_k(x)\,dx = \lim_{k \to \infty} \int f_k(x)\,dx$$

Monotone convergence (2)



Dominated convergence

Theorem 152 (Dominated convergence)

Let $f_n \colon B \to \mathbb{R}$ with $|f_n(x)| \leq g(x)$ on B, where $g(x)$ is integrable. Assume that the pointwise limit exists:

$$\forall x \in B : \quad f(x) := \lim_{n \to \infty} f_n(x)$$

Then:

$$\int f(x)\,dx = \lim_{n \to \infty} \int f_n(x)\,dx$$

Partial derivatives

Functions on $\mathbb{R}^n$

We now consider functions $f : \mathbb{R}^n \to \mathbb{R}$.

- Input space: n-dimensional.
- Output space: 1-dimensional.

(The standard object in machine learning!)

Generalizing derivatives from 1 to n dimensions

Two obvious "ideas":

(1) Consider the function $f(x) = f(x_1, \ldots, x_n)^T$ as a function of one coordinate only and keep the others fixed; then apply 1-dimensional intuition.

$\Rightarrow$ partial derivatives

(2) Approximate f locally by something linear.

$\Rightarrow$ total derivative

As we will see, leads to similar but not identical notions.

Partial derivatives on $\mathbb{R}^n$

Consider $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 153 (Partial derivatives on $\mathbb{R}^n$)

A function f is called partially differentiable with respect to variable x_j at point $\xi \in \mathbb{R}^n$ if the 1-dimensional (!) function:

$$g : \mathbb{R} \to \mathbb{R}, \quad g(x_j) := f(\xi_1, \xi_2, \ldots, \xi_{j-1}, x_j, \xi_{j+1}, \ldots, \xi_n)$$

is differentiable at $\xi_j \in \mathbb{R}$.

Intuition: Fix all arguments except the j-th.

Partial derivatives on $\mathbb{R}^n$ (2)

Notation:

point at which we evaluate the derivative.

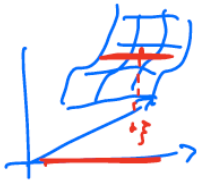
$$\frac{\partial f}{\partial x_j}(\bar{x}) := \lim_{h \to 0} \frac{f(\bar{x} + e_j \cdot h) - f(\bar{x})}{h}$$

j-th unit vector: $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$

constant > 0

"round" delta-sign

variable wrt which we compute the derivative



Gradient

If all partial derivatives exist, then the vector of all partial derivatives is called the gradient:

$$\operatorname{grad}(f(\xi)) = \nabla f(\xi) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\xi) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\xi) \end{pmatrix} \in \mathbb{R}^n.$$

Jacobian matrix

If $f : \mathbb{R}^n \to \mathbb{R}^m$, we decompose f into its m component functions $f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$. We define the Jacobian matrix:

$$Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

$\to$ The first row is $\operatorname{grad}(f_1)^\top$

Existence of Gradient $\not\Rightarrow$ Continuity of f!

For functions $f : \mathbb{R} \to \mathbb{R}$, we know that if f is differentiable, then it is continuous.

Note that in the n-dimensional case, the existence of a gradient is not enough for this:

⚠

Even if all partial derivatives exist at ξ, we do not know whether f is continuous at ξ!

Need stronger notions ... total derivative.

Example

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, defined as:

$$f(x,y) = \begin{cases} \frac{x \cdot y}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } x = y = 0. \end{cases}$$

One can prove that the two partial derivatives at $(0,0)$ exist and are both 0, but the function f is not continuous at 0.

Exercise!

Total derivative

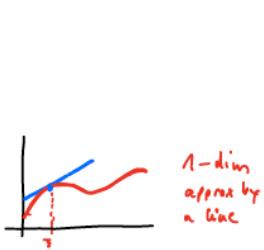
Differentiable function

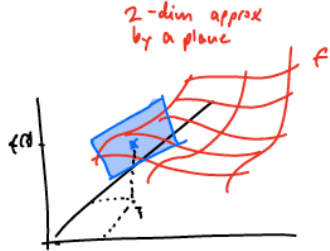
Consider $f : \mathbb{R}^n \to \mathbb{R}^m$, $\xi \in U$. f is differentiable at ξ if there exists a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ such that for $h \in \mathbb{R}^n$:

$$f(\xi + h) - f(\xi) = L(h) + r(h),$$

with

$$\lim_{h \to 0} \frac{r(h)}{|h|} \to 0.$$





Differentiable function (2)

⚠ "Differentiable" refers to the existence of the total derivative, not the partial ones.

(As we will see, total derivative $\Longrightarrow$ partial derivatives, but not vice versa.)

Differentiable, Continuous, Gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at ξ. (Holds for all of the following theorems)

Theorem 154 (Continuous)

Then f is continuous at ξ.

Differentiable, Continuous, Gradient (2)

Theorem 155 (Gradient and linear function)

The linear functional L coincides with the gradient:

$$f(\xi + h) - f(\xi) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\xi) \cdot h_j + r(h),$$

or equivalently:

$$f(\xi + h) - f(\xi) = \langle \nabla f(\xi), h \rangle + r(h),$$

where $h = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$.

Differentiable, Continuous, Gradient (3)

Special case:

$$h = e_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \Rightarrow \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) \cdot h_j = \frac{\partial f}{\partial x_j} = \text{j-th partial derivative}$$

Differentiable, Continuous, Gradient (4)

Theorem 156 (Differentiability via Partial Derivatives)

If $f : \mathbb{R}^n \to \mathbb{R}^m$, it is differentiable iff all coordinate functions $f_1, \ldots, f_m$ are differentiable. Then all partial derivatives exist, and:

$$L(h) = (\text{Jacobi matrix}) \cdot h.$$

⚠ The theorem says that *if* total derivative exists, then we can use the partial derivatives to compute it.

❓ Is there a theorem that starts with the partial derivatives and then derives the total one?

Yes:

Continuous Partial Derivatives $\Longrightarrow$ Differentiable

Theorem 157 (Continuous Partial Derivatives $\Longrightarrow$ Differentiable)

If all partial derivatives exist and they are all continuous, then f is differentiable.

Directional Derivatives

Directional Derivatives

Definition 158 (Directional Derivatives)

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $v \in \mathbb{R}^n$ with $\|v\| = 1$.

The directional derivative of f at ξ in the direction of v is defined as:

$$D_v f(\xi) := \lim_{h \to 0} \frac{f(\xi + h \cdot v) - f(\xi)}{h}.$$

Observe: Partial derivatives are directional derivatives in the direction of the unit vectors.

Differentiable Directional $\Longrightarrow$ Derivatives and Gradient

Theorem 159 (Differentiable Directional $\Longrightarrow$ Derivatives and Gradient)

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable in ξ, then all the directional derivatives exist, and we can compute them by:

$$D_v f(\xi) = (\nabla f(\xi))^\top v = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}(\xi),$$

where $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$.

Largest ascent along gradient

Proposition 160 (Largest ascent along gradient)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at ξ. Then the largest value $|D_v f(\xi)|$ among the directional derivatives is attained in the direction of the gradient:

$$v = \frac{\nabla f(\xi)}{\|\nabla f(\xi)\|}.$$

Largest ascent along gradient (2)

Proof intuition:

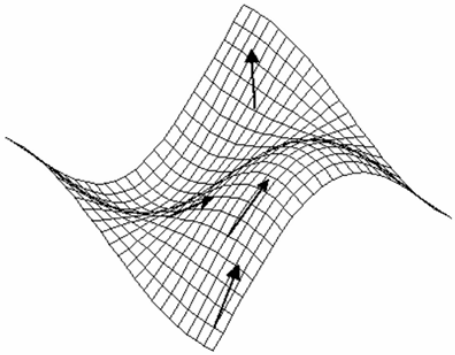
$$\max_{\{v;\|v\|=1\}} |D_v f(\xi)| = \max_{\{v;\|v\|=1\}} |\langle \nabla f(\xi), v\rangle| \overset{\text{Cauchy-Schwarz}}{\leq} \max_{\{v;\|v\|=1\}} \|\nabla f(\xi)\| \cdot \|v\| = \|\nabla f(\xi)\|.$$

In particular, for $v = \frac{\nabla f(\xi)}{\|\nabla f(\xi)\|}$, we have:

$$D_v f(\xi) = \langle \nabla f(\xi), \frac{\nabla f(\xi)}{\|\nabla f(\xi)\|} \rangle = \|\nabla f(\xi)\|.$$

Thus, the maximum is obtained for this choice of v, and the inequality becomes an equality. □

Illustration of the gradient



Gradient Vectors Shown at Several Points on the Surface of cos(x) sin(y)

Higher-Order Derivatives

Higher-Order Derivatives

Consider $f : \mathbb{R}^n \to \mathbb{R}$, assume it is differentiable, so all partial derivatives $\frac{\partial f}{\partial x_j} : \mathbb{R}^n \to \mathbb{R}$ exist. If the partial derivatives are differentiable themselves, we can take their derivatives:

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) := \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

These are called second-order partial derivatives.

Attention: Order Matters!

⚠ In general, we cannot change the order of derivatives:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \neq \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Example

Let $f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = \frac{x \cdot y}{x^2+y^2}$.

$$\text{Gradient: } \nabla f(x,y) = \begin{pmatrix} \frac{y^3(y^2-x^2)}{(x^2+y^2)^2} \\ \frac{xy^2(3x^2+y^2)}{(x^2+y^2)^2} \end{pmatrix}.$$

- At $(0,y)$:

$$\frac{\partial f}{\partial x}(0,y) = y \quad \forall y,$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \boxed{1}.$$

Example (2)

- At $(x, 0)$:

$$\frac{\partial f}{\partial y}(x, 0) = 0 \quad \forall x,$$

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \boxed{0}.$$

Conclusion: The two derivatives do not agree at the point $(0, 0)$, demonstrating that the order of differentiation matters.

Hessian

Definition 161 (Hessian)

Let $f : \mathbb{R}^n \to \mathbb{R}$. Then we define the Hessian of f at a point x by

$$(Hf)_{ij}(x) := \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \quad i, j = 1, \ldots, n.$$

Be Aware of Different Dimensionality

- $f : \mathbb{R}^n \to \mathbb{R}$: function
- $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$: first derivative, n partial derivatives $\frac{\partial f}{\partial x_i}$
- $Hf : \mathbb{R}^n \to \mathbb{R}^{n \times n}$: second derivative, n^2 partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$

Continuously differentiable functions

Definition 162 (Continuously differentiable functions)

We say that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, if all partial derivatives $\frac{\partial f}{\partial x_i}$ exist and are continuous.

We say that f is twice continuously differentiable if f is continuously differentiable and all partial derivatives $\frac{\partial f}{\partial x_i}$ are again continuously differentiable.

Analogously: k times continuously differentiable.

Notation.

$$C^k(\mathbb{R}^n, \mathbb{R}^m) = \{f : \mathbb{R}^n \to \mathbb{R}^m \mid f \text{ is } k \text{ times cont. diff.}\},$$

$$C^\infty(\mathbb{R}^n, \mathbb{R}^m) = \{f : \mathbb{R}^n \to \mathbb{R}^m \mid f \text{ is infinitely often cont. diff.}\}.$$

Continuously Differentiable $\Rightarrow$ can change order of derivatives

Theorem 163 (Schwartz)

Assume that f is twice continuously differentiable. Then we can exchange the order in which we take partial derivatives:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Analogously: k times continuously differentiable $\Rightarrow$ can exchange order of first k partial derivatives.

Minima/Maxima

Critical point

Definition 164 (Critical point)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. If

$$\nabla f(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

we call x a critical point.

⚠ A critical point can have many different geometric meanings:

Minimum

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- f has a local minimum at x_0 if there exists $\varepsilon > 0$ such that

$$\forall x \in B_\varepsilon(x_0) : f(x) \geq f(x_0).$$

- f has a strict local minimum at x_0 if there exists $\varepsilon > 0$ such that

$$\forall x \in B_\varepsilon(x_0) : f(x) > f(x_0).$$

- f has a global minimum at x_0 if

$$\forall x \in \mathbb{R}^n : f(x) \geq f(x_0).$$

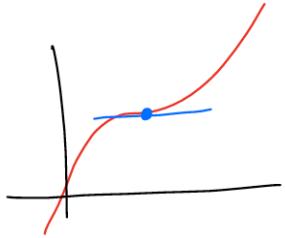
- f has a strict global minimum at x_0 if

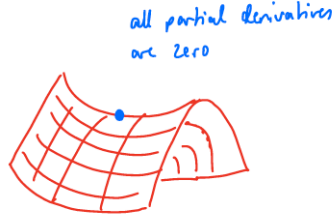
$$\forall x \in \mathbb{R}^n : f(x) > f(x_0).$$

Minimum (2)

Saddle point

If f is differentiable and x_0 is a critical point that is neither a local minimum nor a maximum, then we call it a saddle point:

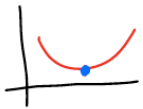




How Can We Find Out Which Case We Have?

Intuition in 1-dimensional case: second derivatives might help.

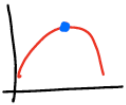
Strict local min:



$f'(x) = 0$

$f''(x) > 0$

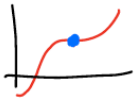
Strict local max



$f'(x) = 0$

$f''(x) < 0$

saddle pt:



$f'(x) = 0$

$f''(x) = 0$

Critical Points and the Hessian

Theorem 165 (Critical Points and the Hessian)

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathbb{R}^n)$. Assume that x_0 is a critical point, i.e., $\nabla f(x_0) = 0$. Then:

(i) If x_0 is a local minimum (maximum), then the Hessian $Hf(x_0)$ is positive semi-definite (negative semi-definite).

(ii) If $Hf(x_0)$ is positive definite (negative definite), then x_0 is a strict local minimum (maximum).

(iii) If $Hf(x_0)$ is indefinite (has both positive and negative eigenvalues), then x_0 is a saddle point.

Critical Points and the Hessian (2)

- In (i), we can have a strict local maximum, yet the Hessian is only semi-definite. Example: $f(x) = x^4$ at $x = 0$.

- In (ii), if the Hessian is semi-definite, no statement can be made.

Derivatives of popular matrix/vector functions

Example: Linear least squares

Given training points $x_1, \ldots, x_n \in \mathbb{R}^d, y_1, \ldots, y_n \in \mathbb{R}$. We want to approximate this data by a linear least squares problem: find a linear function

$$f : \mathbb{R}^d \to \mathbb{R}$$

that minimizes the least squares error.

Example: Linear least squares (2)

In matrix notation:

$$X = \begin{pmatrix} \cdots x_1 \cdots \\ \cdots x_2 \cdots \\ \cdots \vdots \cdots \\ \cdots x_n \cdots \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

$$f : \mathbb{R}^d \to \mathbb{R}, \quad f(x) = \langle w, x \rangle \quad \text{with parameter vector } w,$$

determine w as

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} |\underbrace{\langle x_i, w \rangle}_{f(x_i)} - y_i|^2 = \min_{w \in \mathbb{R}^d} \underbrace{\boxed{\|Xw - Y\|^2}}_{=:g(x)}.$$

Denote $g(w)$ as the $\boxed{objective function}$.

Solution "by foot"

To optimize for w, we need to take derivatives of the objective function to 0:

$$\frac{\partial g}{\partial w} \stackrel{!}{=} 0.$$

To compute the gradient "by foot" is pretty cumbersome:

- Write function coordinate-wise:

$$g(w) = \sum_{j=1}^{n} \left(y_j - \sum_{k=1}^{d} x_{jk} w_k \right)^2.$$

- Take partial derivatives:

$$\frac{\partial g}{\partial w_i} = \sum_{j=1}^{n} (-x_{ji}) \cdot 2 \left(y_j - \sum_{k=1}^{d} x_{jk} w_k \right).$$

Observe that we can write the result using matrices

$$\frac{\partial g}{\partial \omega_i} = \sum_{j=1}^{n}(-x_{ji})\cdot 2\cdot\left(y_j - \sum_{k=1}^{m} x_{jk}\omega_k\right) = -2\cdot\sum_{j=1}^{n} x_{ji}\cdot(y - X\omega)_j = \left(-2X^t(y - X\omega)\right)_i$$

Thus,

$$\nabla g(w) = -2X^t(Y - Xw).$$

Observe: This is structurally close to the 1-dimensional case:

$$g(w) = (y - xw)^2,$$

$$g'(w) = -x(y - xw)\cdot 2 = -2x(y - xw).$$

The matrix cookbook

Lookup table ("cookbook") for gradients of many important functions

Examples for functions of vectors: $f : \mathbb{R}^n \to \mathbb{R}$

- Linear function:

$$f(x) = a^t x = \langle a, x \rangle \quad (a \in \mathbb{R}^n) \quad \Rightarrow \quad \frac{\partial f}{\partial x} = a \quad \in \mathbb{R}^n.$$

- Quadratic function:

$$f(x) = x^t A x \quad \Rightarrow \quad \frac{\partial f}{\partial x} = (A + A^t)x \quad \in \mathbb{R}^n.$$

Examples for functions of matrices: $f : \mathbb{R}^{n\times m} \to \mathbb{R}$

- Linear form:

$$f(X) = \underbrace{a^t}_{1\times n}\underbrace{X}_{n\times m}\underbrace{b}_{m\times 1} \quad \text{for } a \in \mathbb{R}^n, b \in \mathbb{R}^m \quad \Rightarrow \quad \frac{\partial f}{\partial X} = \underbrace{\underbrace{a}_{n\times 1}\cdot\underbrace{b^t}_{1\times m}}_{n\times m} \in \mathbb{R}^{n\times m}.$$

- Quadratic form:

$$\underbrace{f(X)}_{n\times m} = \underbrace{a^t}_{1\times m}\underbrace{X^t}_{m\times n} C \underbrace{X}_{n\times m}\underbrace{b}_{m\times 1} \quad \text{for } a \in \mathbb{R}^m, b \in \mathbb{R}^n, C \in \mathbb{R}^{m\times m} \quad \Rightarrow$$

$$\frac{\partial f}{\partial X} = CXab^t + CXba^t.$$

Examples for functions of matrices: $f : \mathbb{R}^{n\times n} \to \mathbb{R}$

- Trace:

$$f(X) = \operatorname{tr}(X) \quad \Rightarrow \quad \frac{\partial f}{\partial X} = I \quad \in \mathbb{R}^{n\times n}.$$

$$f(X) = \operatorname{tr}(A \cdot X) \quad \Rightarrow \quad \frac{\partial f}{\partial X} = A.$$

$$f(X) = \operatorname{tr}(X^t A X) \quad \Rightarrow \quad \frac{\partial f}{\partial X} = (A + A^t)X.$$

- Determinant:

$$f(X) = \det(X) \quad \Rightarrow \quad \frac{\partial f}{\partial X} = \det(X) \cdot (X^{-1})^t.$$

$$\frac{\partial \det}{\partial a_{sr}} = \det(A) \cdot (A^{-1})_{rs}.$$

Examples for functions of matrices: $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$

- Inverse:

$$f(A) = A^{-1}, \quad f_{ij} := (A^{-1})_{ij} \quad \Rightarrow \quad \frac{\partial f_{ij}}{\partial a_{uv}} = -(a_{iu})^{-1}(a_{vj})^{-1}.$$

Part III
Optimization

Convex sets and functions

Convex set

Definition 166 (Convex set)

Consider a set $\Omega \subset V$ in a vector space V. The set Ω is called convex if for all $t \in [0,1]$ and for all $x_1, x_2 \in \Omega$,

$$tx_1 + (1-t)x_2 \in \Omega.$$

Intuition: The line connecting x_1, x_2 is completely inside Ω.

Convex set (2)

Intersection of convex sets are convex

Easy to see from the definition:
If A, B are two convex sets, then also $A \cap B$ is convex.

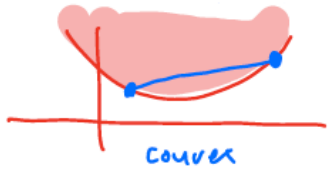
Convex function

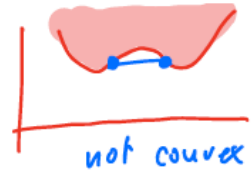
Consider a vector space V, a set $\Omega \subset V$, and a function $f : \Omega \to \mathbb{R}$. The function f is called convex (or strictly convex ("$<$" instead of "$\leq$")) if:

- Ω is a convex set.
- $\forall t \in [0, 1]$:

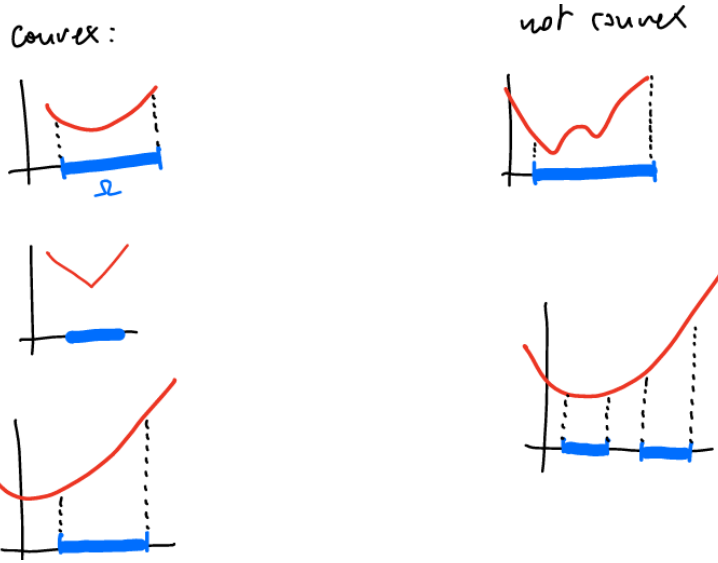
$$f(tx + (1-t)y) \overset{<}{\leq} tf(x) + (1-t)f(y).$$

Intuition: The line connecting $(x, f(x))$ and $(y, f(y))$ is "above" the graph of the function.





Examples

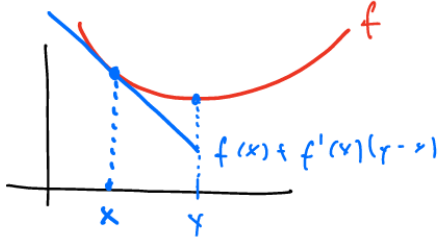


Differentiable, convex functions

Proposition 167 (Differentiable, convex functions)

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex $\iff$ $\forall x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x).$$

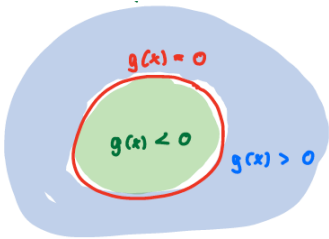


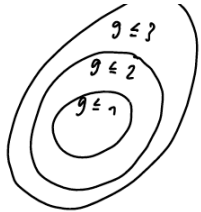
Convex $\Longrightarrow$ sublevel sets convex

Observation: For a convex function g, the sublevel sets

$$S_a := \{x \mid g(x) \leq a\}$$

are convex.





(Funnily: It is not true the other way around: A function can have all sublevel sets convex while the function itself is not convex.)

Strongly convex functions

Consider a function f defined on a convex domain. For simplicity, let us assume that it is differentiable. We say that f is μ-strongly convex if and only if for all $x, y \in \mathbb{R}^d$,

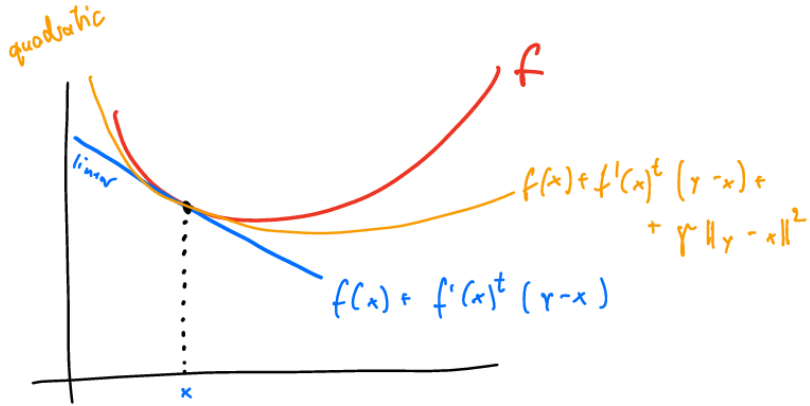
$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu}{2}\|y - x\|_2^2.$$

Intuition: "Lower bound μ on the curvature."

Intuition on $\mathbb{R}$:

- A linear function is convex but not strongly convex.
- If f is twice differentiable, then f is convex if $f'' \geq 0$.
- f is μ-strongly convex if $f'' > 0$ with $f'' \geq \mu$.

Strongly convex functions (2)

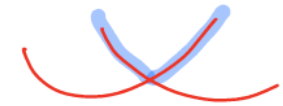


function f is "above" the quadratic approximation of f

Operations that preserve convexity of functions

- Weighted sums: If $f_1, \dots, f_n$ are convex, $w_1, \dots, w_n > 0$, then $f := \sum w_i f_i$ is convex.

- Pointwise max: If f_1, f_2 are convex, then $g(x) := \max\{f_1(x), f_2(x)\}$ is convex.



L-Lipschitz and L-smooth functions

- A function f is called L-Lipschitz if there exists a constant $L > 0$ such that

$$|f(u) - f(v)| \leq L \cdot \|u - v\| \quad \text{for all } u, v \in \Omega.$$

  If f is differentiable, this is equivalent to $\|\nabla f\| \leq L$.
  Intuition: "Upper bound on steepness."

- A function f is called L-smooth if its gradient (!) ∇f is L-Lipschitz.
  Intuition: "Upper bound L on the curvature."

Convexity and second derivatives

Proposition 168 (Convexity and second derivatives)

Let $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d$ open, convex. Assume that f is twice continuously differentiable (Hessian symmetric). Then:

- f is convex if the Hessian of f is positive semi-definite $\forall x$.
- f is strictly convex if the Hessian is positive definite.
- f is strongly convex with parameter μ if the smallest eigenvalue $\lambda_{\min}$ of the Hessian satisfies

$$\lambda_{\min}(H_f(x)) \geq \mu > 0 \quad \text{for all } x \in \Omega.$$

Proof.

Exercise. □

Convexity and first derivatives

Proposition 169 (Convexity and first derivatives)

Let $f : \Omega \to \mathbb{R},\ \Omega \subset \mathbb{R}^d$ be continuously differentiable and Ω an open convex set, then:

- f is convex $\iff \forall x, y \in \Omega$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

- f is strictly convex $\iff$ we have strict inequality.
- f is strongly convex with parameter $\mu \iff \forall x, y \in \Omega$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2$$

Condition number

If f is μ-strongly convex and β-smooth, and is twice differentiable, then all eigenvalues of the Hessian are in the interval $[\mu, \beta]$.

We then denote by

$$\kappa := \frac{\beta}{\mu}$$

the condition number.

(Often it is also defined with the Hessian directly, as the ratio of the largest to smallest eigenvalue of the Hessian.)

Will see: It's important for gradient descent.

Jensen's inequality

Intuition: Let f be convex.

- Convex: if $\lambda_1 + \lambda_2 = 1,\ \lambda_1, \lambda_2 \geq 0$, we have

$$f\left(\sum_{i=1}^{2} \lambda_i x_i\right) \leq \sum_{i=1}^{2} \lambda_i f(x_i)$$

- Can extend this to finite sums: if $\sum_{i=1}^{m} \lambda_i = 1,\ \lambda_i \geq 0$, then

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$$

Jensen's inequality (2)

- Can extend this to integrals:

$$f\left(\int_S x\, dP(x)\right) \leq \int_S f(x)\, dP(x) \quad \text{, (for an arbitrary measure P)}$$

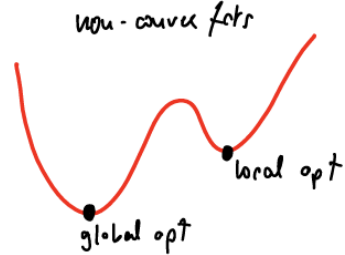
[or $f\big(\mathbb{E}(x)\big) \leq \mathbb{E}\big(f(x)\big)$ where $\mathbb{E}$ is the expectation]

This is called Jensen's inequality for convex functions.

Convex functions and global optima

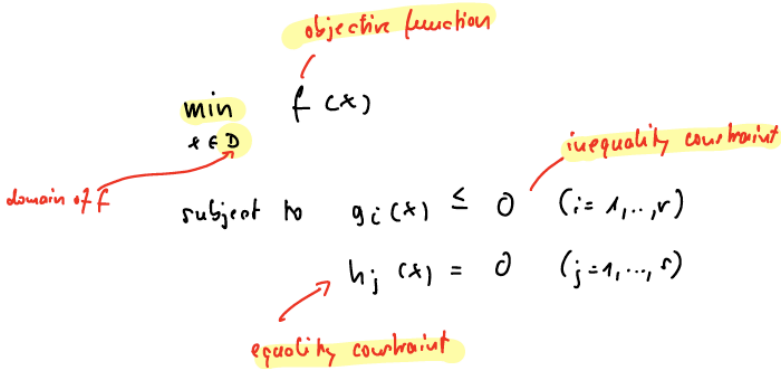
Theorem 170 (Convex functions and global optima)

Any local minimum of a convex function is also a global minimum!



Optimization problems

General optimization problem



- A point $x \in \mathcal{D}$ that satisfies all constraints is called feasible.
- A minimizer is typically denoted by x^*, the optimal value as f^*:

$$f^* = f(x^*)$$

Convex optimization problem

An optimization problem is called convex if:

- The objective f is convex (and defined on a convex domain).
- The inequality constraints g_i are convex.
- The equality constraints h_j are linear:

$$\langle a_j, x \rangle - b_j = 0$$

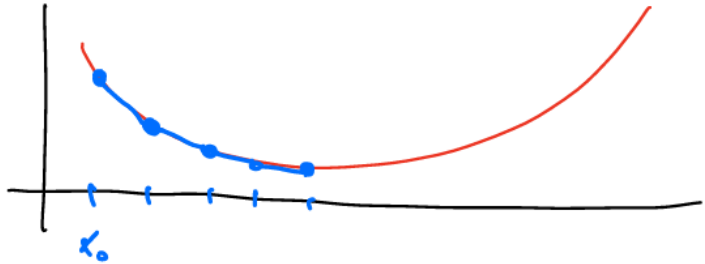
Feasible set is convex

The set of feasible points of a convex optimization problem is convex:

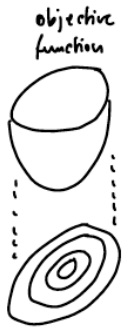
- Domain $\mathcal{D}$ is convex.
- For each inequality constraint $g_i(x) \leq 0$, the set $\{x \mid g_i(x) \leq 0\}$ is convex.
- For each equality constraint $h_j(x) = 0$, the set $\{x \mid h_j(x) = 0\}$ is convex.
- The intersection of convex sets is convex.

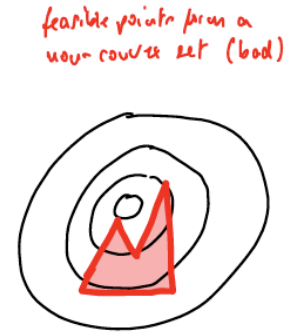
Standard algorithms to solve convex problems

... some variant of gradient descent, see next lecture ...



What if the set of feasible points is not convex?





Local, global, unique solutions

Theorem 171 (Optimality of convex problems)

Consider a convex optimization problem. Then:

- Any locally optimal point is also globally optimal.
- If the objective function f is strictly convex and there exists a global optimum x^*, then it is unique.

Linear optimization problem

Given $c \in \mathbb{R}^d,\ A \in \mathbb{R}^{m \times d},\ b \in \mathbb{R}^d$, a linear program in standard form looks as follows:

$$\min_{x \in \mathbb{R}^d} c^t x \quad \text{subject to} \quad Ax - b \leq 0$$

- The objective is a linear objective function.
- The constraints are linear constraints (read the inequality component-wise).

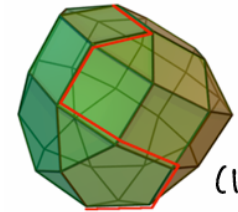
A linear program is a particular case of a convex optimization problem.

Standard algorithm to solve linear programs

The simplex algorithm: it exploits the geometric structure of the domain.

- The domain is a simplex (because of linear constraints).
- The optimum sits in one of the corners.

The simplex algorithm jumps from corner to corner in a clever way.



(Image: Wikipedia)

Quadratic optimization problem

Given $c \in \mathbb{R}^d,\ Q \in \mathbb{R}^{d\times d}$ symmetric, $A \in \mathbb{R}^{m\times d},\ b \in \mathbb{R}^m,\ B \in \mathbb{R}^{k\times d},\ r \in \mathbb{R}^k$, a quadratic program in standard form is given as:

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}x^tQx + c^tx \quad \text{sbuject to} \quad Ax \leq b, \quad Bx = y$$

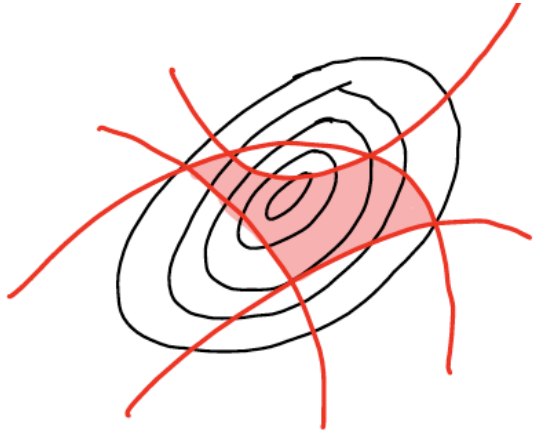
- The objective is a quadratic objective.
- The constraints are linear constraints.

If Q is positive definite, then the problem is convex.

Quadratic optimization with quadratic constraints

⚠ If the constraints are quadratic, the problem is not always convex.

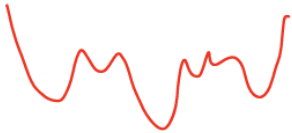
QCQP
quadratically
constraint
quadratic
program



Different kinds of optimization problems

We often distinguish different kinds of optimization problems:

- (discrete or not)
- Differentiable: Do first derivatives exist? (Gradient descent, first-order methods)
- Twice differentiable: Do second derivatives exist? (Hessian, second-order methods)
- Convex / non-convex
- Does the function have just one (global) minimum or do local minima exist?



Equivalent optimization problems

Consider the optimization problem:

$$\min_{x \in \mathbb{R}} f(x) \quad \text{subject to} \quad x \in C$$

Let $h : \mathbb{R} \to \mathbb{R}$ be a monotone function and consider:

$$\min_{x \in \mathbb{R}} h(f(x)) \quad \text{s.t.} \quad x \in C$$

The two are equivalent optimization problems: they have the same solutions.

Example:

- $\sqrt{x}$ is not convex, but
- $(\sqrt{x})^4 = x^2$ is convex.

Change of variables

Consider a bijective mapping $\Phi : \mathbb{R} \to \mathbb{R}$, and assume that its image covers the set $C \subset \mathbb{R}$. Then the two problems

$$\min_x f(x) \quad \text{s.t. } x \in C$$

$$\min_y f\big(\Phi(y)\big) \quad \text{s.t. } \Phi(y) \in C$$

are equivalent (have the same solutions).

Example:

$$\min x + y \quad \text{s.t. } x^2 + y^2 = 1$$

Use polar coordinates: $x = \cos\theta,\ y = \sin\theta.$
Then the problem is

$$\min \cos\theta + \sin\theta, \qquad \text{without any constraint!}$$

(The constraint is automatically satisfied, we got rid of it.)

Equivalent optimization problems

There are many other ways to transform optimization problems:

- Eliminating equality constraints
- Introducing slack variables
- ...

In practice, it can make a big difference!

→ See machine learning lectures for more examples...

First vs. second order methods

First order methods exploit gradient information to find a direction of descent:

- Gradient descent (GD)
- Stochastic gradient descent (SGD)

Second order methods Second order methods additionally exploit the second derivatives (Hessian) to determine the step size:

- Newton
- BFGS

Constraint convex optimization: primal, dual, Lagrangian

Literature: Boyd: Convex optimization

Lagrangian: intuitive point of view

Lagrange multiplier for equality constraints

Consider the following convex optimization problem:

$$\text{minimize } f(x) \quad \text{subject to } g(x) = 0$$

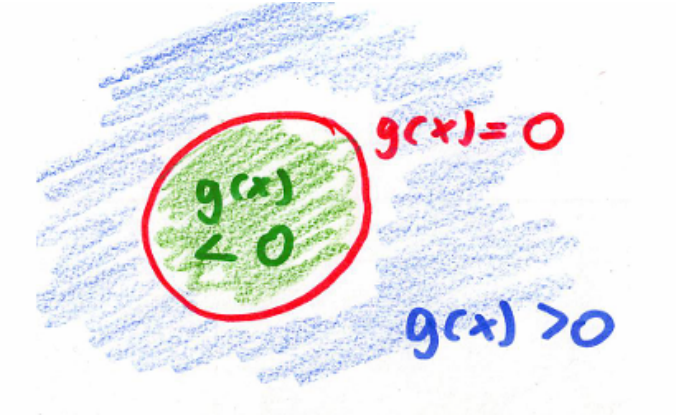
where f and g are convex.

We make several simplifying assumptions for now (e.g., convexity), in order to get an intuition for the Lagrange approach.

Later we can prove everything formally, without many of the assumptions.

Lagrange multiplier for equality constraints (2)

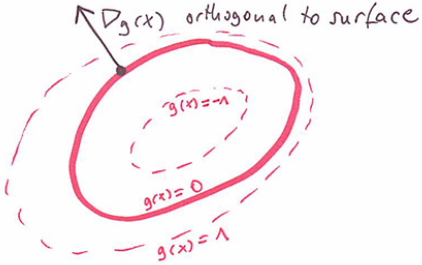
Recall: If g is convex, then its sublevel sets are convex:



Sublevel set: $\{x \mid g(x) \leq c\}$ (the green set in the figure).

Lagrange multiplier for equality constraints (3)

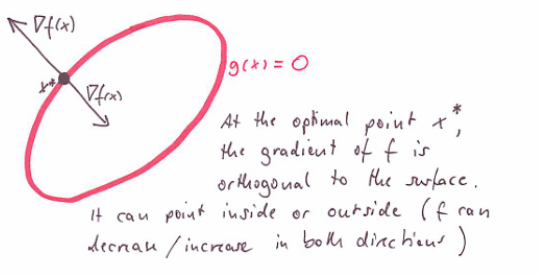
Gradient (equality constraint): For any point x on the "surface" $\{g(x) = 0\}$, the gradient $\nabla g(x)$ is orthogonal to the surface itself.



Intuition: To increase/decrease $g(x)$, you need to move away from the surface, not walk along the surface.

Lagrange multiplier for equality constraints (4)

Gradient (objective function): Consider the point x^* on the surface $\{g(x) = 0\}$ for which $f(x)$ is minimized. This point must have the property that $\nabla f(x)$ is orthogonal to the surface.

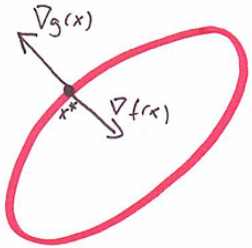


Intuition: Otherwise, we could move a little along the surface to decrease $f(x)$.

Lagrange multiplier for equality constraints (5)

Consequence: At the optimal point, $\nabla g(x)$ and $\nabla f(x)$ are parallel, that is, there exists some $\nu \in \mathbb{R}$ such that:

$$\nabla f(x) + \nu \nabla g(x) = 0$$



At the optimal point x^*, ∇g and ∇f are parallel to each other (and can point either in the same or the opposite direction).

Lagrange multiplier for equality constraints (6)

We now define the Lagrangian function:

$$L(x, \nu) = f(x) + \nu g(x)$$

where $\nu \in \mathbb{R}$ is a new variable called the Lagrange multiplier. Now observe:

- The condition $\nabla f(x) + \nu \nabla g(x) = 0$ is equivalent to $\nabla_x L(x, \nu) = 0$.
- The condition $g(x) = 0$ is equivalent to $\nabla_\nu L(x, \nu) = 0$.

To find an optimal point x^*, we need to find a saddle point of $L(x, \nu)$, that is, a point such that both $\nabla_x L(x, \nu)$ and $\nabla_\nu L(x, \nu)$ vanish.

Simple example

Consider the problem to minimize $f(x)$ subject to $g(x) = 0$, where $f, g : \mathbb{R}^2 \to \mathbb{R}$ are defined as:

$$f(x_1, x_2) = x_1^2 + x_2^2 - 1$$
$$g(x_1, x_2) = x_1 + x_2 - 1$$

Observe: It is hard to solve this problem by naive methods because it is unclear how to take care of the constraints.

Solution by the Lagrange approach:

Write it in the standard form:

$$\text{minimize } x_1^2 + x_2^2 - 1 \quad \text{subject to } x_1 + x_2 - 1 = 0.$$

Simple example (2)

The Lagrangian is:

$$L(x, \nu) = \underbrace{x_1^2 + x_2^2 - 1}_{f(x_1,x_2)} + \nu \underbrace{(x_1 + x_2 - 1)}_{g(x_1,x_2)}.$$

Now compute the derivatives and set them to 0:

$$\nabla_{x_1} L = 2x_1 + \nu \stackrel{!}{=} 0$$
$$\nabla_{x_2} L = 2x_2 + \nu \stackrel{!}{=} 0$$
$$\nabla_{\nu} L = x_1 + x_2 - 1 \stackrel{!}{=} 0$$

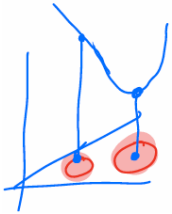
If we solve this linear system of equations, we obtain $(x_1^*, x_2^*) = (0.5, 0.5)$.

Lagrange multiplier for inequality constraints

Consider the following convex optimization problem:

$$\text{minimize } f(x) \quad \text{subject to } g(x) \leq 0,$$

where f and g are convex.



We now distinguish two cases: the constraint is *active* or *inactive*.

Lagrange multiplier for inequality constraints (2)

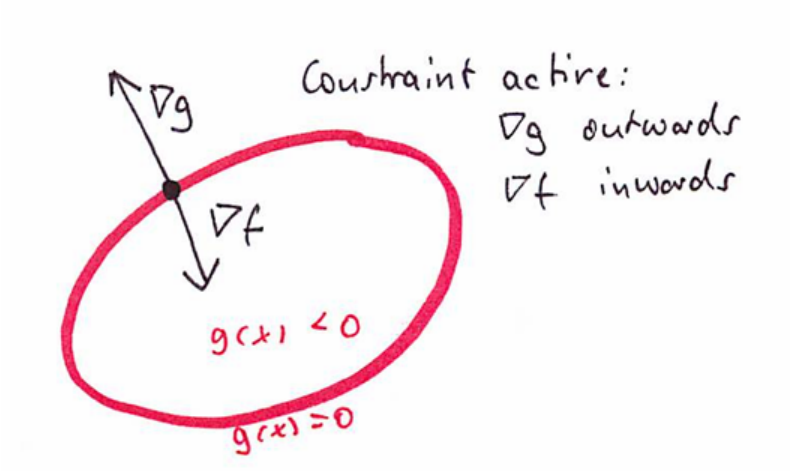
Case 1: Constraint is "active", that is the optimal point is *on* the surface $g(x) = 0$.

Again, ∇f and ∇g are parallel at the optimal point.

But furthermore, the direction of derivatives matters:

- The derivative of g points outwards (at any point on the surface $g = 0$). This is always the case if g is convex.
- Then the derivative of f is directed inwards (otherwise we could decrease the objective by walking inside).

Lagrange multiplier for inequality constraints (3)

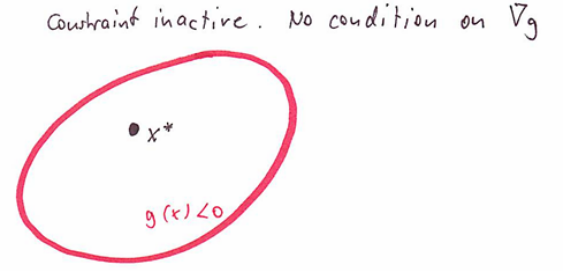


So we have $\nabla f(x) = -\lambda \nabla g(x)$ for some value $\lambda > 0$.

Lagrange multiplier for inequality constraints (4)

Case 2: Constraint is "inactive", that is the optimal point is not on the surface $g(x) = 0$, but somewhere in the interior.

- Then we have $\nabla f = 0$ at the solution (otherwise we could decrease the objective value).
- We do not have any condition on ∇g (it is as if we would not have this constraint).



Lagrange multiplier for inequality constraints (5)

We can summarize both cases using the Lagrangian again. We now define the Lagrangian

$$L(x, \lambda) = f(x) + \lambda g(x)$$

where the Lagrange multiplier has to be positive: $\lambda \geq 0$.

- Case 1: Constraint active, $\lambda > 0$.
  - Need to find a saddle point: $\nabla_x L(x, \lambda) = \nabla_\lambda L(x, \lambda) = 0$.
- Case 2: Constraint inactive, $\lambda = 0$.
  - Then $L(x, \lambda) = f(x)$. Hence $\nabla_x L(x, \lambda) = \nabla_x f(x) \stackrel{!}{=} 0, \ \nabla_\lambda L(x, \lambda) \equiv 0$.

In both cases, we have again a saddle point of the Lagrangian.

Lagrange multiplier for inequality constraints (6)

Also in both cases, we have $\lambda g(x^*) = 0$:

- Constraint active: $\lambda > 0,\ g(x^*) = 0.$
- Constraint inactive: $\lambda = 0,\ g(x^*) \neq 0.$

This is called the Karush-Kuhn-Tucker (KKT) condition.

Simple example

What are the side lengths of a rectangle that maximize its area, under the assumption that its perimeter is at most 1?

We need to solve the following optimization problem:

$$\text{maximize } x \cdot y \quad \text{subject to } 2x + 2y \leq 1.$$

Bring the problem in standard form:

$$\text{minimize } (-x \cdot y) \quad \text{subject to } 2x + 2y - 1 \leq 0.$$

Form the Lagrangian:

$$L(x, y, \lambda) = -xy + \lambda(2x + 2y - 1).$$

Simple example (2)

Saddle point conditions / derivatives:

$$\frac{\partial L}{\partial x} = -y + 2\lambda \stackrel{!}{=} 0,$$
$$\frac{\partial L}{\partial y} = -x + 2\lambda \stackrel{!}{=} 0,$$
$$\frac{\partial L}{\partial \lambda} = 2x + 2y - 1 \stackrel{!}{=} 0.$$

Solving this system of three equations gives $x = y = 0.25$.

Simple example (3)

Now need to see: when does this approach work, when does it not work, what can we prove about it?

Lagrangian: formal point of view

Lagrangian and dual: formal definition

Consider the primal optimization problem:

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0 \ (i = 1, \ldots, m) \\ &\qquad\qquad h_j(x) = 0 \ (j = 1, \ldots, k) \end{aligned}$$

Denote by x^* a solution of the problem and by $p^* := f_0(x^*)$ the objective value at the solution.

Lagrangian and dual: formal definition (2)

Define the corresponding Lagrangian as follows:

- For each equality constraint j, introduce a new variable $\nu_j \in \mathbb{R}$. For each inequality constraint i, introduce a new variable $\lambda_i \geq 0$. These variables are called Lagrange multipliers.
- Then define

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{k} \nu_j h_j(x) \qquad , \lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}, \quad \nu = \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_k \end{pmatrix}.$$

Define the dual function $g : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ by:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu).$$

Dual function as lower bound on primal

Proposition 172 (Dual function is concave)

No matter whether the primal problem is convex or not, the dual function is always concave in (λ, ν).

Proof. For fixed x, $L(x, \lambda, \nu)$ is linear in λ and ν and thus concave. The dual function as a pointwise infimum over concave functions is concave as well.

Note that concave is good, because we are going to maximize this function later on. □

Dual function as lower bound on primal (2)

Proposition 173 (Dual function as lower bound on primal)

For all $\lambda_i \geq 0$ and $\nu_j \in \mathbb{R}$, we have $g(\lambda, \nu) \leq p^*$.

Proof.

- Let x_0 be a feasible point of the primal problem (that is, a point that satisfies all constraints).
- For such a point we have:

$$\sum_{i=1}^{m} \underbrace{\lambda_i}_{\geq 0} \underbrace{f_i(x_0)}_{\leq 0} + \sum_{j=1}^{k} \nu_j \underbrace{h_j(x_0)}_{=0} \leq 0$$

Dual function as lower bound on primal (3)

- This implies:

$$L(x_0, \lambda, \nu) = f_0(x_0) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(x_0) + \sum_{j=1}^{k} \nu_j h_j(x_0)}_{\leq 0} \leq f_0(x_0)$$

  Note that this property holds in particular when $x_0 = x^*$.

- Moreover, for any x_0 (and in particular for $x_0 := x^*$) we have:

$$\underbrace{\inf_x L(x, \lambda, \nu)}_{g(\lambda,\nu)} \leq L(x_0, \lambda, \nu).$$

Dual function as lower bound on primal (4)

- Combining the last two properties gives:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) \leq L(x^*, \lambda, \nu) \leq f_0(x^*).$$

□

Dual optimization problem

Have seen: the dual function provides a lower bound on the primal value. Finding the highest such lower bound is the task of the dual problem.

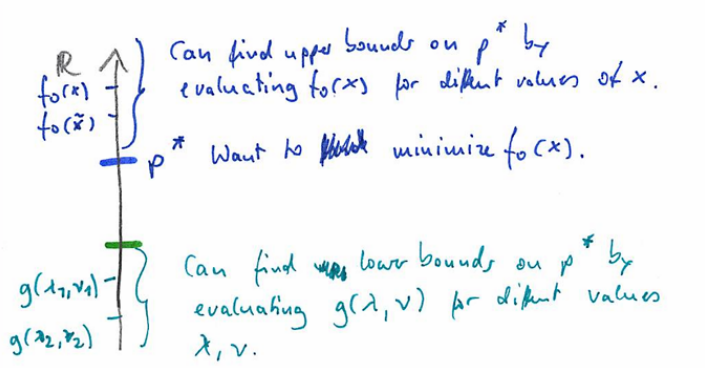
We define the dual optimization problem as:

$$\max_{\lambda,\nu} g(\lambda, \nu) \quad \text{subject to } \lambda_i \geq 0,\ \nu_j \in \mathbb{R}.$$

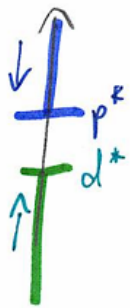
Denote the solution of this problem by λ^*, ν^* and the corresponding objective value $d^* := g(\lambda^*, \nu^*)$.

Dual optimization problem (2)

Dual vs. Primal, some intuition:



Dual optimization problem (3)



Primal problem: find best upper bound on p^*
(= find p^*)

Dual problem: find best lower bound on p^*
(= find d^*)

Weak duality

Proposition 174 (Weak duality)

The solution d^* of the dual problem is always a lower bound for the solution of the primal problem, that is: $d^* \leq p^*$.

Proof. Follows directly from Proposition 2 above. □

We call the difference $p^* - d^*$ the duality gap.

Strong duality

- We say that strong duality holds if $p^* = d^*$.
- This is not always the case, just under particular conditions. Such conditions are called constraint qualifications in the optimization literature.
- Convex optimization problems often satisfy strong duality, but not always.

Strong duality (2)

Examples:

- Linear problems have strong duality.
- Quadratic problems have strong duality.
- There exist many convex problems that do not satisfy strong duality. Here is an example:

$$\begin{aligned} &\text{minimize}_{x,y}\ \exp(-x) \\ &\text{subject to}\ \frac{x}{y} \leq 0 \\ &\qquad\qquad y \geq 0. \end{aligned}$$

One can check that this is a convex problem, yet $p^* = 1$ and $d^* = 0$.

Strong duality: how to convert the solution of the dual to the one of the primal

By strong duality: $p^* = d^*$, that is, we get the same objective values. But how can we recover the primal variables x^* that lead to this solution, if we just know the dual variables λ^*, ν^* of the optimal dual solution?

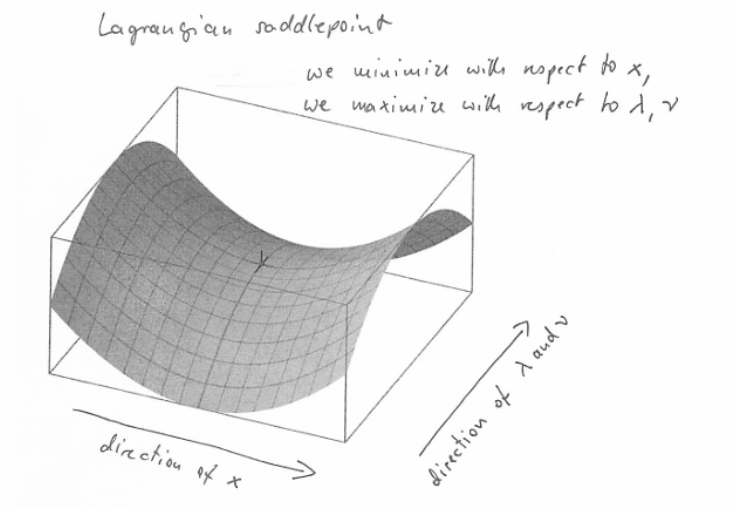
EXERCISE!

Strong duality implies saddle point

Proposition 175 (Strong duality implies saddle point)

Assume strong duality holds, let x^* be the solution of the primal and (λ^*, ν^*) the solution of the dual optimization problem. Then (x^*, λ^*, ν^*) is a saddle point of the Lagrangian.

Strong duality implies saddle point (2)



Strong duality implies saddle point (3)

Proof.

- We first show that x^* is a minimizer of $L(x, \lambda^*, \nu^*)$:
  - By the strong duality assumption we have $f_0(x^*) = g(\lambda^*, \nu^*)$.
  - With this we get:

$$f_0(x^*) = g(\lambda^*, \nu^*) \overset{\text{def}}{=} \inf_x L(x, \lambda^*, \nu^*) \leq L(x^*, \lambda^*, \nu^*) \leq f_0(x^*)$$

    (last inequality follows from Proposition 3)
  - Because we have the same term on the left and right side, we have equality everywhere.
  - So in particular,

$$\inf_x L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$$

Strong duality implies saddle point (4)

- Then we show that (λ^*, ν^*) are maximizers of $L(x^*, \lambda, \nu)$.
  - This follows from the definition of (λ^*, ν^*) as solutions of

$$\max_{\lambda,\nu} \underbrace{\min_{x} L(x, \lambda, \nu)}_{g(\lambda,\nu)}$$

- Taken together we get:

$$L(x^*, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*)$$

That is, (x^*, λ^*, ν^*) is a saddle point of the Lagrangian:

- It is a minimum for x (with fixed λ^*, ν^*).
- It is a maximum for (λ, ν) (with fixed x^*).

Saddle point always implies primal solution

Proposition 176 (Saddle point always implies primal solution)

If (x^*, λ^*, ν^*) is a saddle point of the Lagrangian, then x^* is always a solution of the primal problem.

Proof. Not very difficult, but we skip it. □

Remarks:

- This proposition always holds (not only under strong duality).
- This proposition gives sufficient conditions for optimality. Under additional assumptions (constraint qualifications) it is also a necessary condition.

Why is this whole approach useful?

- Whenever we have a saddle point of the Lagrangian, we have a solution of our constraint optimization problem. This is great, because otherwise we would not know how to solve it.
- If strong duality holds, we even know that any solution must be a saddle point. So if we don't find a saddle point, then we know that no solution exists.
- If your original minimization problem is not convex, at least its dual is a concave maximization problem (or, by changing the sign, a convex minimization problem). If the duality gap is small, then it might make sense to solve the dual instead of the primal (you will not find the optimal solution, but maybe a solution that is close).

Gradient descent

Literature:

- Francis Bach: Learning theory from first principles. Forthcoming book, pdf online.
- Bottou, Curtis, Nocedal: Optimization methods for large scale machine learning, 2018.

Gradient descent (vanilla version)

Assume we want to solve an optimization problem:

$$\min_{w \in \Omega} f(w)$$

where $f : \Omega \to \mathbb{R},\ \Omega \subset \mathbb{R}^d,\ f$ is differentiable.

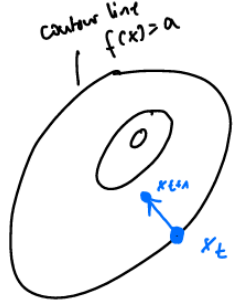
Gradient descent starts with a randomly chosen start point w_0 and then makes a small step in the direction opposite of the gradient $\nabla f(w_t)$, of step size γ_t:

$$\boxed{w_{t+1} = w_t - \gamma_t \nabla f(w_t)}$$

Intuition: Gradient descent and contour lines

Observe:

- The *gradient* of a function is orthogonal to its contour lines.
- Gradient descent steps are orthogonal to contour lines.



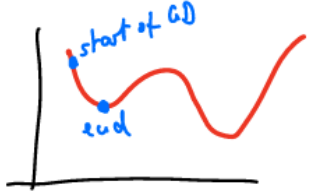
Why does it help if f is differentiable?

- Well, we want to compute gradients...
- If we don't have gradients, our life really becomes hard.
- In ML, we always construct our problems in such a way that the loss f is differentiable in the end. We might sacrifice many things when modeling a problem, but not differentiability.

($\rightarrow$ surrogate loss functions)

Why does convexity help?

- For non-convex f, gradient descent (GD) typically finds local optima, but not the global one.
- If f is convex, we know we found the global optimum.



Why can it help if we have strong convexity?

Functions that are strongly convex cannot have long, "arbitrarily flat" parts. On such parts, GD would take forever. But if f is strongly convex, this does not exist.

Mathematically, for strongly convex f, we can estimate how far we are at most from the optimal point.

In ML, we can sometimes achieve strong convexity through regularization.

Why does smoothness help?

- GD makes steps in the direction of the gradient.
- But if the gradient itself changes wildly, then already after a small step we can be in trouble.
- If f is L-smooth, the gradient only changes slowly, so gradient steps work better.
- The smoother f, the larger steps we can dare to take. For example, in the theorem below we choose a constant step size $\gamma_t = \frac{1}{L}$.

(In practice, one often decreases the step size over time, see below.)

Recap: Condition number

If f is μ-strongly convex and β-smooth, and is twice continuously differentiable, then all eigenvalues of the Hessian are in the interval $[\mu, \beta]$.

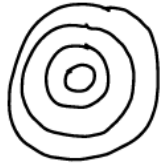
We then denote by $\kappa := \frac{\beta}{\mu}$ the condition number.

(Often it is also defined with the Hessian directly, or as the ratio of the largest to smallest eigenvalue of the Hessian.)

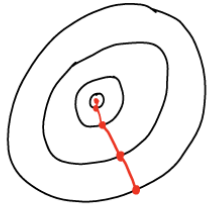
It always holds that $\mu \leq \beta$. Thus $\kappa \geq 1$.

Intuition for condition number

- Condition number small $\Rightarrow \mu \approx \beta$
  - Contour lines of the function are close to a circle.
- Condition number large
  - Contour lines are very "elongated."



Intuition: Speed of convergence



K small, GD converges fast

Steps orthogonal to the contour lines take us pretty straight to the center.

K large, GD converges slowly

Steps orthogonal to the contour lines jump wildly

Choosing a starting point

- Typically, a random point. Parameters often initialized with Gaussian noise of "the correct size."
- Sometimes one uses a warm start: use a first heuristic to guess a good starting point. (Note that this is not what fine tuning is; see next slide.)

Fine tuning

In complex models: Somebody trains, say, an image classifier or a language model on huge amounts of data. We take the trained model and would like to run GD directly on the model – but typically, we don't have access to the original architecture and parameter space.

Then we just use the representation that has been learned by the original model and train a second classifier on top.

Stopping conditions

- Once you observe that the objective doesn't change a lot... a bit unclear in practice.
- In deep learning, people often continue training even though the training error is pretty much 0... representation still might change.

⚠ : As opposed to "traditional" scenarios, we are not interested to minimize the objective f. We want a small test error resp. a "good" representation of the data.

Convergence of GD (smooth, convex)

- smooth: upper bound on curvature
- convex: no lower bound on curvature

Theorem 177 (Convergence of GD (smooth, convex))

Assume that f is L-smooth and convex with a global minimizer η^*. Choosing step size $r_t = 1/L$, the iterates $(\theta_t)_t$ of GD satisfy:

$$f(\theta_t) - f(\eta^*) \leq \frac{1}{t} \cdot \frac{L}{2} \|\theta_0 - \eta^*\|^2$$

With $\frac{1}{t}$: "linear convergence".

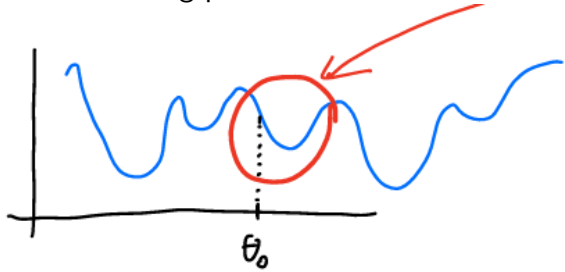
Some more intuition for this theorem

- Convex but not strongly convex. Could have a situation like this, need to make sure that a minimum does exist (assumption on η^*):

- If we would walk from θ_0 to η^* on a direct line with steps of size $\frac{1}{L}$, we would need $\|\theta_0 - \eta^*\|L$ many steps ($\sim$ constant in the bound).

Some more intuition for this theorem (2)

- Note that in ML we often do not have globally convex problems. So this bound might tell us something about how fast we converge to the local optimum in the basis of attraction of our starting point.



No guarantee about convergence to global optimum!!!

Convergence of GD (smooth, strongly convex)

Theorem 178 (Convergence of GD (smooth, strongly convex))

Assume that f is L-smooth and μ-strongly convex. Denote by $\kappa = \frac{L}{\mu}$ the condition number. Let η^* be the global minimizer.
Choosing step size $r_t = \frac{1}{L}$, the iterates $(\theta_t)_t$ of GD on f satisfy:

$$f(\theta_t) - f(\eta^*) \leq \left(1 - \frac{1}{\kappa}\right)^t \cdot (f(\theta_0) - f(\eta^*))$$

$$\leq \exp\left(-\frac{t}{\kappa}\right)(f(\theta_0) - f(\eta^*)) \approx \exp(-t)$$

With $\exp\left(-\frac{t}{K}\right)$: "exponential convergence".

Remarks about this theorem

- κ is always ≥ 1, so $\left(1 - \frac{1}{\kappa}\right) \in [0, 1[$, so $\left(1 - \frac{1}{\kappa}\right)^t \to 0$ as $t \to \infty$.
- Constant $(f(\theta_0) - f(\eta^*))$ on the rhs now measures the "distance" between start and end in the objective f, not the orig space. Can do this because of strong convexity.
- If we don't have restrictions on the domain of f, strong convexity implies the existence of global minimizer η^*.
- Convergence speed is surprisingly fast!

Convex vs strongly convex

Convergence for strongly convex case is *much faster*!

Issues with the step size

- If the step size is too small, convergence of GD can take forever.
- If the step size is too large, GD might never converge because we always miss the optimum.
  ($\rightarrow$ The algorithm could even diverge.)

- In ML, step size is called the learning rate.

Learning rate decay

Unless one does something more clever, one typically uses a learning rate decay, for example:

$$\alpha_t = \alpha_0 \exp(-\kappa \cdot t) \approx \exp(-t) \quad \text{(exponential decay)}$$

$$\alpha_t = \frac{\alpha_0}{(1 + \kappa \cdot t)} \approx \frac{1}{1+t} \quad \text{(inverse decay)}$$

The parameter κ is the decay rate.

Line search to determine step size

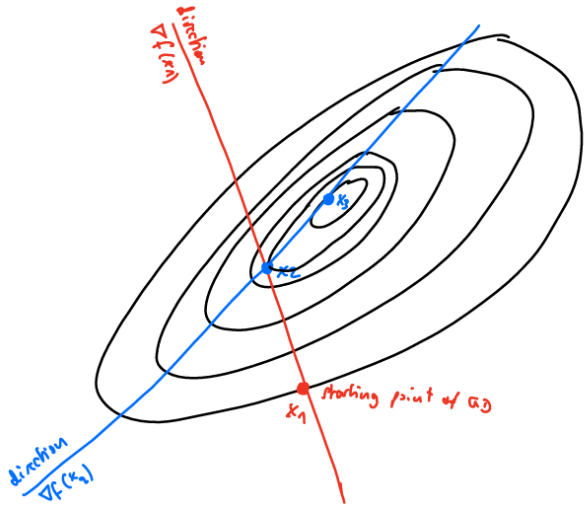
Want to perform a gradient step $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$, where we choose the step size α_t such that we minimize $f(x)$ in this direction:

$$\alpha_t = \operatorname*{argmin}_{\alpha} f(x_t - \alpha \nabla f(x_t))$$

Can for example use binary search to approximate α_t.

Computationally expensive, not so often used in ML.

Line search intuition



search along the whole line in direction of $\nabla f(x_0)$ for the point on which obj. fct f is minimized

Then again search along line in direction of $\nabla f(x_2)$ for point on which obj. fct f is minimized

etc

Using momentum: idea

Consider a situation where we zig-zag slowly towards our destination:

Idea: Let the descent direction "inherit" some part from the previous one, such that we get more of an "average" feeling:

- Traditional:

$$w_{t+1} = w_t - \alpha_t \cdot \nabla f(w_t)$$

- Momentum:

$$w_{t+1} = w_t - [\beta \cdot \nabla f(w_{t-1}) + \alpha \cdot \nabla f(w_t)]$$

  - β: momentum (friction)
  - $\nabla f(w_{t-1})$: previous gradient
  - $\nabla f(w_t)$: current gradient

Using momentum: idea (2)

Properties:

- It avoids the zig-zag.
- It gains speed in steep areas that might help to travel faster over flat parts.
- Might "overshoot" at the solution.

"Like a marble that races on the loss surface."

Vanishing gradient problem

- Occurs if the gradients corresponding to some parameters get so small that they barely change.
- Particularly the case in deep networks, where the products of the network parameters get multiplied during backpropagation. Partial derivatives of parameters in the first layers can get very small.
- Particularly for loss functions with e.g., tanh, (not for ReLU).
- Lots of suggestions in DNN literature, e.g., batch normalization.

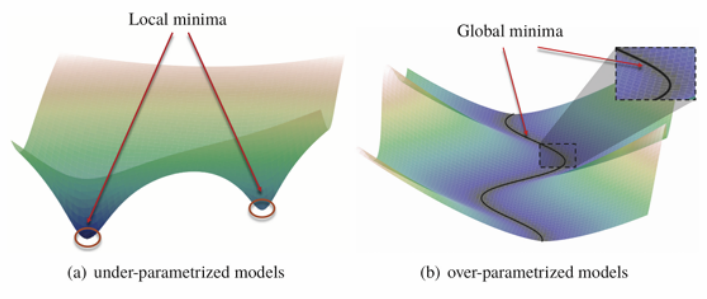
Exploding gradient problem

- "Opposite" of vanishing gradients: in early layers, gradients get too large $\rightarrow$ uncontrollable behavior.
- Reason either if weights are too large (initialization!) or gradients too large ($|\nabla w_i| > 1$) $\rightarrow$ gradient clipping.

- Lots of solutions discussed in the DNN literature (e.g., skip connections).

Just one little teaser: loss landscape in deep learning

- The "loss landscape" in deep learning (when we try to optimize parameters of a neural network) is highly non-convex.
- However, it seems that "many" of the local optima one finds are already global optima!



→ Belkin: Fit without fear: remarkable properties of deeplearning, 2021.

Stochastic gradient descent

Literature:

- Francis Bach: Learning theory from first principles. Forthcoming book, pdf online.
- Bottou, Curtis, Nocedal: Optimization methods for large scale machine learning. 2018.

Motivation

In ML we typically minimize the training loss of the form

$$g(w) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_w(x_i), y_i)$$

where $(x_i, y_i)_{i=1,\dots,n}$ are our training points and $\mathcal{L}$ is a loss (for example the squared loss, the logistic loss), and w are the parameters we optimize.

The gradient $\nabla g(w)$ is

$$\nabla g(w) = \frac{1}{n} \underbrace{\sum_{i=1}^{n}}_{\text{n terms}} \nabla \mathcal{L}(f_w(x_i), y_i).$$

Idea: "sample" the gradient

Consider the gradient

$$\nabla g(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(f_w(x_i), r_i).$$

If n is large:

- It is really costly to compute it.
- The data set might have redundancies, so we see similar info often.
- From a statistical point of view, whenever you see a large sum of random quantities, you guess that:
  - The sum is close to its expected value.
  - This might still be true if you subsample terms.

Also, because we are interested in the test error in the end, a lot of stochasticity might, perhaps, prevent some overfitting.

Stochastic gradient descent (vanilla)

Given training points $(x_i, y_i)_{i=1,\dots,n}$ and training loss

$$g(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_w(x_i), y_i),$$

with gradient

$$\nabla g(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i).$$

In each step of the algorithm:

- Sample one training point (x_{i0}, y_{i0}) and compute the simplified gradient $\nabla \ell(f_{w_t}(x_{i0}), y_{i0}) =: \nabla \ell_i(w_t)$.
- Make gradient step:

$$w_{t+1} = w_t - \alpha_t \cdot \nabla \ell_i(w_t)$$

Until convergence.

Stochastic gradient descent (mini-batch)

Instead of subsampling one point at each time, one typically subsamples a "mini-batch" consisting of a number k of data points.

Then apply the same principle:

- Sample $(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$.
- Simplified gradient $\nabla \sum_{s=1}^{k} \ell_{i_s}(w_t) = \nabla g_k(w_t)$.
- Gradient step

$$w_{t+1} = w_t - \alpha_t g_k(w_t).$$

To implement this, choose a random permutation of the data set. Then pick one "batch" after the other, until one has made one full pass through the data.

Typically, one uses several passes through the data set.

Stochastic gradient descent (mini-batch) (2)

Note that both variants (vanilla and mini-batch) use the same principle: They replace the full gradient by a random estimate of the gradient.

- Introduces noise.
- Larger batches $\Rightarrow$ less noise, smaller variance of the estimates.

Stochastic gradient descent (general view)

More generally, one can consider any estimate of the gradient $\hat{g}$ and use it in gradient descent:

$$w_{t+1} = w_t - \alpha_t \hat{g}(w_t).$$

Typically, one would require that this estimate is unbiased $\rightarrow$ "on average correct" (see later in the statistics part of the lecture).

First remarks

- As the name suggests, SGD is not a deterministic algorithm: noisy updates!
- Some of the steps might even make a "wrong" step such that f is increasing. We hope though that "on average" it will be fine.
- We save compute and storage, compared to the standard GD.
- All guarantees for SGD would need to be statements "in expectation" or "with high probability."
- Subsampling is an unbiased estimate of the gradient, but the random errors do not decrease as we go along!
  Instead, we will use a decreasing step size, which will make the errors smaller and smaller.

Convergence of SGD (convex)

Theorem 179 (Convergence of SGD (convex))

Assume that F is convex, β-Lipschitz, and has a minimizer $\Theta_\star$ that satisfies $\|\Theta_\star - \Theta_0\|^2 \leq D$. Assume that the gradients of the SGD iterates are all bounded by a constant B, that is $\|g_t(\Theta_{t-1})\|^2 \leq B^2$ for all $t > 1$, and that the estimate $\hat{g}$ of the gradient is unbiased. Choose the step size $\alpha_t = \frac{D}{B} \cdot \frac{1}{\sqrt{t}}$.

Then the iterates of SGD on F satisfy

$$\underbrace{\mathbb{E}(F(\bar{\Theta}_t)}_{\text{"average iterate"}} - F(\Theta_\star)) \leq \frac{DB^2}{2} \cdot \frac{2 + \log(t)}{\sqrt{t}}, \qquad \text{slower than for GD}$$

where

$$\bar{\Theta}_t = \frac{\sum_{s=1}^{t} \gamma_s \Theta_{s-1}}{\sum_{s=1}^{t} \gamma_s}.$$

Digesting this theorem

- No assumptions about strong convexity (could have really flat parts), no assumptions on smoothness. Instead, only that F is Lipschitz.
- Step size is decreasing (otherwise SGD would not converge because variance does not decrease).
- This bound is expressed in terms of the "average iterate," which is a form of stabilizing the result.

Convergence of SGD (strongly convex)

Consider a regularized problem of the form

$$G(\Theta) = F(\Theta) + \frac{\mu}{2}\|\Theta\|^2.$$

Theorem 180 (Convergence of SGD (strongly convex))

Assume F is convex, β-Lipschitz, μ-strongly convex. Consider the regularized problem G and assume that it admits a unique minimizer $\Theta_\star$. Then, under the same assumptions as before (unbiased, bounded $\|g_t(\Theta)\|^2$) and choosing the step size $\gamma_t = \frac{1}{\mu t}$ (steps decrease faster),

$$\mathbb{E}\left(G(\bar{\Theta}_t) - G(\Theta_\star)\right) \leq \frac{2B^2(1+\log t)}{\mu \cdot t} \qquad (\rightarrow \text{faster converence})$$

Remarks

- Smoothness does not help to improve a lot in the SGD case.
- SGD converges slower than GD in terms of accuracy, but needs much less computation (in terms of n, the number of training points). So if the computational budget is limited, it is the method of choice.
- More refined bounds are needed to understand the behavior of different SGD versions (e.g., sampling one gradient vs using mini-batches).

Comparing the bounds

Resulting error $|F(\Theta) - F(\Theta_\star)|$ as a function of the number t of steps performed:

	Convex	Strongly Convex
Nonsmooth	Deterministic: $1/\sqrt{t}$ Stochastic: $1/\sqrt{t}$	Deterministic: $B^2/(\mu t)$ Stochastic: $B^2/(\mu t)$
Smooth	Deterministic: $1/t^2$ Stochastic: $1/\sqrt{t}$ Finite sum: n/t	Deterministic: $\exp(-t\sqrt{\mu/L})$ Stochastic: $L/(\mu t)$ Finite sum: $\exp(-\min\{1/n, \mu/L\}t)$

Final Remarks on SGD

To run SGD in practice, people use millions of tricks, and the choice of parameters and their scaling as learning proceeds makes a lot of difference!

Studying the behavior of SGD for all these scaling laws is challenging in practice and in theory!

... beyond this lecture ...

Why is SGD so popular in ML?

- Large scale problems $\Rightarrow$ computational issues
- Redundant data
- "Stochasticity" introduced by SGD acts as a regularizer: in the end we don't care about the training error (so it does not matter if we don't optimize it perfectly), but we want to have a small test error. The SGD noise might help to prevent overfitting.

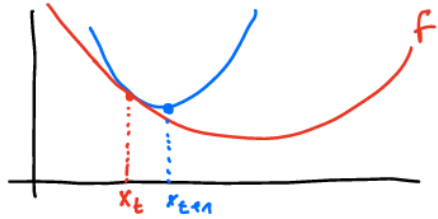
Second order methods Newton

Intuition (1-dim)

Gradient descent only considers the first derivative of the function.

Newton's method also looks at the second derivative and exploits it to choose a good step size:

Intuitively: fit not just a line but a parabola to the function at the current data point (given slope and curvature). Then proceed to the point where this parabola is minimal.



Newton method (1-dim)

Consider the Taylor expansion of the objective f to the second order:

$$f(w_t + \varepsilon) \approx f(w_t) + \varepsilon f'(w_t) + \frac{1}{2}\varepsilon^2 f''(w_t).$$

Now search for ε that minimizes $f(w_t + \varepsilon)$:

$$\frac{d}{d\varepsilon}\left(f(w_t + \varepsilon)\right) = \cdots = f'(w_t) + \varepsilon f''(w_t) = 0.$$

$$\Rightarrow \varepsilon = -\frac{f'(w_t)}{f''(w_t)}.$$

Set update rule:

$$w_{t+1} = w_t - \frac{f'(w_t)}{f''(w_t)}.$$

Newton method (d-dim)

Can derive a similar argument in the d-dim case, resulting in the update:

$$w_{t+1} = w_t - H^{-1} \left(\nabla f(w_t) \right),$$

where H is the Hessian and $\nabla f(w_t)$ is the gradient.

- Particularly useful on convex fs with positive definite Hessians.
- Convergence on smooth functions might be faster than for standard GD.
- Trouble if Hessians are indefinite (saddle points) or not even invertible.

Computational costs

- Computationally costly (Hessian and its inverse!).
- Approximation algorithms to the inverse of the Hessian exist:
  - Conjugate gradient.
  - Quasi-Newton methods.
  - BFGS algorithm.

Part IV
Probability Theory

- Recommended text book: Jacod/Protter: Probability essentials.
- If you want to see all the math behind it, I recommend Shiryaev: Probability.

σ-Algebra

Motivation

Given some set X, we want to "measure" the size of sets:

- Define notion of a "volume," e.g., to define the Lebesgue integral.
- Define notion of a "probability" of a set.

Ideally, we would like to assign a "measure" to each subset of X. Thus the measure might be a mapping

$$m : \mathcal{P}(X) \to \mathbb{R}^+,$$

where $\mathcal{P}(X)$ is the power set (all subsets).

However, for many domains, including $X = \mathbb{R}$, this leads to mathematical problems.

We need to restrict the set of "measurable" subsets.

σ-Algebra

Definition 181 (σ-Algebra)

Let X be a non-empty set. A σ-algebra on X is a non-empty collection $\mathcal{F}$ of subsets of X such that:

(i) $\mathcal{F}$ is closed under taking complements:

$$A \in \mathcal{F} \implies X \setminus A \in \mathcal{F}.$$

(ii) $\mathcal{A}$ is closed under countable unions:

$$A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

(iii) $X \in \mathcal{F}$.

Examples

- Trivial σ-algebras: Given X, we can define two (pretty useless) σ-algebras:
  - $\mathcal{F}_1 = \{\emptyset, X\}$,
  - $\mathcal{F}_2 = \mathcal{P}(X) = \{A \mid A \subseteq X\} \qquad (A \subset B :\iff a \in A \implies a \in B)$.

- Given X, let $\mathcal{G}$ be any collection of subsets of X. The σ-algebra generated by $\mathcal{G}$ is the smallest σ-algebra $\mathcal{A}$ such that $\mathcal{G} \subseteq \mathcal{A}$.
  Notation: $\sigma(\mathcal{G})$.

  Remark: Existence can be proved easily by an explicit construction:

$$\sigma(\mathcal{G}) = \bigcap \{\Sigma \mid \Sigma \text{ is a } \sigma\text{-algebra that contains } \mathcal{G}\}.$$

Borel σ-algebra

Consider a metric space (X, d) and let $\mathcal{G}$ be the collection of open subsets of X. Then the Borel σ-algebra is defined as $\sigma(\mathcal{G})$.

Measures

Measurable space

Definition 182 (Measurable space)

A measurable space consists of a set X and a σ-algebra $\mathcal{F}$ over X.

Notation: $(X, \mathcal{F})$. The sets in $\mathcal{F}$ are called measurable.

Measure

Definition 183 (Measure)

Given a measurable space $(X, \mathcal{F})$, a measure is a map $\mu : \mathcal{F} \to [0, \infty]$ such that:

(i) $\mu(\emptyset) = 0$.

(ii) For any countable collection of disjoint subsets $(S_i)_{i \in \mathbb{N}}$ with $S_i \in \mathcal{F}$, we have

$$\mu\left(\bigcup_{i \in \mathbb{N}} S_i\right) = \sum_{i \in \mathbb{N}} \mu(S_i).$$

⚠ A measure is a function on $\mathcal{F}$, not on X!

Measure space

Definition 184 (Measure space)

A measurable space $(X, \mathcal{F})$ endowed with a measure μ is called a measure space $(X, \mathcal{F}, \mu)$.

Example: Discrete measure

Given a finite (or countable) space

$$X = \{x_1, x_2, x_3, \ldots\},$$

define the σ-algebra $\mathcal{F}$ to be the set of all subsets of X. Consider a sequence $(m_i)_{i\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^{\infty} m_i$ is finite.

(Example: $m_i = \frac{1}{i^2}$)

To define $\mu : \mathcal{F} \to \mathbb{R}$, proceed as follows:

- Define μ on all "elementary" sets:

$$\mu(\{x_i\}) := m_i.$$

- For all other sets $A \in \mathcal{F}$, deduce the measure (due to countability) by:

$$\mu(A) = \sum_{x_i \in A} \mu(\{x_i\}) = \sum_{x_i \in A} m_i.$$

Lebesgue measure on $\mathbb{R}$

Consider the space $\mathbb{R}$ with its Borel-σ-algebra: the σ-algebra $\mathcal{B}$ induced by the open sets $]a, b[\subset \mathbb{R}, a < b$.

To each such interval, assign the volume:

$$\text{vol}(]a, b[) = b - a.$$

One can prove that this notion of volume can be extended to the whole σ-algebra $\mathcal{B}$. The resulting measure on $(\mathbb{R}, \mathcal{B})$ is called the Lebesgue measure and is often denoted with the letter λ.

Lebesgue measure on $\mathbb{R}^d$

Similarly to the 1-dim case, one can also construct a measure on $\mathbb{R}^d$ with the Borel-σ-algebra $\mathcal{B}$:

- Consider sets of the form $]a_1, b_1[\quad \times \quad]a_2, b_2[\times \ldots \times]a_d, b_d[$, and assign them the volume:

$$(b_1 - a_1) \cdot \ldots \cdot (b_d - a_d).$$

One can extend this consistently to the whole σ-algebra:

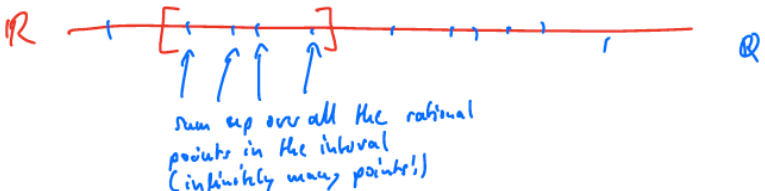
$$\Rightarrow \text{multi-dim Lebesgue measure } \lambda_d.$$

A funny measure on $\mathbb{R}$

Let $\mathcal{A}$ be the Borel-σ-algebra on $\mathbb{R}$. We want to define a measure that just assigns mass to rational numbers. Let $q_1, q_2, q_3, \ldots$ be all rational numbers ($\mathbb{Q}$, countably many). Consider $(m_i)_{i\in\mathbb{N}}$ as before: $m_i = \frac{1}{i^2}$.

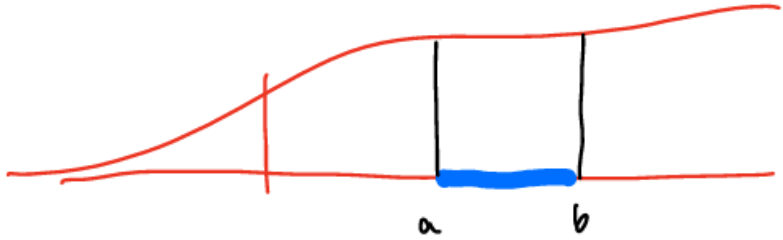
- Define: $\mu(\{q_i\}) = m_i$.
- For intervals $[a, b]$, define:

$$\mu([a,b]) = \sum_{q_i \in [a,b]} \mu(\{q_i\}) = \sum_{q_i \in [a,b]} m_i.$$



More general measures on $\mathbb{R}$

Let $X = \mathbb{R}$, $\mathcal{F} =$ Borel-σ-algebra. Let $F : \mathbb{R} \to \mathbb{R}$ be monotonically increasing and continuous.



For an interval $]a, b]$, define its measure as:

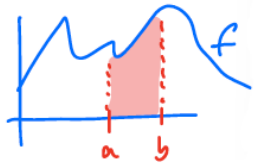
$$\mu(]a, b]) = F(b) - F(a).$$

One can prove that this "measure" can be extended to the whole σ-algebra, making it a well-defined measure on $\mathbb{R}$.

Measure with a density

Consider a function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that is integrable. For intervals $[a, b]$, define:

$$\nu([a, b]) := \int_a^b f(x)\, dx.$$



This can be extended to a proper measure on $(\mathbb{R}, \mathcal{B})$.

Then ν is the measure with density f with respect to the Lebesgue measure.

More keywords: Radon–Nikodym Theorem.

Carathéodory extension theorem

We have seen several instances where we defined a "measure" on intervals $[a, b]$ or $[a, b[$ or $]a, b]$ and concluded that it "extends" to the whole σ-algebra.

The mathematical basis for this approach is the Carathéodory theorem.

(Skipped here.)

Measures with continuous and discrete parts

Observe: Measures can have "continuous and discrete" parts.

Example:

If λ is the Lebesgue measure on $[0,1]$ with Borel-σ-algebra, let δ_0 be the discrete measure that assigns mass 1 to the set $\{0\}$.

Then we can define a measure $\nu = \lambda + \delta_0$ (whereδ_0 ist the dirac measure):

$$A \subset [0,1]; \qquad \nu(A) = \lambda(A) + \delta_0(A) = \begin{cases} \lambda(A), & \text{if } 0 \notin A, \\ \lambda(A) + 1, & \text{if } 0 \in A. \end{cases}$$

More keywords: Lebesgue decomposition theorem of measures.

Null sets

Consider a measure space $(\mathcal{X}, \mathcal{F}, \mu)$.

A subset $N \in \mathcal{A}$ is called a null set if $\mu(N) = 0$.

We say that a property holds almost everywhere if it holds for all $x \in \mathcal{X}$ except for x in a null set N.

(In probability theory, we say "almost surely.")

Measurable mappings

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and $(\mathcal{W}, \mathcal{A})$ be a measurable space.

A mapping $f : \mathcal{X} \to \mathcal{W}$ is called measurable if:

$$\forall A \in \mathcal{A}, f^{-1}(A) \in \mathcal{F} \quad (\text{"pre-images of measurable sets are measurable"}).$$

The mapping f induces a measure ν on $(\mathcal{W}, \mathcal{A})$ via:

$$\nu(A) = \mu(f^{-1}(A)).$$

TODO

TODO, add slides for:

- integrating wrt measure $\int f d\mu$
- trafo formula

Probability measure

Probability measure

- $\mathcal{A} = \{s_1, s_2, \dots\} \subset \Omega$
- $\Omega \in \mathcal{A}$
- $P : \mathcal{A} \to \mathbb{R}$

A measure P on a measurable space $(\Omega, \mathcal{A})$ is called a probability measure if $P(\Omega) = 1$.

The elements in $\mathcal{A}$ are called events.

$(\Omega, \mathcal{A}, P)$ is called a probability space.

Example: Throw a die

- Let $\Omega = \{1, 2, \ldots, 6\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ (the σ-algebra generated by the "elementary events" $\{1\}, \{2\}, \ldots, \{6\}$).
- P can be defined uniquely by assigning:

$$P(\{1\}) = P(\{2\}) = \ldots = P(\{6\}) = \frac{1}{6}.$$

- For example:

$$P(\{1, 5\}) = P(\{1\}) + P(\{5\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Example: Throw a die (2)

- Throw two dice:

$$\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} = \{(1,1), (1,2), (1,3), ..\}$$

where the first coordinate represents the first die and the second coordinate represents the second die.

$\mathcal{A} = \mathcal{P}(\Omega)$, and:

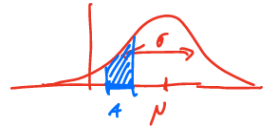
$$P(\{(i,j)\}) = \frac{1}{36}.$$

Example: Normal distribution

$\Omega = \mathbb{R}$, $\mathcal{A} =$ Borel-σ-algebra.

Let $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}_+$ be defined as:

$$x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



Define $P : \mathcal{A} \to [0,1]$ by:

$$P(A) := \int_A f_{\mu,\sigma}(x)\, dx.$$

P is the probability measure on $(\mathbb{R}, \mathcal{A})$ with density $f_{\mu,\sigma}$.

More examples

... See later ...

Cumulative distribution function (cdf)

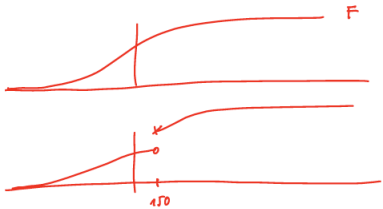
Cumulative cistribution function (cdf)

Definition 185 (Cumulative cistribution function (cdf))

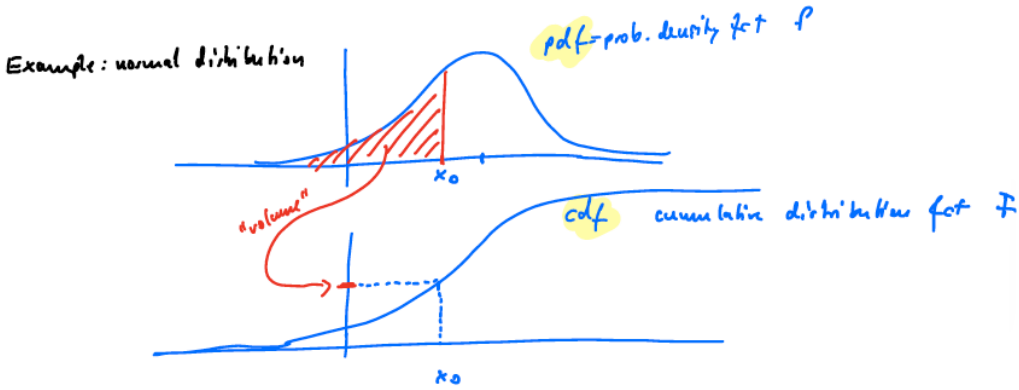
Let P be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define the function $F : \mathbb{R} \to \mathbb{R}$ by:

$$F(x) \mapsto P(]-\infty, x]).$$

We say that F is a cumulative distribution function (cdf).



cdf vs. pdf


Example: normal distribution
pdf = prob. density fct f
x_0
"volume"
cdf cumulative distribution fct F
x_0

Properties of cdf

A cdf satisfies the following properties:

(i) F is monotonically increasing with:

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F(x) = 1.$$

(ii) F is continuous from the right: For any sequence $(x_n)_n$ with $x_n \searrow x$ (i.e. $x_n \geq x_{n+1}$ and $x_n \to x$), then also $F(x_n) \to F(x)$.

The other way round? Is it true that a function F satisfying (i) and (ii) is always the cdf of a probability measure? $\to$Yes.

"Cdf" gives rise to a unique probability measure

Proposition 186 (Measure from Distribution Function)

Let $F : \mathbb{R} \to \mathbb{R}$ be a function with properties (i) and (ii). Then there exists a unique probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that:

$$P(]-\infty, x]) := F(x).$$

Random variable

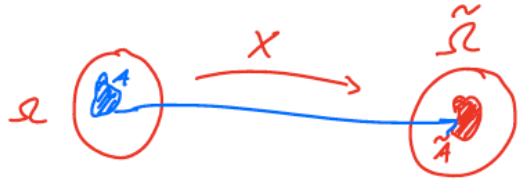
Random variable

Definition 187 (Random variable)

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ be another measurable space. A mapping $X : \Omega \to \tilde{\Omega}$ is called a random variable if X is measurable, i.e.,

$$\forall \tilde{A} \in \tilde{\mathcal{A}} : X^{-1}(\tilde{A}) := \{\omega \in \Omega \mid X(\omega) \in \tilde{A}\} \in \mathcal{A}.$$

Random variable (2)



people in the world

height

$\tilde{A}$ = "at least 170"

A = "person that are at least 170 cm tall"

$P(A) = 0.1$

Example

Example: sum of two dice

- Let:
  - $\Omega = \{(i,j) \mid i,j \in \{1,\dots,6\}\}$, $\quad \tilde{\Omega} = \{2,\dots,12\}$
  - $\mathcal{A} = \mathcal{P}(\Omega)$, $\quad \tilde{\mathcal{A}} = \mathcal{P}(\tilde{\Omega})$
  - $P(\{(i,j)\}) = \frac{1}{36}$
- Define X: "sum of the two values":

$$X : \Omega \to \underbrace{\{2,\dots,12\}}_{\tilde{\Omega}}, \quad (i,j) \mapsto i+j.$$

X is measurable.

$$P(\text{sum} = 12) = P(\{(i,j) \mid i+j = 12\}) = P(\{(6,6)\}) = \frac{1}{36}.$$

Distribution of a random variable

Definition 188 (Distribution of a random variable)

A random variable $X : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$ induces a measure on the target space:

$$\forall \tilde{A} \in \tilde{\mathcal{A}}, \quad P_X(\tilde{A}) := P(X^{-1}(\tilde{A})).$$

This is a probability measure on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ and is called the distribution of X.

Induced σ-algebra

$$(\Omega, \mathcal{A}, P), \quad (\tilde{\Omega}, \tilde{\mathcal{A}}), \quad X : \Omega \to \tilde{S}$$

Definition 189 (Induced σ-algebra)

Let $X : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$. Then the family:

$$\sigma(X) := \{X^{-1}(\tilde{A}) \mid \tilde{A} \in \tilde{\mathcal{A}}\}$$

is a σ-algebra on Ω, and it is called the σ-algebra induced by X.

(It is the smallest σ-algebra on Ω that makes X measurable.)

$$X^{-1}(\tilde{A}) = \{\omega \in \Omega \mid X(\omega) \in \tilde{A}\}.$$

Expectation

Expectation (finite case)

Consider a finite random variable $X : \Omega \to \mathbb{R}$ (that is, $X(\Omega)$ is finite).

Definition 190 (Expectation (finite case))

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $S \subset \mathbb{R}$ finite, and $X : \Omega \to S$ a random variable. Then

$$\mathbb{E}(X) := \sum_{r \in S} r \cdot P(X = r)$$

is called the expectation of X.

(Sometimes written as EX, $\mathbb{E}X$, or $E(X)$).

A random variable is called "centered" if $\mathbb{E}(X) = 0$.

Examples

- Toss a coin:
  - $\Omega = \{\text{head}, \text{tail}\}, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad P(\text{head}) = p, \quad P(\text{tail}) = 1 - p$, where $0 \leq p \leq 1$.
  - $X : \Omega \to \{0, 1\}, \quad \text{head} \mapsto 1, \quad \text{tail} \mapsto 0.$
  - Then:

$$\mathbb{E}(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot (1 - p) + p = p.$$

In ML:

- Train error: Expected error w.r.t. empirical distribution, which assigns probability $\frac{1}{n}$ to all n training points.
- Test error of a classifier: Expected error w.r.t. the (unknown) underlying data distribution.

Expectation is linear

Proposition 191 (Expectation is linear)

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be two random variables. Then:

$$\mathbb{E}(aX + bY) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y), \quad \text{for } a, b \in \mathbb{R}.$$

Expectation is linear (2)

Proof. (discrete case)

$$\mathbb{E}(aX + bY) = \sum_{x,y} (ax + by) \cdot P(X = x, Y = y).$$

$$= a \sum_{x,y} x \cdot P(X = x, Y = y) + b \sum_{x,y} y \cdot P(X = x, Y = y).$$

Using the law of total probability:

$$= a \sum_{x} x \cdot \underbrace{\sum_{y} P(X = x, Y = y)}_{P(X=x)} + b \sum_{y} y \cdot \underbrace{\sum_{x} P(X = x, Y = y)}_{P(Y=y)}.$$

$$= a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y).$$

□

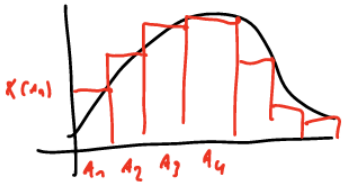
Expectation of a general random variable

Motivation

Let $(\Omega, \mathcal{A}, P)$ be a measure space, and $X : \Omega \to \mathbb{R}$ a random variable. We want to define $\mathbb{E}(X) := "\int X \, dP"$.

If P has a probability density ρ and $X : \mathbb{R} \to \mathbb{R}$, we might do something similar to the Riemann integral:

$$\mathbb{E}(X) \approx \sum_i X(A_i) \cdot P(A_i) \xrightarrow{\lim} \int x \, p(x) \, dx$$



Motivation (2)

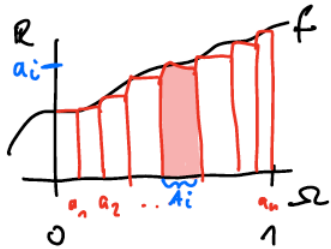
Consider a probability space $(\Omega, \mathcal{A}, P)$ and a measurable function $f : X \to \mathbb{R}$. We want to compute "$\int f(x)\, dP$", the integral of the function w.r.t. the probability measure P.

Motivation (3)

(1): If $\Omega = [0, 1]$ with the uniform distribution, this is the same as computing the "normal integral". Intuitively, partition the space $\Omega = \mathbb{R}$ into sets $A_1, A_2, \ldots, A_n$ and compute the approximate integral:

$$\sum_{i=1}^{n} f(a_i) \cdot \underline{\text{length}}(A_i).$$

Then, as $n \to \infty$, this (hopefully) converges to: $\int_0^1 f(x)\, dx.$

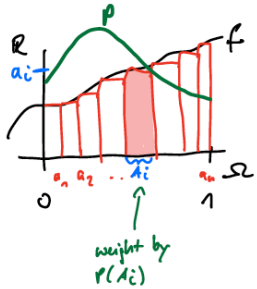


Motivation (4)

(2): Now consider Ω with a probability distribution with density ρ. We now want to pursue a similar approach. Instead of using the length of A_i as volume, use the probability of A_i as notion of volume:

$$\sum_{i=1}^{n} f(a_i) \cdot P(A_i).$$

Letting $n \to \infty$, this (hopefully) converges to:



$$\int_0^1 f(x) \cdot p(x)\, dx.$$

with $p(x)=$ weight by the probability density

Motivation (5)

(3): But what do we do if we don't have a density p? Or a more general input space Ω?

We cannot describe the "weight" by a function p, so we need a more general construction.

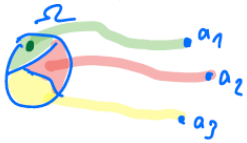
As it will turn out, we'll turn this process upside down:

Instead of starting with a partition of the input space Ω and refining it (which might be difficult because Ω does not have any "structure" beyond a σ-algebra), we use a partition of the target space $\mathbb{R}$:

Construction of an integral: Simple random variable

Step 1: Assume the random variable $X : \Omega \to \mathbb{R}$ only takes finitely many values: $X \in \{a_1, \ldots, a_k\}$. That is, there exist some sets $A_1, \ldots, A_k$ such that:

$$X(\omega) = \sum_{i=1}^{k} a_i \cdot \mathbb{1}_{A_i}(\omega),$$



where

$$\mathbb{1}_{A_i}(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i, \\ 0 & \text{otherwise.} \end{cases}$$

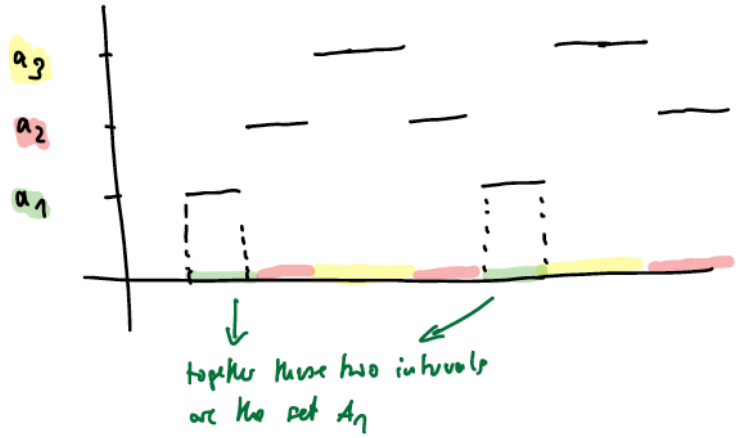
Construction of an integral: Simple random variable (2)

Note that because X is a random variable, the sets A_i need to be in $\mathcal{A}$. We call such a random variable "simple". We now define:

$$\mathbb{E}(X) := \sum_{i=1}^{k} a_i \cdot P(A_i).$$

(Note: This coincides with the previous definition in the discrete case.)

Construction of an integral: Simple random variable (3)



Construction of an integral: Non-negative random variable

Step 2: Assume the random variable X takes values in $[0, \infty]$. We define:

$$''\mathbb{E}(X)'' := \sup\{\mathbb{E}(Y) \mid Y \text{ simple r.v. with } 0 \leq Y \leq X\}.$$

If the supremum exists and is finite, we call the resulting number the expectation $\mathbb{E}(X)$ of the non-negative random variable X.

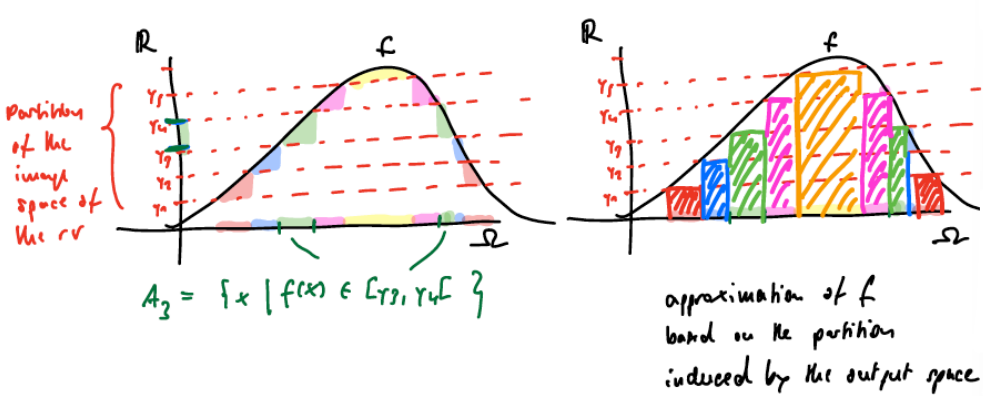
Construction of an integral: Non-negative random variable (2)

Intuitively:

- We "discretize" the output of the random variable into finitely many values.
- We look at the corresponding partitions $A_1, \ldots, A_k$ of the input space Ω.
- We use these partitions to "approximate" X from below.
- By taking the supremum over all such partitions, we implicitly use finer and finer partitions (without explicitly defining what "refinements" of partitions are).

Construction of an integral: Non-negative random variable (3)

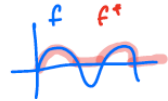
Illustration of Step 2: Consider the case of a random variable $f : \mathbb{R} \to \mathbb{R}$.



Construction of an integral: General case

Step 3: Let X be a general random variable $X : \Omega \to \mathbb{R}$. Define:

$$X^+ := \max\{X, 0\} \quad \text{and} \quad X^- := -\min\{X, 0\}.$$



Then $X = X^+ - X^-$, and both X^+ and X^- are non-negative random variables. Now assume that both X^+ and X^- have finite expectations. We then define:

$$\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Notation:

$$\mathbb{E}(X) = \int_\Omega X(\omega)\, dP(\omega) = \int X\, dP.$$

Important properties of the expectation

We introduce the notation $L^1(\Omega, \mathcal{A}, P)$ to denote all random variables for which a finite expectation exists.

- Consider two random variables X, Y on Ω such that $X(\omega) < Y(\omega)$ for all ω and $X, Y \in L^1$. Then:
$$\mathbb{E}(X) < \mathbb{E}(Y).$$
- $X \in L^1(\Omega, \mathcal{A}, P) \iff |X| \in L^1(\Omega, \mathcal{A}, P)$.
In this case:
$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|).$$
- Any bounded random variable possesses an expectation:
$$X \text{ bounded} \iff \forall \omega \in \Omega : |X(\omega)| \leq C \quad \text{for some } C \in \mathbb{R}.$$

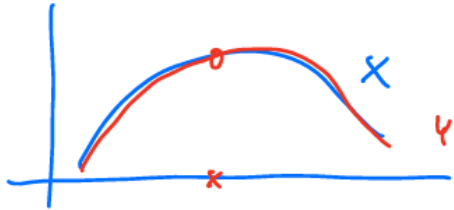
(Remark: Any bounded random variable has a well-defined, finite $\mathbb{E}(X)$).

Important properties of the expectation (2)

- $X = Y$ a.s. $\implies \mathbb{E}(X) = \mathbb{E}(Y)$.

Recall:
$X = Y$ a.s. $\iff \exists N \in \mathcal{A}$ with $P(N) = 0$ such that $\forall \omega \in \Omega \setminus N : X(\omega) = Y(\omega)$.

"Functions agree everywhere except on a set of measure 0"



Conditional probability

Unions and intersections

Notation:

$$P(A \cap B) = P(\text{“A and B”})$$

$$P(A \cup B) = P(\text{“A or B”})$$

Conditional probability

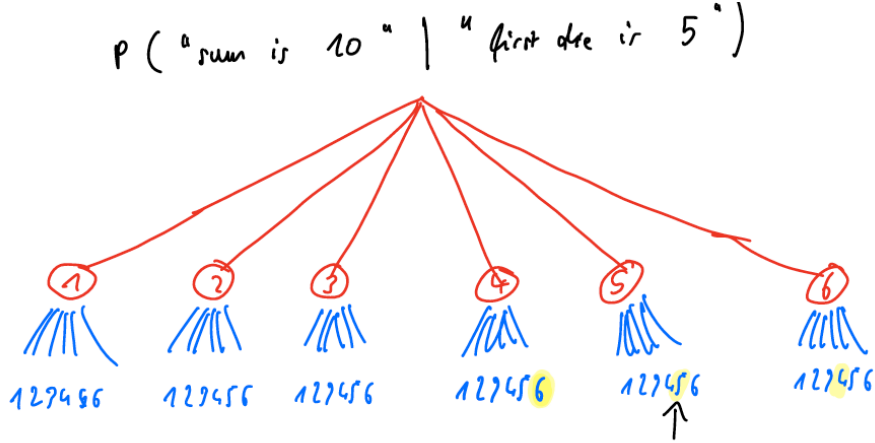
Definition 192 (Conditional probability)

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $A, B \in \mathcal{A}$ with $P(B) > 0$. Then:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

is called the conditional probability of A given B.

Example with two dice



Example with two dice (2)

$$P(\text{“sum is 10”} \mid \text{“first die is 5”}) = \frac{P(\text{“sum is 10” and “first die is 5”})}{P(\text{“first die is 5”})}$$

$$= \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

Makes sense: When we know the first die is 5, then the second one has to be 5 as well to achieve sum $= 10$, which happens with probability $1/6$.

Note: of course the "ordering" matters

- $\Omega =$ "all persons on earth" (a finite set)
- $\mathcal{A} = \mathcal{P}(\Omega)$
- $P =$ "uniform"

Event $A :=$ "person has been vaccinated"
Event $B :=$ "person has disease"

Two different conditional distributions:

$$\underbrace{P(\text{disease} \mid \text{vaccinated}) \quad \text{and} \quad P(\text{vaccinated} \mid \text{disease})}$$

Not the same!!!

Conditional Distribution (positive condition)

Theorem 193 (Conditional Distribution (positive condition))

Let $B \in \mathcal{A}$ with $P(B) > 0$.

The mapping

$$P_B : \mathcal{A} \to [0,1], \quad A \mapsto P(A \mid B)$$

is a probability measure on $(\Sigma, \mathcal{A})$; it is called the conditional distribution of P with respect to B.

B is fixed, $P(A \mid B)$ as a function of A.

Regular Conditional Distribution

In the definition of the conditional distribution, we require $P(B) > 0$.

But: In ML we often want to condition on events with probability 0:

$$P(Y = 1 \mid X = 0.1)$$

where X is a training point drawn from normal distribution.

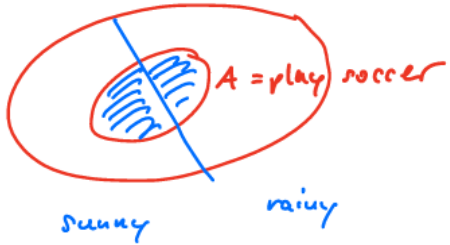
Under certain assumptions, the resulting "conditional distribution" exists and is well-defined (needs heavy math...). In ML we can typically take these assumptions for granted... see later.

Law of total probability and Bayes formula

Law of total probability

Let $B_1, B_2, \ldots, B_k$ be a disjoint partition of Ω with $B_i \in \mathcal{A}, P(B_i) > 0$ for all i, and $A \in \mathcal{A}$. Then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i) = \sum_{i=1}^{n} P(A \cap B_i).$$



Bayes formula

Let $B_1, B_2, \ldots, B_k$ be a disjoint partition of Ω with $B_i \in \mathcal{A}$ for all i, and $A \in \mathcal{A}$ with $P(A) \neq 0$. Then

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_i P(A \mid B_i) \cdot P(B_i)}.$$

Example: breast cancer screening

Assume:

- 1% of all women above 40 have breast cancer.
- 90% of women with breast cancer will be test positive. (true positives)
- 8% of women without breast cancer will receive a positive result as well. (false positives)

Given that a woman receives a positive test result, what is the likelihood that she has breast cancer?

$$
\begin{aligned}
P(\mathsf{cancer} \mid \mathsf{positive}) &= \frac{P(\mathsf{p} \mid \mathsf{c}) \cdot P(\mathsf{c})}{P(\mathsf{p} \mid \mathsf{c}) \cdot P(\mathsf{c}) + P(\mathsf{p} \mid \neg\mathsf{c}) \cdot P(\neg\mathsf{c})} \\
&= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.08 \cdot 0.99} \approx 10\%.
\end{aligned}
$$

Independence

Two Independent Events

Definition 194 (Two Independent Events)

Consider a probability space $(\Omega, \mathcal{A}, P)$. Two events $A, B \in \mathcal{A}$ are called independent if

$$P(A \cap B) = P(A) \cdot P(B).$$

Notation: $A \perp\!\!\!\perp B$

Observation: A is independent of $B \iff P(A \mid B) = P(A)$.

Many Independent Events

Definition 195 (Many Independent Events)

A family of events $(A_i)_{i \in I}$ is called independent if for all finite subsets $J \subseteq I$ we have

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i).$$

Note: A family is called pairwise independent if $\forall i, j \in I$:

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j).$$

$\rightarrow$ This does not imply independence!

Independent Random Variables

Proposition 196 (Independent Random Variables)

Two random variables $X : \Omega \to \Omega_1$, $Y : \Omega \to \Omega_2$ are called independent if their induced σ-algebras $\mathcal{G}(X), \mathcal{G}(Y)$ are independent:

$$\forall A \in \mathcal{G}(X), B \in \mathcal{G}(Y) : \quad P(A \cap B) = P(A) \cdot P(B).$$

Notation: $X \perp\!\!\!\perp Y$

Expectation of Independent Product

Proposition 197 (Expectation of Independent Product)

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be two random variables. Then:

$$X, Y \text{ independent} \implies \mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

Expectation of Independent Product (2)

Proof. (Discrete case)

$$\sum_{i,j} x_i y_j \, P(X = x_i, Y = y_j) \overset{\text{ind.}}{=} \sum_{i,j} x_i y_j \, P(X = x_i) \cdot P(Y = y_j)$$

$$= \left(\underbrace{\sum_i x_i \, P(X = x_i)}_{<\infty} \right) \cdot \left(\underbrace{\sum_j y_j \, P(Y = y_j)}_{<\infty} \right)$$

$$= \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

□

Variance and convariance

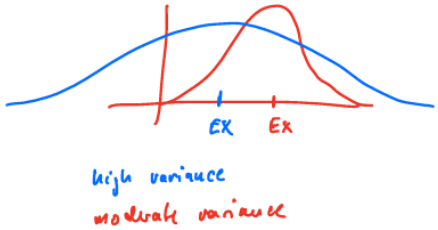
Variance & Standard Deviation

Definition 198 (Variance & Standard Deviation)

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be random variables with $\mathbb{E}(X^2) < \infty, \mathbb{E}(Y^2) < \infty$. Then

$$\mathrm{Var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

is called the variance of X, and $\sqrt{\mathrm{Var}(X)} := \sigma_X$ is called the standard deviation.



Covariance & Correlation

Definition 199 (Covariance & Correlation)

$$\mathrm{Cov}(X,Y) := \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

is called the covariance of X and Y.

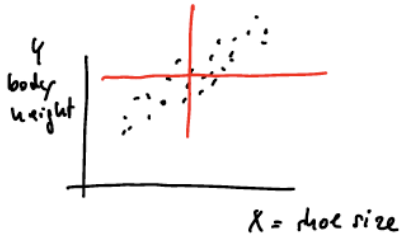
$$\rho_{XY} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,1]$$

is called the correlation coefficient.

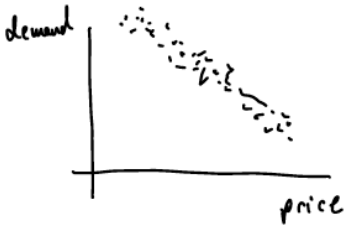
If $\mathrm{Cov}(X,Y) = 0$, then X and Y are called uncorrelated.

Intuition about covariance

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

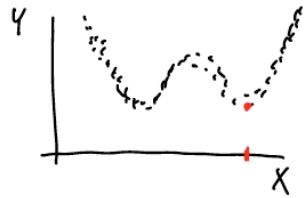


Positive, large covariance:
$\rho \approx 0.9$: Variables increase together.

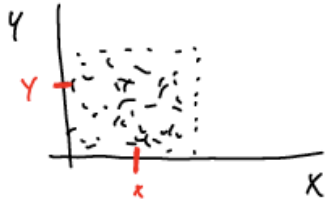


Negative covariance:
$\rho \approx -0.9$: Variables are inversely related.

Intuition about covariance (2)



Uncorrelated: $\mathrm{Cov} \approx 0$: Variables are not linearly related, but may still be dependent..



Independence

⚠ Uncorrelated $\rightleftharpoons$ independence.

Properties of variance and covariance

- $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
- $\mathrm{Cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- $\mathbb{E}(aX + b) = a \cdot \mathbb{E}(X) + b$
- $\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X)$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \, \mathrm{Cov}(X, Y)$
- If X, Y are independent, $\mathrm{Cov}(X, Y) = 0$ (but not vice versa).
- If X, Y are independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

k-th moment

$L^k(\Omega, \mathcal{A}, P) := \{X : \Omega \to \mathbb{R} \mid X \text{ measurable and } \int_\Omega |X|^k \, dP < \infty\}$

If $X^k \in L^1(\Omega, \mathcal{A}, P)$, then

$$\mathbb{E}(X^k) = \int X^k \, dP$$

is called the k-th moment of X, and

$$\mathbb{E}\left((X - \mathbb{E}(X))^k\right)$$

the k-th centered moment.

Markov and Chebyshev Inequality

Cauchy-Schwarz Inequality

Proposition 200 (Cauchy-Schwarz Inequality)

If $X, Y \in L^2(\Omega, \mathcal{A}, P)$, then:

$$\mathbb{E}(X \cdot Y)^2 \leq \mathbb{E}(X^2) \cdot \mathbb{E}(Y^2)$$

Markov inequality

Proposition 201 (Markov inequality)

Let $\varepsilon > 0$, $f : [0, \infty[\to [0, \infty[$ be monotonically increasing. Then:

$$P(|Y| > \varepsilon) \leq \frac{\mathbb{E}(f(|Y|))}{f(\varepsilon)}$$

In particular,

$$P(|Y| > \varepsilon) \leq \frac{\mathbb{E}(|Y|)}{\varepsilon}$$

Chebyshev Inequality

Proposition 202 (Chebyshev Inequality)

Let $\varepsilon > 0$, $X \in L^2(\Omega, \mathcal{A}, P)$. Then:

$$\underbrace{P(|X - \mathbb{E}(X)| > \varepsilon)}_{\text{key quantity in learning theory!}} \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

Different types of probability measures SKIPPED

Discrete measure

- Let $\Omega = \{x_1, x_2, \dots\}$ be finite or at most countable.
- Let $\mathcal{A} = \mathcal{P}(\Omega)$.
- We define a probability measure $P\colon \mathcal{A} \to [0,1]$ by assigning probabilities to the "elementary events":

$$P(\{x_i\}) := p_i$$

  Where: $0 \le p_i \le 1, \quad \sum_i p_i = 1.$
- For $A \in \mathcal{A}$ we define:

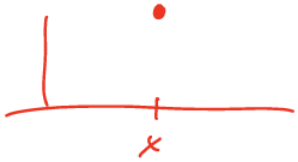
$$P(A) = \sum_{\{i | x_i \in A\}} p_i$$

Example: Toss a coin; distribution on $\mathbb{Q}$.

Dirac measure

For $x \in \mathbb{R}$, we define the Dirac measure δ_x on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by setting

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$



Sometimes this is called a point mass at a point x.

A discrete measure on $\mathbb{R}$ can be written as a sum of Dirac measures. For example, throwing a die can be described as

$$\frac{1}{6}(\delta_1 + \delta_2 + \cdots + \delta_6)$$

Measures with a density

Consider $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and the Lebesgue measure λ. Consider a function $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that is measurable and satisfies $\int f \, d\lambda = 1$. $(= \int f(x) \, dx)$

Then we define a measure ν on $\mathbb{R}^n$ by setting, for all $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\nu(A) := \int_A f(x) \, dx.$$

Then ν is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with density f.

Notation: $\nu = f \cdot \lambda$

Measures with a density (2)

Question: Can we describe every probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ in terms of a density?

Answer: No!

Counterexample: δ_0 Dirac measure

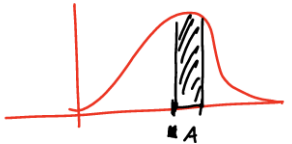
Absolutely continuous measures

Definition 203 (Absolutely continuous measures)

A probability measure ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called absolutely continuous with respect to another measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ if every μ-null set is also a ν-null set:

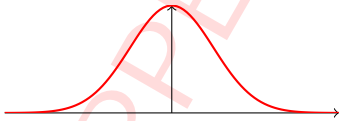
$$\forall B \in \mathcal{B}(\mathbb{R}^n): \quad \mu(B) = 0 \;\Rightarrow\; \nu(B) = 0.$$

Notation: $\nu \ll \mu$



Examples

- $\mathcal{N}(0,1) \ll \lambda$



- $\delta_0 \not\ll \lambda$, because

$$\lambda(\{0\}) = 0 \quad \text{but} \quad \delta_0(\{0\}) = 1.$$

Theorem of Radon-Nikodym

Theorem 204 (Radon-Nikodym)

Consider two probability measures ν, μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Then the following two statements are equivalent:

(1) ν has a density w.r.t. μ.

(2) ν is absolutely continuous w.r.t. μ.

Theorem of Radon-Nikodym (2)

Proof.

- (1) $\Rightarrow$ (2) is easy.
- (2) $\Rightarrow$ (1): We need to construct a density!

Consider the set $\mathcal{G}$ of all functions g with the following properties:

$$(*) \quad \begin{cases} \bullet \quad g \text{ is measurable}, \quad g \geq 0 \\ \bullet \quad g \cdot \mu \leq \nu, \text{ that is:} \\ \qquad \forall A \in \mathcal{B}(\mathbb{R}^n): \quad \int_A g \, d\mu \leq \nu(A) \end{cases}$$

- Observe: $g \equiv 0$ satisfies $(*)$, so $\mathcal{G}$ is not empty.
- If g, h both satisfy $(*)$, then $\sup(g, h)$ also satisfies $(*)$.

Theorem of Radon-Nikodym (3)

- Define $\gamma := \sup_{g \in \mathcal{G}} \int g \, d\mu$ and construct a sequence $(g_n)_{n \in \mathbb{N}}$ such that $\lim \int g_n \, d\mu = \gamma$.
- Define "density" $f := \sup g_n$.
- Now prove: f does the job.

□

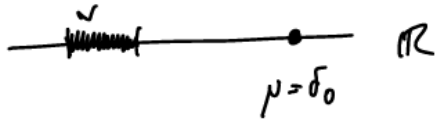
Theorem of Radon-Nikodym (4)

Definition 205 (Singular measures)

Let μ, ν be measures on $(\Omega, \mathcal{F})$. ν is called singular w.r.t. μ if there exists $A \in \mathcal{F}$ such that

$$\mu(A) = 0 \quad \text{but} \quad \nu(A^c) = 0$$

Notation: $\mu \perp \nu$



Example: $\lambda \perp \delta_0$

Lebesgue decomposition

Theorem 206 (Decomposition by Lebesgue)

Let μ, ν be probability measures on $(\Omega, \mathcal{F})$. Then there exists a unique decomposition

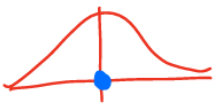
$$\nu = \nu_1 + \nu_2$$

such that

$$\nu_1 \ll \mu \quad \text{and} \quad \nu_2 \perp \mu.$$

Example: $\nu = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\delta_0$

$$\nu = \nu_1 + \nu_2 \quad \text{where} \quad \nu_1 = \frac{1}{2}\mathcal{N}(0,1), \quad \nu_2 = \frac{1}{2}\delta_0$$



Lebesgue decomposition (2)

Proof. Let $\mathcal{N}_\mu$ be the set of all null-sets w.r.t. μ, i.e. $\mathcal{N}_\mu \subset \mathcal{F}$.

$$\alpha := \sup \{\nu(A) \mid A \in \mathcal{N}_\mu\}$$

Construct a countable sequence $(A_n)_{n\in\mathbb{N}}$, $A_n \in \mathcal{N}_\mu$, such that $\nu(A_n) \nearrow \alpha$.
By countable additivity, we get:

$$\nu\underbrace{\left(\bigcup_{n\in\mathbb{N}} A_n\right)}_{:=\mathcal{N}} = \alpha$$

Define:

$$\nu_1(A) := \nu(A \cap \mathcal{N}^c), \quad \nu_2(A) := \nu(A \cap \mathcal{N})$$

Cantor distribution: non-trivial distribution that is singular wrt λ

1. Construct the Cantor set:

- Start with $C_0 := [0, 1]$
  "Remove middle part"

- $C_1 := [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$
  "Remove middle parts from all intervals"

- $C_2 = \ldots$

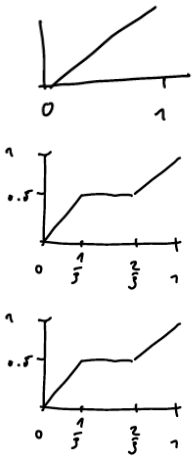
The Cantor set is the limit of this process. It is compact, non-empty, and has empty interior.

Cantor distribution: non-trivial distribution that is singular wrt λ (2)

2. Now construct a probability distribution:

Consider the cdfs of the sets $C_0, C_1, C_2, \ldots$

- On C_0: uniform on $[0, 1]$
- On C_1: uniform on $[0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$
- On C_2: ... take the limit and call the resulting measure γ.



Cantor distribution: non-trivial distribution that is singular wrt λ (3)

3. Properties

- The cdf of "γ" is continuous.
- ν is a probability measure.
- $\lambda(C) = 0$ (Lebesgue measure)

$$\Rightarrow \quad \lambda \perp \gamma$$

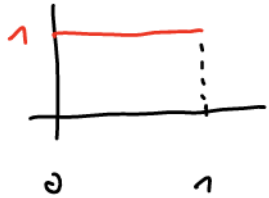
Important examples of probability distributions

Uniform distribution

Uniform distribution on $\{1, \ldots, n\}$:

$$P(\{i\}) = \frac{1}{n}.$$

Uniform distribution on $[0, 1]$: Distribution with constant density:



$$f(x) = \begin{cases} 1 & \text{if } x \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Uniform distribution on $[0, 1]^d \to$ independence!

Binomial distribution

Binomial distribution on $\{0, \ldots, n\}$:

Toss a coin n times, independently, each time with probability p of observing heads. Denote heads as 1, tails as 0. Let $X := \#$ heads. Then:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Poisson distribution

Poisson distribution on $\mathbb{N}$: Parameter $\lambda > 0$:

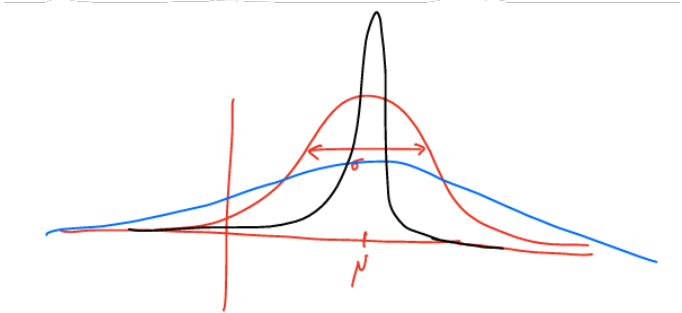
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Intuition: Number of incoming calls at a hotline.

Normal Distribution on $\mathbb{R}$

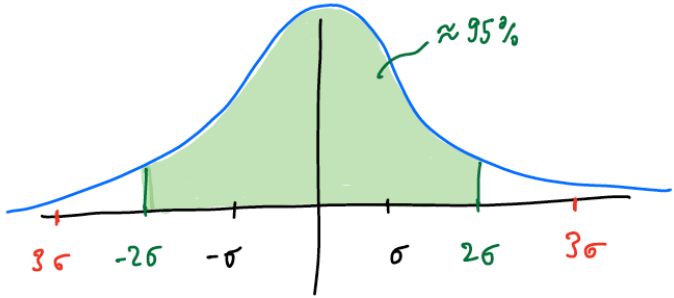
Density: Parameter μ (mean), σ (std. deviation):

$$f_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



Notation: $\mathcal{N}(\mu, \sigma^2)$

A rule of thumb for normal distributions



The area under the normal distribution:

$[-\sigma, \sigma] \rightsquigarrow \approx 70\%$

$[-2\sigma, 2\sigma] \rightsquigarrow \approx 95\%$

$[-3\sigma, 3\sigma] \rightsquigarrow \approx 99\%$

Sum of Independent Normals is Normal

Proposition 207 (Sum of Independent Normals is Normal)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $X \perp Y$. Then:

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Proof.

It is elementary to see that the mean of $X + Y$ is $\mu_1 + \mu_2$ and the variance is $\sigma_1^2 + \sigma_2^2$.

However, it is not trivial to prove that the resulting distribution is again normal. The standard approach requires convolution and characteristic functions. □

Multivariate normal distribution

Let $X : \Omega \to \mathbb{R}^n$, $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, $\mu_i = \mathbb{E}(X_i)$, $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$.

$\Sigma \in \mathbb{R}^{n \times n}$ with $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, called the covariance matrix.

The density is given by:

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Notation: $X \sim \mathcal{N}(\mu, \Sigma)$.

The Covariance Matrix is Positive Semi-Definite

Proposition 208 (The Covariance Matrix is psd)

Let C be the covariance matrix of any set of random variables, i.e., $C_{ij} = \mathrm{Cov}(X_i, X_j)$. Then C is symmetric and positive semi-definite.

Proof.

Symmetry is clear because $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.

To prove positive semi-definiteness, we show that for all $a_1, \ldots, a_n \in \mathbb{R}$:

$$a^T C a = \sum_{i,j=1}^{n} a_i a_j C_{ij} \geq 0.$$

The Covariance Matrix is Positive Semi-Definite (2)

Using the definition of C_{ij}:

$$a^T C a = \sum_{i,j=1}^{n} a_i a_j C_{ij} \stackrel{\text{def}}{=} \sum_{i,j=1}^{n} a_i a_j \, \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

By linearity of expectation:

$$a^T C a = \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i (X_i - \mu_i) \right)^2 \right].$$

Since the square of any real number is non-negative:

$$a^T C a \geq 0.$$

PCA & Contour Lines

Because Σ is symmetric and positive semi-definite, it has real-valued eigenvalues ≥ 0.

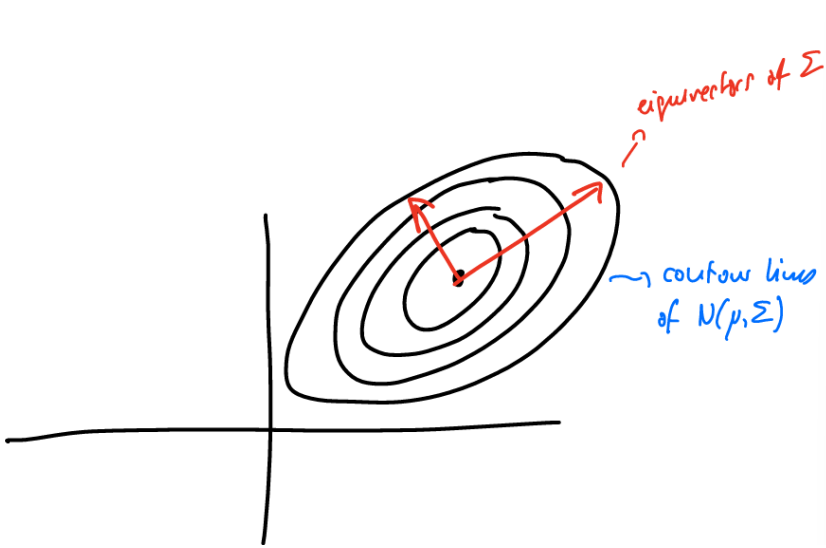
The contour lines of a multivariate normal distribution are ellipses:

$$\text{Contour line: set } \{x \mid f_{\mu,\sigma}(x) = c\}$$

$$f_{\mu,\sigma}(x) = c \iff (x-\mu)^T \Sigma^{-1} (x-\mu) = \tilde{c}.$$

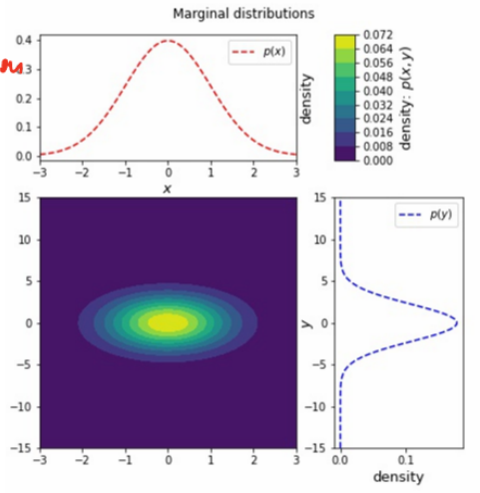
This is the ellipse equation.

PCA & Contour Lines (2)



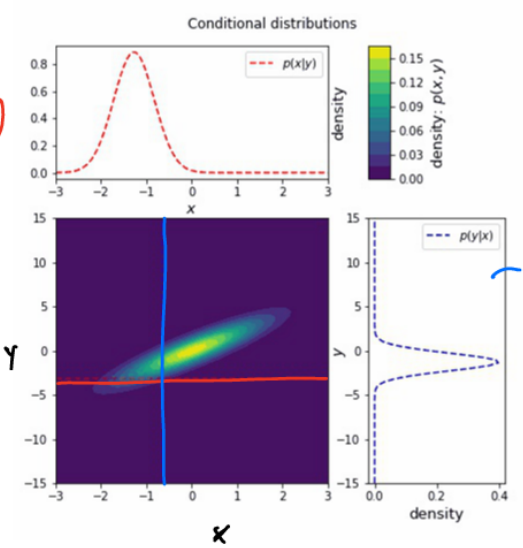
Marginal distributions of Gaussians are Gaussians

Marginal distribution of x



Marginal distribution of y

Conditional distributions of Gaussians are Gaussians



conditional
$p(x \mid y = -4)$

conditional
$p(y \mid x = -0.7)$

Mixture of Gaussians

Consider "mixing weights" $a_1, \ldots, a_k \in [0,1]$ such that $\sum_{i=1}^{k} a_i = 1$.

Fix $\mu_1, \ldots, \mu_k$ and $\Sigma_1, \ldots, \Sigma_k$. Define the new density:

$$f(x) = \sum_{i=1}^{k} a_i \, f_{\mu_i, \Sigma_i}(x).$$

This is called a mixture of Gaussians.

Examples of Distributions with "Infinite" Expectation

Cauchy Distribution on $\mathbb{R}$: Density:

$$f(x) = \left(\pi r \left(1 + \left(\frac{x - x_0}{r}\right)^2\right)\right)^{-1}.$$

Power Law Distributions on $\mathbb{R}$: A family of distributions that satisfy:

$$P(X > x) = c \cdot x^{-d}, \quad d \text{ is a parameter.}$$

- If $d \leq 3$, no variance exists.
- If $d \leq 2$, no mean exists.

ML Keyword: Preferential attachment model for social networks.

Heavy-Tailed vs Sub-Gaussian Distributions

"Tail Behavior" of a Gaussian:

$$P(|X| > t) = 2 \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Heavy-Tailed Distributions: Distributions where $P(|X| > t)$ is larger than for a Gaussian.

Sub-Gaussian: Distributions where $P(|X| > t)$ is smaller than for a gaussian

Sub-Gaussian distributions are particularly popular in learning theory because they exhibit strong concentration (as seen later).

Convergence of random variables

Almost Sure Convergence

Definition 209 (Almost Sure Convergence)

Consider random variables $X_i : \Omega \to \mathbb{R}$, $i \in \mathbb{N}$, $X : \Omega \to \mathbb{R}$, on a probability space $(\Omega, \mathcal{A}, P)$. The sequence $(X_i)_{i\in\mathbb{N}}$ converges to X almost surely if:

$$P\left(\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\}\right) = 1.$$

Notation: $X_i \longrightarrow X$ a.s.

⚠ To check that this definition is well-defined, we must show that the event

$$\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\}$$

is an element of $\mathcal{A}$:

Well-defined

Proposition 210 (Well-defined)

$$\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\} \in \mathcal{A}.$$

Proof. (Sketch) The limit $\lim_{i\to\infty} X_i(\omega) = X(\omega)$ means:

$$\forall k \in \mathbb{N}, \exists N \in \mathbb{N}, \forall n > N : |X_n(\omega) - X(\omega)| < \frac{1}{k} \quad \Rightarrow \varepsilon = \frac{1}{k}$$

Thus, we write:

$$\{\omega \mid X_i(\omega) \longrightarrow X(\omega)\} = \bigcap_{k\in\mathbb{N}} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N} \{\omega \mid |X_n(\omega) - X(\omega)| < \frac{1}{k}\} \in \mathcal{A}.$$

Since X_i, X are measurable, $|X_n - X|$ is measurable, so this set is in $\mathcal{A}$. □

Convergence in probability

Definition 211 (Convergence in probability)

The sequence $(X_i)_{i\in\mathbb{N}}$ converges to X in probability $:\Longleftrightarrow$

$$\forall \varepsilon > 0 : P(\{\omega \in \Omega \mid |X_i(\omega) - X(\omega)| > \varepsilon\}) \longrightarrow 0.$$

Convergence in L_p

Definition 212 (Convergence in L_p)

The sequence $X_n \to X$ in L^p ("in the p-th mean") $:\Longleftrightarrow$

$$X_n, X \in L^p \text{ and } \|X_n - X\|_p \longrightarrow 0.$$

$$\|X_n - X\|_p^p = \int |X_n - X|^p \, dP(x) \longrightarrow 0$$

Note: $\|z\|_p^p = \int |z|^p \, d\mu$

Weak convergence

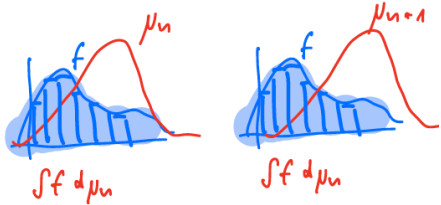
Definition 213 (Weak convergence)

Let $\mathcal{M}^1(\mathbb{R}^n)$ be the set of all probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Assume $(\mu_n) \subset \mathcal{M}^1(\mathbb{R}^n)$, $\mu \in \mathcal{M}^1(\mathbb{R}^n)$. $C_b(\mathbb{R}^n)$ is defined as the space of bounded continuous functions.

$\mu_n \to \mu$ weakly $:\iff$

$$\forall f \in C_b(\mathbb{R}^n) : \int f \, d\mu_n \longrightarrow \int f \, d\mu.$$

⚠ Weak convergence is defined for measures, not for random variables.



(Excursion: Weak Convergence in Functional Analysis)

In functional analysis, a sequence (x_n) in a Banach space B converges weakly if for all bounded linear functionals f, we have:

$$f(x_n) \longrightarrow f(x).$$

(i.e., for all $f \in B'$).

The space $\mathcal{M}^1(\mathbb{R}^n)$ itself is not a Banach space, but it is contained in $\mathcal{M}(\mathbb{R}^n)$, the space of all bounded measures. The dual space of $\mathcal{M}(\mathbb{R}^n)$ is $C_b(\mathbb{R}^n)$.

Thus, the weak convergence defined above coincides with the notion of weak convergence on $\mathcal{M}(\mathbb{R}^n)$.

Convergence in distribution

Definition 214 (Convergence in distribution)

Let $X_n, X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$. The sequence X_n converges in distribution to X: $\iff$

the distributions P_{X_n} converge to P_X weakly.

Relationship between notions of convergence

We have the following implications, and none of the "missing directions" hold in general:

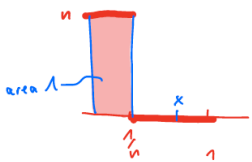
$$\text{almost surely} \implies \text{in probability} \implies \text{in distribution}$$

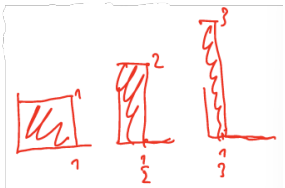
$$\text{in } L^p (p > 1) \implies \text{in } L^1 \implies \text{in probability} \implies \text{in distribution}$$

Example (Convergence a.s., in probability, but not in L^1)

Consider $[0,1]$ with the uniform distribution.
Define $X_n : [0,1] \to \mathbb{R}$ as:

$$X_n(x) = \begin{cases} n & \text{for } 0 \leq x \leq \frac{1}{n}, \\ 0 & \text{otherwise.} \end{cases}$$





$\forall x > 0 : X_n(x) \to 0.$
Can formally see: converges a.s., in prob. But: no convergence in L^1.

Example (Convergence in Probability, in L^1, but not a.s.)

"Sliding blocks of smaller and smaller volume."

$f_1 = \mathbb{1}_{[0,1]}$

$f_2 = \mathbb{1}_{[0,1/2]}$, $f_3 = \mathbb{1}_{[1/2,1]}$

$f_4 = \mathbb{1}_{[0,1/3]}$, $f_5 = \mathbb{1}_{[1/3,2/3]}$, $f_6 = \mathbb{1}_{[2/3,1]}$

Example (Convergence in Distribution, but not in Probability)

- $X_n : [0,1] \to \mathbb{R}$, $\forall i : X_i(\omega) = \mathbb{1}_{[0,\frac{1}{2}[}$

- $X = \mathbb{1}_{[\frac{1}{2},1]}$

Obviously, $X_n \not\longrightarrow X$ in probability, but:

$$\underline{P_{X_1}} = \frac{1}{2}(\delta_0 + \delta_1) = P_{X_2} = P_{X_3} = \cdots = P_X.$$

So $X_n \longrightarrow X$ in distribution.

Let $\Omega = \mathbb{R}$. The measure δ_0 is the measure on $\mathbb{R}$ called the Dirac measure at 0:

$$P(A) = \begin{cases} 1 & \text{if } 0 \in A \\ 0 & \text{otherwise} \end{cases}$$

Example (Convergence in distribution, but not in probability)

Space $[0,1]$ with uniform distribution.

- $X_i : [0,1] \to \mathbb{R}$, $\forall i : X_i(\omega) = \mathbb{1}_{[0,\frac{1}{2}[}$

- $X = \mathbb{1}_{[\frac{1}{2},1]}$

Obviously, $X_n \not\longrightarrow X$ in probability, but:

$$P_{X_1} = \text{Ber}\left(\frac{1}{2}\right) = P_{X_2} = P_{X_3} = \cdots = P_X$$

Bernoulli distribution on $\{0,1\}$

So $X_n \longrightarrow X$ in distribution!

Theorem of Borel-Cantelli

Definition "A_n infinitely often"

Definition 215 (A_n infinitely often)

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and (A_n) a sequence of events in $\mathcal{A}$.

$$P(A_n \text{ infinitely often}) := P(A_n\, i.o.) = P\left(\{\omega \in \Omega \mid \omega \in A_n \text{ for infinitely many } n\}\right).$$

Almost Sure Convergence $\iff$ "$> \varepsilon$ infinitely Often"

Proposition 216 (Almost Sure Convergence $\iff$ "$> \varepsilon$ infinitely Often")

Let X_n, X be random variables on $(\Omega, \mathcal{A}, P)$. Then:

$$X_n \longrightarrow X \; a.s. \iff \forall \varepsilon > 0,\; P\left(\underbrace{\{|X_n - X| > \varepsilon}_{A_n} \underbrace{\text{ infinitely often}\}}_{\text{w.r.t. } n} \right) = 0.$$

Almost Sure Convergence $\iff$ "$> \varepsilon$ infinitely Often" (2)

Proof.

$$(\bigstar) = \{\lim X_n = X\} = \left\{\forall k : |X_n - X| > \frac{1}{k} \text{ at most finitely often}\right\}$$

$$= \bigcap_{k\in\mathbb{N}} \left\{|X_n - X| > \frac{1}{k} \text{ at most finitely often}\right\}$$

$$= \left(\underbrace{\bigcup_{k\in\mathbb{N}} \left\{|X_n - X| > \frac{1}{k} \text{ infinitely often}\right\}}_{(\bigstar\bigstar)} \right)^{\text{complement}}$$

Thus $P(\bigstar) = 1 \iff P(\bigstar\bigstar) = 0$

□

Borel-Cantelli

Theorem 217 (Borel-Cantelli)

Consider a sequence of events $(A_n) \subset \mathcal{A}$:

(1) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \, i.o.) = 0$.

(2) If $\sum_{n=1}^{\infty} P(A_n) = \infty$, and if (A_n) are independent, then $P(A_n \, i.o.) = 1$.

Application in Learning Theory

Assume that $\underbrace{P(|X_n - X| > \frac{1}{n}) < \delta_n}_{\text{conv. in prob.}}$, and assume that $\sum_{n=1}^{\infty} \delta_n < \infty$. Then, you can use Borel-Cantelli to prove that:

$$P(|X_n - X| > \frac{1}{n} \, i.o.) = 0,$$

thus $X_n \longrightarrow X \, a.s.$ (See assignments.)

The law of large numbers (LLN)

i.i.d.

... is an abbreviation for

"independent and identically distributed."

A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is i.i.d. if they are independent and all follow the same distribution.

Being i.i.d. is one of the standard assumptions in learning theory.

Strong vs Weak Law of Large Numbers

- Strong law: implies convergence almost surely (a.s.).
- Weak law: implies convergence in probability.

Weak Law of Large Numbers

Theorem 218 (Weak Law of Large Numbers)

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with $\mathrm{Var}(X_i) = C < \infty$ and $\mathbb{E}(X_i) = \mu$. Denote $S_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:

$$P(|S_n - \mu| > \varepsilon) \leq \frac{C}{n\varepsilon^2}.$$

In particular:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \longrightarrow 0 \quad \text{in probability.}$$

(Observe: i.i.d. $\implies$ that all X_i have the same mean and variance.)

Weak Law of Large Numbers (2)

Proof. Wlog assume $\mu = 0$ (otherwise consider the centered r.v. $X_i - \mu$).
The weak law then follows directly from the Chebyshev inequality:

$$P\left(|S_n - \mu| > \varepsilon\right) \overset{\mu=0}{=} P\left(|S_n| > \varepsilon\right) \leq \frac{\mathbb{E}(S_n^2)}{\varepsilon^2}$$

Now exploit that

$$\mathbb{E}(S_n^2) = \mathbb{E}\left(\left(\frac{1}{n}\sum X_i\right)^2\right) = \frac{1}{n^2}\mathbb{E}\left(\sum_{i,j} X_i X_j\right)$$

$$\overset{\text{ind.}}{=} \frac{1}{n^2}\sum_i \underbrace{\mathbb{E}(X_i^2)}_{\mathsf{Var}(X_i)} + \frac{1}{n^2}\sum_{i\neq j} \underbrace{\mathbb{E}(X_i)\mathbb{E}(X_j)}_{=0} = \frac{1}{n^2}\cdot n \cdot \underbrace{\mathsf{Var}(X_i)}_{=C} = \frac{C}{n}$$

Plugging this in above immediately gives the desired result.

□

Strong law of large numbers (vanilla version)

Theorem 219 (Strong law of large numbers (vanilla version))

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with $\mathrm{Var}(X_i) \leq C < \infty$ and $\mathbb{E}(X_i) = \mu < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu \quad \text{almost surely.}$$

Strong law of large numbers (vanilla version) (2)

Proof.

We prove the theorem under a slightly stronger condition: $\mathbb{E}(X_i^4) < \infty$. Without loss of generality we assume that $\mu = 0$ (otherwise we replace X_i by $X_i - \mu$). To simplify notation, introduce $S_n := \sum_{i=1}^n X_i$.

General idea:

- We want to apply Borel–Cantelli to events of the form

$$\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \delta_n \right\}$$

- Need to find events s.th. $P(\{\cdots > \delta_n\}) \leq \varepsilon_n$ and $\sum_{n=1}^{\infty} \varepsilon_n < \infty$
- For those individual events we are going to use the Markov inequality.

Strong law of large numbers (vanilla version) (3)

Proof Step 1: Under the given assumptions, there exists a constant $K < \infty$ such that

$$\mathbb{E}(S_n^4) \leq Kn^2 :$$

→Proof:

$$\mathbb{E}(S_n^4) = \mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^4\right)$$

- Multiply out
- Exploit that all X_i have the same distribution
- And that $\mathbb{E}(X_i) = 0$

$$= n \cdot \mathbb{E}(X_1^4) + 3n(n-1) \cdot \mathbb{E}(X_1^2 X_2^2) \leq K \cdot n^2$$, for some constant K

Strong law of large numbers (vanilla version) (4)

Proof Step 2:

Markov inequality for function $f(x) = x^4$:

$$\left(\text{Recall Markov inequality: } f \text{ monotone increasing } \Rightarrow P(|Y| > \varepsilon) \leq \frac{\mathbb{E}(f(|Y|))}{f(\varepsilon)}\right)$$

For all $\gamma > 0$,

$$P\left(\frac{1}{n}|S_n| > n^{-\gamma}\right) \leq \frac{\mathbb{E}\left(\left(\frac{1}{n}|S_n|\right)^4\right)}{n^{-4\gamma}} \overset{\text{Step 1}}{\leq} \frac{Kn^2}{n^{-4\gamma}} = K \cdot n^{-2+4\gamma}$$

Strong law of large numbers (vanilla version) (5)

Proof Step 3:

Borel–Cantelli

Fix some $\gamma \in \left]0, \frac{1}{4}\right[$ and define the events

$$A_n := \left\{ \frac{1}{n} |S_n| \geq n^{-\gamma} \right\}$$

Then

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} K \cdot n^{-2+4\gamma} < \infty \quad \text{(since } -2+4\gamma < -1\text{)}$$

Borel–Cantelli $\Rightarrow P(A_n \text{ i.o.}) = 0$. Thus "high deviations" can only happen at most finitely often $\Rightarrow$ convergence a.s. $\square$

Many more variants of the LLN exist

There exist many, many versions of the LLN under all kinds of assumptions.

Let us try to get some intuition for what is really needed.

The Conditions in the Theorem

Independence: The theorem does not hold if we allow for arbitrary dependencies.

Example:

$$X_1 \sim \text{Ber}\left(\frac{1}{2}\right) \quad (\text{"fair coin"})$$

$$X_2 = X_3 = \cdots =: X_1 \quad (\text{all other coins get the same result as the first one})$$

Here, X_i are identically distributed with mean $\mu = \frac{1}{2}$.

But: $S_n = \begin{cases} 0 & \text{if } X_1 = 0 \\ 1 & \text{if } X_1 = 1 \end{cases}$, hence $S_n \to \begin{cases} 0 \\ 1 \end{cases}$ but not to $\mathbb{E}(X_n) = \frac{1}{2}$.

One can weaken independence "a bit" (e.g., stationary sequences, martingale differences, ...).

The Conditions in the Theorem (2)

i.i.d. distributed: This condition is not really necessary.
For example, a popular version of the theorem states that the variance of all the X_i needs to be bounded by the same constant:

$$\forall i \in \mathbb{N} : \mathsf{Var}(X_i) \leq C.$$

Bounded variance: In cases where the X_i are identically distributed, we do not need a bounded variance.

Bounded expectation: This is obviously needed, otherwise we would not even be able to state the result...

Machine Learning Question

Consider a training set $(x_i, y_i)_{i=1,\dots,n}$ drawn i.i.d. $\sim P$. Consider a loss function ℓ and a classifier f_n that has been trained on this data.

Training error:

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_n(x_i), y_i).$$

Test error:

$$R(f) = \mathbb{E}\big[\ell(f_n(x_i), y_i)\big].$$

Does the LLN now state that $R_n \longrightarrow R$?
No! Why?

The central limit theorem (CLT)

Central limit theorem (vanilla version)

Theorem 220 (Central limit theorem (vanilla version))

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Consider the random variable $S_n := \sum_{i=1}^{n} X_i$. We normalize it to:

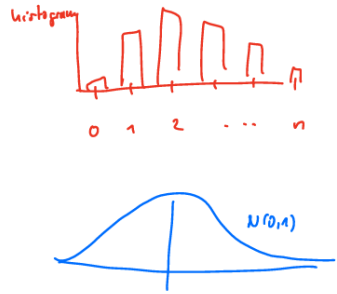
$$Y_n := \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad \text{(which has mean } 0 \text{ and standard deviation } 1\text{)}.$$

Then $Y_n \longrightarrow Y$ in distribution, where $Y \sim \mathcal{N}(0,1)$.

Illustration

Let X_i represent coin tosses, where head $\triangleq 1$ and tail $\triangleq 0$.

$$S_n = \sum X_i \in [0, n].$$



Which Conditions Are Needed?

Also here, many more versions...

- Independence is really important.
- We don't need identical distributions at all.
- Bounded variance helps but can be weakened. In the end, we need to establish conditions that assert that each individual random variable "is negligible in the limit" and does not dominate the resulting sum.

CLT in d dimensions

Theorem 221 (CLT in d dimensions)

Let $(X_i)_{i\geq 1}$ be i.i.d. random variables with values in $\mathbb{R}^d$. Let μ be the (d-dimensional) mean and C the covariance matrix between the d variables. Then:

$$\lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}} = Z$$

where Z is a d-dimensional random variable with distribution $\mathcal{N}(0, C)$.

Discussion

Really impressive!

- We don't need any assumptions on the exact distribution of the X_i.
- Whenever something is a sum of many i.i.d. random variables, it is going to look like a normal distribution.
- Example: Measurement errors.
- The CLT is "the basis" for statistical testing!

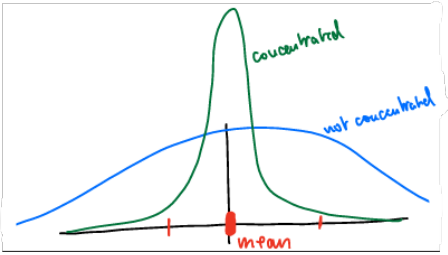
Concentration inequalities

First motivation

In ML, we very often replace individual statistics by their expectations, e.g., the training error.

The LLN says that the mean converges to the expected value. But the speed of convergence can be really slow.

Concentration inequalities tell us that with high probability, an event is very close to its mean.



Hoeffding inequality

Theorem 222 (Hoeffing inequality)

Let $X_1, \ldots, X_n$ be random variables, independent, assume that $X_i \in [a_i, b_i]$ a.s. for $i = 1, \ldots, n$.

Let $S_n := \sum_{i=1}^{n} (X_i - \mathbb{E}(X_i))$. Then for any $t > 0$,

$$P(S_n > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

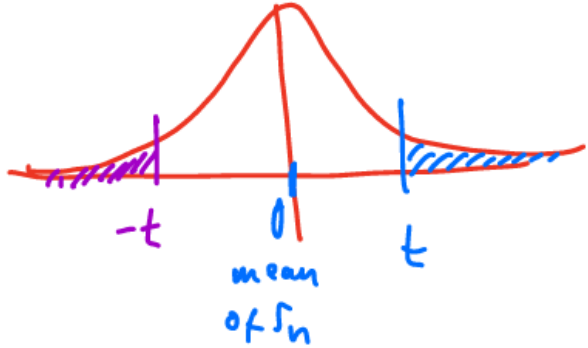
Hoeffding inequality (2)

Corollary 223

If the $X_1, \ldots, X_n$ are i.i.d. with $X_i \in [a, b]$, then

$$P\left(\frac{1}{n}|S_n| > t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Illustration



Application of Hoeffding: LLN

Consider the following variant of the LLN:

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables, $a \leq X_i \leq b$, and let X have the same distribution as the X_i. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}(X) \quad \text{a.s.}$$

Note that we do not make any assumptions on $\mathbb{E}(X_i^4)$ as in our earlier proof of the LLN. We now use Hoeffding to prove it.

Proof of the LLN with Hoeffding

Proof. Step 1: Hoeffding implies:

$$P\left(\frac{1}{n}\sum X_i - \mathbb{E}(X) > t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

$$P\left(\frac{1}{n}\sum X_i - \mathbb{E}(X) < -t\right) = P\left(\frac{1}{n}\sum(-X_i) - \mathbb{E}(-X) > t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Combining both, we get:

$$P\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}(X)\right| > t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Proof of the LLN with Hoeffding (2)

Step 2: Now we want to apply Borel–Cantelli to get a.s. convergence of

$$Z_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

Define event A_n from Borel–Cantelli:

$$\sum_{n=0}^{\infty} P(Z_n - \mathbb{E}(X) > t) \leq 2 \cdot \underbrace{\sum_{n=0}^{\infty} \exp\left(-\frac{2nt^2}{(b-a)^2}\right)}_{=: \text{ sum}} \overset{(\star)}{\leq} \infty$$

Proof of the LLN with Hoeffding (3)

(★) Substitute: $r := \exp\left(-\frac{2t^2}{(b-a)^2}\right)$, observe that $r \in]0,1[$ if $t > 0$.

Observe: $\exp\left(-\frac{2nt^2}{(b-a)^2}\right) = r^n$

Thus the sum becomes:

$$2 \cdot \sum_{n=0}^{\infty} r^n = 2 \cdot \frac{1}{1-r} < \infty$$

Now Borel–Cantelli gives almost sure convergence. □

Hoeffding for ML training error?

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. training points.
Let f be an arbitrary, fixed function, and ℓ a loss function.

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \quad \text{(empirical risk)}$$

$$R(f) = \mathbb{E}(\ell(f(X), Y)) \quad \text{(true risk)}$$

Do we have $R_n(f) \longrightarrow R(f)$ as $n \to \infty$?

Hoeffding for ML training error? (2)

(1) If f is independent of the training points, then the terms

$$\ell(f(X_1), Y_1), \ldots, \ell(f(X_n), Y_n)$$

are independent *(needs a proof)*, and we can apply Hoeffding to assert that $R_n(f) \longrightarrow R(f)$.

(2) But in ML, we typically deal with a function f_n that has been chosen by an algorithm that looks at all training points!
So the terms

$$\ell(f_n(X_1), Y_1), \ldots, \ell(f_n(X_n), Y_n)$$

are not independent, and we cannot simply use Hoeffding to assert that $R_n(f_n) \longrightarrow R(f)$.
$\Rightarrow$ see lecture on statistical ML next term!!!

Hoeffding is tight without further assumptions

Hoeffding is tight (cannot be improved without further assumptions). For fair coin tosses it is tight.

But: Not tight if the coin is biased, as we need other inequalities.

Might want to use this instead:

Bernstein inequality

Theorem 224 (Bernstein inequality)

Let $x_1, \ldots, x_n$ be independent with 0 mean, $|x_i| < 1$ a.s. Let

$$\sigma^2 := \frac{1}{n} \sum_{i=1}^{n} \mathsf{Var}(x_i).$$

Then for all $t > 0$,

$$P\left(\frac{1}{n} \sum_{i=1}^{n} x_i \geq t \right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + t/3)} \right).$$

Concentration inequality for functions with bounded differences

Consider a function $g : \mathbb{R}^n \to \mathbb{R}$ (or more generally $g : \mathcal{X}^n \to \mathbb{R}$ for some arbitrary space $\mathcal{X}$).
We say that g has the bounded differences property if there exist constants $c_1, \ldots, c_n$ such that:

$$(\star) \quad \sup_{\substack{X_1,\ldots,X_n \in \mathcal{X} \\ \tilde{X}_i \in \mathcal{X}}} \left| g(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n) - g(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n) \right| \leq$$

Example: $g(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^n X_i$, and $a \leq X_i \leq b \, \forall i$, then g satisfies $(\star)$ with $c_i = b - a$.

Theorem of McDiarmid

Theorem 225 (Theorem of McDiarmid)

Let $X_1, \ldots, X_n$ be independent random variables, $X_i \in \mathcal{X}_i$, and $g : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ a function with bounded differences property. Then, for any $t > 0$,

$$P\left(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Cool observation

Consider again the ML scenario with i.i.d. training pts (x_i, y_i), a loss ℓ, and a function f_n that has been chosen based on the training points. Consider again the term

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f_n(x_i), y_i) =: g(x_1, \dots, x_n).$$

If we can establish a bounded difference statement for g, then we can apply McDiarmid and get concentration!!

Observe that g involves the construction fun as well:

$$\underbrace{(x_1, \dots, x_n) \xrightarrow{\text{learning algorithm}} f_n \xrightarrow{\text{risk}} \frac{1}{n}\sum_{i=1}^{n} \ell(f_n(x_i), y_i)}_{g(x_1,\dots,x_n)}.$$

Applications of concentration inequalities

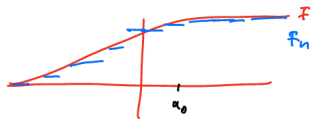
- In ML: everywhere!!!
- Stability in ML
- Standard theoretical CS, randomized algorithms, e.g., Johnson-Lindenstrauss
- Largest eigenvalue of a random symmetric matrix

Glivenko-Cantelli Theorem SKIPPED

Glivenko-Cantelli Theorem

F cdf: $F(a) = P(X \leq a)$, $X_1, \ldots, X_n \sim F$, i.i.d. , $F_n : \mathbb{R} \to [0, 1]$

$$F_n(a) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq a\}}$$



Now fix any particular $a_0 \in \mathbb{R}$.

$$F_n(a_0) \longrightarrow F(a_0) \text{ by the law of large numbers.}$$

Because $\mathbb{1}_{\{X_i \leq a_0\}}$ is a Binomial random variable with

$$p = P(X_i \leq a_0).$$

So it is clear that $F_n \to F$ pointwise a.s. (i.e. for all a_0)
Now let's look at uniform convergence.

Glivenko-Cantelli Theorem (2)

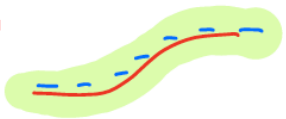
Theorem 226 (Theorem of Glivenko-Cantelli)

$X_1, \ldots, X_n$ i.i.d. random variables with cdf F. Let F_n be the empirical cdf induced by the sample. Then:

$$P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F(a)| > \varepsilon\right) \leq 8 \cdot (n+1) \cdot \exp\left(-\frac{n\varepsilon^2}{32}\right).$$

In particular,

$$\sup |F_n - F| \to 0 \text{ a.s.}, \quad \text{i.e. } F_n \to F \text{ uniformly a.s.}$$



Proof

Observe: LLN $\Rightarrow P\left(|F_n(a_0) - F(a_0)| > \varepsilon\right) \to 0$ for any fixed a_0.

Problem: Need to look at

$$P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F(a)| > \varepsilon\right)$$

Difficult to evaluate the supremum because $\mathbb{R}$ is uncountable.

If we take a supremum over a finite set, it is easier:

$$P\left(\max_{i=1,\dots,n} |U_i| > \varepsilon\right) = P(|U_1| > \varepsilon \underline{\vee} |U_2| > \varepsilon \underline{\vee} \cdots \vee |U_n| > \varepsilon) \leq \sum_{i=1}^{n} P(|U_i| > \varepsilon)$$

Proof (2)

Trick of the proof: Convert $\sup_{a \in \mathbb{R}}$ to something "finite" directly.

How could we achieve this?

$$|\text{red} - \text{green}| = |\text{red} - \text{blue}| \leq 2 \cdot |\text{green} - \text{blue}|$$

Proof (3)

Step 1: Symmetrization by ghost sample

Assume $X'_1, \ldots, X'_n \sim F$ independently ("ghost sample"),

Denote by F'_n the empirical cdf induced by the ghost sample.

Now it is easy to prove:

$$P\left(\sup_a \left|F_n(a) - F(a)\right| > \varepsilon\right) \leq 2\, P\left(\sup_a \left|F_n(a) - F'_n(a)\right| > \frac{\varepsilon}{2}\right)$$

Proof (4)

Step 2: Want to split this in two terms

$$|F_n(a) - F_n'(a)| = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{\{X_i \leq a\}} - \mathbb{1}_{\{X_i' \leq a\}} \right) \right| \qquad (\bigstar)$$

Introduce Rademacher random variables $\sigma_1, \ldots, \sigma_n$:

$$\sigma_i(\{-1\}) = \sigma_i(\{1\}) = 1/2$$

Distribution of $(\bigstar)$ is the same as the distribution of:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{\{X_i \leq a\}} - \mathbb{1}_{\{X_i' \leq a\}} \right) \right| = (\bigstar\bigstar)$$

Proof (5)

Now we have:

$$2\,P\left(\sup_a |F_n(a) - F'_n(a)| > \frac{\varepsilon}{2}\right) = 2\,P\left(\sup_a \left|\frac{1}{n}\sum_i \sigma_i \left(\mathbb{1}_{\{X_i \le a\}} - \mathbb{1}_{\{X'_i \le a\}}\right)\right| > \frac{\varepsilon}{2}\right)$$

$$\le 2\,P\left(\sup_a \underbrace{\left|\frac{1}{n}\sum_i \sigma_i \mathbb{1}_{\{X_i \le a\}}\right|}_{u} > \frac{\varepsilon}{4}\right) + 2\,P\left(\sup_a \underbrace{\left|\frac{1}{n}\sum_i \sigma_i \mathbb{1}_{\{X'_i \le a\}}\right|}_{v} > \frac{\varepsilon}{4}\right)$$

Observe: $P(|U - V| > \frac{\varepsilon}{2}) \le P(|U| > \frac{\varepsilon}{4}$ or $|V| > \frac{\varepsilon}{4})$ (right side is necessary for left side)

$$= 4\,P\left(\sup_a \left|\frac{1}{n}\sum_i \sigma_i \mathbb{1}_{\{X_i \le a\}}\right| > \frac{\varepsilon}{4}\right)$$

Proof (6)

Step 3: Exploit "finite structure"

Fix $X_1, \ldots, X_n$ (i.e. condition on $X_1, \ldots, X_n$)

We look at $\mathbb{1}_{\{X_i \le a\}}$.

The random variables $\mathbb{1}_{\{X_1 \le a\}}, \ldots, \mathbb{1}_{\{X_n \le a\}}$ have at most $n+1$ realizations for fixed a.

$$P\left(\sup_a \frac{1}{n}\left|\sum_i \sigma_i \mathbb{1}_{\{X_i \le a\}}\right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n\right)$$

$$\le (n+1) \sup_a \underbrace{P\left(\frac{1}{n}\left|\sum_i \sigma_i \mathbb{1}_{\{X_i \le a\}}\right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n\right)}_{\text{use Hoeffding (\#)}}$$

Proof (7)

Step 4: Apply Hoeffding to (#)

$$\forall a : P\left(\frac{1}{n}\left|\sum_i \sigma_i \mathbb{1}_{\{X_i \le a\}}\right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n\right) \le 2\exp\left(-\frac{n\varepsilon^2}{32}\right)$$

Combining everything gives the theorem.

Product space, joint distributions

Product space, joint distribution

Consider two measurable spaces $(\Omega_1, \mathcal{A}_1), (\Omega_2, \mathcal{A}_2)$.

Define the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with:

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\},$$

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \{A_1 \times A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}.$$

Consider two random variables:

$$X_1 : (\Omega, \mathcal{A}, P) \to (\Omega_1, \mathcal{A}_1),$$

$$X_2 : (\Omega, \mathcal{A}, P) \to (\Omega_2, \mathcal{A}_2).$$

Product space, joint distribution (2)

Define

$$X := (X_1, X_2) : (\Omega, \mathcal{A}, P) \to (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2),$$
$$(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega)).$$

The distribution $P_{(X_1,X_2)}$ on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ is called the joint distribution of X_1 and X_2.

Example in ML: (X, Y) where X is the input data, Y is the label.

Product measure

Definition 227 (Product measure)

Consider $(\Omega_1, \mathcal{A}_1, P_1)$, $(\Omega_2, \mathcal{A}_2, P_2)$ two probability spaces.
We define the product measure $P_1 \otimes P_2$ on the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ as:

$$(P_1 \otimes P_2)(A_1 \times A_2) := P_1(A_1) \cdot P_2(A_2).$$

Product $\sim$ Independence

Theorem 228 (Product $\sim$ Independence)

Two random variables X_1, X_2 are independent if and only if their joint distribution coincides with the product distribution:

$$P_{(X_1,X_2)} = P_1 \otimes P_2.$$

Marginal distributions

Marginal distribution

Definition 229 (Marginal distribution)

Consider the joint distribution $P_{(X_1,X_2)}$ of two random variables $X := (X_1, X_2)$. The marginal distribution of X with respect to X_1 is the original distribution of X_1 on $(\Omega_1, \mathcal{A}_1)$, namely P_{X_1}.

Similarly for P_{X_2}.

Example in discrete case

$Y \backslash X$	x_1	x_2	x_3	Σ
y_1	p_1	p_2	p_3	$p_1 + p_2 + p_3 = P(Y = y_1)$
y_2	p_4	p_5	p_6	$p_4 + p_5 + p_6 = P(Y = y_2)$
Σ	$p_1 + p_4 = P(X = x_1)$	$\dots$	$\dots$	

Marginal with respect to X *marginal with respect to Y*

Marginal distribution in the case of densities

Let $X, Y : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $Z := (X, Y)$. Assume that the joint distribution of Z has a density f on $\mathbb{R}^2$. Then the following statements hold:

(1) Both X and Y have densities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by:

$$f_X(x) = \int_{-\infty}^{\infty} \underbrace{f(x,y)}_{\text{joint density}} \, dy \rightsquigarrow \text{sum over } y$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) \, dx \rightsquigarrow \text{sum over } x$$

(2) X and Y are independent if and only if:

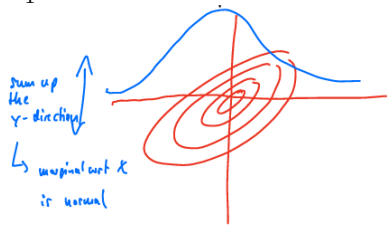
$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \text{a.s.}$$

Special case: Marginals of multivariate normal distributions

2-dim: Consider a 2-dimensional normal random variable $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with mean

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2, \quad \text{and covariance matrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

Then the marginal distribution of X with respect to X_1 is again a normal distribution with mean μ_1 and variance σ_1^2.



Special case: Marginals of multivariate normal distributions (2)

n-dim: Let $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$. Group the variables:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} =: \tilde{X}, \quad \begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix} =: X^{\#}.$$

Denote the mean:

$$\mu = \begin{pmatrix} \tilde{\mu} \\ \mu^{\#} \end{pmatrix}, \quad \tilde{\mu} \in \mathbb{R}^k, \mu^{\#} \in \mathbb{R}^{n-k}.$$

Special case: Marginals of multivariate normal distributions (3)

Denote the covariance matrix:

$$\mathbb{R}^{n\times n} \ni \Sigma_i = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \text{ with } \Sigma_{11} \in \mathbb{R}^{k\times k},\ \Sigma_{22} \in \mathbb{R}^{(n-k)\times(n-k)},\ \text{etc.}$$

The marginal of X with respect to $\tilde{X}$ is a normal distribution on $\mathbb{R}^k$ with mean $\tilde{\mu}$ and covariance Σ_{11}.

Conditional distributions

Conditional distributions in the discrete case

Discrete case:

We know conditional probabilities:

$$P(A \mid B)$$

defined for events $A, B \in \mathcal{A}$ and $P(B) > 0$.

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be discrete random variables, $y \in \mathbb{R}$, such that $P(Y = y) > 0$.

Then we can define the conditional probability measure:

$$P_{X|Y=y} : A \mapsto P(X \in A \mid Y = y).$$

This is a probability measure.

Conditional distributions in the case of densities

Assume $Z := (X, Y)$ has a joint density $f : \mathbb{R}^2 \to \mathbb{R}$, and marginal densities $f_X, f_Y : \mathbb{R} \to \mathbb{R}$. Then the function

$$f_{X|Y=y}(x) := \frac{f(x, y)}{f_Y(y)}$$

is then also a density on $\mathbb{R}$, called the conditional density of X given $Y = y$.

General case (not discrete, no density)

For general random variables this is surprisingly complicated!

$\rightsquigarrow$ "regular conditional probabilities" $\rightsquigarrow$ Skipped

Example: normal distributions

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \tilde{\mu} \\ \mu^{\#} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

If $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, then the conditional distribution

of $\tilde{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$ w.r.t. $x^{\#} = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix}$ is given by

$$p_{\tilde{x}|x^{\#}} \sim \mathcal{N}\left(\mu + \Sigma_{12}\Sigma_{22}^{-1}(x^{\#} - \tilde{\mu}),\ \Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Conditional expectations

Conditional expectation in the discrete case (conditional on event)

Definition 230 (Conditional expectation in the discrete case)

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$. Assume X takes finitely (countably) many values $x_1, \ldots, x_n \in \mathbb{R}$, and Y takes finitely (countably) many values $y_1, \ldots, y_m \in \mathbb{R}$, always with a positive probability. Then we define the conditional expectation

$$\mathbb{E}(Y \mid X = x_i) := \sum_{j=1}^{m} y_j \underbrace{P(Y = y_j \mid X = x_i)}_{\text{well-defined}}$$

Example

Example: Consider two dice:

$$X = \text{first die}, \quad Y = \text{second die}, \quad \text{independent}$$

We compute:

$$\mathbb{E}(\text{sum} \mid X = 1) = \sum_{i=1}^{12} i \cdot P(\text{sum} = i \mid X = 1).$$

Expanding:

$$\mathbb{E}(\text{sum} \mid X = 1) = \sum_{k=1}^{6} (1 + k) \cdot P(Y = k \mid X = n).$$

Example (2)

Since X and Y are independent:

$$\mathbb{E}(\text{sum} \mid X = 1) = \sum_{k=1}^{6} (1 + k) \cdot \frac{1}{6} = \frac{1}{6} \sum_{k=1}^{6} (1 + k) = 1 + 3.5 = 4.5$$

So far, we have defined $\mathbb{E}(Y \mid X = x_i)$. But often we want to consider the function $\mathbb{E}(Y \mid X)(\omega)$, which is a random variable:

$$\mathbb{E}(Y \mid X) : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}).$$

Leads to the following:

Conditional expectation with respect to a random variable

Definition 231 (Conditional expectation with respect to a random variable)

Definition (discrete case): Let X, Y as before. Then the conditional expectation is defined as follows:

$$\mathbb{E}(Y \mid X) := f(X),$$

with

$$f(x) = \begin{cases} \mathbb{E}(Y \mid X = x) & \text{if } P(X = x) > 0, \\ \text{arbitrary, say } 0 & \text{otherwise.} \end{cases}$$

⚠ $\mathbb{E}(Y \mid X)$ is only defined a.s.

General case?

Problem: $P(X = x)$ might be 0 all the time...

Case of densities is still straightforward:

Case of joint densities

- Let $X, Z \colon \Omega \to \mathbb{R}$ have a joint density $f(x, z)$.
- Let $g \colon \mathbb{R} \to \mathbb{R}$ be a bounded function and define $Y := g(Z)$.
- Assume we want to compute the conditional expectation

$$\mathbb{E}(Y \mid X) = \mathbb{E}(\underbrace{g(Z)}_{=Y} \mid X)$$

- The marginal density of X is given by:

$$f_X(x) = \int f(x, z)\, dz$$

Case of joint densities (2)

- The conditional density of Z given $X = x$ is:

$$f_{X=x}(z) = \frac{f(x,z)}{f_X(x)} \qquad (\text{if } f_X(x) \neq 0)$$

- Now consider

$$h(x) := \int \underbrace{g(Z)}_{=Y} f_{X=x}(z)\, dz$$

- Define

$$\mathbb{E}(Y \mid X) = h(X)$$

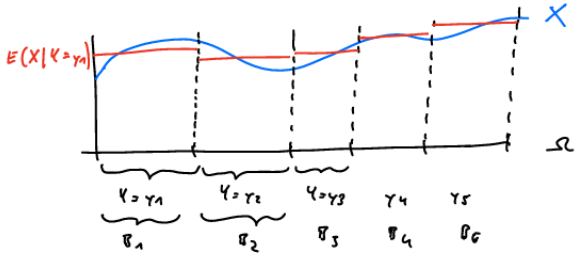
Idea of the more general case

Consider:

- X: continuous random variable,
- Y: discrete random variable taking values $y_1, \ldots, y_5$.

Idea of the more general case (2)

We want to analyze $\mathbb{E}(X \mid Y)$.



Want to "define"

$$\mathbb{E}(X \mid Y) := \sum_{i=1}^{5} \mathbb{E}(X \mid Y = y_i) \cdot \underbrace{\mathbb{1}_{B_i}(\omega)}_{rv}.$$

But need to make sure that $\mathbb{E}(X \mid Y)$ is measurable with respect to $\sigma(Y)$, the σ-algebra generated by Y. This ensures that $\mathbb{E}(X \mid Y)$ is a valid random variable.

Conditional expectation on L_1

Definition 232 (Conditional expectation on L_1)

Let $X : (\Omega, \mathcal{F}_0, P) \to \mathbb{R}$, $X \in L_1(\Omega, \mathcal{F}_0, P)$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}_0$ (Intuition: $\mathcal{F}_0$ will be the σ-algebra generated by the variable we want to condition on).

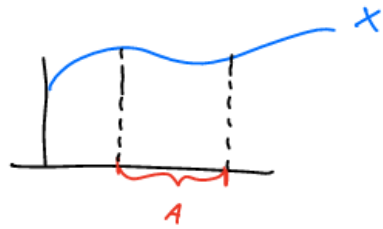
We now define the conditional expectation of X given $\mathcal{F}$: $\mathbb{E}(X \mid \mathcal{F})$ as any random variable Z that satisfies:

(1) Z is measurable with respect to $\mathcal{F}$,

(2) For all $A \in \mathcal{F}$, we have:

$$\int_A X \, dP = \int_A Z \, dP.$$

Conditional expectation on L_1 (2)



Existence: The existence of $\mathbb{E}(X \mid \mathcal{F})$ is not clear a priori, it needs to be proved.

$$\mathbb{E}(X \mid Y) := \mathbb{E}(X \mid \sigma(Y))$$

Examples:

- If $X = Y$, then $\mathbb{E}(X \mid Y) = X$ (a.s.).
- If $X \perp\!\!\!\perp Y$, then $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$ (a.s.).

Part V
Statistics

Literature: Wassermann: all of statistics

Point estimates, bias & variance, consistency

Standard setup in parametric statistics

We assume that data is generated by a particular family of distributions, for example,

$$\mathcal{F} = \left\{ \mathcal{N}(\mu, \sigma^2) \mid \underbrace{\mu \in \mathbb{R}, \sigma^2 > 0}_{\Theta} \right\}.$$

The family $\mathcal{F}$ is called the statistical model.

More generally,

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \},$$

with θ being one particular parameter of the function f and Θ the space of all possible parameters.

We are given a sample $X_1, \ldots, X_n \sim f_\theta$ (typically, i.i.d.), but the true underlying θ is unknown.

Conventions

- Parameter space Θ (*"capital theta"*).
- True (unknown) parameters θ (*"lower case theta"*).
- $P_\theta, \mathbb{E}_\theta$ refer to the probability and expectation under the distribution f_θ.
- Estimates typically get a "hat": $\hat{\theta}, \hat{\mu}, \ldots$.

Point estimation

The goal of point estimation is to estimate θ.

Definition 233 (Point estimation)

Given a statistical model

$$\mathcal{F} = \{f_\theta \mid \theta \in \Theta\},$$

and a sample $X_1, \ldots, X_n \sim f_\theta \in \mathcal{F}$, a point estimator $\hat{\theta}_n$ of parameter θ is a function

$$\hat{\theta}_n := g(X_1, \ldots, X_n).$$

Bias of an estimator

Definition 234 (Bias of an estimator)

The bias of such an estimator is defined as

$$\operatorname{bias}(\hat{\theta}_n) := \mathbb{E}_\theta(\hat{\theta}_n) - \theta.$$

- With $\mathbb{E}_\theta$ being the expectation w.r.t. the (true) distribution of f_θ.
- $\hat{\theta}_n$ being the estimate.
- θ the true parameter.

Intuition: repeat the procedure very often (infinitely often) and average over the estimate $\hat{\theta}_n$.

An estimate is unbiased if its bias is zero.

Variance and standard error

Definition 235 (Variance and standard error)

The variance of an estimator is defined as

$$\mathrm{Var}_\theta(\hat{\theta}_n).$$

The corresponding standard deviation is called the standard error (se).

Typically, se is unknown, but it can be estimated: $\hat{se}$.

Example

Let $X_1, \ldots, X_n \sim \underbrace{\text{Bernoulli}(p)}_{\{=0,1\}}$ with parameter $\underbrace{p}_{\theta} \in \underbrace{[0,1]}_{\Theta}$. Define

$$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

as an estimate of p.

$$E_p(\hat{p}_n) = E_p\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E_p(X_i)}_{p} = p.$$

Thus, $\hat{p}_n$ is unbiased because

$$E_p(\hat{p}_n) - p = p - p = 0.$$

Example (2)

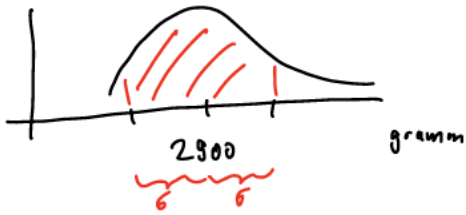
The standard error of this estimate is

$$se = \sqrt{\mathrm{Var}_p(\hat{p}_n)} = \sqrt{\frac{1}{n}\,\mathrm{Var}_p(X_1)} = \sqrt{\frac{p(1-p)}{n}}.$$

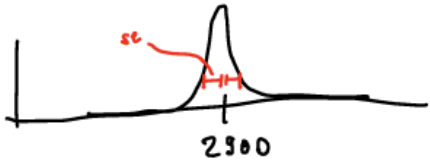
We can, for example, estimate it by

$$\hat{se} = \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}.$$

Example: Weight of baby



Distribution of individual data points: $\mu = 2900$, $\sigma = 500$.



Distribution of the estimator $\hat{\mu}$.

Mean squared error

Definition 236 (Mean squared error)

The mean squared error (MSE) of an estimator is the quantity

$$\mathrm{MSE}(\hat{\theta}_n, \theta) = E_\theta\left((\hat{\theta}_n - \theta)^2\right)$$

Bias-Variance decomposition

Theorem 237 (Bias-Variance decomposition)

$$\mathrm{MSE}(\hat{\theta}_n, \theta) = \mathrm{bias}^2(\hat{\theta}_n) + \mathrm{Var}_\theta(\hat{\theta}_n).$$

$\rightarrow$ "how good is our estimate"

Bias-Variance decomposition (2)

Proof.

$$\begin{aligned}\mathbb{E}_\theta\left(\left(\hat{\theta}_n-\theta\right)^2\right) &= \mathbb{E}_\theta\left(\left(\hat{\theta}_n-\mathbb{E}\hat{\theta}_n+\mathbb{E}\hat{\theta}_n-\theta\right)^2\right)\\ &= \mathbb{E}_\theta\left(\left(\hat{\theta}_n-\mathbb{E}\hat{\theta}_n\right)^2\right)\\ &\quad + 2\mathbb{E}_\theta\left(\left(\hat{\theta}_n-\mathbb{E}\hat{\theta}_n\right)\left(\mathbb{E}\hat{\theta}_n-\theta\right)\right)\\ &\quad + \mathbb{E}_\theta\left(\left(\mathbb{E}\hat{\theta}_n-\theta\right)^2\right)\end{aligned}$$

Note:

$$2\left(\mathbb{E}\hat{\theta}_n-\theta\right)\cdot\mathbb{E}_\theta\left(\hat{\theta}_n-\mathbb{E}\hat{\theta}_n\right)=2\left(\mathbb{E}\hat{\theta}_n-\theta\right)\cdot\left(\mathbb{E}_\theta\hat{\theta}_n-\mathbb{E}_\theta\hat{\theta}_n\right)=0$$

Bias-Variance decomposition (3)

Thus:

$$\mathbb{E}_\theta\left(\left(\hat{\theta}_n-\theta\right)^2\right)=\underbrace{\mathbb{E}_\theta\left(\left(\hat{\theta}_n-\mathbb{E}\hat{\theta}_n\right)^2\right)}_{\mathsf{Var}(\hat{\theta}_n)}+\underbrace{\left(\mathbb{E}\hat{\theta}_n-\theta\right)^2}_{=(\mathsf{bias}(\hat{\theta}_n))^2}$$

□

Example

$$\mathcal{F} = \left\{ \mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

Sample: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ, σ^2, i.i.d.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ is an unbiased estimator of } \mu.$$

$$\hat{\sigma}_1^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \qquad \text{"first estimator"}$$

$$\hat{\sigma}_2^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \qquad \text{"second estimator"}$$

Example (2)

$$E(\hat{\sigma}_1^2) = \frac{n-1}{n}\sigma^2 \quad \text{so the bias is } \frac{1}{n}\sigma^2$$

$$E(\hat{\sigma}_2^2) = \sigma^2 \quad \text{(unbiased!)}$$

$$\mathrm{Var}(\hat{\sigma}_1^2) = \frac{2(n-1)}{n^2}\sigma^4$$

$$\mathrm{Var}(\hat{\sigma}_2^2) = \frac{2}{n-1}\sigma^4$$

Example (3)

$$\mathrm{MSE}(\hat{\sigma}_1^2) = \mathrm{bias}^2 + \mathrm{var} = \left(\frac{2(n-1)}{n^2}\right)\sigma^4$$

$$\mathrm{MSE}(\hat{\sigma}_2^2) = \cdots = \frac{2}{n-1}\sigma^4$$

$$\Rightarrow \mathrm{MSE}(\hat{\sigma}_1^2) < \mathrm{MSE}(\hat{\sigma}_2^2)$$

Consistent estimator

Definition 238 (Consistent estimator)

A point estimator $\hat{\theta}_n$ of θ is consistent (strongly consistent) if

$$\hat{\theta}_n \to \theta \text{ in probability (a.s.) as } n \to \infty.$$

Theorem 239

If an estimator satisfies $\mathrm{bias} \to 0$ and $se \to 0$ as $n \to \infty$, then the estimator is consistent.

Confidence sets

Confidence sets

Definition 240 (Confidence sets)

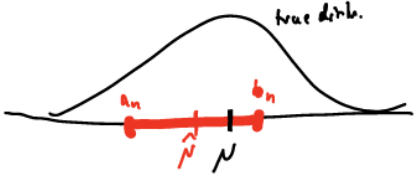
A $(1-\alpha)$-confidence interval for a parameter $\theta \in \mathbb{R}$ is an interval $C_n = (a_n, b_n)$ where $a_n = a(X_1, \ldots, X_n)$, $b_n = b(X_1, \ldots, X_n)$ are functions of the sample $X_1, \ldots, X_n$ such that

$$\underbrace{P_\theta}_{\text{true parameters}} \Big(\underbrace{\theta}_{\text{deterministic}} \in \underbrace{C_n}_{\text{random}} \Big) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.$$

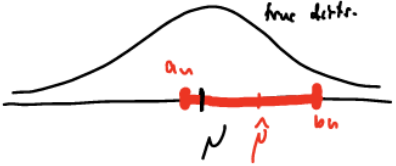
$(1-\alpha)$ is called the coverage of the confidence interval.

Illustration

First experiment: $X_1, \ldots, X_n$ and $\hat{\mu}$. First confidence set:



Second experiment: $X'_1, \ldots, X'_n$ and $\hat{\mu}'$. Second confidence set:



... in $(1-\alpha)$ of the repetitions, the true μ is inside the red interval.

Example

Consider a coin flip with

$$P(X = 1) = p, \quad P(X = 0) = 1 - p,$$

where $p \in [0, 1]$ is unknown. We want to estimate it.

We observe $X_1, \ldots, X_n \sim f_p$.

Define

$$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Choose a confidence level α and define the confidence interval $C_n = (a_n, b_n)$. To this end, define

$$\varepsilon_n^2 := \frac{\log(2/\alpha)}{2n}.$$

Example (2)

Proposition 241

$$C_n := (\hat{p}_n - \varepsilon_n, \hat{p}_n + \varepsilon_n)$$

is a confidence interval with coverage $1 - \alpha$.

Example (3)

Proof. By Hoeffding inequality, for any t we have

$$\mathbb{P}\left(|\hat{p}_n - p| > t\right) \leq \underbrace{2\exp(-2nt^2)}_{\alpha}$$

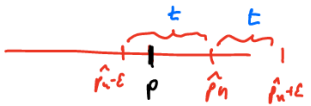
Set $\alpha := 2\exp(-2nt^2)$

and solve for t:

$$\log\left(\frac{\alpha}{2}\right) = -2nt^2$$

$$\Rightarrow \quad t^2 = \frac{-\log\left(\frac{\alpha}{2}\right)}{2n} = \frac{\log\left(\frac{2}{\alpha}\right)}{2n}$$

Choose $\varepsilon_n = t$.



□

Maximum likelihood estimator

Motivating example

$\mathcal{F} = \{A \mid A \text{ symmetric, } n \times n \text{ matrix, } a_{ij} \in \{0, 1\}\}$ (adjacency matrices of graphs)

We observe random walks from the graph of length 10.

Goal: reconstruct (estimate) A.

Idea: Among all adjacency matrices $A \in \mathcal{F}$, select the one that has the highest likelihood to have produced the random walks we have observed.

$\Rightarrow$ Maximum likelihood approach.

Likelihood

More formally: Consider a parametric family

$$\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}.$$

We observe i.i.d. points $X_1, \ldots, X_n \sim f_\theta \in \mathcal{F}$.

The likelihood of the data given a parameter θ_0 is

$$P_{\theta_0}(X_1, \ldots, X_n) \overset{\text{notation!}}{=} P(X_1, \ldots, X_n \mid \theta_0).$$

$$= \prod_{i=1}^{n} P(X_i \mid \theta_0).$$

Maximum likelihood

To estimate the true parameter θ, we select θ such that the likelihood is maximized:

$$\hat{\Theta} = \operatorname*{argmax}_{\theta \in \Theta} P(X_1, \ldots, X_n \mid \theta).$$

$$\overset{\text{ind.}}{=} \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta).$$

This is equivalent to the problem

$$\hat{\Theta} = \operatorname*{argmax}_{\theta} \log \left(\prod_{i=1}^{n} P(X_i \mid \theta) \right)$$

$$= \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \underbrace{\log \underbrace{P(X_i \mid \theta)}_{\in [0,1]}}_{<0}.$$

Maximum likelihood (2)

which is equivalent to minimizing the negative log-likelihood:

$$\hat{\Theta} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \underbrace{-\log P(X_i \mid \theta)}_{>0}.$$

$\Rightarrow$ Maximum-likelihood estimator (MLE).

Solving maximum likelihood problems

Sometimes this optimization problem is easy:

- We might be able to solve it analytically (rare).
- If we are lucky, it is convex.
- Most typically, it is not convex.

Example for an analytic solution

Model: $X \sim \text{Poisson}(\lambda)$, meaning that

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{with } E(X) = \lambda, \quad \text{Var}(X) = \lambda.$$

We observe $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and want to construct the MLE estimator for λ.

Compute the likelihood:

$$L(\lambda) = P(X_1, \ldots, X_n \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}.$$

Example for an analytic solution (2)

Log-likelihood function:

$$\log L(\lambda) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right)$$

$$= \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log(X_i!)) = f(\lambda).$$

Example for an analytic solution (3)

Optimization for λ: Take the derivative w.r.t. λ and set to zero:

$$f'(\lambda) = \sum_{i=1}^{n} \left(\frac{X_i}{\lambda} - 1 \right) = \frac{1}{\lambda} \left(\sum_{i=1}^{n} X_i \right) - n = 0.$$

Solving for λ:

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Thus, the ML estimator is:

$$\hat{\lambda} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

□

MLE properties

From the theory side, the MLE often (but not always) has nice properties:

(1) If the model $\mathcal{F}$ consists of "nice" functions, then the MLE based on an i.i.d. sample is consistent.

(2) If $\mathcal{F}$ consists of "nice" functions, the MLE estimate $\hat{\Theta}_{MLE}$ is asymptotically normal:

$$\frac{\hat{\Theta}_{MLE} - \theta}{se} \xrightarrow{\text{distr.}} N(0,1).$$

$$\frac{\hat{\Theta}_{MLE} - \theta}{\hat{se}} \xrightarrow{\text{distr.}} N(0,1).$$

MLE properties (2)

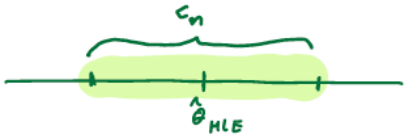
(3) This can be used to construct approximate confidence intervals:

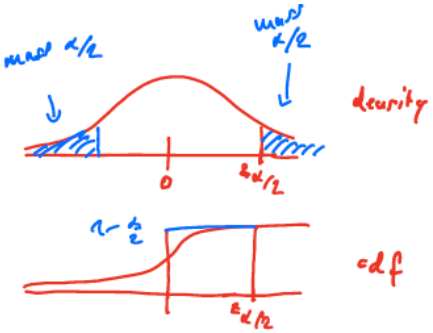
$$C_n := (\hat{\Theta}_{MLE} \underbrace{-z_{\alpha/2}\hat{se}}_{-\varepsilon}, \quad \hat{\Theta}_{MLE} \underbrace{+z_{\alpha/2}\hat{se}}_{+\varepsilon}),$$

where

$$z_{\alpha/2} := \Phi^{-1}(1-\alpha/2), \qquad \text{with } \Phi^{-1} \text{ cdf of } N(0,1)$$

MLE properties (3)





C_n is an "approximate confidence interval" in the sense that

$$P_\theta\left(\theta \in C_n\right) \to 1-\alpha \quad \text{as } n \to \infty.$$

Sufficiency & Identifiability

Sufficiency

Intuition: Given a sample $X_1, \ldots, X_n \sim f_\theta \in \mathcal{F}$, we typically convert the (large) sample to a statistic

$$T(X_1, \ldots, X_n)$$

(in the extreme case, one number).

Question: Can we recover the true parameter θ from that statistic?

If yes, we would like to call the statistic $T(X_1, \ldots, X_n)$ *"sufficient"*.

Sufficiency (2)

Which properties would we need to assert sufficiency?

- When we observe two samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_n$, and

$$T(X_1, \ldots, X_n) = T(X'_1, \ldots, X'_n),$$

  then we would infer the same θ.
- When we know $T(X_1, \ldots, X_n)$, then we would need some way to estimate θ just based on $T(X_1, \ldots, X_n)$.

Formal definition is technical, skipped.

Identifiability

Sometimes families of distributions can be described in different ways with different sets of parameters.

Definition 242 (Identifiability)

A parameter θ for a family

$$\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$$

is identifiable if distinct values of θ correspond to distinct pdfs in $\mathcal{F}$:

$$\theta \neq \theta' \Rightarrow f_\theta \neq f_{\theta'}.$$

(Identifiability is a property of the model itself, not of the data.)

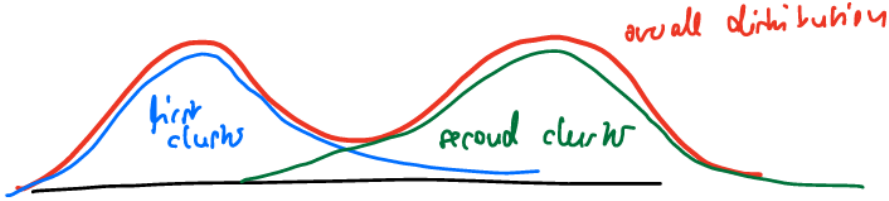
Example (Identifiability)

The problem of clustering consists in finding "groups" in data.

From a statistical point of view, this is equivalent to the problem of decomposing the underlying distribution of a mixture, say:

$$p = \alpha_1 p_1 + \alpha_2 p_2$$

(where α_1, α_2 are the mixing weights)



Example (Identifiability) (2)

Depending on our statistical model, such patterns might be identifiable or not.

Often, parametric models are identifiable, for example, mixtures of Gaussians.

Non-parametric models, for example implicit kernel algorithms, are often not identifiable.

Hypothesis testing

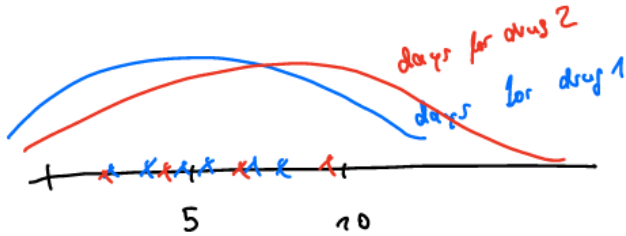
Motivation

Example:

Two drugs D_1, D_2. We measure the number of days to recovery for both drugs:

$$X_1, \ldots, X_n \quad \text{treated with } D_1$$

$$X_1', \ldots, X_n' \quad \text{treated with } D_2$$

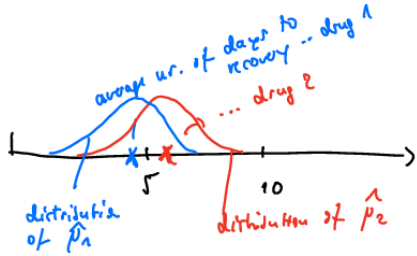


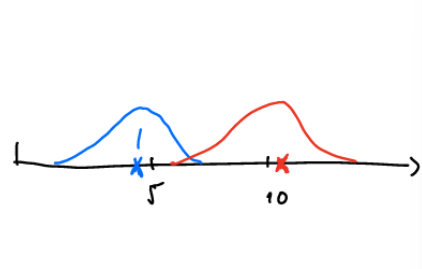
Question: Is Drug 1 better than Drug 2?

General idea

The idea is to consider the distributions of the estimates $\hat{\mu}_1, \hat{\mu}_2$.

If they are "far apart", we would say that they are different.





But how do we know what "far apart" is?

Example

We want to test whether a coin is fair.

Null hypothesis: H_0: Coin is fair.

Alternative hypothesis: H_1: Coin is unfair.

We sample many coin flips and estimate

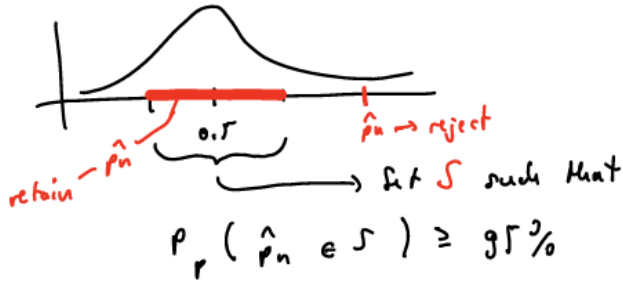
$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We want to reject H_0 if $\hat{p}_n$ is "far away" from 0.5.

Question: What does "far away" mean?

Example (2)

We look at the distribution of $\hat{p}_n$ under the null hypothesis.



If $\hat{p}_n$ falls outside this set, we reject H_0.

More formal setup

Statistical model:

$$\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}.$$

Assume that

$$\Theta_0 \subset \Theta, \quad \Theta_1 \subset \Theta, \quad \Theta_0 \cap \Theta_1 = \emptyset, \quad \Theta_0 \cup \Theta_1 = \Theta.$$

We want to test

$$H_0 : \theta \in \Theta_0 \quad \rightarrow \text{null hypothesis}$$

against

$$H_1 : \theta \in \Theta_1 \quad \rightarrow \text{alternative hypothesis}$$

More formal setup (2)

We sample data from the unknown f_θ and compute a test statistic $T(X_1, \ldots, X_n)$.

Now, we construct a rejection region R_n such that:

$$T(X_1, \ldots, X_n) \in R_n \quad \Rightarrow \quad \text{reject } H_0.$$

$$T(X_1, \ldots, X_n) \notin R_n \quad \Rightarrow \quad \text{retain } H_0.$$

More formal setup (3)

Typical hypotheses are of the form:

- $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta \geq \theta_0$.

Two types of errors can occur:

	Test retains H_0	Test rejects H_0
H_0 true	☺	Type I error
H_1 true	Type II error	☺

Role of H_0 vs. H_1

⚠ The role of the two hypotheses is not "symmetric".

It is like a legal trial: we believe H_0 unless there is strong evidence against it.

Thus, if a test does not reject H_0, it does not imply that H_0 is correct.

Power of a test, β

Definition 243 (Power of a test, β)

The power function of a test with rejection region R is the function

$$\beta(\theta) := P_\theta\left(T(X) \in R\right).$$

- If $\theta_0 \in \Theta_0$, then $T(X)$ should not end up in R. For such θ_0,

$$\beta(\theta_0) = P(\text{Type I error}).$$

  Ideally, $\beta(\theta_0)$ should be small.
- If $\theta_1 \in \Theta_1$, then we hope that $T(X) \in R$. So,

$$\beta(\theta_1) = 1 - P(\text{Type II error}).$$

  Here, $\beta(\theta_1)$ should be large.

Level of a test

Definition 244 (Level of a test)

We say that a test is of level α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

Intuition: Worst-case guarantee: no matter which $\theta \in \Theta_0$ we pick, the Type I error is not larger than α.

(Intuition to remember: $\alpha \mathrel{\hat{=}}$ Type I error.)

Power of a test

We always specify the power of a test against a fixed alternative parameter $\theta_1 \in \Theta_1$.

The power of the test against alternative θ_1 is given by:

$$1 - \beta(\theta_1).$$

(Intuition: $\beta \mathrel{\hat{=}} 1 - P(\text{Type II error}).$)

Standard approach for testing

Standard procedure: We fix the level α of a test in advance, for example 0.05 or 0.01.

Then we can also look at the Type II error. For example, among several tests of level α, you might choose the one that has the smallest Type-II error.

Uniformly most powerful test

Definition 245 (Uniformly most powerful test)

Let $\mathcal{T}$ be a set of tests of level α for testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0.$$

A test in $\mathcal{T}$ with power function $\beta(\theta)$ is uniformly most powerful (UMP) if

$$\beta(\theta) \geq \beta'(\theta) \quad \text{for all } \theta \in \Theta_0^c$$

and for all β' that are power functions for other tests in $\mathcal{T}$.

Remark: In practice, it is often impossible to find a UMP test.

Example: summer or winter?

Assume we want to test whether it is currently summer or winter.

Say, we typically believe it is summer unless evidence is against it:

$$H_0 = \{\text{summer}\}, \quad H_1 = \{\text{winter}\}.$$

Now we construct several tests based on the average temperature of the current day (one measurement).

Example: summer or winter? (2)

Test 1 Decision rule: Look at the temperature outside. If it is below $5°C$, reject H_0 and accept H_1.

Type-I error: How often does it get below $5°C$ in summer? Say, only in **5%** of all summer days.

$$P(\text{Test rejects } H_0 \mid H_0 \text{ is true}) = 5\%.$$

$\Rightarrow$ Good. ☺

Type-II error: How often does it get above $5°C$ in winter? Say, on **50%** of the days.

$$P(\text{Test rejects } H_1 \mid H_1 \text{ is true}) = 50\%.$$

$\Rightarrow$ Not so great. ☹

Example: summer or winter? (3)

Test 2: Maybe $0°C$ is a better threshold. Reject H_0 if $T \leq 0°C$.

- Reduces Type-I error.
- Still increases Type-II error.

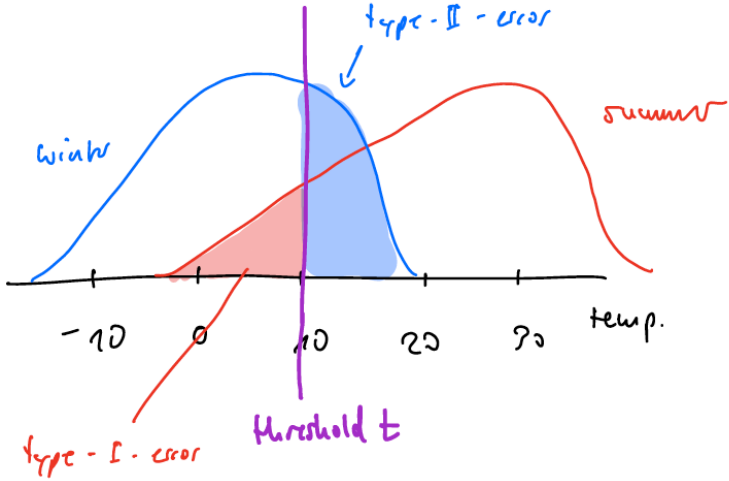
Test 3: Reject if $T \leq 10°C$.

- Increases Type-I error.
- Reduces Type-II error.

Conclusion: We cannot really construct a good test, because:

- Temperature distributions are overlapping.
- Only one measurement is used.

Example: summer or winter? (4)

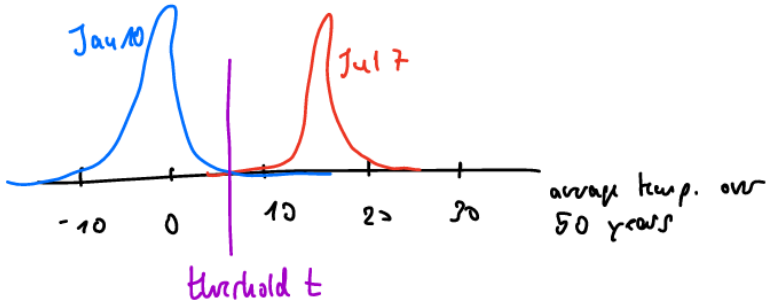


Example: summer or winter? (5)

Improving the test: To increase the power of a test, we need more sample points. They need to be independent!

For example, fix a date (e.g., July 7) and average the temperatures on this day over the last 50 years.

The same can be done for winter (e.g., January 10).



Example: summer or winter? (6)

Test 4: Alternative approach

Decision rule: Look outside the window. If somebody with skis is walking by, reject H_0 (summer) and accept H_1 (winter).

Type-I error: Say, **1%** (maybe someone just bought skis in summer).

$$P(\text{Test rejects } H_0 \mid H_0 \text{ is true}) = 1\%. \quad \text{☺}$$

Type-II error: **99%** or more! (Even if it is winter, I might not see skiers in front of my window.)

$$P(\text{Test rejects } H_1 \mid H_1 \text{ is true}) = 99\%.$$

→ This test has low power! ☹

Binomial example

We have 5 coins with $P(\text{"1"}) = \theta \in \{0, 1\}$.

We want to test

$$H_0 : \theta \leq \frac{1}{2} \quad \text{vs} \quad H_1 : \theta > \frac{1}{2}$$

Test 1

Decision rule: Reject H_0 if we observe 5 times "1":

$$1 \quad 1 \quad 1 \quad 1 \quad 1$$

Power function:

$$\beta(\theta) = P_\theta(\text{reject}) = \theta^5$$

Binomial example (2)

Test 2

Decision rule: Reject H_0 if we observe at least 3 times "1":

$$00111 \quad \text{or} \quad 01101 \quad \text{or} \quad \dots$$

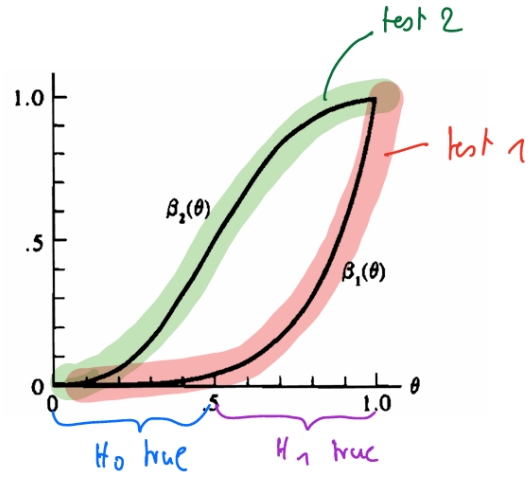
Power function:

$$\beta(\theta) = P_\theta(3, 4, 5 \text{ times } 1) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \theta^5$$

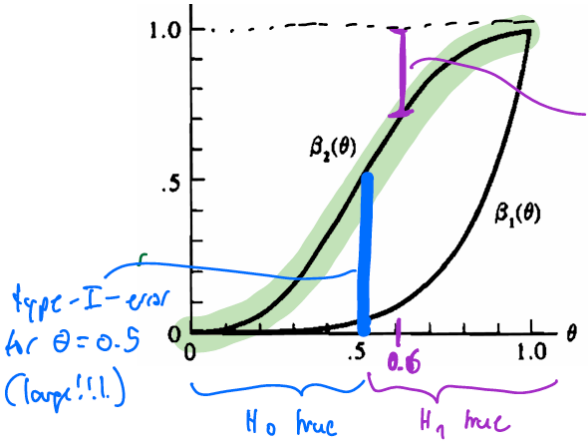
Power functions of both tests:

$$\beta(\theta) = P_\theta(\text{reject } H_0)$$

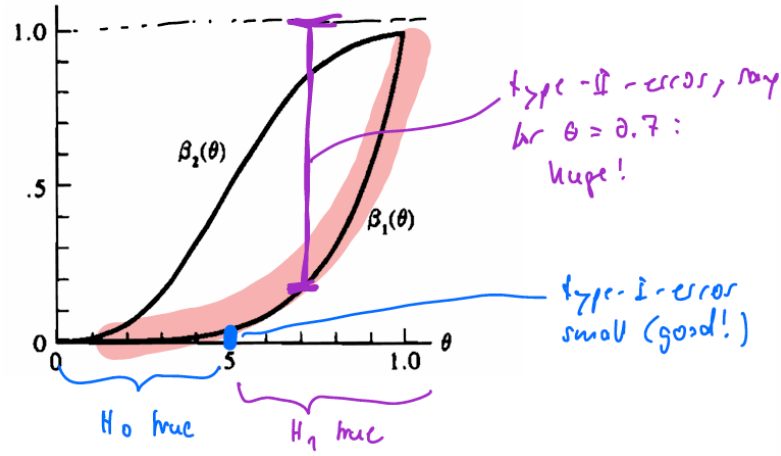
Binomial example (3)



Binomial example (4)

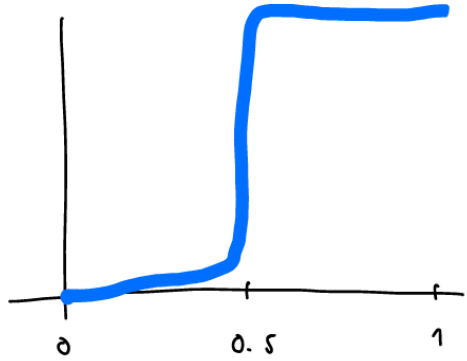


Binomial example (5)



Binomial example (6)

Would like to have such a power function:

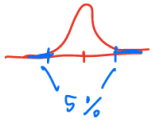


How could we achieve this? Increase the sample size...

z-test (wald test)

If the distribution of the test statistic is asymptotically normal:

$$\frac{\hat{\theta} - \theta_0}{\hat{\text{se}}} \longrightarrow N(0,1)$$



For the hypothesis:

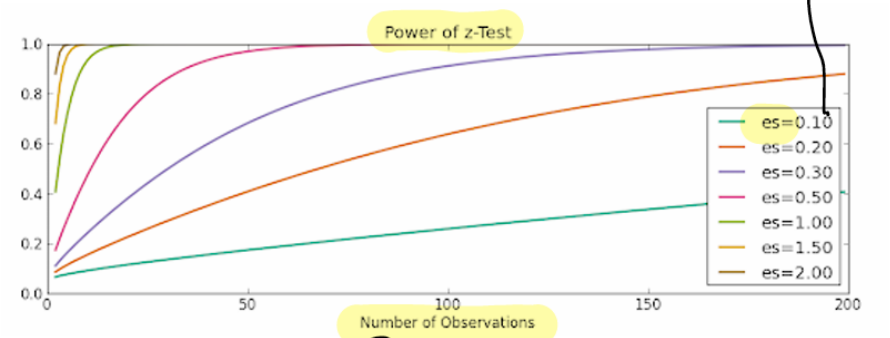
$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

Decision rule: Reject H_0 when:

$$\left| \frac{\hat{\theta} - \theta_0}{\hat{\text{se}}} \right| > z_{\alpha/2}$$

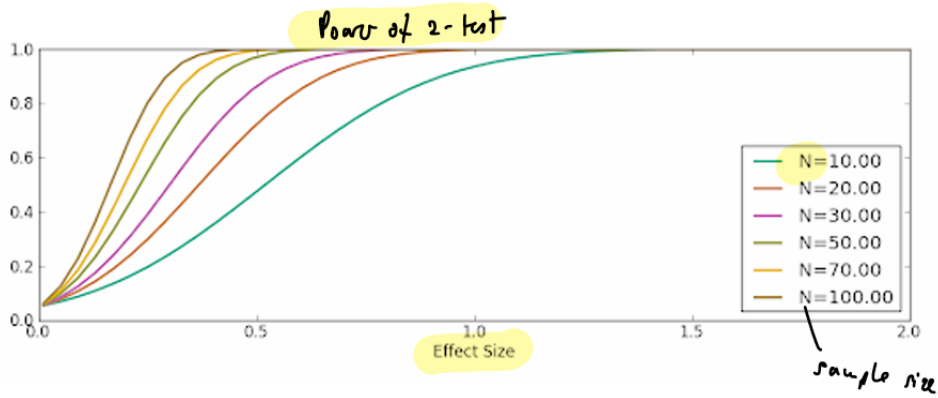
where $z_{\alpha/2}$ is the real number such that the normal cdf attains value $\alpha/2$.

z-test (wald test) (2)



Observation: The power of the test increases with the sample size.

z-test (wald test) (3)



Observation: The power of the test increases with effect size $|\theta - \theta_0|$.

Neyman - Pearson - Lemma & Likelihood ratio tests

Neyman - Pearson

Theorem 246 (Neyman - Pearson)

Suppose we test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Consider

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i \in \mathcal{I}} f(x_i|\theta_1)}{\prod_{i \in \mathcal{I}} f(x_i|\theta_0)} \qquad \rightarrow \text{Likelihood ratio}$$

Assume we reject H_0 if $T > k$ (for some k). If we choose k such that

$$P_{\theta_0}(T > k) = \alpha,$$

then this is the most powerful level-α-test.

More general likelihood-ratio test

Parameter space $\Theta_1, \Theta_0 \subset \Theta_1, \Theta_1 = \Theta_0^c$.

Then we consider the test statistic

$$\tilde{T} = \frac{\sup\limits_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup\limits_{\theta \in \Theta_1} \mathcal{L}(\theta)}$$

or even simpler

$$T = \frac{\sup\limits_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup\limits_{\theta \in \Theta} \mathcal{L}(\theta)}$$

and we determine a parameter λ such that the rejection region is of the form

$$R = \{T \leq \lambda\}.$$

More general likelihood-ratio test (2)

In practice the difficulties are

- compute the suprema (in practice)
- fix R, fix λ (in theory)

p-values

p-values

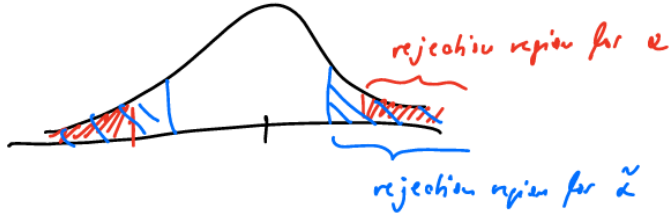
Consider a test at level α, and denote its rejection region as R_α.

Recall:

$$\alpha = P(\text{Type-I-Error}).$$

The smaller α, the more difficult does it get to reject H_0.

(We often even have that $\alpha < \tilde{\alpha} \Rightarrow R_\alpha \subset R_{\tilde{\alpha}}$)



p-values (2)

Definition 247 (p-value)

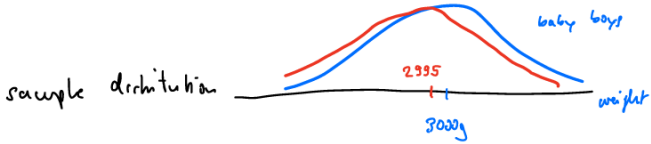
The p-value is defined as

$$p = \inf\{\alpha \mid T(x_1, \ldots, x_n) \in R_\alpha\}$$

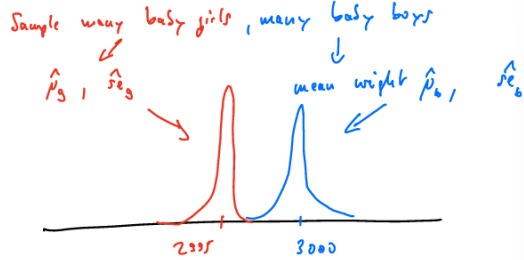
i.e., the smallest α for which the level-α-test would reject the null hypothesis.

Intuition: smaller p-values are "better", more evidence for rejecting the null.

Example: baby boys and girls



distribution of test statistics



for n large test will find a statistically significant difference. ⇒ small p

Statistically significant

Convention:

$$p \leq 0.05 \quad \rightarrow \quad \text{"significant"}$$

$$p \leq 0.01 \quad \rightarrow \quad \text{"highly significant"}$$

⚠ Even though an effect might be "statistically significant", it might not be "scientifically significant. For example, the effect size might be tiny.

Statistically significant (2)

⚠

- p-value does not give the probability that the null hypothesis is true (Θ is not a random variable in the frequentist setup).
- We can have small p-values, yet the "effect size" can be tiny.
- A large p-value is not necessarily strong evidence for H_0. It could just mean that our test has little power.

Multiple testing

Motivation

Example: gene expression data

m	patients with cancer ($n = 20$)	control group ($n = 20$)
gene 1		
gene 2		
⋮		
gene 17	0.5, 0.2, 0.9, 0.8, 0.5	0.01, 0.05, 0.1, 0.02
⋮		
gene 105		
⋮		
gene $\underbrace{1000}_{=:m}$		

$\rightarrow \alpha$ % of the tests will "ring a bell"

Motivation (2)

Assume we run, for each gene, a test of level α:

$$P(\text{Test i makes Type-I-Error}) = 5\%.$$

Now we have m tests.

Probability of at least one Type-I error

$$P(\text{at least one of the tests makes a Type-I-Error}) =$$

$$P(t_1 \text{ makes error or } t_2 \text{ makes error or } \ldots \text{ or } t_m \text{ makes error})$$

$$= 1 - P(\text{no error in } t_1 \text{ and no error in } t_2 \text{ and } \ldots)$$

$$= 1 - \prod_{i=1}^{m} P(\text{no error in } t_i) = \underbrace{1 - (0.95)^m}_{(\bigstar)} \underset{m\to\infty}{\longrightarrow} 1$$

Motivation (3)

- $m = 1 \Rightarrow (\bigstar) = 0.05$
- $m = 10 \Rightarrow (\bigstar) = 0.40$
- $m = 50 \Rightarrow (\bigstar) = 0.92$

Many "wrong" tests!

Family-wise error rate (FWER)

Definition 248 (Family-wise error rate (FWER))

Consider a family of m tests. The family-wise error rate (FWER) is the probability that at least one Type-I-Error occurs in the family:

$$\begin{aligned}\text{FWER} = P(&t_1 \text{ makes Type-I-Error}\\ &\text{or } t_2 \text{ makes Type-I-Error}\\ &\text{or } \ldots\\ &\text{or } t_m \text{ makes Type-I-Error})\end{aligned}$$

Bonferroni correction

Assume we run m tests, and we want to achieve a FWER α (e.g., $\alpha = 0.05$). Then we run the individual tests with level

$$\frac{\alpha}{m} =: \alpha_{\text{single}}.$$

Then

$$\text{FWER} = P(\text{at least one Type-I-Error})$$

$$= P(t_1 \text{ error or } t_2 \text{ error or } \dots)$$

$$\leq \sum_{i=1}^{m} \underbrace{P(t_i \text{ makes error})}_{\alpha_{\text{single}}} = m \cdot \alpha_{\text{single}} = m \cdot \frac{\alpha}{m} = \alpha.$$

Bonferroni discussion

Bonferroni controls the FWER.

Advantage: simple, correct.

Disadvantage: too conservative, low power (high Type-II-Error).
$\rightarrow$ The test barely discovers anything!

False discovery rate, FDR

Definition 249 (False discovery rate, FDR)

Assume we have a family of m tests. We call

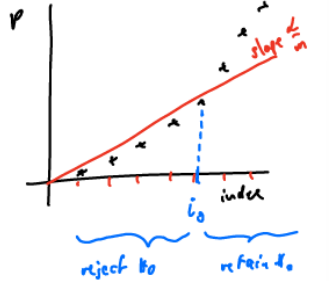
$$\mathbb{E}\left(\frac{\#\text{ false rejections}}{\#\text{ all rejections}}\right) =: \text{FDR}$$

the false discovery rate.

Benjamini/Hochberg: Controlling FDR.

Benjamini/Hochberg (1998) approach

- Fix FDR α in advance.
- Run the m individual tests and evaluate their p-values.
- Sort p-values increasingly: $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \cdots \leq p_{(m)}$
- Define thresholds $l_i := i \cdot \frac{\alpha}{m}$
- Find the largest index i_0 such that $p_{(i_0)} \leq l_{i_0}$ (below the red line)
- Reject the hypotheses for $i = 1, \ldots, i_0$, retain all the others.



Theorem (Benjamini-Hochberg)

Theorem 250 (Benjamini-Hochberg)

If the Benjamini-Hochberg procedure is applied (and the tests are independent), then regardless of how many null hypotheses are true and regardless of the distribution of p-values where the null is false, we obtain

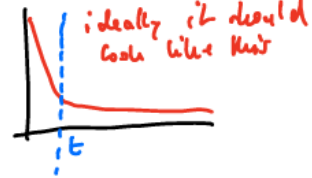
$$\text{FDR} \leq \alpha.$$

Remark: Similar approach also works without independence assumption, many modifications exist.

Intuition

- Under the null hypothesis, the p-values always have a uniform distribution on $[0, 1]$.

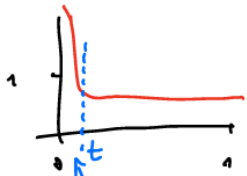
density of p-values under H_0



density of p-values under H_1

Intuition (2)

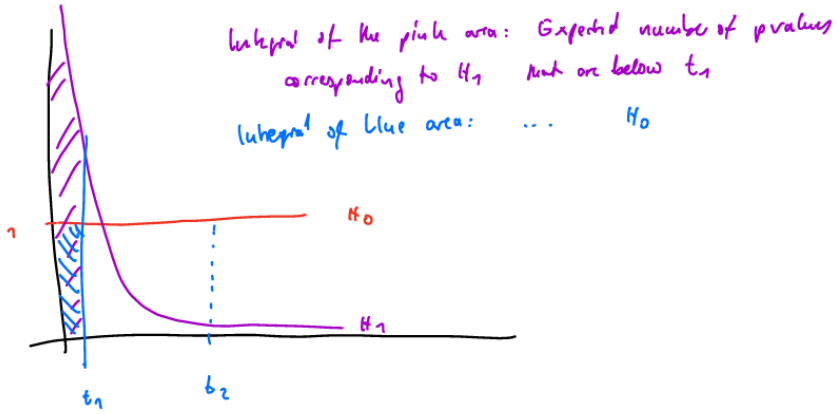
- If we have some H_0 and some H_1 being true, then the density would maybe look like this:



here we have (hopefully) many of the H_1s but we also have some H_0s.

Goal: set threshold t such that FDR satisfies what we want.

Intuition (3)



By varying t from 0 to 1 we control the FDR:

- For t_1, the FDR is small.
- For t_2, the FDR is large.

General remarks

- BH tends to have more power than Bonferroni.
- BH controls FDR, not FWER (overall Type-I-Error)!
- BH works best in sparse regimes where only few tests reject the null.
- BH gives guarantees on FDR, but in general does not minimize it.
- When all the H_0 are true, BH $\approx$ Bonferroni.

Non-parametric tests

Non-parametric tests

Standard (parametric scenario):

- Statistical model: $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$

- Observe data, compute a test statistic, for example the mean $\bar{X}$.
- Need to know the distribution of the test statistic T_n under the null distribution.

Goodness-of-fit test (gof)

Goodness-of-fit tests:

Goal is to test whether a data set comes from a particular distribution F_0.

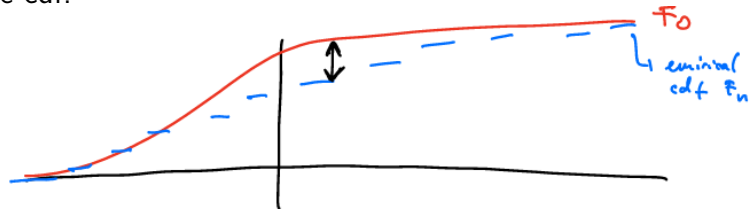
$$H_0 : F(x) = F_0(x)$$

$$H_1 : F(x) \neq F_0(x),$$

where $F(x)$ is the true distribution that generated the data.

Kolmogorov-Smirnov test for gof

We consider the cdf:



- $F_0 =$ cdf of the given distribution
- $F_n =$ empirical cdf of the data

Kolmogorov-Smirnov test for gof (2)

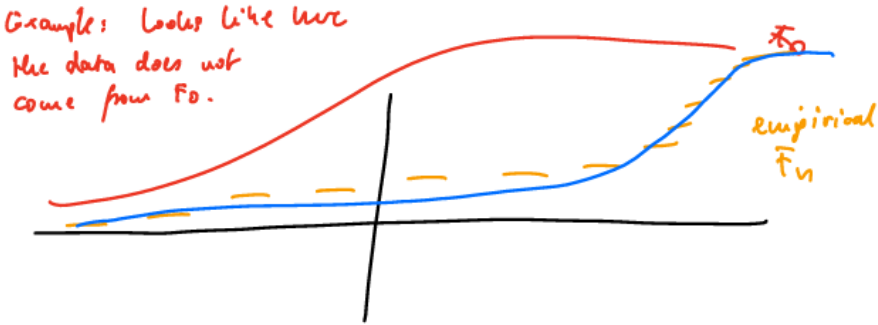
$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

By the Glivenko-Cantelli theorem, we know that under the null hypothesis, $F_n \longrightarrow F_0$ uniformly a.s.

It is possible to compute the distribution of D_n, and it is independent of F_0, it just depends on n.

From this, we can compute rejection thresholds and design a test.

Kolmogorov-Smirnov test for gof (3)



Two sample test

Two sample test:

$X_1, \ldots, X_n \sim F_1$ a first sample distributed according to F_1,

$Y_1, \ldots, Y_m \sim F_2$ a second sample distributed according to F_2.

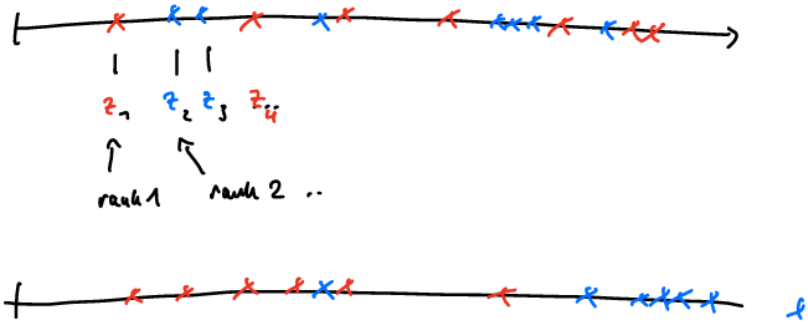
Question: $F_1 = F_2$?

$$H_0 : F_1 = F_2 \quad , \quad H_1 : F_1 \neq F_2$$

Wilcoxon-Mann-Whitney test (based on ranks)

Test:

- "Pool the sample": $X_1, \dots, X_n, Y_1, \dots, Y_m \in \mathbb{R}$.
- Sort the pooled sample in increasing order and retrieve the rank of all points $\rightarrow$ rank(X_i), rank(Y_i).



Wilcoxon-Mann-Whitney test (based on ranks) (2)

Compute the rank sums for both groups:

$$W_{\text{red}} = \sum_{i \in \text{red population}} \text{rank}(X_i)$$

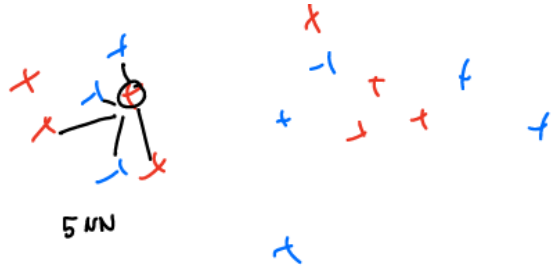
$$W_{\text{blue}} = \sum_{i \in \text{blue population}} \text{rank}(Y_i)$$

Decision:

- If $|W_{\text{red}} - W_{\text{blue}}|$ is small, we retain H_0.
- If $|W_{\text{red}} - W_{\text{blue}}|$ is large, we reject H_0.

Extension to a multivariate setting using k-nearest neighbors

- Two samples, we pool them:



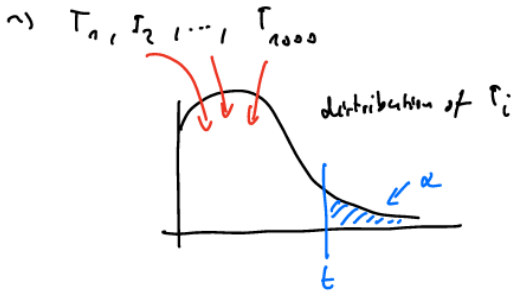
- For each point, we look at the colors of the k nearest neighbors.
- Under the null hypothesis, we expect that the number of red neighbors $\approx$ number of blue neighbors.

Permutation (randomization) tests

- Sample $X_1, \ldots, X_n$ (group A) and $Y_1, \ldots, Y_n$ (group B).
- Compute observed statistic $T_{\text{observed}} = \text{mean}(\text{red}) - \text{mean}(\text{blue})$.
- Pool the sample.
- For $k = 1, \ldots, 10^3$: shuffle the group membership ("colors").
- Compute the difference

$$T_k = \text{mean}(\text{red}) - \text{mean}(\text{blue}).$$

Permutation (randomization) tests (2)



- Find α-quantile to determine rejection threshold.
- Check whether the observed T_{observed} on the true data is $\leq t$.

Bootstrap test

Motivation

Given $X_1, \ldots, X_n \sim F$, with no knowledge on F, we want to estimate a parameter $\theta = t(F)$.

We generate an estimate $\hat{\theta}$ based on $X_1, \ldots, X_n$ and want to know how reliable $\hat{\theta}$ is.

The first thing to look at is the standard error se.

- If we have assumptions on F, we can empirically compute the distribution of $\hat{\theta}$, the $\hat{se}$, ... (this is rare!).

Motivation (2)

- We could try to obtain many samples:

$$X_1^{(1)}, \ldots, X_n^{(1)}$$

$$X_1^{(2)}, \ldots, X_n^{(2)}$$

$$\vdots$$

$$X_1^{(M)}, \ldots, X_n^{(M)}$$

and then estimate the distribution of $\hat{\theta}$:

Problem: need too many samples.

Then we could maybe build a test on this.

Idea of the Bootstrap

- Given the sample $X_1, \ldots, X_n$, estimate $\hat{\theta}_{\text{orig}}$.
- Draw a subsample of $X_1, \ldots, X_n$, compute $\hat{\theta}^*$, repeat very often.

"Hope": The histogram (3) of $\hat{\theta}^*$ is "close" to the histogram of $\hat{\theta}$ (2), which is close to the true distribution (1).

Example: Estimate the standard error of an estimate $\hat{\theta}$.

Algorithm in pseudo code

Input: $X_1, \ldots, X_n \rightarrow n =$ number of original sample points. $\rightarrow B =$ number of bootstrap replications.

Procedure:

- For $b = 1, \ldots, B$:
  - Sample $X_1^*, \ldots, X_n^*$ uniformly with replacement from $X_1, \ldots, X_n$.
  - Estimate the parameter $\hat{\theta}_b^*$.

Output: Gives us $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$

Algorithm in pseudo code (2)

Estimate the standard error $\hat{se}$ of the original estimate $\hat{\theta}$ by the standard deviation of the bootstrap replicates:

$$\hat{se}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \left(\underbrace{\frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*}_{\text{mean of replicates}} \right) \right)^2 \right)^{1/2}$$

Does it always work?

Consistency result for bootstrap

Theorem 251 (Consistency result for bootstrap)

- Assume that $X_1, \ldots, X_n \sim F$, i.i.d., and

$$\mathbb{E}(\|X_1\|^2) < \infty.$$

- Let $\hat{\theta} = g(X_1, \ldots, X_n)$ be the parameter that we estimate. Assume that g is continuously differentiable in a neighborhood of $\mu = \mathbb{E}X_1$, with a non-zero gradient.

Then the bootstrap estimate of the standard error is strongly consistent.

Example where it goes wrong

- $X_1, \ldots, X_n \sim \text{Uniform}[0, \theta]$, where $\theta \in [0, 1]$ is unknown.
- Want to estimate θ. The MLE estimate of θ is simply the largest number we observe:

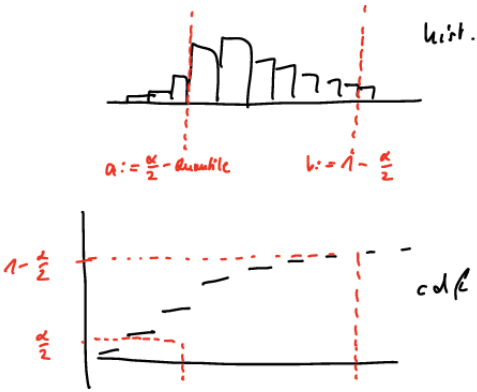
$$\hat{\theta} = \max_{i=1,\ldots,n} X_i.$$

- Estimating the se by bootstrap is going to fail.
- Estimating tails or extreme values by bootstrap is problematic.

Confidence sets by Bootstrap

Bootstrap-percentile-method:

- Given sample $X_1, \ldots, X_n$, estimate $\hat{\theta}$.
- Generate bootstrap replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
- Look at the histogram of the $\hat{\theta}_b^*$.



Confidence sets by Bootstrap (2)

Confidence Interval:

$$CI = [a, b]$$

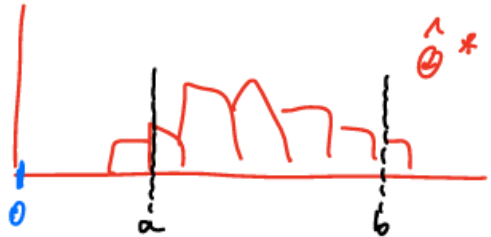
It has coverage $1 - \alpha$ because

$$P_\theta \left(\hat{\theta} \in CI \right) \geq 1 - \alpha.$$

Remark: Approximately, because n, B finite.

Conclusion: Subsequently, you can construct bootstrap tests in the obvious way.

Confidence sets by Bootstrap (3)



$H_0: \hat{\theta} = 0 \quad \text{vs.} \quad H_1: \hat{\theta} \neq 0$

Bayesian statistics

Frequentist vs. Bayesian statistics

Frequentist statistics:

- Probability = limiting frequency.
- Parameters θ are constants, we cannot assign probabilities to them.
- Statistics behave well when repeated often.

Bayesian statistics:

- Probability = degree of belief.
- Parameters do have probabilities.
- Have a prior belief about the world, update it based on observed data.

Bayesian statistics: the model

Assume a statistical model $\{f_\theta \mid \theta \in \Theta\}$, as in the frequentist approach.

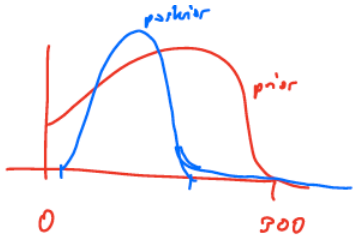
It encodes our prior assumptions on the data-generating process in general.

θ is unknown, and we want to estimate it.

Bayesian approach: prior distribution

We assume that we have a prior belief about the parameters θ:

$f(\theta)$ prior distribution



Bayesian statistics: The likelihood

Observe data $X_1, \ldots, X_n$ i.i.d. from some f_θ (θ unknown).

We call

$$f(X \mid \theta) \quad \text{the likelihood of the data given the parameter } \theta,$$

where:

- f is the density,
- X is the data
- θ the parameter

(In the frequentist world, we could now use MLE to select the parameter that maximizes the likelihood.)

Bayesian statistics: posterior

Now we update our belief and compute the posterior using Bayes' rule:

$$\underbrace{f(\theta \mid X_1, \ldots, X_n)}_{\text{posterior}} = \frac{f(X_1, \ldots, X_n \mid \theta) \cdot f(\theta)}{\int f(X_1, \ldots, X_n \mid \theta) f(\theta) d\theta}$$

- Likelihood: $f(X_1, \ldots, X_n \mid \theta)$.
- Prior: $f(\theta)$.
- Normalizing constant: denominator, does not depend on θ anymore.

$\rightarrow$ The posterior is a distribution.

Statistics derived from posterior

Now you can make statements based on the posterior.

- If you want to return one "best guess" for θ, you could use:
  - max of posterior (MAP)
  - mean of posterior

- You can construct confidence intervals:

Find a, b such that

$$P(\theta \in [a, b]) = 95\%.$$

Discussion

Advantages:

- Easy to interpret.
- Natural way to incorporate prior knowledge.

Disadvantages:

- Analytic solutions are rare; typically, you have to solve computationally hard problems.
- Need to choose a prior.

High-dimensional probability and statistics

Literature:

Books:

- Wainwright: High-dimensional statistics.
- Vershynin: High-dimensional probability.

Papers:

- PCA in high dimensions (Johnstone, Paul, 2018).
- Practice, Theory and Theorems for random matrix theory in ML (Mahoney, 2022).

High-dimensional statistics is different!

General setting:

- High-dimensional data (images, feature-based, genetic data, ...).
  dimension d is easily 100 or 1000.
- In some applications, one also has very few data points (e.g., medical case: $n = 20$ patients, $d = 1000$ genetic markers).
- ML models tend to have many parameters, m.

Classic vs. high-dimensional statistics

"Classic statistics:" d fixed, $n \to \infty$: LLN, CLT, ...

"Modern regime:" $d \to \infty$, $n \to \infty$.

Different ways to model this, the question is always how to "couple" d and n and in which way.

Example: $d/n \to$ const.

Things behave very differently and unexpectedly in this regime.

Concentration phenomena in high-dimensional spaces

Norms of high-dimensional vectors are concentrated

Consider $X_1, \dots, X_d$ independent, $\mathbb{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$, and let $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$

Expected norm of X:

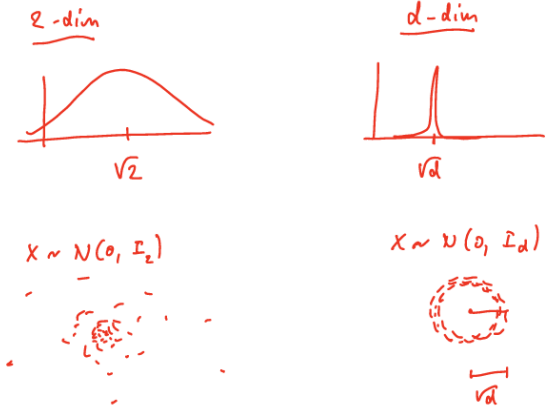
$$\mathbb{E}(\|X\|^2) = \mathbb{E}\left(\sum_{i=1}^{d} X_i^2\right) = \sum_{i=1}^{d} \underbrace{\mathbb{E}X_i^2}_{n} = d$$

Concentration of the norm: By Bernstein inequality, we can prove:

$$P(|\|X\| - \sqrt{d}| > t) \leq 2\exp(-ct^2)$$

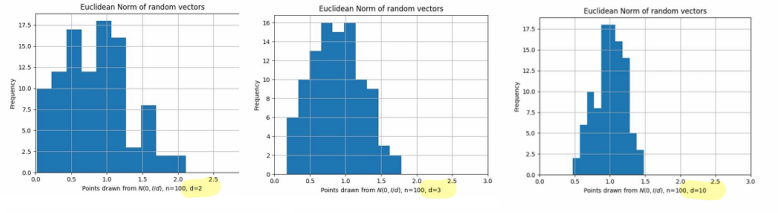
Norms of high-dimensional vectors are concentrated (2)

Illustration:



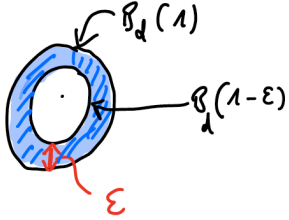
Norms of high-dimensional vectors are concentrated (3)

Simulations: points drawn from d-dim normal distribution $\mathcal{N}\left(0, \frac{1}{d} \cdot I_d\right)$ in increasing dimensions $(d = 2, 3, 10, 100, 1000)$



Volume of balls is concentrated along "surface"

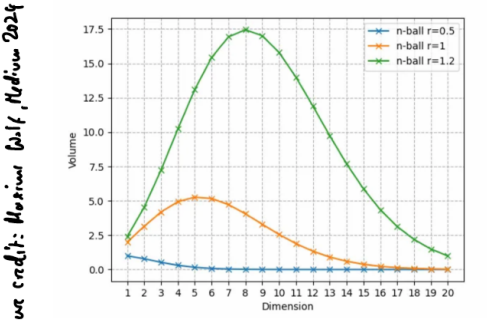
Concentration can be seen from the point of view of sampling (previous slides), but also from the point of view of geometry:

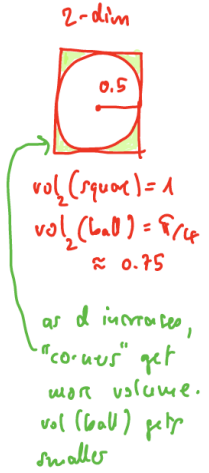


$$
\begin{aligned}
&\text{vol}\left(B_d(1) - B_d(1-\varepsilon)\right) \\
&= 1^d \cdot \text{vol}(B_d(1)) - (1-\varepsilon)^d \cdot \text{vol}(B_d(1)) \\
&= \underbrace{\left(1 - (1-\varepsilon)^d\right)}_{\longrightarrow 1 \text{ as } d \longrightarrow \infty} \cdot \text{vol}(B_d(1))
\end{aligned}
$$

Volume of unit balls gets tiny

The Euclidean volume of the unit ball in $\mathbb{R}^d$ goes to 0 as $d \to \infty$.



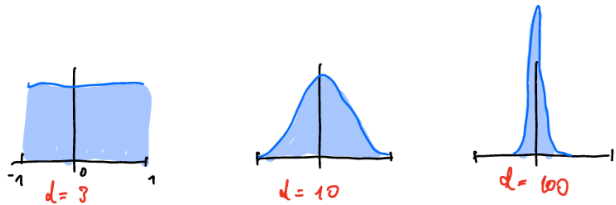


Random vectors on the sphere are nearly orthongonal

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} \quad \text{drawn uniformly from sphere}$$

$$\mathbb{E}\left(|\langle x, y\rangle|\right) = \frac{1}{d} \quad (\approx 0 \text{ if } d \text{ is large!})$$

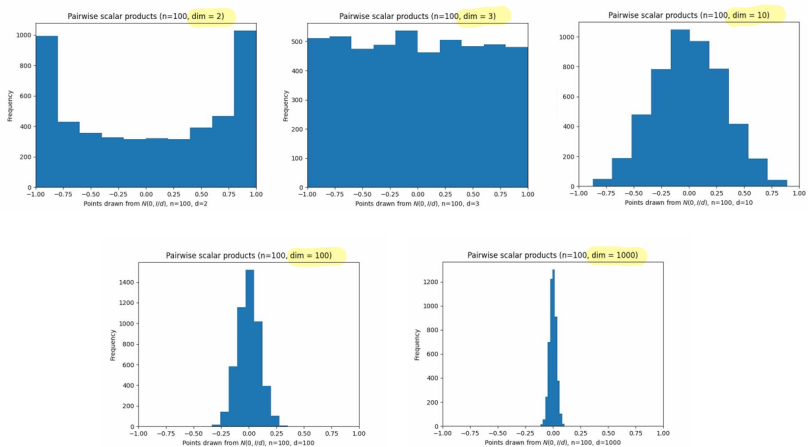
Distribution of $\langle x, y\rangle$ for two independent random vectors on the sphere:



$\rightarrow$ PCA gets problematic, see later.

Random vectors on the sphere are nearly orthongonal (2)

Simulations: points drawn from d-dim normal distribution $\mathcal{N}\left(0, \frac{1}{d} \cdot I_d\right)$ in increasing dimensions $(d = 2, 3, 10, 100, 1000)$



Distances in high-dimensional spaces

If we are lucky... pairwise distances are still meaningful
$\rightarrow$ Johnson-Lindenstrauss

Can work "OK" if we are in a low dimension.

Theorem 252 (Johnson-Lindenstrauss)

Consider n points in $\mathbb{R}^d$. One can find a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ with $k \approx \log n/\varepsilon^2$ such that the distances do not change by more than ε:

$$(1-\varepsilon)\|x_i - x_j\| \leq \|f(x_i) - f(x_j)\| \leq (1+\varepsilon)\|x_i - x_j\|$$

$\rightarrow$ "random projections"

But often we are not lucky: random projections look like (simple) Gaussians

$X \sim \text{Unif}(S_{d-1}\sqrt{d})$. Project X on a fixed vector v:

In practice, this often also holds for non-uniform data.

We don't see any class or cluster structure!

Distances in high-dim spaces are dominated by noise

Let

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2 I_d) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_2, \sigma_2^2 I_d)$$

Then

$$\begin{aligned}
\mathbb{E}\left(\|X - Y\|^2\right) &= \mathbb{E}\left(\sum_{i=1}^{d} |X_i - Y_i|^2\right) \\
&= \sum_{i=1}^{d} \left(\mathrm{Var}(X_i - Y_i) + \left(\mathbb{E}(X_i - Y_i)\right)^2\right) \\
&= \underbrace{d \cdot (\sigma_1^2 + \sigma_2^2)}_{\text{noise}} \quad + \quad \underbrace{\|\mu_1 - \mu_2\|^2}_{\text{signal}}
\end{aligned}$$

Distances in high-dim spaces are dominated by noise (2)

→ Classification is difficult!

Within-class-distances $\ll$ between-class-distances?!

Mixture of Gaussians with different variances:

$$X, X' \sim \mathcal{N}(\mu_1, \sigma_1^2 I_d), \quad Y, Y' \sim \mathcal{N}(\mu_2, \sigma_2^2 I_d).$$

Within-class distance:

$$\mathbb{E}(\|X - X'\|^2) = 2d\sigma_1^2 \quad \text{(class 1)}$$

$$\mathbb{E}(\|Y - Y'\|^2) = 2d\sigma_2^2 \quad \text{(class 2)}$$

Within-class-distances $\ll$ between-class-distances?! (2)

Between-class distance:

$$\mathbb{E}(\|X - Y\|^2) = \|\mu_1 - \mu_2\|^2 + d(\sigma_1^2 + \sigma_2^2).$$

If $\sigma_1 \ll \sigma_2$, then $\underbrace{d \cdot (\sigma_1^2 + \sigma_2^2)}_{\text{between}} \ll \underbrace{d \cdot 2\sigma_2^2}_{\text{within-class-2}}$

$\rightarrow$ clustering is difficult!

ML Consequences

- Different classes overlap heavily and are hard to distinguish based on individual distances.

- Might hope for assumptions that tell us that actually our space does not have such a high dimension:
  - sparsity ($\rightarrow$ book by Wainwright)
  - manifold assumption

Let's take a look at matrices...

Estimating covariance matrices in high dimension

Assume we have n data points in d dimensions $\Rightarrow n \cdot d$ "measurements.
Covariance matrix has d^2 entries ... obvious trouble if $n \ll d$.
The default estimate for the covariance matrix, the sample cov. matrix:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i'$$

might not converge to the true underlying cov. matrix.

If n is low and d is high, (n/d low), the entries of the sample covariance matrix are unreliable and do not converge to the true values.
$\rightarrow$ If $n \ll d$, there is no way to fix that!!!

Estimating covariance matrices in high dimension (2)

Alternatively, if $d \gg n$, one needs to make structural assumptions, for example:

- Spiked covariance model where the "signal" just lives in few data dimensions
- Sparsity assumptions
- Block-matrix assumptions
- ...

Eigenvalues in high dimensions

Consider a $n \times d$-matrix with iid random entries (e.g., Gaussian $\mathcal{N}(0,1)$): n points, d dimensions each.

The true covariance matrix is the identity.

Can we use the sample cov. matrix to estimate the eigenvalues of the true cov. matrix? (d numbers to estimate)

Different things happen, depending on the ratio $\frac{d}{n}$:

Eigenvalues in high dimensions (2)

If $d \ll n$, then eigenvalues concentrate:

Theorem 253 (Concentration theorem (Vershynin book Thm. 4.7.1))

Let x subgaussian vector in $\mathbb{R}^d$, denote by Σ the true covariance matrix, and by Σ_n the empirical covariance matrix based on n sample points. Then

$$\mathbb{E}\left(\|\Sigma_n - \Sigma\|\right) \leq \text{const} \cdot \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) \|\Sigma\|$$

with the norm being the ℓ_2-operator norm (i.e. distance between largest eigenvalues). If $\frac{d}{n} \to$ const, then bound does not vanish any more!

Subgaussian: $\mathbb{P}(|X| > t) \leq 2\exp(-t^2 \cdot \text{const})$

Eigenvalues in high dimensions (3)

If $d/n \to$ const, then the distribution of eigenvalues is known:

Consider random vectors whose entries are iid with variance 1.

Theorem 254 (The Marchenko–Pastur theorem)

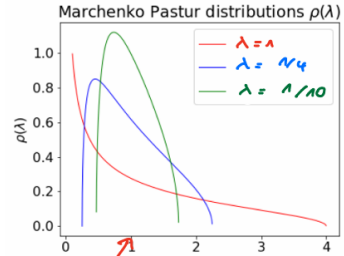
The Marchenko–Pastur theorem characterizes the distribution of eigenvalues of the sample covariance matrix as $\frac{d}{n} \to \lambda$:

(It has a closed-form expression that I did not want to put on the slide.)

General behavior: The larger λ, the more the eigenvalues "spread", see next slide.

Eigenvalues in high dimensions (4)

$\lambda = \frac{d}{n}$



True cov. matrix is I_d, so all true eigs are 1.

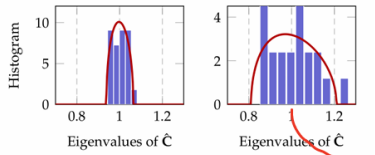


Figure by M. Mahoney: empirical eigs vs. limit distribution for a matrix with standard normal entries

$d = 20$
$n = 1000$

$d = 20$
$n = 100$

true eigenvalue is 1

Eigenvectors (PCA) in high dimension

PCA on high-dimension data is problematic, to say the least.

Even the leading eigenvector is typically inconsistent!!!

Why should we care?

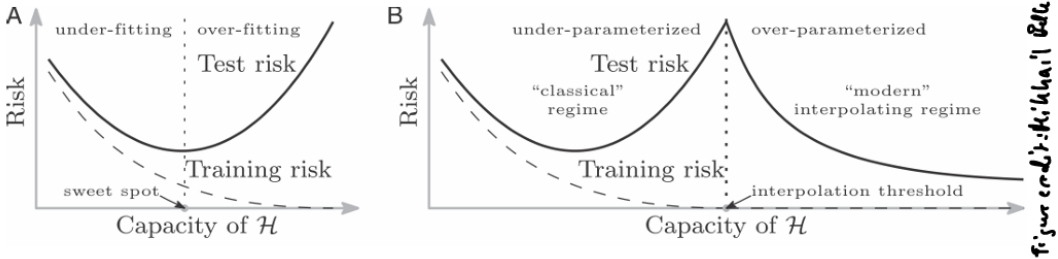
Weights in a DNN can behave funny aw well if you try to analyze their statistics

(see for example papers by M. Mahoney)

Outlook: double descent

Finally, let's look at "model size" (in terms of the number of parameters) as well.

Relation between model size in ML and overfitting:



Long debate ... see statistical ML lecture next term ...