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...

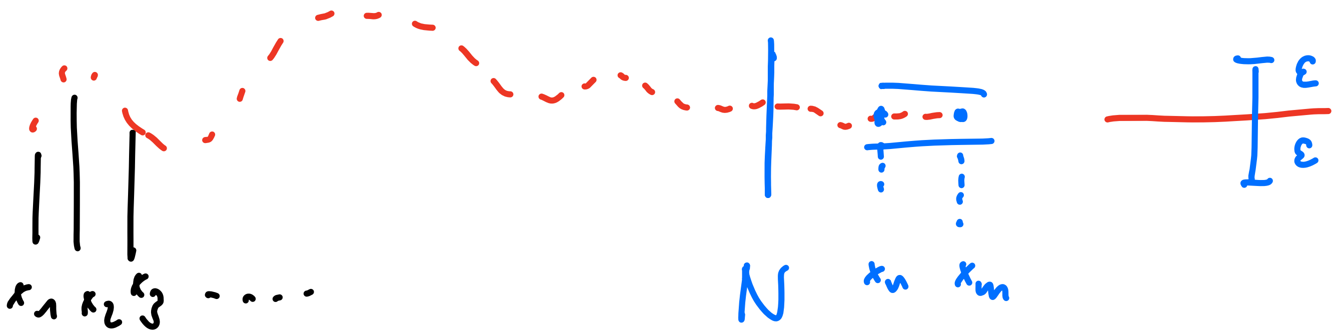
Sequences and convergence

ML keyword: convergence of a learning algorithm!

Cauchy sequence

Def: $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^d$ is called a <u>Cauchy sequence</u> if

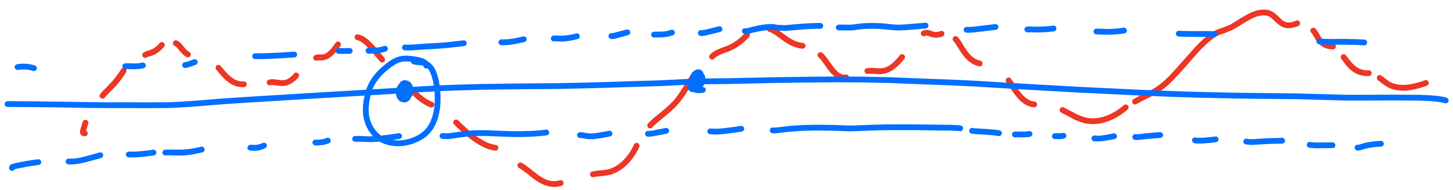
$$\forall \varepsilon > 0 \;\; \exists N \in \mathbb{N} \;\; \forall n,m > N : \; |x_n - x_m| < \varepsilon$$



Accumulation point

A point $x \in \mathbb{R}$ is called a 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

$$\forall \, \varepsilon > 0 \;\; \exists \, N \;\; \forall \, n > N : \;\; |x_n - x| < \varepsilon$$

<u>Notation</u>: $\lim_{n \to \infty} x_n = x \;, \quad x_n \xrightarrow[n \to \infty]{} x$

First observations

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

Example

- $x_n = \frac{1}{n}$ 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 $x_n = \begin{cases} \frac{1}{n} & \text{if } n \text{ even} \\ n & \text{if } n \text{ odd} \end{cases}$

  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, min, 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 $n \in \mathbb{N}$.

<u>Theorem</u> (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 $(x_n)_n$.

  liminf smallest

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

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

Basic concepts in topology

Open and closed sets

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

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

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

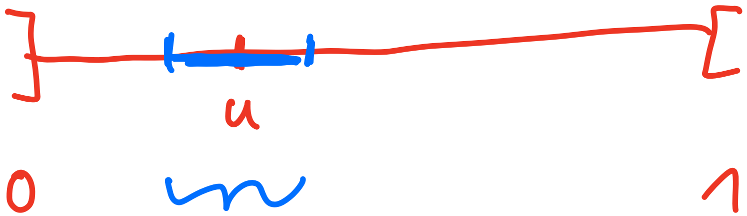
A set $U \subset X$ is called <u>open</u> if

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

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

Examples

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



$B =]u-\varepsilon, u+\varepsilon[$

- A set U can be neither open nor closed:

$[0,1[$

Open vs. closed

<u>Proposition</u>: Complements of open sets are closed.
Complements of closed sets are open.

Interior, closure

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

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

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

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

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

literature not always consistent here ... sometimes one also reads $U \setminus U^{\circ}$ instead of $\bar{U} \setminus U^{\circ}$.

$$X = [0,1[$$

$$\overline{X} = [0,1]$$

$$X^{\circ} =]0,1[$$

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

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

Def A set $U \subset X$ is <u>bounded</u> if there exists $D > 0$ such that $\forall u, v \in U$, $d(u,v) < D$

Dense sets

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

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

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

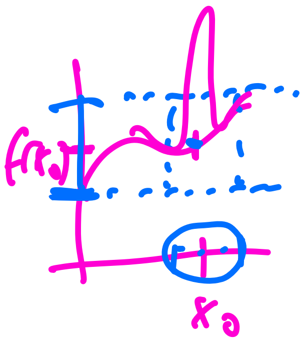
ML keyword: Can we approximate underlying target fct f by the fcts fn that can be constructed by our learning alg?

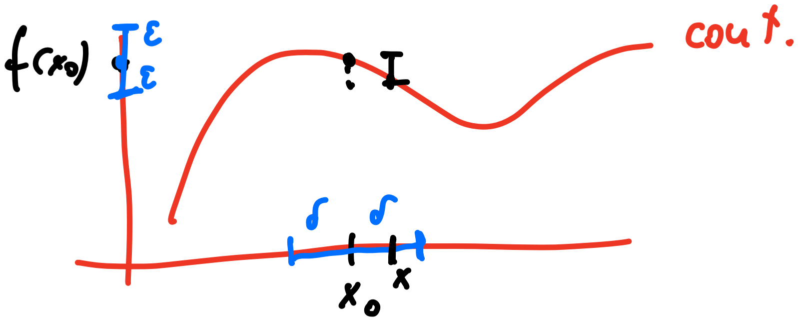
Continuity

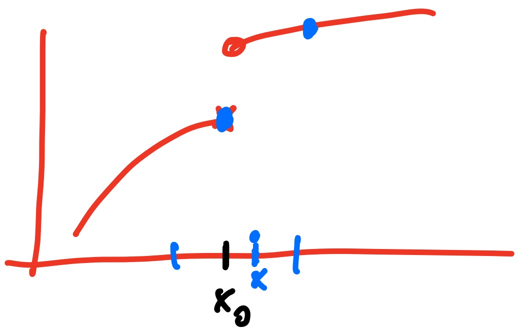
Continuous function

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

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







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

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

Alternative definitions

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

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

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

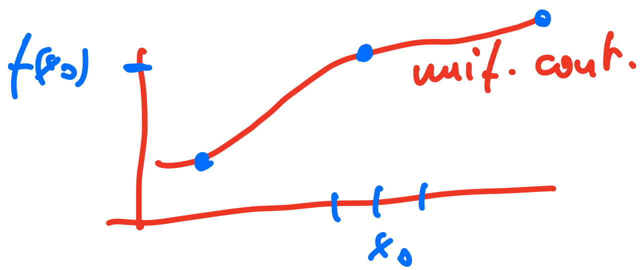
$$\underset{\text{in } Y}{B \subset Y \text{ open}} \quad \Rightarrow \quad f^{-1}(B) := \{x \in X \mid f(x) \in B\} \ \underset{\text{in } X}{\text{open}}$$

$$:= \{x \in X \mid f(x) \in B\}$$

Uniformly continuous

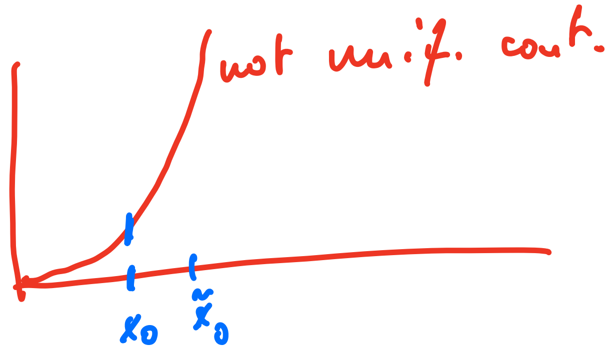
A function $f: X \to Y$ is called uniformly continuous if

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



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

- $f:]0,1] \to \mathbb{R}$, $f(x) = 1/x$ 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 <u>not</u> be uniformly cont. if it were defined on all of $\mathbb{R}$.

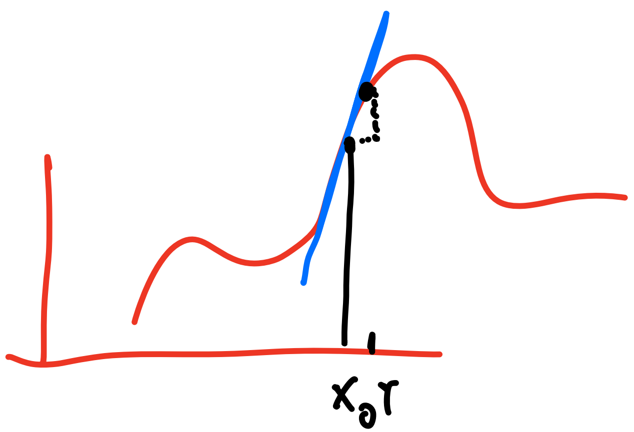
- <u>Proposition</u>: Let $f: [a,b] \to \mathbb{R}$ be continuous. Then it is already uniformly continuous.

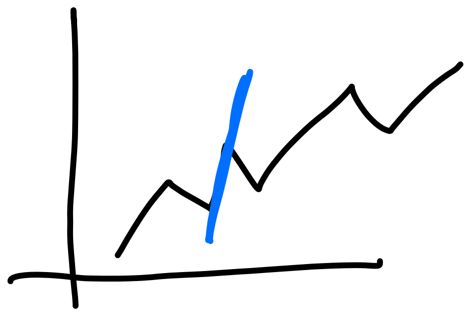
Lipschitz continuous

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

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

Intuition: "bounded derivative"





Lipschitz cont. $\Longrightarrow$ uniformly cont
$\not\Longleftarrow$

Proposition: f Lipschitz continuous $\Rightarrow$ f uniformly continuous.

Proof: easy

⚠ Note that the other way round is not true:

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

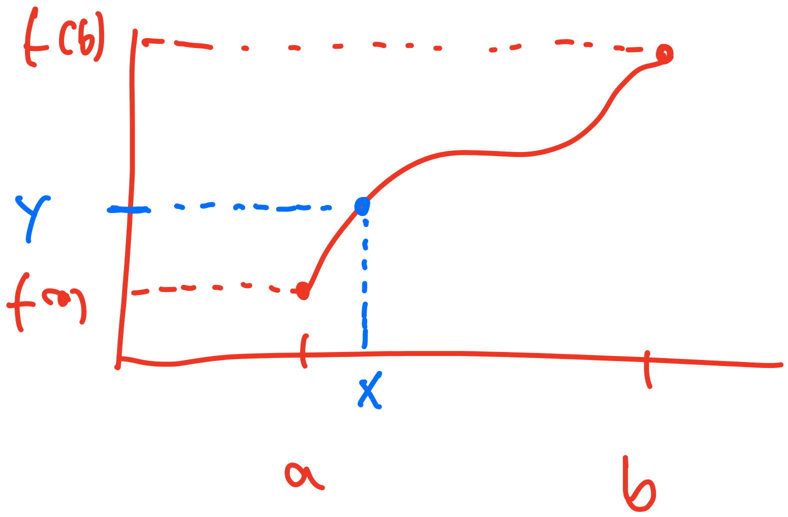
is uniformly continuous, but not Lipschitz continuous.

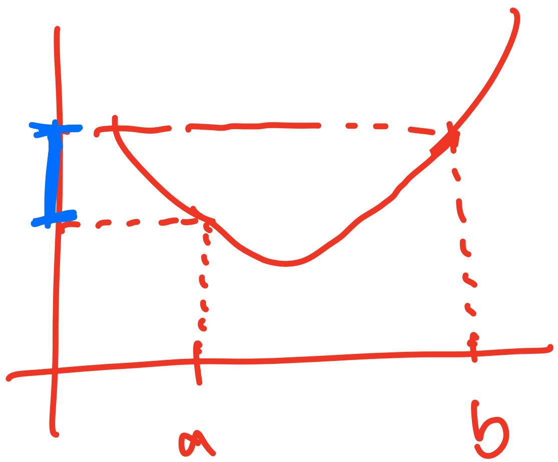
Intermediate value theorem

Theorem:

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

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





Application: finding zero

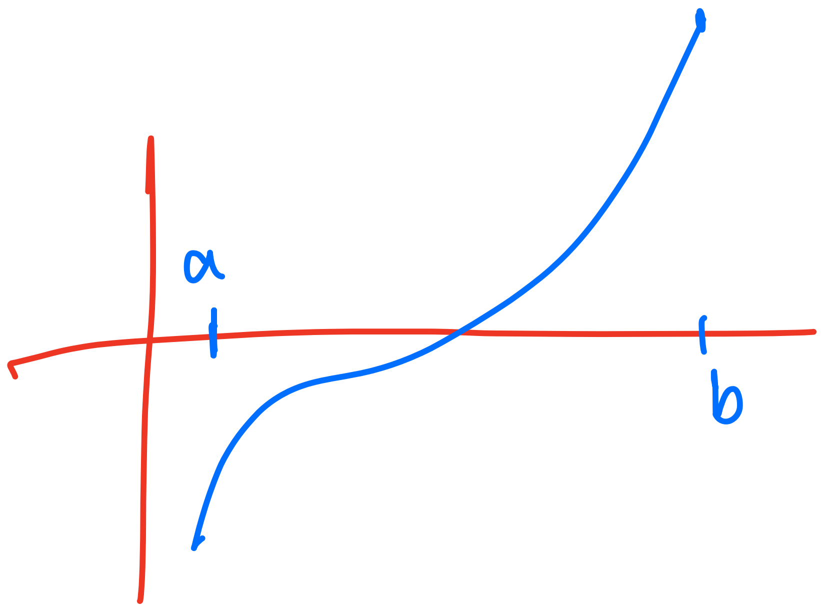
If you want to find x with $f(x)=0$:

find a with $f(a) < 0$,

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 one also uses the notation f^{-1} to denote the pre-image (which does not need to be unique).

Invertible function

<u>Proposition:</u> $D \subset \mathbb{R}$, $f: D \to \mathbb{R}$ continuous, strictly monotone $(a < b \Rightarrow f(a) < f(b))$. Then f is invertible and the inverse is continuous as well.

- Invertible follows from monotonicity

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

Sequence of functions

Pointwise convergence

Def: Consider functions: $f_n : D \to \mathbb{R}$, $D \subset \mathbb{R}$.

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

$$\forall x \in D: \quad f_n(x) \longrightarrow f(x)$$

Example

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

$$f(x) = \begin{cases} 0 & x = 0 \\ 1 & \text{otherwise} \end{cases}$$

This example also shows:

⚠ $f_n \to f$ pointwise, all f_n continuous, this does <u>not</u> imply that f is continuous.

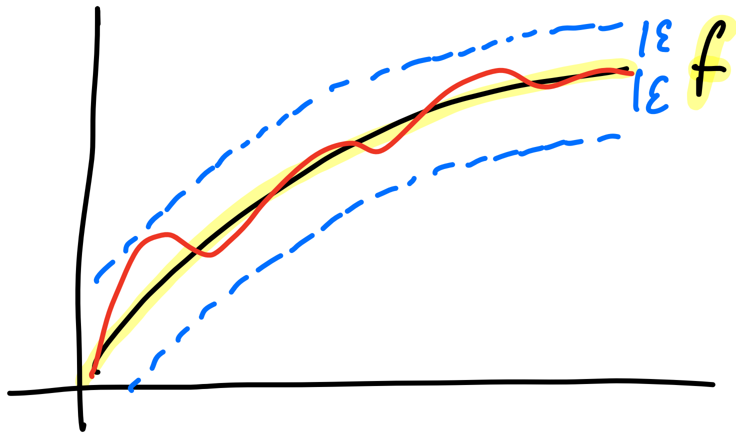
Uniform convergence

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

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

Intuition

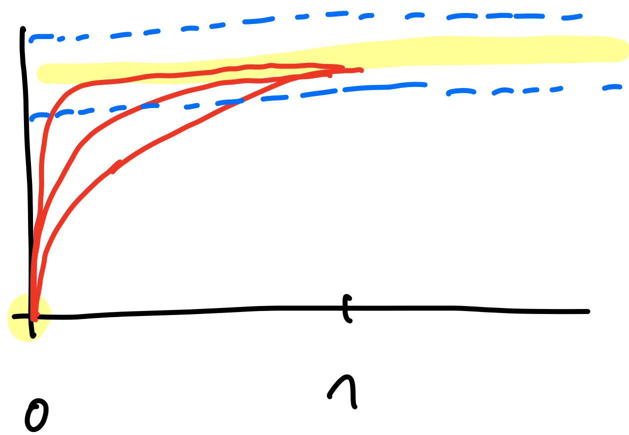
uniform convergence: given ε, there exist N such that all f_n with $n > N$ are contained in ε-tube.



f_n

uniform

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



not uniform, only pointwise

Alternative definition

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

Uniform convergence preserves continuity

Theorem

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

Proof.

Consider $x_0, x \in D$. Suppose some $\varepsilon > 0$ is given.

Observe that for every $n \in \mathbb{N}$,

$$\circledast \quad |f(x_0) - f(y)| \le |f(x_0) - f_n(x_0)| + |f_n(x_0) - f_n(y)| + |f_n(y) - f(y)|$$

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

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

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

Now consider the function f_n. By ass. it is continuous, so there exists $\delta > 0$ such that $|x_0 - x| < \delta \Rightarrow |f_n(x_0) - f_n(x)| < \varepsilon/3$.

Together we then get that for given $\varepsilon > 0$ there exists $\delta > 0$ such that

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

So f is continuous at x_0. ∎

Derivatives (1-dim)

Derivative definition

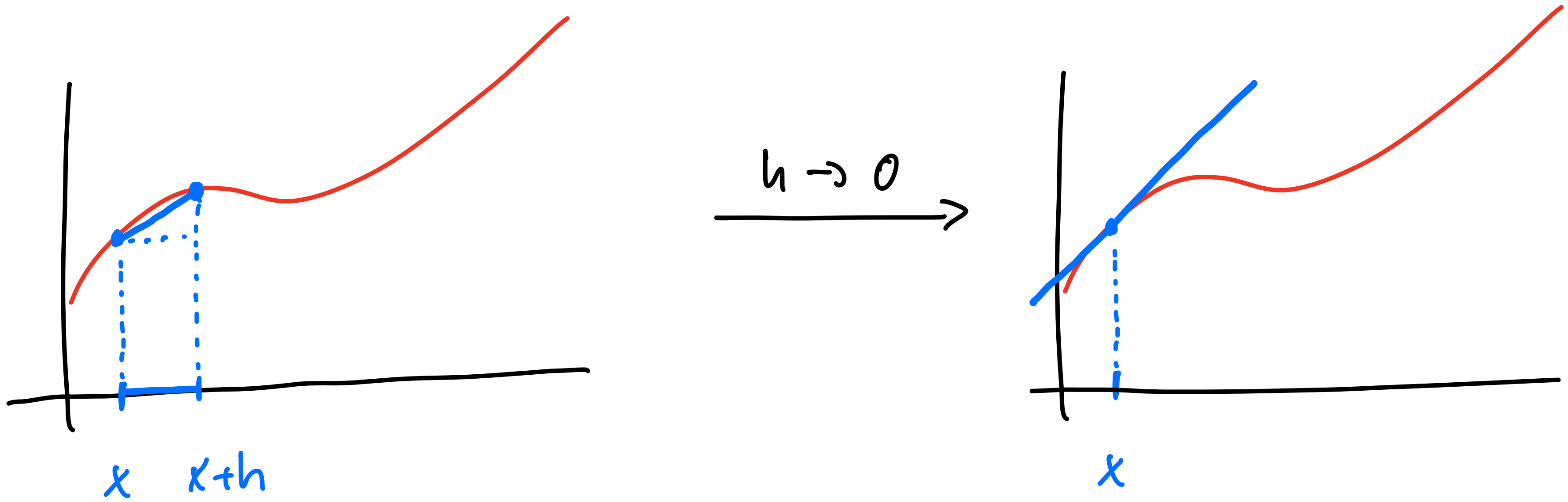
Def $U \subset \mathbb{R}$ an interval, $f : U \to \mathbb{R}$. The function f is called differentiable at $a \in U$ if

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

exists.

We often write $f'(a) = \frac{df}{dx}(a)$

Illustration



Differentiable functions

Def:

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

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

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

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 exists).

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

Differentiable implies continuous

Theorem

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

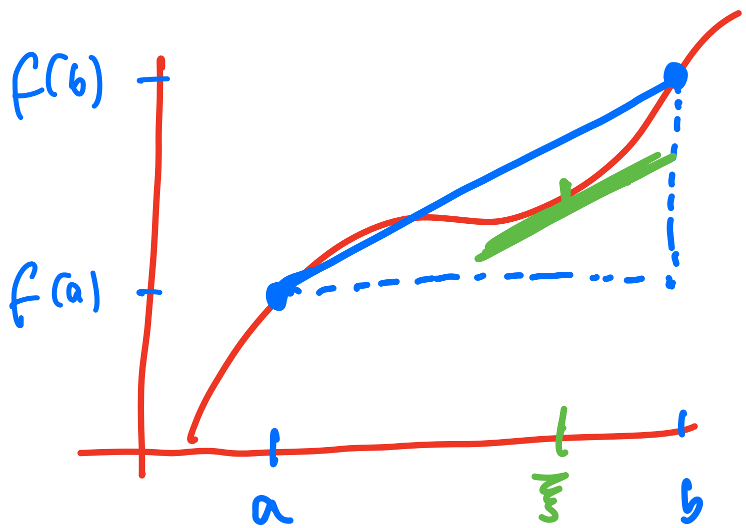
$$|f(x) - f(a)| \leq c_a \cdot |x - a|$$

In particular, f is continuous at a.

Intermediate value thm for derivatives

Theorem $f \in \mathcal{C}^1([a,b])$. Then there exist $\xi \in [a,b]$ such that

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



Exchanging lim and derivative

Theorem

$f_n : [a,b] \to \mathbb{R}$, $f_n \in \mathcal{C}^1[a,b]$. 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 cont. differentiable and

we have

$$(f')(x) = \left(\lim f_n\right)'(x)$$

first take limit of f_n, we obtain f, and then we compute its derivative

$$\overset{!}{=} \left(\lim \left(f_n'\right)\right)(x)$$

first compute all f_n', then take limit of these derivatives

⚠ uniform cont. is really important, otherwise would be wrong!

2024 skipped but

Riemann - integral (1-dim)

... assume that you know this material already!

Construction of the Riemann-integral

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

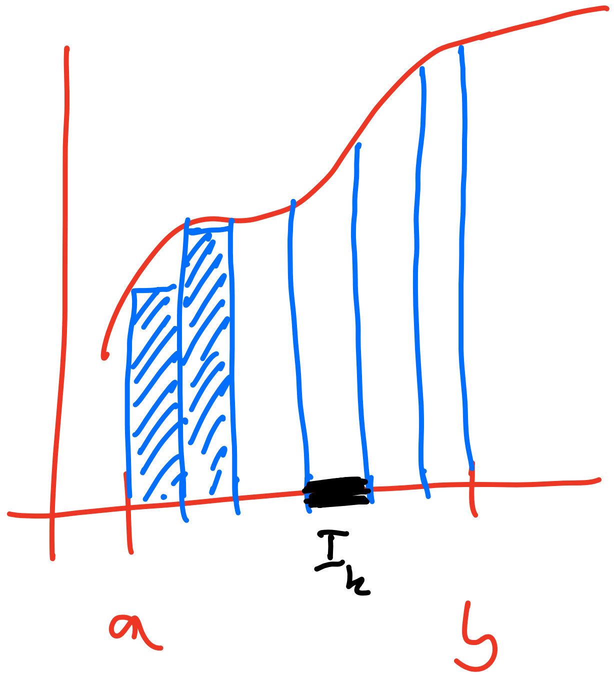
$(\exists\, l, u \in \mathbb{R}\ \forall x \in [a,b]:\ l \le f(x) \le u)$.

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

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

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

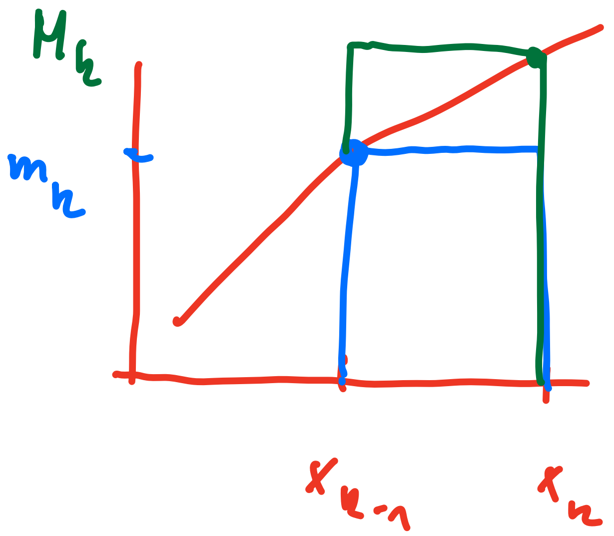
$$I_k := [x_{k-1}, x_k].$$



Define $m_k := \inf(f(I_k))$

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

(exists because f is bounded).



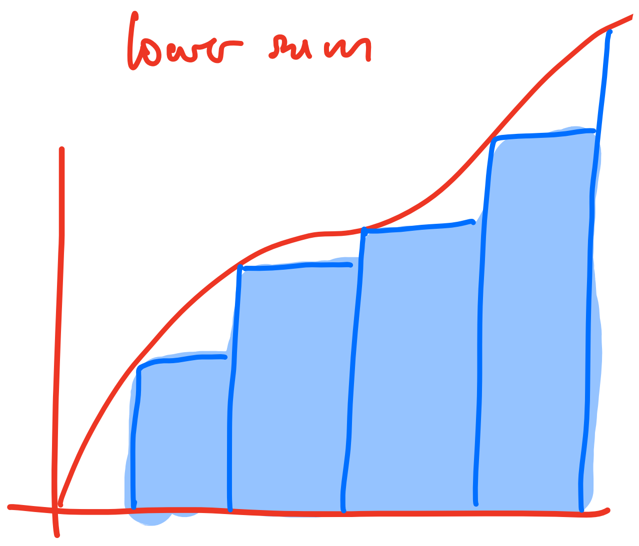
Define the lower sum

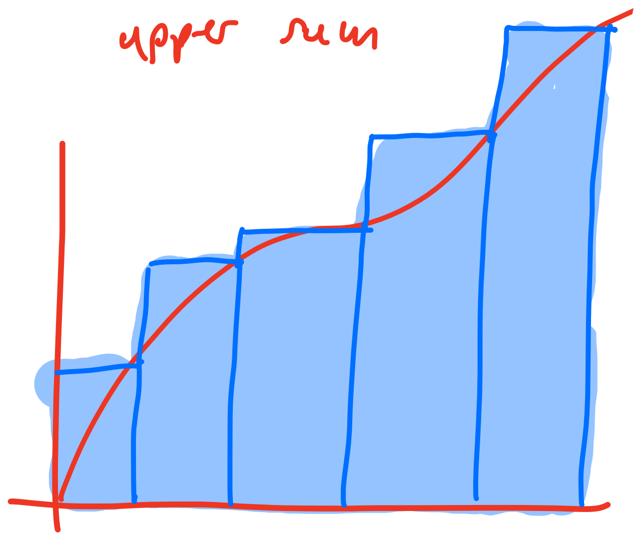
$$s(f, \{x_0, x_1, \ldots, x_n\}) = \sum_{k=1}^{n} |I_k| \cdot m_k$$

$\underbrace{\text{length}}$ of $I_k = x_k - x_{k-1}$

and the upper sum

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





Now define

$$J_* := \sup_{\text{partitions}} \left(s\left(f, \text{partition}\right)\right)$$

$$J^* := \inf_{\text{partitions}} \left(S\left(f, \text{partition}\right)\right)$$

We call f <u>Riemann-integrable</u> if $J_* = J^*$. Then we denote

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

Monotone resp. continuous fcts are Riemann-integrable

<u>Theorem</u>:
- $f : [a,b] \to \mathbb{R}$ monotone $\Rightarrow$ integrable

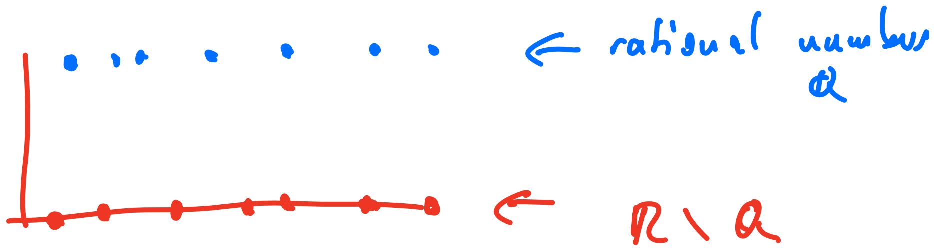
  (i.e. $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$)

- $f : [a,b] \to \mathbb{R}$ continuous $\Rightarrow$ integrable

  (even true if f is continuous everywhere except at finitely many point)

Many fcts are not Riemann-integrable

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



For any interval $I_k = [x_k, x_{k+1}]$,

$$M_k = +1$$

$$m_k = 0$$

Thus $$\underset{\substack{\shortparallel \\ |b-a|\cdot 0}}{J_*} < \underset{\substack{\shortparallel \\ |b-a|\cdot 1}}{J^*}$$

Further shortcomings of the Riemann integral

- One cannot prove theorems about exchanging "integral" with "lim":
$$\lim_{n\to\infty} \int f_n \, dt \overset{?}{=} \int \lim f_n \, dt$$

- Hard to extend to "other spaces".

$\rightsquigarrow$ Lebesgue-integral!

Fundamental theorem of calculus

Fundamental theorem of calculus

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

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

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

If $f \in \mathscr{C}([a,b])$, then $F \in \mathscr{C}^1([a,b])$ and $F'(x) = f(x)$ for all $x \in [a,b]$.

Theorem II: $F : [a,b] \to \mathbb{R}$ continuously differentiable, then

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

Algebraic version of the thm (informal)

Informal, algebraic version:

The integral operator $I : \mathcal{C}[a,b] \longrightarrow \mathcal{C}^1_{\text{"}c\text{"}}([a,b])$

with $\mathcal{C}^1_{\text{"}c\text{"}}[a,b] := \{ f \in \mathcal{C}^1[a,b] : f(c) = 0 \}$

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

Proof Part I

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

Consider $A(h) := \dfrac{F(\xi+h) - F(\xi)}{h}$

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

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

What to prove: converges to $f(\xi)$

as $h \to 0$

Want to prove:

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

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

$$= \frac{1}{h} \int_{\xi}^{\xi+h} \overset{\text{does not depend on } t}{f(\xi)}\, dt$$

$$= \frac{1}{h} \underbrace{(\xi+h - \xi)}_{\substack{\text{length of} \\ \text{interval over} \\ \text{which we} \\ \text{integrate}}} \cdot \underbrace{f(\xi)}_{\substack{\text{constant} \\ \text{in the} \\ \text{integral}}}$$

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

$$= \frac{1}{h} \int_{\xi}^{\xi+h} \underbrace{f(t) - f(\xi)}_{\text{Intuition: small due to continuity of } f \text{ at } \xi}\, dt$$

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

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

Then:

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

$$\leq \frac{1}{h} \int_{\xi}^{\xi+h} \varepsilon\, dt = \frac{1}{h} \cdot \varepsilon \int_{\xi}^{\xi+h} 1\, dt = \frac{1}{h} \cdot \varepsilon \cdot h = \varepsilon.$$

∎ Theorem I

Proof part II

<u>Proof II</u> :

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

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

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

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

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

By (ii) we know that $H'(x) = F'(x) - G'(x) = 0$ for all x

Hence, H is a constant function.

We know that $H(a) = F(a) - \underbrace{G(a)}_{=0 \ (i)} = F(a)$, thus

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

Consider $x = b$.

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

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

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

■ Th. II

Power series

Power series

Def 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 (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 <u>radius of convergence</u> such that

- The series converges (absolutely) for all x with $|x| < r$
- If $|x| < r$, the series even converges uniformly.

⚠ It is unclear what happens for $|x| = r$

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

- $r = \frac{1}{L}$ where $L = \limsup_{n\to\infty} (|a_n|)^{1/n}$
- $r = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|$

if exists

A first example

$$p(x) = \sum_{n=0}^{\infty} \underbrace{n^c}_{a_n} \cdot x^n \qquad \text{for some constant } c$$

Radius of convergence:

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

(independently of c)

A first example (cont.)

<u>Case $c = -1$:</u> $\sum \frac{1}{n} x^n$ has conv. radius $r = 1$

- For $|x| > 1$ it diverges.
- For $|x| < 1$ it converges.
- For $|x| = 1$ no general statement, but we can analyze it more closely:
  - For $x = +1$ the series diverges because
    $$\sum \frac{1}{n} x^n = \sum \frac{1}{n} 1^n = \sum \frac{1}{n} \to \infty$$
  - For $x = -1$ it converges:
    $$\sum \frac{1}{n} (-1)^n = -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4}$$
    $$= -\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \cdots\right) = -\log(2)$$

A first example (cont.)

<u>Case $c=0$</u> : $\sum n^c x^n = \sum x^n$ diverges for $|x| = r$

Convergence radius is still $r=1$.

- For $|x| < 1$ series converges.
- For $|x| > 1$ series diverges.
- For $|x| = 1$:
  - $x = +1$: $\sum_{i=1}^{N} x^n = \sum_{i=1}^{N} 1 = N \to \infty$ diverges.
  - $x = -1$: $\sum x^n = -1 + 1 - 1 + 1 - 1 + \ldots$

    does not converge

More examples

- **Exponential series:**

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

because

$$\left|\frac{a_n}{a_{n+1}}\right| = \frac{1/n!}{1/(n+1)!} = \frac{(n+1)!}{n!} = n+1 \to \infty$$

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 $N \to \infty$.

- $\sum_{n=0}^{\infty} n!\, x^n$ has $r = 0$:

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

From power series to Taylor series, intuition

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

Let's take its derivative:

$$f'(x) = \left(a_0 + a_1(x-a) + a_2(x-a)^2 + a_3(x-a)^3 + \dots \right)'$$

would need to prove this

$$= a_1 + 2a_2(x-a) + 3a_3(x-a)^2 + \dots$$

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

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

$$f^{(k)}(x) = \sum_{\underline{n=k}}^{\infty} a_n \left(n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \right) (x-a)^{n-k}$$

In particular, we have for $x = a$

$$f^{(k)}(a) = a_k k! \quad \text{or, stated otherwise} \quad \boxed{a_k = \frac{f^{(k)}(a)}{k!}}$$

From power series to Taylor series, formally

Theorem: 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 neat formula of above that expresses the coeff. in terms of derivatives. $\square$

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 $\sum \frac{f^{(n)}}{n!}(x-a)^n$ and "hope" that it converges to the function?

$\overset{?}{=} f(x)$???

Taylor series

Taylor series

Theorem: $J \subset \mathbb{R}$ open interval, $f: J \to \mathbb{R}$, $f \in \mathcal{C}^{n+1}(J)$, $a, x \in J$. Define

$$T_n(x,a) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

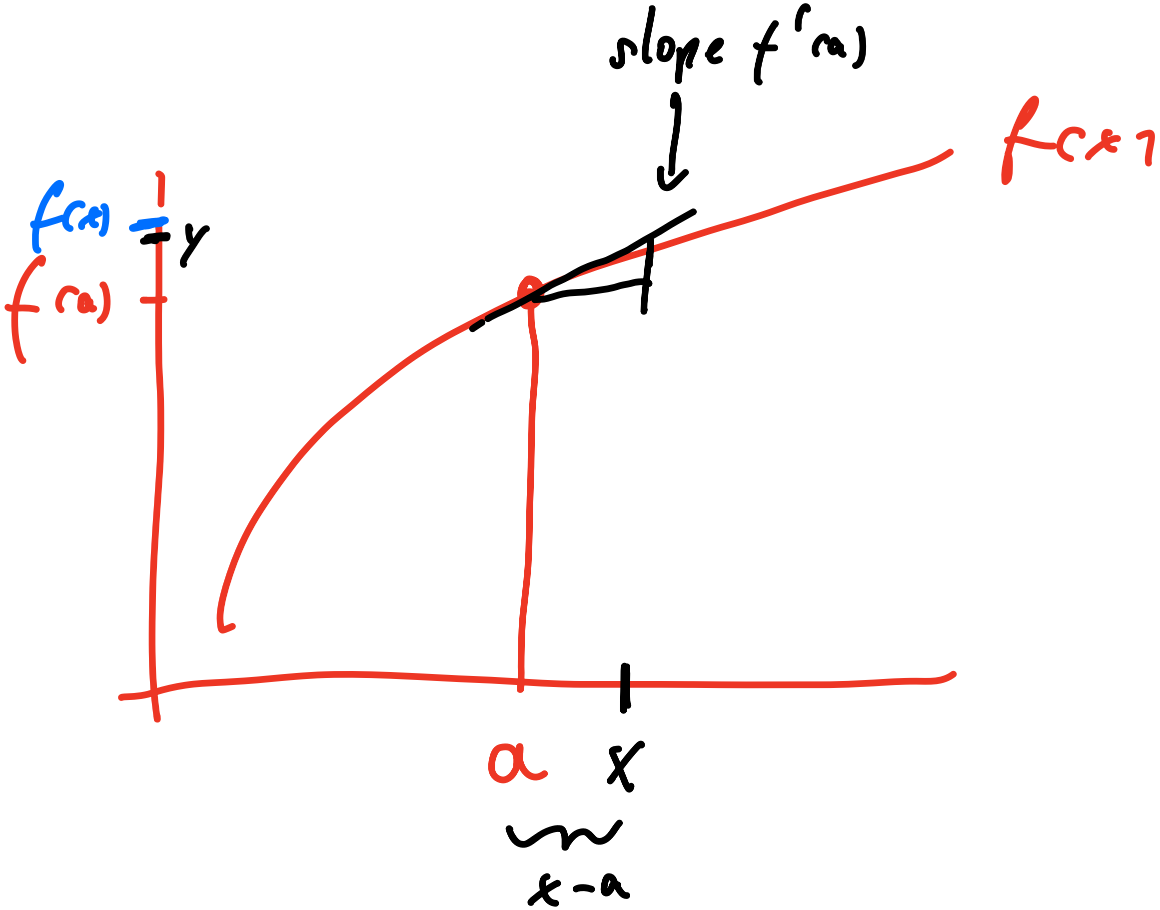
Taylor series up to degree n

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

Remainder term

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

Intuition about Taylor series



$$f(x) \approx \underbrace{f(a) + f'(a)(x-a)}_{y} + \ldots$$

Proof

Proof follows from Fundamental theorem, by induction on n.

Base case n = 0: need to prove

$$f(x) = f(a) + \int_a^x f'(t)\,dt \qquad \hat{=} \text{ Fundam. Theorem}$$

Inductive step $n \rightsquigarrow n+1$:

- Consider $F(t) = \dfrac{(x-t)^{n+1}}{(n+1)!} f^{(n+1)}(t)$
- Take its derivative
- Integrate and exploit fundamental theorem

Taylor with Lagrange remainder

Theorem:

$f \in \mathcal{C}^{n+1}(J)$, $a, x \in J$. 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

Proof Let $I = [a, b]$.

- Consider two functions $F, G \in \mathcal{C}^{n+1}([a,b])$. Assume that $\boxed{F(a) = G(a) = 0, \text{ and } G' \neq 0 \text{ on } [a,b].}$ $(*)$

Now:

$$\frac{F(b) - \overbrace{F(a)}^{=0}}{G(b) - \underbrace{G(a)}_{=0}} = \frac{F(b)}{G(b)} \overset{\text{intermediate value thm}}{=} \frac{F'(\xi)}{G'(\xi)} \quad \text{for some } \xi \in [a,b]$$

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

We would obtain

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

Proof (cont.)

- Now chose $F(x) := f(x) - T_n(x,a) = R_n(x,a)$
  $G(x) := (x-a)^{n+1}$

  - For all k in $0 \leq k \leq n$ we have by construction that
    $f^{(k)}(a) = T_n^{(k)}(a)$, so in particular
    $F^{(k)}(a) = 0$, and we have $G^{(k)}(a) = 0$.

Proof (cont.)

- For $n+1$ we now have

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

By $(\#)$ we obtain

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

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

$\blacksquare$

Taylor convergence

<u>Theorem</u> $f \in \mathcal{C}^\infty(J)$, $x, a \in J$. 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 $R_n(x,a) \xrightarrow{n \to \infty} 0$.

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

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

Follows directly from the Lagrangian 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.

- $f(x) = \log(1+x)$, Taylor series about $a = 0$

Can prove: Convergence radius of Taylor series is $r = 1$

For x outside of $]-1, 1[$ Taylor series does not make sense at all.

Examples (cont.)

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

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

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

All terms will be 0, so $\forall n: T_n(x) = 0$, $r = \infty$

but of course f is not $\equiv 0$, so we get

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

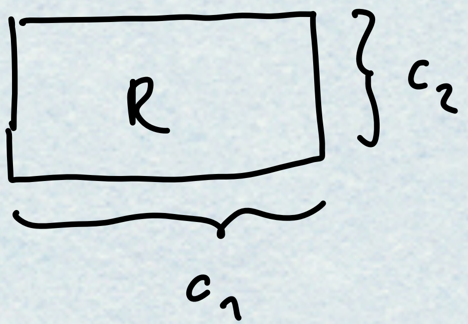
Taylor series converges everywhere, but not to the fct f!

Skipped

Lebesgue measure on $\mathbb{R}^n$

Goal

Want to construct a measure on $\mathbb{R}^n$. Want that rectangles of the form $[a_1, b_1[\times [a_2, b_2[\times \dots \times [a_n, b_n[$ have the "natural volume" given by $\prod_{i=1}^{n} (b_i - a_i)$



$\rightsquigarrow \mathrm{vol}(R) \overset{!}{=} c_1 \cdot c_2$

And want "nice" mathematical properties

Earlier approaches

First approaches (Jordan, Riemann) attempted the following:

"Outer approximation":

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

"Inner approximation:"

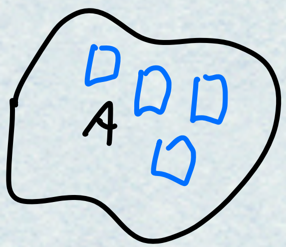
$$\bigcup_{i=1}^{n} \text{rect.}_i \subset A$$

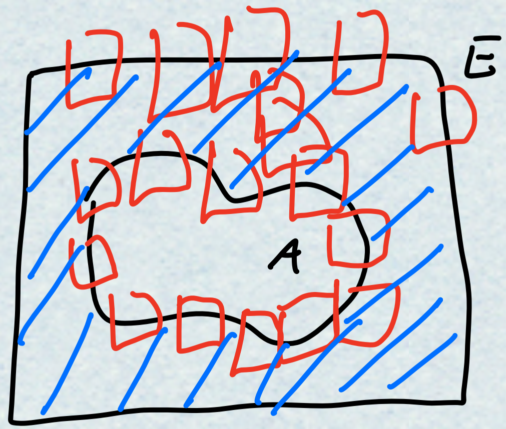
A would be called "measurable" if outer and inner approximation "converge".

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 approx. of the complement:





outer approx. of $\underline{E \setminus A}$

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

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

- Need σ-algebra as underlying structure.

Outer Lebesgue measure

Set the "natural volume" of rectangles:

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

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

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

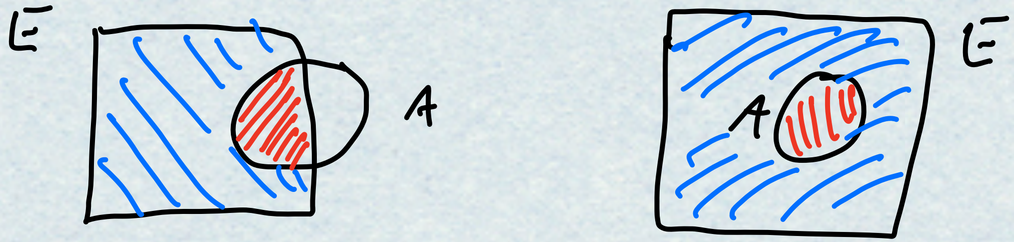
Want to make this into a measure. Problem: if we use $\mathcal{P}(\mathbb{R}^n)$ as σ-algebra, we run into contradictions.

Need to restrict ourselves to a smaller σ-algebra...

Measurable set

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

$$\lambda(E) = \lambda(E \cap A) + \lambda(E \setminus A)$$



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

Outer measure as measure ...

<u>Theorem</u> 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 "~~na~~ natural volume".

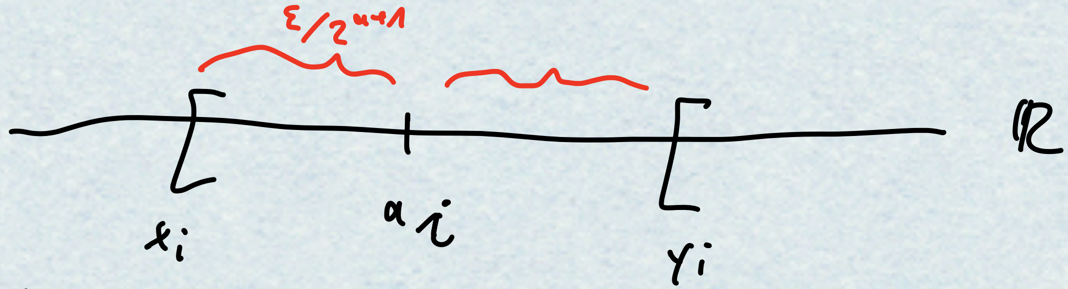
<u>Examples</u>:

- $\lambda(\{x\}) = 0$
- $\lambda(\mathbb{R}) = \infty$
- $A \subset \mathbb{R}$ countable. Then $\lambda(A) = 0$. 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-alg. 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 in $\mathbb{R}^n$ can be written as a countable union of open intervals: $A \subset \bigcup_{i=1}^{\infty} I_i$, I_i open intervals.

(2) For every Lebesgue-measurable set L there exist 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).

SKIPPED

A non-measurable set

Construction (quite abstract!)

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

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

$\frac{\pi}{4}$, $\frac{\pi}{4} + \frac{1}{2}$, $\frac{\pi}{4} + \frac{799}{800}$ would be equivalent

Consider the equivalence classes

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

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

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

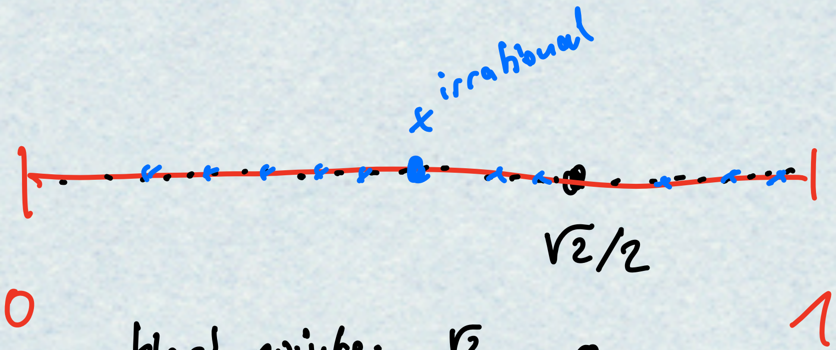
$$\vdots$$

We pick a representative of each of the classes, and denote by N the set of all such representatives.

N is not Lebesgue-measurable!

__Proposition:__ N is __not__ Lebesgue-measurable.

Intuition:



black points: $\frac{\sqrt{2}}{2} + \mathbb{Q}$

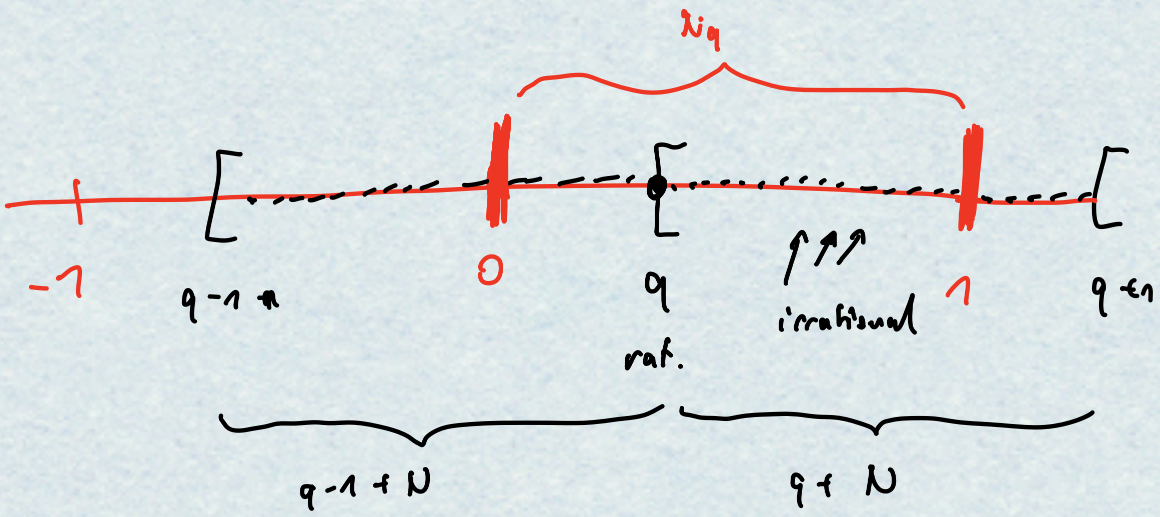
blue crosses: $x + \mathbb{Q}$

Proof

<u>Proof by contradiction:</u>

Assume N is measurable. We now construct the following sets: For $q \in [0,1[$

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



Proof (cont.)

- If N is measurable, then $q+N$ is measurable $\forall q \in [0,1[$

  and $\lambda(N_q) = \lambda(N)$

- $[0,1[= \bigcup_{q \in [0,1] \cap \mathbb{Q}} N_q$

- $N_q \cap N_p \neq \emptyset \Rightarrow N_p = N_q$

  Consequently, $\bigcup N_q$ is disjoint.

- σ-additivity:

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

Proof (cont.)

- Could be that $\lambda(N_q) = 0$. But then

$$\sum_q \lambda(N_q) = 0 \qquad \lightning$$

- Could be that $\lambda(N_q) > 0$. But then

$$\sum_q \lambda(N_q) = \infty$$

$$\lightning$$

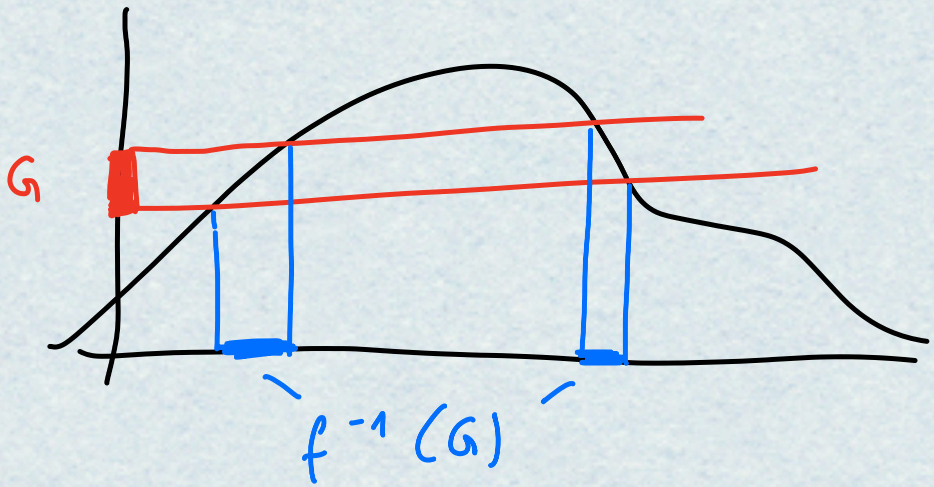
SKIPPED

The Lebesgue-integral on $\mathbb{R}^n$

Intuition: partition Y instead of X

Riemann:

Lebesgue:



Measurable fcts

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

$$\forall G \in \mathcal{G} : \underbrace{f^{-1}(G)}_{=: \{x \in X \mid f(x) \in G\}} \in \mathcal{F}$$

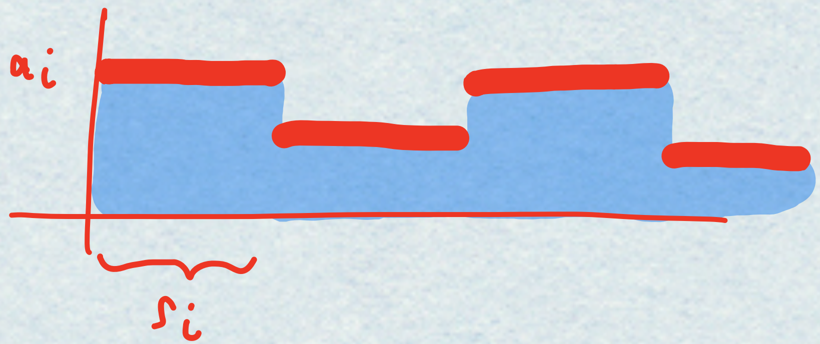
Lebesgue-integral for simple fcts

<u>Def</u> $\phi : \mathbb{R}^n \to \mathbb{R}$ is called a <u>simple function</u> if there exist measurable sets $S_i \subset \mathbb{R}^n$, $a_i \in \mathbb{R}$ such that

$$\phi = \sum_{i=1}^{n} a_i \underbrace{\mathbb{1}_{\{S_i\}}}_{S_i = \phi^{-1}(a_i)}$$

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

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



Lebesgue integral for non-negative fct

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

$$\int f^+ \, d\lambda = \sup \left\{ \int \phi \, d\lambda \;\middle|\; \phi \leq f, \ \phi \text{ simple} \right\}$$

(might be ∞)

approx f by simple fcts

Note: the sets S_i can be complicated sets, not just intervals!

Lebesgue integral for general fcts

- For a general function $f: \mathbb{R}^n \to \mathbb{R}$ we split the function into positive and neg. part: $f = f^+ - f^-$

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

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

- If both f^+ and f^- satisfy $\int f^+ d\lambda < \infty$, $\int f^- d\lambda < \infty$, 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: $\int \mathbb{1}_{\mathbb{Q}} \, d\lambda = 1 \cdot \lambda(\mathbb{Q}) = 0$

Theorem (monotone convergence):

Consider a sequence of functions $f_k : \mathbb{R}^n \to [0, \infty[$

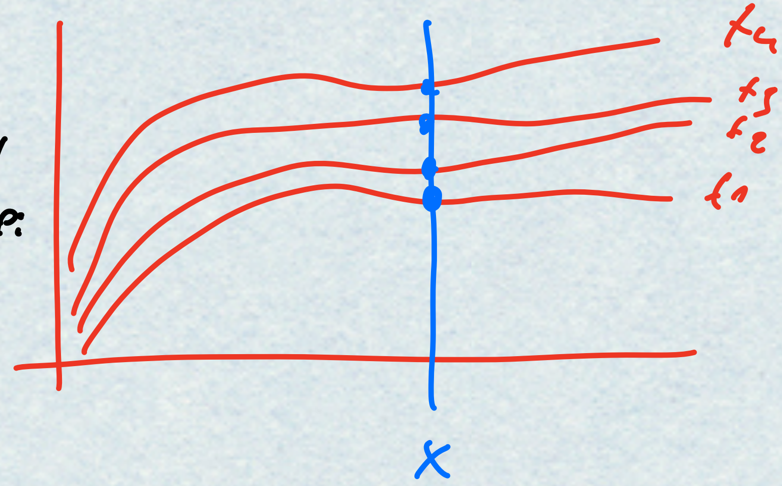
that is pointwise non-decreasing:

$\forall x \in \mathbb{R}^n$: $f_{n+1}(x) \geq f_n(x)$.

Assume that all f_k are measurable,

and that the pointwise limit exists:

$\forall x$: $\lim f_k(x) =: f(x)$



Then:

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

$$\overset{\shortparallel}{\int \lim_{k\to\infty} f_k(x)\, dx}$$

Theorem (dominated convergence):

$f_n : B \to \mathbb{R}$; $|f_n(x)| \leq g(x)$ on B, $g(x)$ is integrable. Assume that the pointwise limit exists: $\forall x \in B$:

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

Then:

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

Partial derivatives

ML Motivation

Gradient descent!

Optimization!

Functions on $\mathbb{R}^n$

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

input space: n-dim

output space 1-dim

(The standard object in machine learning!)

$\mathbb{R}^n \ni x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $f(x) = x_1^2 + x_2^2 \cdot x_1$

$f: \mathbb{R}^2 \to \mathbb{R}$

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

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

Def f is called partially differentiable with resp. to variable x_j at point $\xi \in \mathbb{R}^n$ if the 1-dim (!) function

$g: \mathbb{R} \to \mathbb{R}$,

$$g(x_j) := f(\xi_1, \xi_2, \dots, \xi_{j-1}, x_j, \xi_{j+1}, \dots, \xi_n)$$

intuition: fix all arguments except the jth

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

Notation:

point at which we evaluate the derivative.

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

constant > 0

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

"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

grad f_1

$$Df(x) = \begin{pmatrix} \boxed{\dfrac{\partial f_1}{\partial x_1} \quad \cdots \quad \dfrac{\partial f_1}{\partial x_n}} \\ \vdots \qquad\qquad\qquad \vdots \\ \dfrac{\partial f_m}{\partial x_1} \quad \cdots \quad \dfrac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Gradient $\not\Rightarrow$ continuity of f!

For functions $f: \mathbb{R} \to \mathbb{R}$ we know that if f is differentiable, then the function is continuous. Note that in the n-dim case, the existence of a gradient is <u>not</u> 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

$$f: \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = \begin{cases} \dfrac{x \cdot y}{x^2 + y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } x = y = 0 \end{cases}$$

For $(x,y) \neq (0,0)$

$$\operatorname{grad} f(x,y) = \left(y \cdot \frac{y^2 - x^2}{(x^2+y^2)^2} \,,\; x \cdot \frac{x^2 - y^2}{(x^2+y^2)^2} \right)$$

$$\operatorname{grad} f(0,0) = 0 \quad \text{because} \quad \begin{aligned} f(x,0) &= 0 \quad \forall x \\ f(0,y) &= 0 \quad \forall y \end{aligned}$$

but f is not continuous at 0.

Total derivative

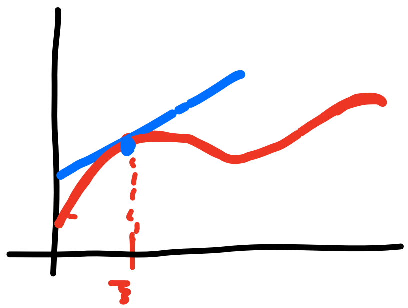
Differentiable fct

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

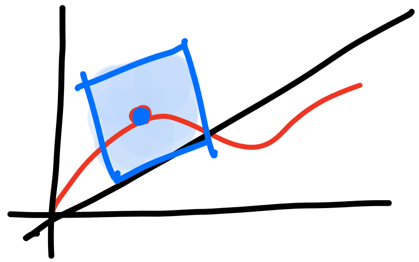
$$f(\xi + h) - f(\xi) = L(h) + r(h)$$

with $\lim\limits_{h \to 0} \frac{r(h)}{|h|} \to 0.$

Intuition: f is "locally linear"



1-dim approx by a line



2-dim approx by a plane

Differentiable, continuous, gradient

<u>Theorem</u> $f : \mathbb{R}^n \to \mathbb{R}$ differentiable at ξ.

- Then f is continuous at ξ.
- The linear functional L coincides with the gradient:

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

$$= \langle \operatorname{grad} f(\xi), h \rangle \quad + r(h)$$

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

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

Continuous partial derivatives ⇒ differentiable

<u>Theorem</u>: If all partial derivatives exist and are all continuous, then f is differentiable.

⚠ If partial derivatives exist, but are not continuous, then f doesn't need to be differentiable.

Directional derivatives

Directional derivatives

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

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

$$D_v f(\xi) := \lim_{t\to 0} \frac{f(\xi + \overset{\in\mathbb{R}}{t}\cdot \overset{\in\mathbb{R}^n,\ \text{direction}}{v}) - f(\xi)}{t}$$

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

Differentiable $\Rightarrow$ directional derivatives and gradient

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

$$D_v f(\xi) = (\operatorname{grad} f)^t \cdot v = \sum_{i=1}^{n} \overset{\in \mathbb{R}}{v_i} \cdot \overset{\text{partial der.}}{\frac{\partial f}{\partial x_i}} (\xi)$$

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

The largest value of all directional derivatives is attained in direction

$$v = \frac{\operatorname{grad} f(\xi)}{\| \operatorname{grad} f(\xi) \|}$$

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:

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

Example

$f: \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = \dfrac{x \cdot y^3}{x^2 + y^2}$

$$\operatorname{grad} f(x,y) = \left(\frac{y^3(y^2 - x^2)}{(x^2+y^2)^2}, \frac{xy^2(3x^2+y^2)}{(x^2+y^2)^2} \right)$$

Have:
- $\dfrac{\partial f}{\partial x}(0,y) = y$ for all y

  $\dfrac{\partial}{\partial y}\left(\dfrac{\partial f}{\partial x}\right) = 1$

- $\dfrac{\partial f}{\partial y}(x,0) = 0$ $\forall$ all x

  $\dfrac{\partial}{\partial x}\left(\dfrac{\partial f}{\partial y}\right) = 0$

Consequently, the two derivatives do not agree on point $(0,0)$.

Hessian

Def Hessian matrix

$f: \mathbb{R}^n \to \mathbb{R}$, then we define the Hessian of f at point x by

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

Take care of different dimensionality

⚠ Caution: dimensions

$$f : \mathbb{R}^n \to \mathbb{R}$$ function

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$ first derivative : n partial deriv. $\frac{\partial f}{\partial x_i}$

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

Continuously differentiable fcts

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

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

Analogously: k times cont. differentiable

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

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

Continuously diff. ⇒ can change order

Theorem (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 cont. diff. ⇒ can exchange order of first k partial derivatives.

Multivariate Taylor Series

To Do!

Minima / maxima

Critical point

Def $f : \mathbb{R}^n \to \mathbb{R}$ differentiable. If $\nabla f(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ then we call x a <u>critical point.</u>

⚠ A critical point can have many different geometric meanings:

Minimum

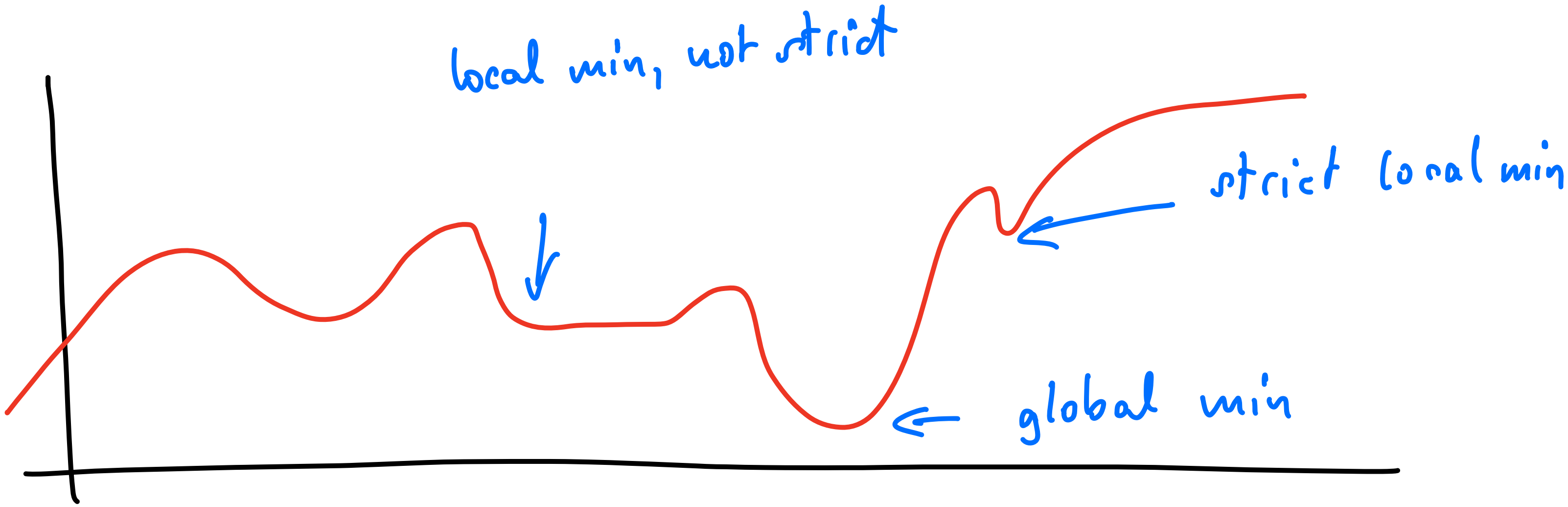
$f: \mathbb{R}^n \to \mathbb{R}$. f has a local minimum at x_0 if there exists $\varepsilon > 0$ such that

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

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

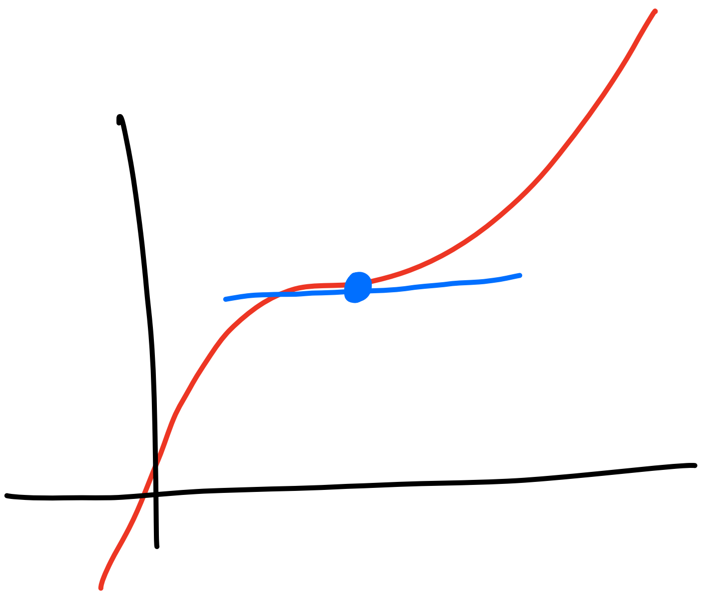
$$\forall x \in B_\varepsilon(x_0) : f(x) > f(x_0)$$

f has a global minimum at x_0 if $\forall x \in \mathbb{R}^n : f(x) \geq f(x_0)$

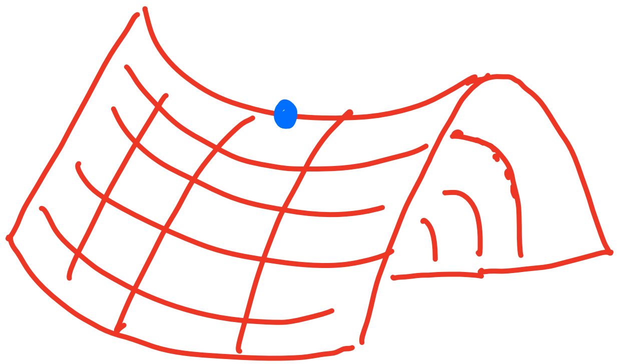


Saddle point

If f is differentiable and x_0 is a critical point that is neither a local minimum or maximum, then we call it a **saddle point**:



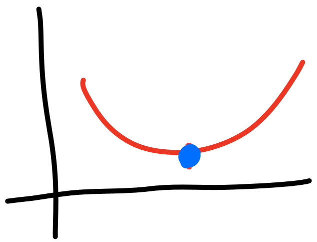
all partial derivatives are zero



How can we find out which case we have?

Intuition in 1-dim case: second derivatives might help:

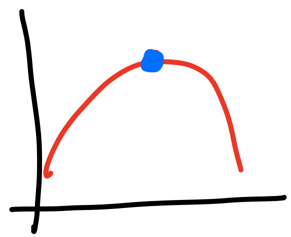
Local min:



$f'(x) = 0$

$f''(x) > 0$

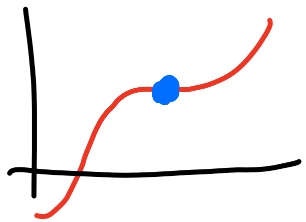
Local max:



$f'(x) = 0$

$f''(x) < 0$

saddle pt:



$f'(x) = 0$

$f''(x) = 0$

Critical points and the Hessian

<u>Theorem</u> $f: \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$. 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 (neg. semi-def.)

(ii) If $Hf(x_0)$ is positive definite (neg. definite), then x_0 is a strict local min (max). If $Hf(x_0)$ is indefinite, then x_0 is a saddle point.

Derivatives of popular matrix/vector functions

Example: Linear least squares

Given training points $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Want to approximate this data by a linear least squares problem: Find a linear function

$f: \mathbb{R}^n \to \mathbb{R}$ that minimizes the least squares error.

In matrix notation: $X = \begin{pmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{pmatrix} \in \mathbb{R}^{n \times d}$,

$f: \mathbb{R}^d \to \mathbb{R}$, $f(x) = \langle w, x \rangle$ with parameter vector w,

determine w as

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^n \left| \langle x_i, w \rangle - y_i \right|^2 = \min_{w \in \mathbb{R}^d} \| X \cdot w - y \|^2$$

objective fct

$=: g(w)$

Solution "by foot"

To optimize for w, need to take derivative of the objective fct to 0:

$$\frac{\partial g}{\partial w} \overset{!}{=} 0 \quad .$$

To compute the gradient by foot is pretty cumbersome:

- Write fct coordinate-wise: $$g\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \sum_{j=1}^{n} \left(y_j - \sum_{k=1}^{m} x_{jk} w_k \right)^2$$

- Take partial derivatives:

$$\frac{\partial g}{\partial w_i} = \sum_{j=1}^{n} \left(- x_{ji} \right) \cdot 2 \cdot \left(y_j - \sum_{k=1}^{m} x_{jk} w_k \right)$$

Observe that we can write result using matrices

$$\frac{\partial g}{\partial w_i} = \sum_{j=1}^{n} (-x_{ji}) \cdot 2 \cdot \Big(y_j - \underbrace{\sum_{k=1}^{m} x_{jk} w_k}_{(Xw)_j}\Big)$$

$$\underbrace{-2 \cdot \sum_{j=1}^{n} x_{ji} \cdot \underbrace{(y - Xw)_j}}_{\left(X^t (y - Xw)\right)_i}$$

$$\nabla g(w) = -2 \, X^t (y - Xw)$$

Observe: "Syntax" close to 1-dim case:

$$g(w) = (y - x \cdot w)^2$$

$$g'(w) = -x(y - xw) \cdot 2 = -2x(y - xw)$$

The matrix cookbook

Lookup table ("cookbook") for gradients of many important functions:

Examples for functions of vectors: $f: \mathbb{R}^n \to \mathbb{R}$

- $\boxed{f(x) = a^t x}$ $\quad (a \in \mathbb{R}^n)$ linear fct

$$= \langle a, x \rangle$$

$$\frac{\partial f}{\partial x} = a \quad \in \mathbb{R}^n$$

- $\boxed{f(x) = x^t A x}$ quadratic fct

$$\Rightarrow \quad \frac{\partial f}{\partial x} = (A + A^t)\, x \quad \in \mathbb{R}^n$$

Examples for functions of matrices: $f : \mathbb{R}^{n \times m} \to \mathbb{R}$

- $f(X) = \overbrace{a^t}^{1\times n}\,\overbrace{X}^{n\times m}\,\overbrace{b}^{m\times 1}$ for $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$

$$\frac{\partial f}{\partial X} = \underbrace{\underset{n\times 1}{a} \cdot \underset{1\times m}{b^t}}_{n\times m} \in \mathbb{R}^{n\times m}$$

- $\underbrace{f(X)}_{X \in n\times m} = \underbrace{a^t}_{1\times m}\,\underbrace{X^t}_{m\times n}\,C\,\underbrace{X}_{n\times m}\,\underbrace{b}_{m\times 1}$ for $a \in \mathbb{R}^m$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{n\times n}$

$$\frac{\partial f}{\partial X} = C^t X a b + C X b a^t$$

Examples for functions of matrices: $f : \mathbb{R}^{n \times n} \to \mathbb{R}$

- $f(X) = tr(X) \Rightarrow \frac{\partial f}{\partial X} = I \in \mathbb{R}^{n \times n}$ trace

- $f(X) = tr(A \cdot X) \Rightarrow \frac{\partial f}{\partial X} = A$

  $f(X) = tr(X^t A X) \Rightarrow \frac{\partial f}{\partial X} = (A + A^t X$

- $f(X) = det(X)$ Determinant

  $$\frac{\partial f}{\partial X} = det(X)\,(X^t)^{-1}$$

  $$\frac{\partial\, det}{\partial a_{sr}} = det(A) \cdot (A^{-1})_{rs}$$

Examples for functions of matrices: $f: \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$

- $f(A) = A^{-1}$, $f_{ij} := (A^{-1})_{ij}$ Inverse

$$\frac{\partial f_{ij}}{\partial a_{uv}} = - (a_{iu})^{-1} (a_{vj})^{-1}$$

? Autodiff ?

Chemp's book