Mathematics for Machine Learning

Ulrike von Luxburg

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

(Version as of November 23, 2026)

Table of contents

Part II

Calculus: Basics . . . 4

Lecture 10

ML Motivation . . . 5
Sequences and convergence . . . 6
Basic concepts in topology . . . 17
Continuity . . . 27

Lecture 11

Sequence of functions . . . 39
Calculus: Derivatives and integrals on $\mathbb{R}$ . . . 47
Derivatives . . . 48
Riemann-integral . . . 56
Lebesgue measure SKIPPED . . . 64
A non-measurable set SKIPPED . . . 75
The Lebesgue-integral . . . 81
Fundamental Theorem of Calculus . . . 95

Lecture 12

Power series . . . 105
Taylor series . . . 117
Calculus: Derivatives and integrals on $\mathbb{R}^d$ . . . 127

Table of contents (2)

Partial derivatives . . . 128
Total derivative . . . 137
Directional Derivatives . . . 145
Higher-Order Derivatives . . . 151
Minima/Maxima . . . 159
Derivatives of popular matrix/vector functions . . . 167
The matrix cookbook . . . 172

Part II
Calculus: Basics

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

Convergence etc

- In the following we are going to establish what it means for a sequence of points to converge, have accumulation points, limits, etc.
- These notions are straight forward if the points live in a normed vector space or a metric space. As soon as we can measure distances, notions of convergence are easy.
- On can also define all these concepts much more abstractly, as it is done in mathematical topology. Very beautiful, but for machine learning not useful.

In this lecture we mainly work in $\mathbb{R}^d$ or general metric spaces.

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

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

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

- Consider the sequence given by $x_n = \frac{1}{n}$. This sequence $(x_n)_n$ is Cauchy and converges to 0. If we consider the sequence as part of the restricted space $]0, 1]$, it is Cauchy, but does not converge (because the limit point is not part of the space).

- 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 1 (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:

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

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

- 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

SKIPPED

Open and closed sets

Definition 2 (Epsilon-ball)

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

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

SKIPPED

Open and closed sets (2)

Definition 3 (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

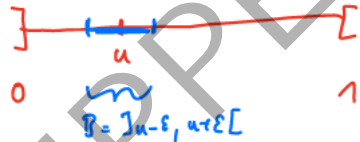
$$\forall u \in U \, \exists \varepsilon > 0 : B_\varepsilon(u) \subset U.$$

SKIPPED

⚠ 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 **neither** open nor closed: $[0, 1[$.

SKIPPED

Open vs. closed

Proposition 4 (Open vs. closed)

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

SKIPPED

Interior, closure

SKIPPED

Interior, closure (2)

Definition 5 (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°

SKIPPED

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

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

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

SKIPPED

Boundary (2)

Definition 6 (Bounded set)

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

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

SKIPPED

Dense sets

Definition 7 (Dense sets)

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

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

SKIPPED

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 8 (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

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

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:

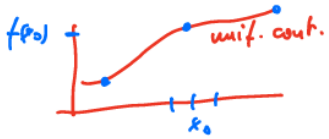
$$B \subset Y \text{ open} \quad \Rightarrow f^{-1}(B) := \{x \in X \mid f(x) \in B\} \text{ is open in } X.$$

Uniformly continuous

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

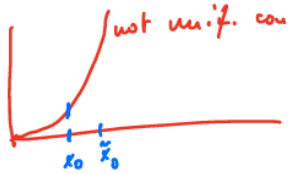
$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x_0, x \in X : d(x, x_0) < \delta \Rightarrow d(f(x), f(x_0)) < \varepsilon.$$

Uniformly continuous (2)



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

Proposition 9 (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 $\implies$ uniform continuity

Proposition 10 (Lipschitz continuous $\implies$ uniform continuous)

If f is Lipschitz continuous, then it is also uniformly continuous.

Proof sketch. 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 11 (Intermediate value theorem)

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

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

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 $f(x) = 0$.

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 12 (Invertible function)

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

Sequence of functions

Pointwise convergence

Definition 13 (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 the function $f : D \to \mathbb{R}$ if

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

Pointwise convergence (2)

Example:

Consider the functions $f_n : [0,1] \to \mathbb{R},\ f_n(x) = x^{1/n}$. The sequence f_n converges pointwise to the function $f : [0,1] \to \mathbb{R},\ f(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{otherwise} \end{cases}$.

Observe that this example also shows that it can happend that $f_n \to f$ pointwise, and all f_n are continuous, but this does not imply that f is continuous.

Uniform convergence

Definition 14 (Uniform convergence)

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

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

Uniform convergence (2)

Alternative defintion:

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

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

Uniform convergence preserves continuity

Theorem 15 (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

Want to show that f is continuous at all points x_0. Choose $x_0 \in D$, $\varepsilon > 0$. Need to show that there exists $\delta > 0$ such that $|x - x_0| < \delta \implies |f(x) - f(x_0)| \leq \varepsilon$.

Step 1: Observe that by the triangle inequality, for every $n \in \mathbb{N}$:

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

Step 2: bound the blue terms:
Because of uniform convergence, there exists $N \in \mathbb{N}$ such that for all $n > N$,

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

Proof (2)

Step 3: bound the green term

Now consider the function f_n. By assumption f_n is continuous, so for the given x_0 and $\varepsilon > 0$ there exists $\delta > 0$ such that

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

Step 4: Combine For given $\varepsilon > 0$, there exists $\delta > 0$ such that for all x,

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

Thus, f is continuous at x_0. □

Calculus: Derivatives and integrals on $\mathbb{R}$

Derivatives

Derivative

Definition 16 (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

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

Differentiable functions

Definition 17 (Differentiable functions)

Let $U \subset \mathbb{R}$ be an interval. The function $f : U \to \mathbb{R}$ 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:

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

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 18 (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.

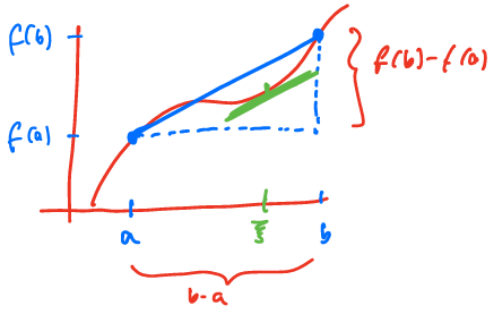
Intermediate value theorem for derivatives

Theorem 19 (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 limit and derivative

Theorem 20 (Exchanging limes and derivative)

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

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

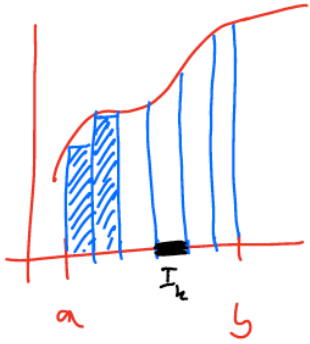
exists for all $x \in [a,b]$, and the derivatives f_n' converge uniformly, then f is continuously differentiable, and we have

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

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

Riemann-integral

Construction of the Riemann integral



Construction of the Riemann integral (2)

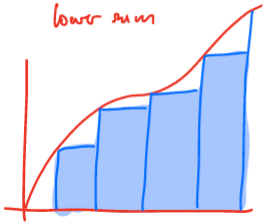
- Consider a function $f : [a, b] \to \mathbb{R}$. Assume that f is bounded.
- Consider $x_0, x_1, \ldots, x_n$ with $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. These points introduce a partition of $[a, b]$ into n intervals $I_k := [x_{k-1}, x_k]$.
- For each k define $m_k := \inf(f(I_k))$ and $M_k := \sup(f(I_k))$. Exist because f is bounded.
- Define the lower and upper sum :

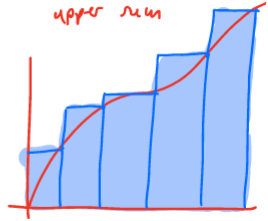
$$s\left(f, \{x_0, x_1, \ldots, x_n\}\right) = \sum_{k=1}^{n} |I_k| \cdot m_k$$

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

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

Construction of the Riemann integral (3)





Construction of the Riemann integral (4)

Now define the upper and lower Riemann integral:

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

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

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

Monotone resp. continuous functions are Riemann-integrable

Theorem 21 (Monotone resp. continuous functions are Riemann-integrable)

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

Shortcomings of the Riemann integral

- Many "simple" functions are not Riemann-integrable.

  Example: the characteristic function $f_{\mathbb{Q}} : [0,1] \to \{0,1\}$ of the rational numbers
  - For any interval $I_k = [x_{k-1}, x_k]$, we have: $M_k = 1, \quad m_k = 0.$
  - Thus: $J_* = (b-a) \cdot 0, \quad J^* = (b-a) \cdot 1.$
  - Because $J_* < J^*$, the function is not Riemann-integrable.

Shortcomings of the Riemann integral (2)

- One cannot prove strong theorems about exchanging "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.

Solution: Lebesgue integral!

Lebesgue measure SKIPPED

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.

SKIPPED

Earlier approaches

First approaches (Jordan, Riemann) attempted the following:

"Outer approximation":

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



SKIPPED

Earlier approaches (2)

"Inner approximation":

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



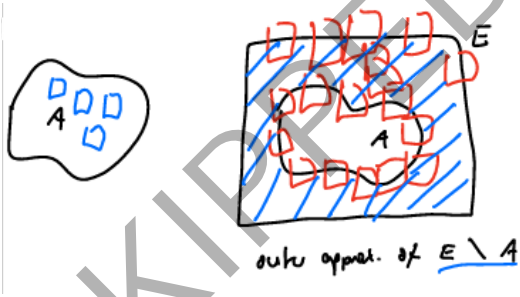
A would be called "measurable" if outer and inner approximation "converge".

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

SKIPPED

Now: generalization of this approach

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



SKIPPED

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

$$\mu(A) = \mu(E) - \mu(E \setminus A)$$

- Need σ-algebra as underlying structure.

Outer Lebesgue measure

Set the "natural volume" of rectangles:

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

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

SKIPPED

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...

SKIPPED

Measurable set

Definition 22 (Measurable set)

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 $\mathbb{R}^n$.

SKIPPED

Outer measure as measure...

Theorem 23 (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".

SKIPPED

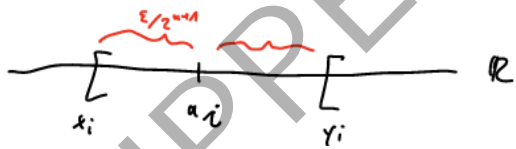
Examples:

- $\lambda(\{x\}) = 0$
- $\lambda(\mathbb{R}) = \infty$
- $A \subset \mathbb{R}$ 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 $\lambda(A) = 0$.

SKIPPED

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).

SKIPPED

A non-measurable set SKIPPED

SKIPPED

Construction (quite abstract!)

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

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

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

SKIPPED

Construction (quite abstract!) (2)

Consider the equivalence classes

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

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

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

$$\vdots$$

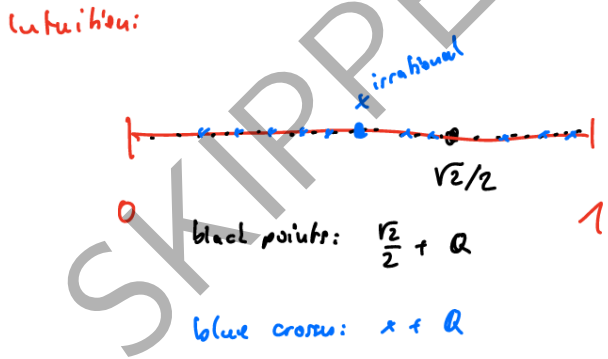
SKIPPED

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

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



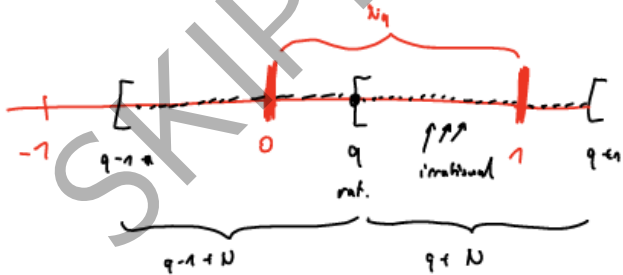
SKIPPED

Proof

Proof by contradiction:

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

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



SKIPPED

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 $\sum_q \lambda(N_q) = \infty$ ↯

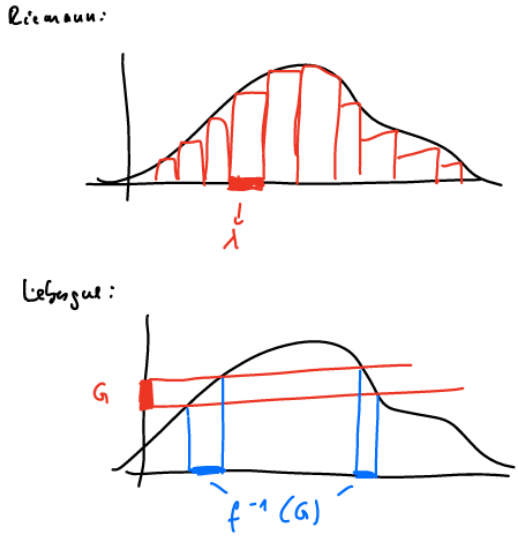
SKIPPED

The Lebesgue-integral

As we did not introduce measure theory, we keep this introduction on an informal level.

Intuition: Partition Y instead of X

Intuition: Partition Y instead of X (2)



Measurable functions

Definition 25 (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:

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

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

SKIPPED

Lebesgue-integral for simple functions

Definition 26 (Lebesgue-integral for simple functions)

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

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

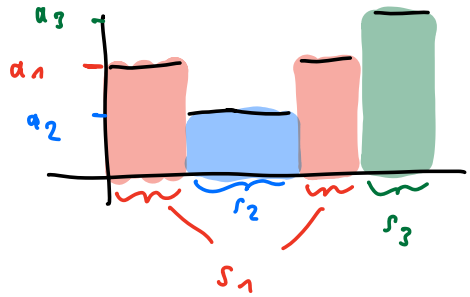
For such a simple function we can define its Lebesgue integral as

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

where λ denotes the Lebesgue measure.

Lebesgue-integral for simple functions (2)

Simple function and its integral:



$$f(x) = \sum_{i=1}^{k} a_i \mathbb{1}_{r_i}$$

$$\int f(x)\,dx = \sum_{i=1}^{k} a_i \lambda(r_i)$$

$$a_1 \lambda(r_1) + a_2 \lambda(r_2) + a_3 \lambda(r_3)$$

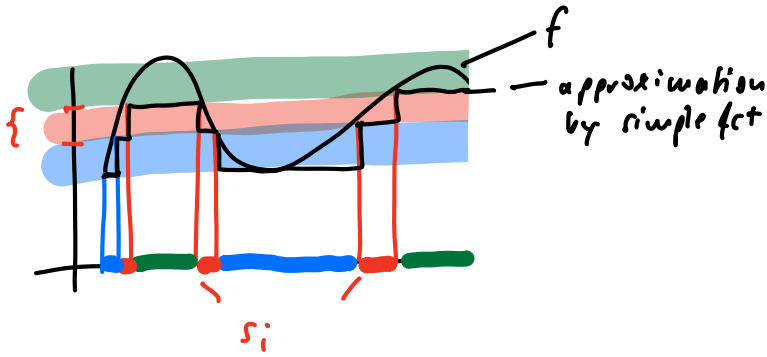
Lebesgue integral for non-negative functions

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

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

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

Lebesgue integral for non-negative functions (2)



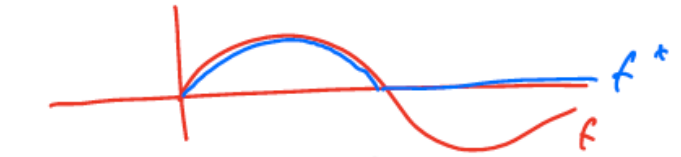
Lebesgue integral for general functions

- For a general function $f\colon \mathbb{R}^d \to \mathbb{R}$ 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

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

Lebesgue integral for general functions (3)

The Lebesgue integral is much more powerful a notion than the Riemann integral. The main reason is that we can approximate functions not just by step functions on intervals, but use much more complicated sets as steps.

Example: because the Lebesgue measure of the set of rational numbers is 0, we can derive:

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

Monotone convergence

Theorem 27 (Monotone convergence)

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

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

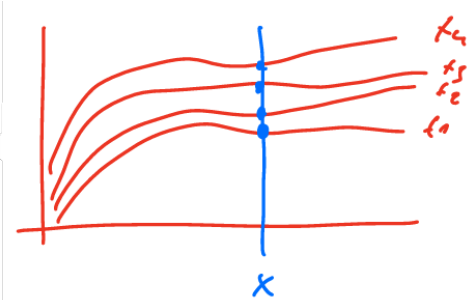
Assume that all f_k are measurable, and that the pointwise limit exists:

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

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 28 (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:

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

Fundamental Theorem of Calculus

Fundamental Theorem of Calculus

Theorem 29 (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

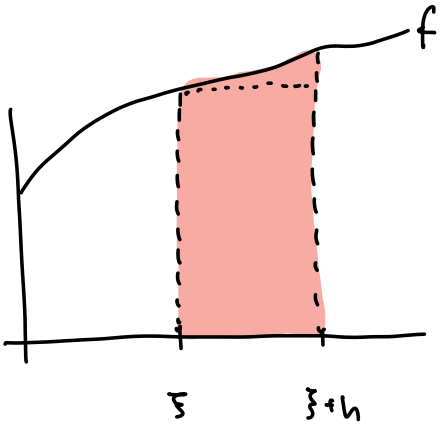
$$F(x) := \int_c^x f(t)\, dt.$$

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)



For small h, the red area is $F(x+h) - F(x)$ and it can be approximated by $h \cdot f(x)$. Thus

$$F(x+h) - F(x) \approx h \cdot f(x) \quad \text{or}$$

$$F'(x) = \frac{F(x+h) - F(x)}{h} \approx f(x)$$

Fundamental Theorem of Calculus (3)

Theorem 30 (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)

Define the space

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

Then the integral operator $I : C([a,b]) \to C_c^1([a,b])$ is an isomorphism (linear, bijective), and its inverse is the differential operator.

Proof Part I

Need to prove that F is differentiable at ξ.

$$
\begin{aligned}
A(h) &:= \frac{F(\xi + h) - F(\xi)}{h} \\
&= \frac{1}{h} \left(\int_c^{\xi+h} f(t)\, dt - \int_c^{\xi} f(t)\, dt \right) \\
&= \frac{1}{h} \int_{\xi}^{\xi+h} f(t)\, dt
\end{aligned}
$$

Want to prove: $A(h)$ converges to $f(\xi)$ as $h \to 0$, that is $A(h) - f(\xi) \overset{!}{\longrightarrow} 0$.

Proof Part I (2)

$$
\begin{aligned}
A(h) - f(\xi) &= \frac{1}{h} \int_{\xi}^{\xi+h} f(t)\, dt - f(\xi) \\
&= \frac{1}{h} \int_{\xi}^{\xi+h} f(t)\, dt - \frac{1}{h} \int_{\xi}^{\xi+h} f(\xi)\, dt \\
&= \frac{1}{h} \int_{\xi}^{\xi+h} f(t) - f(\xi)\ \, dt
\end{aligned}
$$

Intuition says: $f(t) - f(\xi)$ is small due to continuity of f at ξ, and thus the rhs converges to 0.

Proof Part I (3)

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

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

Then:

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

□

Proof Part II

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 definition 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 also know that $H(a) = F(a) - G(a) = F(a)$, thus
(iii) $H(x) \equiv F(a)$

Now 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 31 (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 32 (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

$$p(x) = \sum_{n=1}^{\infty} n^c x^n \quad \text{for some constant } c \in \mathbb{R}.$$

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:

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

A first example (3)

Case $c = 0$: We consider the series $\sum n^c x^n = \sum x^n$, the 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

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

Let's compute its derivatives:

$$\begin{aligned}
f'(x) &= \left(a_0 + a_1(x-a) + a_2(x-a)^2 + a_3(x-a)^3 + \dots\right)' \\
&= a_1 + 2a_2(x-a) + 3a_3(x-a)^2 + \dots \\
&= \sum_{n=1}^{\infty} n \cdot a_n \cdot (x-a)^{n-1}. \\
f''(x) &= \dots \\
f^{(k)}(x) &= \sum_{n=k}^{\infty} a_n \cdot n \cdot (n-1) \cdot \dots \cdot (n-k+1) \cdot (x-a)^{n-k}.
\end{aligned}$$

From power series to Taylor series, intuition (2)

In particular, for $x = a$, we have $f^{(k)}(a) = a_k \cdot k!$

Or, stated otherwise:

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

From power series to Taylor series, formally

Theorem 33 (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 sketch. 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"?

Given any function (possibly with nice assumptions), can we simply build the series

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

and "hope" that it converges to the function?

Taylor series

Taylor polynomial with integral remainder

Theorem 34 (Taylor polynomial of degree n)

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:

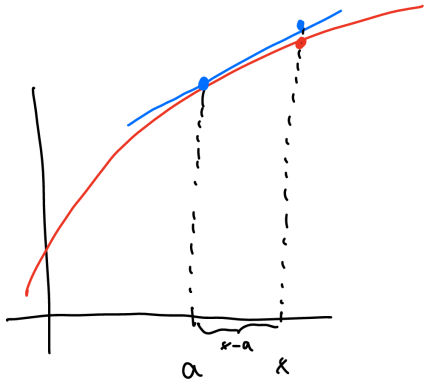
$$T_n(x,a) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \quad \text{(Taylor polynomial of degree } n\text{)},$$

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

Then

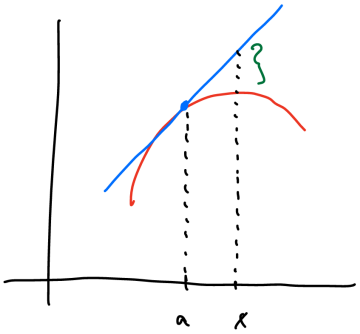
$$f(x) = T_n(x,a) + R_n(x,a).$$

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 $n = 0$:

$$T_0(x, a) = f(a)$$

$$R_0(x, a) = \int_a^x f'(t)\, dt$$

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

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 as again exploit Fundamental Theorem



Taylor with Lagrange remainder

Theorem 35 (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

For simplicity, assume $a < x$. Define:

$$F(t) := f(t) - T_n(t, a)$$
$$G(t) := (t - a)^{n+1}.$$

- By construction, for $k = 0, \ldots, n$:

$$F^{(k)}(a) = G^{(k)}(a) = 0.$$

- By the mean value theorem for the deriviatives, there exists $\xi_1 \in (a, x)$ such that

$$\frac{F(x)}{G(x)} = \frac{F'(\xi_1)}{G'(\xi_1)}.$$

Proof (2)

- Applying the same argument repeatedly (on the intervals (a, ξ_1), (a, ξ_2), etc.), we obtain

$$\frac{F(x)}{G(x)} = \frac{F'(\xi_1)}{G'(\xi_1)} = \cdots = \frac{F^{(n+1)}(\xi)}{G^{(n+1)}(\xi)}.$$

- Since

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

we obtain

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

☺

Taylor convergence

Theorem 36 (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: $f(x) = \log(1 + x)$. Consider Taylor series about $a = 0$:

$$\log(1 + x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

Convergence radius of Taylor series is $r = 1$. For $|x| > 1$ the series diverges.

Examples (2)

- Special example:

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

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

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 for all $x \neq 0$ that

$$T_n(x) \neq f(x)$$

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

Calculus: Derivatives and integrals on $\mathbb{R}^d$

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.

$\rightsquigarrow$ partial derivatives

(2) Approximate f locally by something linear.

$\rightsquigarrow$ 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 37 (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.

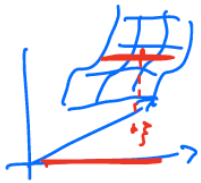
$$\frac{\partial f}{\partial x_j}(\bar{x}) := \lim_{h \to 0} \frac{f(\bar{x} + e_j \cdot h) - f(\bar{x})}{h}$$

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

constant > 0

"round" delta-sign

variable with 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:

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

The first row is $\operatorname{grad}(f_1)^\top$, second row is $\operatorname{grad}(f_2)^\top$, ...

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

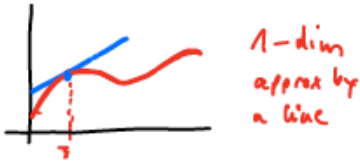
Consider $f : \mathbb{R}^n \to \mathbb{R}^m$, $\xi \in U$. f is differentiable at ξ if there exists a **linear mapping** $L : \mathbb{R}^n \to \mathbb{R}^m$ such that for $h \in \mathbb{R}^n$:

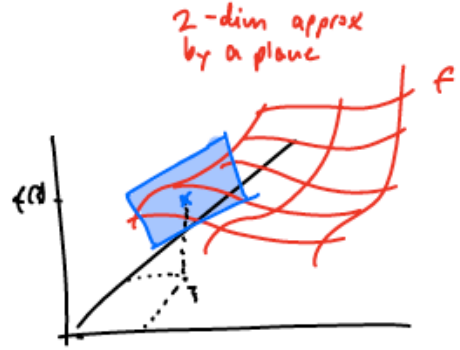
$$f(\xi + h) - f(\xi) = L(h) + r(h),$$

with

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

Differentiable function (2)





Differentiable function (3)

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

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

Differentiable, Continuous, Gradient

On the following slides we always assume that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at ξ.
(Holds for all of the following theorems)

Theorem 38 (Continuous)

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at ξ. Then f is continuous at ξ.

Differentiable, Continuous, Gradient (2)

Theorem 39 (Gradient and linear function)

The linear functional L coincides with the gradient: for $h = (h_1, ..., h_n)^\top$ we have

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

or equivalently,

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

In particular, for $h = e_j$ (the j-th unit vector), we have

$$\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) \cdot h_j = \frac{\partial f(x)}{\partial x_j}$$

Differentiable, Continuous, Gradient (3)

Theorem 40 (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 41 (Continuous Partial Derivatives $\Longrightarrow$ Differentiable)

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

Directional Derivatives

Directional Derivatives

Definition 42 (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 $\implies$ directional derivatives

Theorem 43 (Differentiable Directional $\implies$ 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 44 (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:

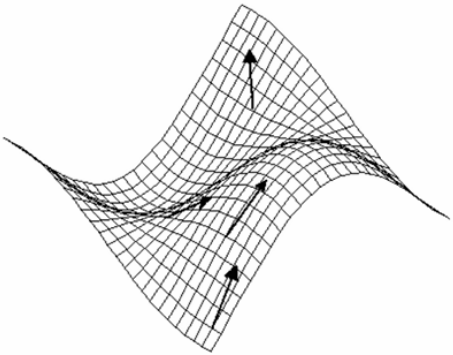
$$\max_{\{v;\|v\|=1\}} |D_v f(\xi)| = \max_{\{v;\|v\|=1\}} |\langle \nabla f(\xi), v\rangle| \overset{\text{Cauchy-Schwarz}}{\leq} \max_{\{v;\|v\|=1\}} \|\nabla f(\xi)\| \cdot \|v\| = \|\nabla f(\xi)\|.$$

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 attained 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 $(0,y)$ (for all y) we have

$$\frac{\partial f}{\partial x}(0,y) = y \text{ and thus } \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = 1.$$

At $(x,0)$ (for all x) we have

$$\frac{\partial f}{\partial y}(x,0) = 0 \text{ and thus } \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = 0.$$

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

Hessian

Definition 45 (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 46 (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.

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

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

Continuously Differentiable $\Rightarrow$ can change order of derivatives

Theorem 47 (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 48 (Critical point)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. We say that x is a critical point if

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

⚠ 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

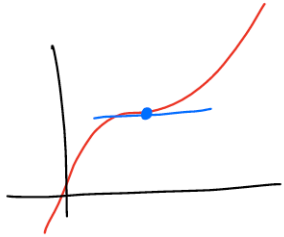
$$\forall x \in \mathbb{R}^n : f(x) > f(x_0).$$

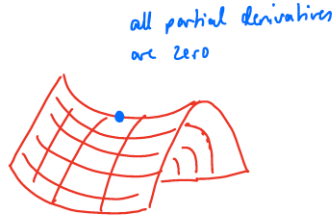
Minimum (2)



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:

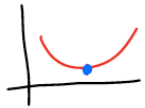




How can we find out which case we have?

Intuition in 1-dimensional case: second derivatives might help.

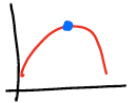
Strict local min:



$f'(x) = 0$

$f''(x) > 0$

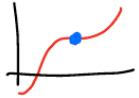
Strict local max



$f'(x) = 0$

$f''(x) < 0$

saddle pt:



$f'(x) = 0$

$f''(x) = 0$

Critical Points and the Hessian

Theorem 49 (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.

In matrix notation:

$$X = \begin{pmatrix} \cdots x_1 \cdots \\ \cdots x_2 \cdots \\ \cdots \vdots \cdots \\ \cdots x_n \cdots \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

Example: Linear least squares (2)

Want to find linear function f with parameter vector w

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

that has minmal least sqares loss. So we want to determine w as the solution of the following optimization problem:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} | \underbrace{\langle x_i, w \rangle}_{f(x_i)} - y_i|^2 = \min_{w \in \mathbb{R}^d} \underbrace{\|Xw - Y\|^2}_{=:g(w)}.$$

We call $g(w)$ the objective function of the optimization problem.

Computing the gradient "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:

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

Writing the gradient in matrix notation

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 that this "looks similar" to the 1-dimensional case: for the objective function $g(w) = (y - xw)^2$ we have

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

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:

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

$$f(X) = \mathrm{tr}(A \cdot X) \quad \Rightarrow \quad \frac{\partial f}{\partial X} = A.$$

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

- 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:

$$f(A) = A^{-1}, \quad f_{ij} := (A^{-1})_{ij} \quad \Rightarrow \quad \frac{\partial f_{ij}}{\partial a_{uv}} = -(a_{iu})^{-1}(a_{vj})^{-1}.$$