A course of analysis for computer scientists

Aleš Pultr

Contents:

1st semester

2nd semester

3rd semester

1st semester

I. Preliminaries

1. Basics

1.1. Logic. The logical connectives "and" and "or" will be as a rule expressed by words, while for the implication we will use the standard symbol "⇒". Negation of a statement A will be expressed by "non$A$". The reader is certainly acquainted with the fact that

$$\text{“}A \;\Rightarrow\; B\text{”} \quad \text{is equivalent with} \quad \text{“non}B \;\Rightarrow\; \text{non}A\text{”}.$$

This is used as a standard trick in proofs.

The quantifier $\exists$ in "$\exists x \in M, A(x)$" indicates that there exists an $x \in M$ such that $A(x)$ holds; often the M is obvious and we write just $\exists x A(x)$. Similarly, the quantifier $\forall$ in "$\forall x \in M, A(x)$" indicates that $A(x)$ holds for all $x \in M$; again, if the range M is obvious we often write just $\forall x A(x)$.

1.2. Sets. $x \in A$ indicates that x is an element of a set A.

We will use the standard symbols for unions:

$$A \cup B, \quad A_1, \cup \cdots \cup A_n, \quad \bigcup_{i \in J} A_i$$

and for intersections:

$$A \cap B, \quad A_1, \cap \cdots \cap A_n, \quad \bigcap_{i \in J} A_i.$$

The difference of sets A, B, that is, the set of all the elements in A that are not in B is denoted by

$$A \smallsetminus B.$$

Recall the DeMorgan formulas

$$A \smallsetminus \bigcup_{i \in J} B_i = \bigcap_{i \in J} (A \smallsetminus B_i) \quad \text{and} \quad A \smallsetminus \bigcap_{i \in J} B_i = \bigcup_{i \in J} (A \smallsetminus B_i).$$

The set of all x that satisfy a condition P is denoted by

$$\{x\,|\,P(x)\}.$$

Thus for instance $A \cup B = \{x\,|\,x \in A \text{ or } x \in B\}$, or $\bigcap_{i\in J} A_i = \{x\,|\,\forall i \in J,\ x \in A_i\}$.

The *cartesian product*

$$A \times B$$

is the set of all pairs (a, b) with $a \in A$ and $b \in B$. We will also work with cartesian products

$$A_1 \times \cdots \times A_n,$$

the systems of n-tuples $(a_1, \dots, a_n)$, $a_i \in A_i$, and later on also with

$$\prod_{i\in l} A_i = \{(a_i)_{i\in J}\,|\,a_i \in A_i\}.$$

The formula $A \subseteq B$ (read "A is a subset of B") indicates that $a \in A$ implies $a \in B$.

The set of all subsets of a set A ("*the powerset of* A") is often denoted by

$$\exp A \quad \text{or} \quad \mathfrak{P}(A).$$

1.3. Equivalence. Decomposition into equivalence classes. An *equivalence* E on a set X is a *reflective, symmetric and transitive* relation $E \subseteq X \times X$, that is, a relation such that

$$\begin{aligned}
&\forall x,\ xEx && \text{(reflexivity)}\\
&\forall x, y,\ xEy \text{ and } yEx \text{ implies } x = y && \text{(reflexivity)}\\
&\forall x, y, z\ xEy \text{ and } yEz \text{ implies } xEz && \text{(transitivity).}
\end{aligned}$$

(We write xEy for $(x, y) \in E$). Set

$$Ex = \{y\,|\,yEx\}.$$

These sets are called the *equivalence classes* of E. We have

1.3.1. Proposition. *Each equivalence on a set X yields a disjoint decomposition into its equivalence classes. On the other hand, each disjoint decomposition*

$$X = \bigcup_{i\in J} X_i$$

gives rise to an equivalence defined by

$$xEy \quad \textit{iff} \quad \exists i,\ x,y \in X_i.$$

Proof. The second statement is obvious. For the first one we have to prove that for any two x, y we have either $Ex = Ey$ or $Ex \cap Ey = \emptyset$. Now if $z \in Ex \cap Ey$ then $xEzEy$, hence xEy, and then, by transitivity again, $z \in Ex$ iff $z \in Ey$. $\square$

Note that in fact we have here a one-to-one correspondence between all equivalences on X and all disjoint decompositions of X.

1.4. Mappings. A *mapping* $f : A \to B$ is the following collection of data:

(1) a set X, called the *domain* of f,

(2) a set Y, called the *range* (or the *codomain*) of f,

(3) and a subset $f \subseteq X \times Y$ such that

- for each $x \in X$ there is a $y \in Y$ such that $(x, y) \in f$, and
- if $(x, y) \in f$ and $(x, z) \in f$ then $x = y$.

The unique y from (3) is usually denoted by $f(x)$ (one sometimes speaks of the value of f in the argument x). It can often be expressed by a formula (for instance $f(x) = x^2$); we have to keep in mind, however, that the domain and range are essential: sending an integer x to the integer x^2 is a different function than sending a real x to the real x^2, and sending a real x to the real x^2 with the range restricted to the non-negative real numbers is yet another one.

A mapping $f : X \to Y$ is *one-to-one* if

$$\forall x, y \in X,\ (x \neq y \ \Rightarrow\ f(x) \neq f(y));$$

it is *onto* if

$$\forall y \in Y \exists x \in X \quad f(x) = y.$$

Note the importance of the information what the range Y is for the latter property.

The *identity mapping* $\mathrm{id}_X : X \to X$ is defined by $\mathrm{id}(x) = x$.

The *image* of a subset $A \subseteq X$ under a mapping $f : X \to Y$, that is, $\{f(x) \,|\, x \in A\}$ will be denoted by $f[A]$, and the preimage $\{x \,|\, f(x) \in B\}$ of $B \subseteq Y$ will be denoted by $f^{-1}[B]$.

1.4.1. Composition of mappings. Given mappings $f : X \to Y$, $g : Y \to X$ we obtain their *composition*

$$g \circ f : X \to Z$$

by setting $(g \circ f)(x) = g(f(x))$.

The *inverse* of a mapping $f : X \to Y$ is a mapping $g : Y \to X$ such that

$$gf = \mathrm{id}_X \quad \text{and} \quad fg = \mathrm{id}_Y.$$

Note that if f has an inverse then it is one-to-one and onto; on the other hand, each one-to-one onto map has a (unique) inverse.

1.4.1. Functions. Mappings $f : X \to Y$ where the range Y is a subset of a system of numbers (natural numbers, integers, rationals, reals, complex numbers – see below) are often called *functions.* We will be in particular concerned with *real functions*, that is, $Y \subseteq \mathbb{R}$. Moreover, in the first months we will have also $Z \subseteq \mathbb{R}$, and speak of *real functions of one real variable.*

2. Numbers.

2.1. Natural numbers. They are supposed to be well known, but let us recall a formal approach (Peano axioms). We have a set

$$\mathbb{N}$$

endowed, first, with a distinguished element 0 and a mapping $\sigma : \mathbb{N} \to \mathbb{N}$ (the *successor function*; we will usually write simply n' for $\sigma(n)$) such that

(1) for each $n \neq 0$ there is precisely one m such that $m' = n$,

(2) 0 is not a successor,

(3) if a statement A holds for 0 (symbolically, $A(0)$) and if $A(n) \Rightarrow A(n')$ then $\forall n A(n)$.

(The last is called the *axiom of induction.*)

Further, there are operations $+$ and $\cdot$ (the latter will be as a rule indicated simply by juxtaposition) such that

$$\begin{aligned} &n+0=n, \quad n+m'=(n+m)', \\ &n\cdot 0=0, \quad nm'=nm+n. \end{aligned}$$

Finally we define an order $n\leq m$ by setting

$$n\leq m \quad \text{iff} \quad \exists k, m=n+k.$$

2.1.1. This results in a system $(\mathbb{N},+,\cdot,0,1,\leq)$ (1 is $0'$, the successor of 0) satisfying

$n+0=n, \quad n\cdot 1=n,$

$m+(n+p)=(m+n)+p, \quad m(np)=(mn)p$ (associativity rules)

$m+n=n+m. \quad mn=nm$ (commutativity rules)

$m(n+p)=mn+mp$ (distributivity)

$n\leq n, \quad m\leq n$ and $n\leq m$ implies $n=m$ (reflexivity and antisymmetry)

$m\leq n$ and $n\leq p$ implies $m\leq p$ (transitivity)

$\forall m,n$ either $n\leq m$ or $m\leq n$

$m\leq n$ implies $n+p\leq m+p$

$m\leq n$ implies $np\leq mp$.

It is an amusing exercise to prove (at least some) of these rules by induction from the axioms above.

2.2. Integers. The set of integers

$$\mathbb{Z}$$

is obtained augmenting $\mathbb{N}$ by negative numbers. The reader can try to find a formal construction (for instance one can add new elements $(n,-)$ with $n\in\mathbb{N}$, $n\neq 0$ and define suitably the operations and order (the only point in which one has to do something not quite obvious is the definition of addition). One obtains a system

$$\mathbb{Z}$$

where all the rules from 1.1 hold with the exception of the last one which has to be replaced by

$$x \leq y \text{ and } z \geq 0 \;\Rightarrow\; xz \leq yz.$$

On the other hand one has one more rule, namely

$$\forall x \; \exists y \;\; \text{such that} \;\; x + y = 0$$

which alows, besides adding and multiplying, also subtracting.

2.3. Rational numbers. We can already add, myltiply and subtract. The arithmetic operation missing is unrestricted division. One cannot have quite unrestricted division (from rules like those above one sees that $0 \cdot x = 0$ hence dividing by 0 does not make much sense. But this will be the only exception in the following system of rational numbers. First take (for instance)

$$X = \{(x, y) \,|\, x, y \in \mathbb{Z}, y \neq 0\}$$

and define

$$(x, y) + (u, v) = (xv + yu, uv) \;\; \text{and} \;\; (x, y)(u, v) = (xu, yv)$$

Then consider the equivalence relation

$$(x, y) \sim (u, v) \;\; \text{if and only if} \;\; xv = uy$$

and set

$$\mathbb{Q} = X / \sim .$$

It is easy tu prove that if

$$(x, y) \sim (x', y') \;\; \text{and} \;\; (u, v) \sim (u', v')$$

then

$$(x, y) + (u.v) \sim (x', y') + (u', v') \;\; \text{and} \;\; (x, y)(u, v) \sim (x'.y')(u', v')$$

(prove it as a simple exercise) and that this allows for defining addition and multiplication on $\mathbb{Q}$, and that we then have, for the equivalence classes (0 is the equivalence class of $(0, n)$ and 1 is the equivalence class of (n, n))

$$\begin{aligned}
&x + 0 = n, \quad x \cdot 1 = x, \\
&x + (y + z) = (x + y) + z, \quad x(yz) = (xy)z && \text{(associativity rules)} \\
&x + y = y + x. \quad xy = yx && \text{(commutativity rules)} \\
&x(y + z) = xz + yz && \text{(distributivity)} \\
&\forall x \exists y,\ x + y = 0 \\
&\forall x \neq 0 \exists y,\ xy = 1.
\end{aligned}$$

Systems satisfying these rules are called *commutative fields.*

Furthermore one can define a relation $\leq$ by

$$(x, y) \leq (u, v) \text{ for } y, v > 0 \ \text{ by } \ xv \leq yu$$

which results in an order on $\mathbb{Q}$ satisfying

$$\begin{aligned}
&x \leq x, \quad x \leq y \text{ and } y \leq x \text{ implies } x = y && \text{(reflexivity and antisymmetry)} \\
&x \leq y \text{ and } y \leq z \text{ implies } x \leq z && \text{(transitivity)} \\
&\forall x, y \text{ either } x \leq y \text{ or } y \leq x \\
&x \leq y \text{ implies } x + z \leq y + z \\
&x \leq y \text{ and } z > 0 \text{ implies } xz \leq yz.
\end{aligned}$$

On speaks of an *ordered (commutative) field.*

It is perhaps not necessary to recall that one stamdardly uses the symbol

$$\frac{p}{q}$$

for the equivalence class containing (p, q).

2.4. Rational numbers are not quite satisfactory. So now we have a system in which we can add, subtract, multiply and divide. Also, it seems to be ordered in a satisfactory way (though it will turn out that improving the order will be the key to solving difficulties).

However, already the old Greeks observed a serious trouble. Suppose you would like to attach lenghts to the segments in natural geometrical constructions. Inevitably you will come to the task to determine square roots. And this one cannot do in the realm of rational numbers.

Suppose $\sqrt{2}$, a number x such that $x^2 = 2$, can be expressed as a rational number, that is we have integers p, q such that

$$\left(\frac{p}{q}\right)^2 = 2.$$

We can assume that the integers p, q are coprime (that is, have no non-trivial divisor).

We have

$$\frac{p^2}{q^2} = 2, \quad \text{that is,} \quad p^2 = 2q^2$$

and hence p has to be even. But then p^2 is divisible by 4, which makes also q even and hence p, q are both divisible by 2, a contradiction.

2.5. Order, suprema and infima. A *linear order* on a set X is a relation $\leq$ satisfying

$$\begin{aligned} &x \leq x && \text{(reflexivity)} \\ &x \leq y \text{ and } y \leq x \text{ implies } x = y && \text{(antisymmetry)} \\ &x \leq y \text{ and } y \leq z \text{ implies } x \leq z && \text{(transitivity)} \\ &\forall x, y \text{ either } x \leq y \text{ or } y \leq x && \text{(linearity)} \end{aligned}$$

If we require just reflexivity, antisymmetry and transitivity we speak of a *partial order*.

An *upper bound* of a subset M of a partially ordered set $(X, \leq)$ is an element b such that

$$\forall x \in M,\ x \leq b;$$

M is said to be bounded (from above) if there is an upper bound of M.

Similarly, we speak of a *lower bound* b of M if

$$\forall x \in M,\ x \geq b,$$

and M is said to be bounded (from below) if there is a lower bound of M.

Very often it is obvious whether the boundedness is required from above or from below and we speak just of a *bounded* set.

A *supremum* of a subset $M \subseteq (X, \leq)$ is the least upper bound of M (needless to say, it does not have to exist). If it exists, it is denoted by

$$\sup M.$$

More explicitly, $s \in X$ is a supremum of M if

(1) for all $x \in M$, $x \leq s$, and

(2) if $x \leq y$ for all $x \in M$ then $s \leq y$.

In a linearly ordered set this is equivalent with

(1) for all $x \in M$, $x \leq s$, and

(2) if $y < s$ then there exists an $x \in M$ such that $y < x$.

The second formulation has its advantages, and we will use it more often than the first one.

Similarly, an *infimum* of M is the greatest lower bound of M. If it exists, it is denoted by

$$\inf M.$$

More explicitly, $i \in X$ is an infimum of M if

(1) for all $x \in M$, $x \geq i$, and

(2) if $x \geq y$ for all $x \in M$ then $i \geq y$

and in a linearly ordered set this is equivalent with

(1) for all $x \in M$, $x \geq i$, and

(2) if $y > i$ then there exists an $x \in M$ such that $y > x$.

Obviously, a supremum resp. infimum is uniquely determined (if it exists).

2.5.1. Example. Recall the trouble with the square root of 2 in 4. Note that in $\mathbb{Q}$ the set $\{x \,|\, 0 \leq x,\ x^2 \leq 2\}$ is bounded (from above) but has no supremum. Similarly, $\{x \,|\, 0 \leq x,\ x^2 \geq 2\}$ is bounded (from below) but has no infimum.

2.5.2. Exercise. Prove that for *linearly* ordered sets the two variants of definitions of supremum resp. infimum are indeed equivalent. How do you use the linearity requirement? Why is it necessary?

2.6. Real numbers. The system of real numbers

$$\mathbb{R}$$

as we will use them, is a completion (in more than one sense of the word) of $\mathbb{Q}$. It is an *ordered commutative field* in which

every non-empty (from above) bounded subset has a supremum. (sup)

In working with reals we will use just the properties listed in 3 and (sup).

2.6.1. Proposition. *In* $\mathbb{R}$ *every non-empty (from below) bounded subset has an infimum.*

Proof. Let M be non-empty and bounded from below. Set

$$N = \{x \mid x \text{ is a lower bound of } M\}.$$

Since M is bounded from below, N is non-empty. Since M is non-empty, N is bounded from above (each $y \in M$ is an upper bound of N). Hence there exists

$$i = \sup N.$$

Now since each $x \in M$ is an upper bound of N, $i \leq x$ for all $x \in M$. On the other hand, if y is a lower bound of M, y is in N and hence $y \leq i = \sup N$. □

3. Real numbers as (Euclidean) line.

3.1. Absolute value. Recall the *absolute value* of a real number

$$|a| = \begin{cases} a \text{ if } a \geq 0, \\ -a \text{ if } a \leq 0 \end{cases}$$

3.1.1. Obviously we have

Observation. $|a + b| \leq |a| + |b|$.

This inequality, called *triangle inequality* will be very often used in proofs, usually without specific mentioning.

3.2. The metric structure of $\mathbb{R}$: the real line. The system of real numbers will be endowed with the *distance*

$$|x - y|.$$

Thus we can view it as (a.o.) a Euclidean line.

Note that this is where the expression "triangle inequality" comes from: setting $a = x - y$ and $b = y - z$ we obtain from 3.1.1

$$|x - z| \leq |x - y| + |y - z|$$

(that is, $\mathrm{dist}(x, z) \leq \mathrm{dist}(x, y) + \mathrm{dist}(y, z)$).

3.3. Note: Summary. Realize that the system $\mathbb{R}$ is quite an involved structure. It is

- a commutative field (algebra with addition, multiplication, subtraction and division),
- a linearly ordered set, and
- a (metric) space.

3.4. Aside: complex (Gauss) plane. The triangle inequality on the line is of course a very simple matter. Let us present a more involved one. We will not need complex nimbers for some time, but let us discuss for a moment their geometric structure. For a complex number $a = x + iy$ we have the *complex conjugate* $\overline{a} = x - iy$ and the absolute value

$$|a| = a \cdot \overline{a} = x^2 + y^2.$$

Note that if we view a complex number $x + iy$ as the point (x, y) in the Euclidean plane we have $|a|$ the standard distance from (0.0), and

$$|a - b|$$

the standard Pythagorean distance of points a and b. The system of complex numbers viewed in this perspective is called the *Gauss plane.* We have

3.4.1. Proposition. *For the absolute value of complex numbers one has*

$$|a + b| \leq |a| + |b|.$$

Proof. Let $a = a_1 + ia_2$ and $b = b_1 + ib_2$. We can assume $b \neq 0$. For any real number λ we have $0 \leq (a_j + \lambda b_j)^2 = a_j^2 + 2\lambda a_j b_j + \lambda^2 b_j$, $j = 1, 2$. Adding these inequalities, we obtain

$$0 \leq |a|^2 + 2\lambda(a_1 b_1 + a_2 b_2) + \lambda^2 |b|^2.$$

Setting $\lambda = -\frac{a_1b_1+a_2b_2}{|b|^2}$ yields

$$0 \le |a|^2 - 2\frac{(a_1b_1+a_2b_2)^2}{|b|^2} + \frac{(a_1b_1+a_2b_2)^2}{|b|^4}|b|^2 = |a|^2 - \frac{(a_1b_1+a_2b_2)^2}{|b|^2}$$

and hence $(a_1b_1+a_2b_2)^2 \le |a|^2|b|^2$. Consequently,

$$\begin{aligned}|a+b|^2 = (a_1+b_1)^2 + (a_2+b_2)^2 = |a|^2 + 2(a_1b_1+a_2b_2) + |b|^2 \le \\ \le |a|^2 + 2|a||b| + |b|^2 = (|a|+|b|)^2. \quad \Box\end{aligned}$$

3.4.2. There are proofs concerning complex numbers that are formally literal repetitions of proofs concerning real ones, but depending on the triangle inequalities. Note that the complex variant proved may be a considerably deeper fact.

II. Sequences of real numbers.

1. Sequences and subsequences

1.1. A(n infinite) *sequence* is an array

$$x_0, x_1, \dots, x_n, \dots .$$

Thus it is, in fact, a mapping $x : \mathbb{N} \to \mathbb{R}$ written as a “table”, that is, a mapping given by the formula $x(n) = x_n$.

Note. Indexing by $0, 1, 2, \dots$ is not essential, the order in the array is. We can have a sequence

$$x_1, x_2, \dots, x_n, \dots$$

or

$$x_1, x_4, \dots, x_{n^2}, \dots$$

etc.; if we wish to see them as tables of mappings as mentioned, we then have, say, $x(n) = x_{n+1}$, or $x(n) = x_{(n+1)^2}$ etc.. See subsequences below that are, of course, themselves sequences.

1.1.1, Our sequences will be mostly infinite but let it be noted that one also speaks of *finite sequences*

$$x_1, x_2, \dots, x_n.$$

and similar.

1.2. Subsequences. A *subsequence* of a sequence

$$x_0, x_1, \dots, x_n, \dots$$

is any sequence

$$x_{k_0}, x_{k_1}, \dots, x_{k_n}, \dots$$

with k_n natural numbers such that

$$k_0 < k_1 < \cdots < k_n < \cdots .$$

Viewing a sequence as a mapping $x : \mathbb{N} \to \mathbb{R}$ as meantioned above we see that a subsequence is a composition $x \circ k$ with $k : \mathbb{N} \to \mathbb{N}$ *increasing*, that is, such that $m < n$ implies $k(m) < k(n)$.

1.2.1. Notation. A sequence $x_1, x_2, \dots$ will be denoted by

$$(x_n)_n;$$

thus the subsequence above will be $(x_{k_n})_n$.

1.3. A sequence $(x_n)_n$ is said to be *increasing, non-decreasing, non-increasing, decreasing*, respectively if

$$m < n \quad \Rightarrow \quad x_m < x_n,\ x_m \leq x_n,\ x_m \geq x_n,\ x_m > x_n \text{ respectively.}$$

2. Convergence. Limit of a sequence

2.1. Limit. We say that a number L is a *limit* of a sequence $(x_n)_n$ and write

$$\lim_n x_n = L$$

if

$$\forall \varepsilon > 0\ \exists n_0 \text{ such that } \forall n \geq n_0,\ |x_n - L| < \varepsilon. \qquad (*)$$

We than say that $(x_n)_n$ *converges* to L, or, without specifying L, that it is *convergent.* Otherwise we speak of a *divergent* sequence.

Using the symbol $\lim_n x_n$ automatically includes stating that the limit exists.

2.1.1. The following formula is obviously equivalent to $(*)$.

$$\forall \varepsilon > 0 \exists n_0 \text{ such that } \forall n \geq n_0,\ L - \varepsilon < x_n < L + \varepsilon.$$

It is easy to visualise (for sufficiently large n, x_n is in an arbitrarily small "ε-neighborhood" of L) and very often easier to work with.

2.1.2. Note. A typical divergent sequence is not a sequence growing over all bounds, like for instance $1, 2, 3, \dots$. Here we can obtain a sort of convergence augmenting the reals by infinites $+\infty$ and $-\infty$ as we will see later. Rather think of sequences like $0, 1, 0, 1, \dots$.

2.2. Observations. 1. *The limit of a constant sequence $x, x, x, \dots$ is x.*
2. *A limit, if it exists is uniquely determined.*

3. *Each subsequence of a convergent sequence converges, and namely to the same limit.*

(Indeed, as for 2, suppose L and K are limits of $(x_n)_n$. For any $\varepsilon > 0$ and sufficiently large n we have $|L-K| = |L-x_n+x_n-K| \le |L-x_n|+|x_n-K| < 2\varepsilon$. For 3 realize that $k_n \ge n$.)

2.2.1. Note. On the other hand, a divergent sequence can have convergent subsequences. Of course, however, if $x_p, x_{p+1}, x_{p+2}, \dots$ (that is, the subsequence with $k_n = p+n$) converges then $(x_n)_n$ converges.

2.3. Proposition. *Let* $\lim a_n = A$ *and* $\lim b_n = B$ *exist. Then* $\lim(\alpha a_n)$, $\lim(a_n + b_n)$, $\lim(a_n \cdot b_n)$ *and, if all* b_n *and* B *are non-zero, also* $\lim \frac{a_n}{b_n}$ *exist and we have*

(1) $\lim(\alpha a_n) = \alpha \lim a_n$,

(2) $\lim(a_n + b_n) = \lim a_n + \lim b_n$,

(3) $\lim(a_n \cdot b_n) = \lim a_n \cdot \lim b_n$,

(4) $\lim \frac{a_n}{b_n} = \frac{\lim a_n}{\lim b_n}$.

Notes before the proof. 1. Realize that the role of the $\varepsilon > 0$ in the definition of limit above is that of an "arbitrary small positive real number" the precise value of which is not quite so important. Thus it suffices, for instance, to prove that for each $\varepsilon > 0$ there is an n_0 such that for $n \ge n_0$ you have $|x_n - L| < 100\varepsilon$ (you could have determined the n_0 for $\frac{1}{100}\varepsilon$ instead of ε, to begin with.

2. Remember the trick of adding 0 in the form of $x - x$ (here, $x = a_n B$) in proving (3). It will be used more often.

Proof. (1): We have $|\alpha a_n - \alpha A| = |\alpha||a_n - A|$. Thus, if $|a_n - A| < \varepsilon$ we have $|\alpha a_n - \alpha A| < |\alpha|\varepsilon$.

(2) If $|a_n - A| < \varepsilon$ and $|b_n - B| < \varepsilon$ then $|(a_n + b_n) - (A+B)| = |a_n - A + b_n - B| \le |a_n - A| + |b_n - B| < 2\varepsilon$.

(3) If $|a_n - A| < \varepsilon$ and $|b_n - B| < \varepsilon$ then

$$\begin{aligned}|a_n b_n - AB| &= |a_n b_n - a_n B + a_n B - AB| \le \\ &\le |a_n b_n - a_n B| + |a_n B - AB| = |a_n||b_n - B| + |B||a_n - A| < \\ &< (|A|+1)|b_n - B| + |B||a_n - A| < (|A|+|B|+1)\varepsilon\end{aligned}$$

(we have used the obvious fact that if $\lim a_n = A$ then, for sufficiently large n, $|a_n| < |A| + 1$).

(4) In view of (3) it suffices to prove that $\lim \frac{1}{b_n} = \frac{1}{\lim b_n}$. Let $|b_n - B| < \varepsilon$. Then

$$\left|\frac{1}{b_n} - \frac{1}{B}\right| = \left|\frac{b_n - B}{b_n B}\right| = \left|\frac{1}{b_n B}\right| |b_n - B| \leq \left|\frac{2}{BB}\right| |b_n - B| < \left|\frac{2}{BB}\right| \varepsilon.$$

since obviously if $\lim b_n = B \neq 0$ then, for sufficiently large n, $|b_n| > \frac{1}{2}|B|$. □

2.4. Proposition. *Let* $\lim a_n = A$ *and* $\lim b_n = B$ *exist and let* $a_n \leq b_n$ *for all* n. *Then* $A \leq B$.

Proof. Suppose not. Then $\varepsilon = A - B > 0$. Choose n such that $|a_n - A| < \frac{1}{2}\varepsilon$ and $|b_n - B| < \frac{1}{2}\varepsilon$; then $a_n > A + \frac{\varepsilon}{2}$ and $b_n < B - \frac{\varepsilon}{2}$, and hence $a_n > b_n$, a contradiction. □

2.5. Proposition. *Let* $\lim a_n = A = \lim b_n$ *and let* $a_n \leq c_n \leq b_n$ *for all* n. *Then* $\lim c_n$ *exists and is equal to* A.

Proof. Choose n_0 such that for $n \geq n_0$ we have $|a_n - A| < \varepsilon$ and $|b_n - A| < \varepsilon$. Then

$$A - \varepsilon < a_n \leq c_n \leq b_n < A + \varepsilon.$$

Use 2.1.1. □

2.6. Proposition. *A bounded (from above) non-decreasing sequence of real numbers converges to its supremum.*

Proof. As $\{x_n \,|\, n \in \mathbb{N}\}$ is non-empty and bounded, it indeed has a supremum s. If ε is greater than zero there has to be an n_0 such that $s - \varepsilon < x_{n_0}$ and then for all $n \geq n_0$,

$$s - \varepsilon < x_{n_0} \leq x_n \leq s.$$

Use 2.1.1. □

2.7. Theorem. *Let* a, b *be real numbers and let* $a \leq x_n \leq b$ *for all* n. *Then there is a subsequence* $(x_{k_n})_n$ *of* $(x_n)_n$ *convergent in* $\mathbb{R}$, *and* $a \leq \lim_n x_{k_n} \leq b$.

Proof. Set

$$M = \{x \,|\, x \in \mathbb{R},\ x \leq x_n \text{ for infinitely many } n\}.$$

M is non-empty since $a \in M$ and b is an upper bound of M. Hence there exists $s = \sup M$ and we have $a \leq s \leq b$.

For every n the set

$$K(n) = \{n \,|\, s - \frac{1}{n} < x_n < s + \frac{1}{n}\}$$

is infinite. Indeed, by 2.5 (second formulation of the definition of supremum) we have an $x > s - \varepsilon$ such that $x_n > x$ for infinitely many n, while by the definition of the set M there are only finitely many n such that $x_n \geq s + \varepsilon$.

Choose k_1 such that

$$s - 1 < x_{k_1} < s + 1$$

. Let us already have $k_1 < k_2 < \cdots < k_n$ such that for $j = 1, \dots, n$

$$s - \frac{1}{j} < x_{k_j} < s + \frac{1}{j}.$$

Since $K(n+1)$ is infinite there is a $k_{n+1} > k_n$ such that

$$s - \frac{1}{n+1} < x_{k_{n+1}} < s + \frac{1}{n+1}.$$

Thus chosen subsequence $(x_{k_n})_n$ of $(x_n)_n$ obviously converges to s. □

3. Cauchy sequences

3.1. A sequence $(x_n)_n$ is said to be *Cauchy* if

$$\forall \varepsilon > 0 \; \exists n_0 \text{ such that } \forall m, n \geq n_0, \; |x_m - x_n| < \varepsilon.$$

3.1.1. Observation. *Every convergent sequence is Cauchy.*
(Indeed, if $|x_n - L| < \varepsilon$ for $n \geq n_0$ then for $m, n \geq n_0$,

$$|x_n - x_m| = |x_n - L + L - x_m| \leq |x_n - L| + |L - x_m| < 2\varepsilon.)$$

3.2. Lemma. *If a Cauchy sequence has a convergent subsequence then it converges itself.*

Proof. Suppose $(x_n)_n$ is Cauchy and $\lim x_{k_n} = x$. Let $\varepsilon > 0$.

Choose n_1 such that for $m, n \geq n_1,\ |x_m - x_n| < \varepsilon$ and n_2 such that for $n \geq n_2,\ |x_{k_n} - x| < \varepsilon$. Set $n_0 = \max(n_1, n_0)$.

Now if $n \geq n_0$

$$|x_n - x| = |x_n - x_{k_n} + x_{k_n} - x| \leq |x_n - x_{k_n}| + |x_{k_n} - x| \leq 2\varepsilon$$

since $k_n \geq n \geq n_1$. □

3.3. Lemma. *Every Cauchy sequence is bounded.*

Proof. Choose an n_0 such that $|x_n - x_{n_0}| < 1$ for all $n \geq n_0$. Then we have

$$a = \min\{x_j \,|\, j = 1, 2, \dots, n_0\} - 1 \leq x_n \leq b = \max\{x_j \,|\, j = 1, 2, \dots, n_0\} + 1$$

for all n. □

3.4. Theorem. (Bolzano-Cauchy Theorem) *A sequence of real numbers is convergent if and only if it is Cauchy.*

Proof. A Cauchy sequence is by Lemma 3.3 bounded and hence, by Theorem 2.7 has a convergent subsequence. Apply Lemma 3.2.

The other implication has been already observed in 3.1.1. □

3.4.1. Remarks. 1. The proof was very short, but this was because we have had already prepared the essence in Theorem 2.7.

2. Bolzano-Cauchy Theorem is extremely important. Realize that it is a criterion of convergence that can be used without any previous knowledge of the value of the limit, or of values from which it could have been computed.

4. Countable sets: the size of sequences as the smallest infinity

This section is about general sequences, not just about sequences of real numbers.

4.1. Comparing cardinalities. Two sets X, Y are equally large (we say that they have the same cardinality and write

$$\mathrm{card} X = \mathrm{card} Y\ )$$

if there is an invertible (that is, one-to-one onto) mapping $f : X \to Y$. One writes

$$\mathrm{card}X \le \mathrm{card}Y$$

if there is a one-to-one mapping $f : X \to Y$. This means that Y is at least as large as X.

Note. The question naturally arises whether $\mathrm{card}X \le \mathrm{card}Y$ and $\mathrm{card}Y \le \mathrm{card}X$ implies $\mathrm{card}X = \mathrm{card}Y$. This is obvious for finite sets and not quite so obvious for infinite ones, but it is true, by Cantor-Bernstein Theorem.

4.2. Proposition. *The size of the set of natural numbers is the smallest infinite one. Formally, if X is infinite then $\mathrm{card}\mathbb{N} \le \mathrm{card}X$.*

Proof. We can construct a one-to-one mapping $f : \mathbb{N} \to X$ inductively as follows. Choose $f(0) \in X$ arbitrarily. Suppose $f(0), \dots, f(n)$ have been chosen. Since X is infinite, $X \smallsetminus \{f(0), \dots, f(n)\}$ is non-empty and we can choose $f(n+1) \in X \smallsetminus \{f(0), \dots, f(n)\}$. □

4.3. Countable sets. A set is said to be *countable* if $\mathrm{card}X = \mathrm{card}\mathbb{N}$. In other words, a set is countable if there is a one-one onto map $f : \mathbb{N} \to X$, hence iff the set can be ordered into a one-to-one sequence

$$X : \ x_0, x_1, \dots, x_n \dots$$

(set $x_n = f(n)$).

If we want to say that X is finite or countable we say that it is *at most countable.*

Note that

4.3.1. *checking that a set is countable it suffices to know it is infinite and order it into any sequence: the possible repetitions can be deleted and we still have an (infinite) sequence.*

4.4. Proposition. *Let X_n, $n \in \mathbb{N}$, be at most countable. Then*

$$X = \bigcup_{n=0}^{\infty} X_n$$

is at most countable.

Proof. Let us order the sets X_n into sequences

$$X_n : \ x_{n0}, x_{n1}, \dots, x_{nk}, \dots \ .$$

Now we can order X into the sequence

$$x_{00},\quad x_{01}, x_{10},\quad x_{02}, x_{11}, x_{20},\quad x_{03}, x_{12}, x_{21}, x_{30}, \dots,$$
$$\dots x_{0,k}, x_{1,k-1}, x_{2,k-2}, \dots, x_{k-2,2}, x_{k-1,1}, x_{k,0}, \dots .$$

□

4.5. Corollary. *Let X be countable. Then $X \times X$ is countable.* (Indeed, $X \times X = \bigcup_{x \in X} X \times \{x\}$.)

4.6. Corollary. *The set $\mathbb{Q}$ of all rational numbers is countable.*

4.7. Corollary. *Let X be countable. Then any finite cartesian power X^n is countable, and hence also*

$$\bigcup_{n=0}^{\infty} X^n$$

is countable.

Consequently, the set of all finite subsets of X is countable.

4.8. Fact. *The set $\mathbb{R}$ of all real numbers is* not *countable.*

Proof. Represent a real number between zero and one in a decadic expansion

$$r:\ 0.r_1 r_2 \cdots r_n \cdots\ .$$

Now order all such numbers ina sequence (vertically)

$$\begin{aligned} r_1 &:\ 0.r_{11}r_{12}r_{13}\cdots r_{1n}\cdots \\ r_2 &:\ 0.r_{21}r_{22}r_{23}\cdots r_{2n}\cdots \\ r_3 &:\ 0.r_{31}r_{32}r_{33}\cdots r_{3n}\cdots \\ &\dots \\ r_k &:\ 0.r_{k1}r_{k2}r_{k3}\cdots r_{kn}\cdots \\ &\dots \end{aligned}$$

Now set

$$x_n = \begin{cases} 1 \text{ if } r_{nn} \neq 1, \\ 2 \text{ if } r_{nn} = 1. \end{cases}$$

The real number $r = 0.x_1x_2\cdots x_n\cdots$ has not appeared in the sequence above – a contradiction. $\square$

4.9. Cantor Diagonalization Theorem. The procedure in 4.8 is a special case of the famous Cantor diagonalization.

Theorem. (Cantor) *The cardinality of the set $\mathfrak{P}(X)$ of all subsets of a set X is strictly bigger than that of X itself.*

Proof. Suppose $\mathrm{card}X = \mathrm{card}\mathfrak{P}(X)$. Then we have a one-to-one onto mapping $f : X \to \mathfrak{P}(X)$. Set

$$A = \{x \,|\, x \in X,\ x \notin\ f(x)\}$$

and consider the $a \in X$ such that $A = f(a)$. We cannot have $a \notin A = f(a)$ because then $a \in A$ by the definition of A. But we cannot have $a \in A$ either, because then, for the same reason $a \notin A$. $\square$

.

III. Series.

1. Summing a series as a limit of partial sums

1.1. Let $(a_n)_n$ be a sequence of real numbers. An associated *series*

$$\sum_{n=0}^{\infty} a_n \quad \text{or} \quad a_0 + a_1 + a_2 + \cdots$$

is the limit $\lim_n \sum_{k=0}^{n} a_k$, provided it exists.

More precisely, if the limit exists we speak of a *convergent series*; otherwise we speak of a *divergent series.*

1.2. A series that is easy to sum: the geometric one. Let q be a real number, $0 \le q < 1$. Consider the finite sums

$$s(n) = 1 + q + q^2 + \cdots + q^n.$$

We have

$$q \cdot s(n) = q + q^2 + \cdots + q^{n+1} = s(n) - 1 + q^{n+1}$$

so that

$$s(n) = \frac{1 - q^{n+1}}{1 - q}$$

and since $\lim_n q^n = 0$ (else we had $a = \inf_n q^n > 0$ and then $\frac{a}{q} > a$ and hence for some k, $q^k < \frac{a}{q}$ and $q^{k+1} < a$ – a contradiction) we have

$$\sum_{n=0}^{\infty} q^n = \lim_n s(n) = \frac{1}{1-q}. \qquad \square$$

1.3. Proposition. *Let a series* $\sum_{n=0}^{\infty} a_n$ *converge. Then* $\lim_n a_n = 0$.

Proof. Suppose not. Then there is a $b > 0$ such that for every n there is a $p_n > n$ such that $|a_{p_n}| \ge b$. Hence

$$\left| \sum_{k=0}^{p_n} a_k - \sum_{k=0}^{p_n - 1} a_k \right| = |a_{p_n}| \ge b$$

and the sequence $(\sum_{k=0}^{n} a_k)_n$ is not even Cauchy. □

1.4. A divergent case: the harmonic series. The necessary condition from 1.3 is not sufficient. Here is a example (which has also other uses), the *harmonic series*

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots .$$

Consider the finite sums

$$S_n = \sum_{k=10^n+1}^{10^{n+1}} \frac{1}{k}$$

(hence,

$$S_0 = \frac{1}{2} + \cdots + \frac{1}{10},\ S_1 = \frac{1}{11} + \cdots + \frac{1}{100},\ S_2 = \frac{1}{101} + \cdots + \frac{1}{1000},\ \text{ etc.}).$$

S_n has $9 \cdot 10^n$ summands all of them $\geq \frac{1}{10^{n+1}}$ so that $S_n \geq \frac{9}{10}$ and hence

$$\sum_{k=0}^{10^{n+1}} \frac{1}{k} = 1 + S_0 + \cdots S_n \geq 1 + n\frac{9}{10}.$$

1.4.1. For the same reasons we have divergent series

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \cdots \quad \text{and} \quad 1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots$$

2. Absolutely convergent series

2.1. A series $\sum_{n=1}^{\infty} a_n$ is *absolutely convergent* if

$$\sum_{n=1}^{\infty} |a_n|$$

converges.

2.2. Proposition. *An absolutely convergent series converges.*

More generally, if $|a_n| \leq b_n$ *for all* n *and if* $\sum_{n=1}^{\infty} b_n$ *converges then* $\sum_{n=1}^{\infty} a_n$ *converges.*

Proof. Set

$$s_n = \sum_{k=1}^{n} a_k \quad \text{and} \quad \overline{s}_n = \sum_{k=1}^{n} b_k.$$

Recall II.3. The sequence $(\overline{s}_n)_n$ converges and hence it is Cauchy. Now for $m < n$

$$|s_n - s_m| = |\sum_{k=m+1}^{n} a_k| \leq \sum_{k=m+1}^{n} |a_k| \leq \sum_{k=m+1}^{n} b_k = |\overline{s}_n - \overline{s}_m|;$$

thus the sequence $(s_n)_n$ is Cauchy, and hence convergent. □

Remark. This is an example of a very important consequence of Bolzano-Cauchy Theorem. Note that we have here an existence of a sum about the value of which we have no information.

2.3. Theorem. *The series $\sum_{n=0}^{\infty} a_n$ converges absolutely if and only if for every $\varepsilon > 0$ there is an n_0 such that for every finite $K \subseteq \{n \,|\, n \geq n_0\}$ we have $\sum_{k \in K} |a_k| < \varepsilon$.*

Proof. For the sequence $(x_n)_n$ with $x_n = \sum_{k=0}^{n} |a_k|$ and $n_0 \leq n \leq m$ we have $|x_n - x_m| = \sum_{m \leq k \leq n} |a_k|$. Hence the condition on the finite sets K (recalling again that all the summands are non-negative), is just another way of stating that $(x_n)_n$ is Cauchy. □

2.3.1. Note. By Theorem 2.3 we see that the sum of an absolutely convergent series can be viewed as arbitrarily well approximated by sums over finite subsets of $\mathbb{N}$: for any ε we have a finite subset of $\mathbb{N}$ so that no finite subset of the $|a_k|$ in the residual part adds to more than ε. In the following theorem we will see another aspect of this fact: an absolutely convergent series can be arbitrarily reshuffled and the sum does not change.

For the non-absolutely conergent series this is not at all the case. There, the sum is just the limit of the sums of the segments over the sets $\{1, 2. \dots, n\}$, and heavily depends on the order $a_1, a_2, a_3, \dots$ as we will see in the next section.

2.4. Theorem. *Let $s = \sum_{n=1}^{\infty} a_n$ converge absolutely. Then the value of the sum does not depend on the order of the a_n in the sequence. More precisely, for any $p : \mathbb{N} \to \mathbb{N}$ that is one-to-one and onto, $\sum_{n=1}^{\infty} a_{p(n)}$ converges to the same sum s.*

Proof. For $\varepsilon > 0$ choose, first, by 2.3 an n_1 such that for every finite $K \subseteq \{n \,|\, n \geq n_1\}$ we have $\sum_{k\in K} |a_k| < \varepsilon$. Further, choose an $n_2 \geq n_1$ such that $|\sum_{k=1}^{n_2} a_k - s| < \varepsilon$. Finally choose an $n_0 \geq n_2$ such that for $n \geq n_0$,

$$\{p(1), \dots, p(n)\} \supseteq \{1, 2, \dots, n_2\}.$$

Now let $n \geq n_0$. Set $K = \{p(1), \dots, p(n)\} \smallsetminus \{1, 2, \dots, n_2\}$. We have

$$|\sum_{k=1}^{n} a_{p(k)} - s| = |\sum_{k=1}^{n_2} a_k + \sum_{k\in K} a_k - s| =$$

$$= |\sum_{k=1}^{n_2} a_k - s + \sum_{k\in K} a_k| \leq |\sum_{k=1}^{n_2} a_k - s| + \sum_{k\in K} |a_k| < 2\varepsilon. \quad \square$$

2.5. Two criteria of absolute convergence. The summability of geometric series (see 1.2) and Proposition 2.2 lead to the following easy criteria of convergence.

2.5.1. Proposition. (D'Alembert Criterion of Convergence) *Let there be a $q < 1$ and n_0 such that for all $n \geq n_0$,*

$$\left|\frac{a_{n+1}}{a_n}\right| \leq q.$$

Then $\sum_{n=1}^{\infty} a_n$ absolutely converges. If there is an n_0 such that for $n \geq n_0$

$$\left|\frac{a_{n+1}}{a_n}\right| \geq 1$$

Then $\sum_{n=1}^{\infty} a_n$ diverges.

Proof. If the first holds we have for $n \geq n_0$, $|a_{n+1}| \leq q|a_n|$ so that $|a_{n+k}| \leq |a_{n_0}| \cdot q^k$.

The second statement is trivial. □

2.5.2. Proposition. Cauchy Criterion of Convergence) *Let there be a $q < 1$ and n_0 such that for all $n \geq n_0$,*

$$\sqrt[n]{|a_n|} \leq q.$$

Then $\sum_{n=1}^{\infty} a_n$ absolutely converges. If there is an n_0 such that for $n \geq n_0$

$$\sqrt[n]{|a_n|} \geq 1$$

Then $\sum_{n=1}^{\infty} a_n$ *diverges.*

Proof. This is even more straightforward: if we have $\sqrt[n]{|a_n|} \leq q$ then $|a_n| \leq q^n$. $\square$

2.5.3. These criteria are often presented in a weaker, but transparent form:

If $\lim_n \left|\frac{a_{n+1}}{a_n}\right| < 1$ *resp.* $\lim_n \sqrt[n]{|a_n|} < 1$ *then the series* $\sum_{n=1}^{\infty} a_n$ *converges absolutely, If* $\lim_n \left|\frac{a_{n+1}}{a_n}\right| > 1$ *resp.* $\lim_n \sqrt[n]{|a_n|} > 1$ *then the series* $\sum_{n=1}^{\infty} a_n$ *does not converge at all.*

In this formulation one sees the apparent gap: what happens if the limit is 1? In fact, anything; such a series can then still be absolutely convergent, or convergent but not absolutely so, or not convergent at all (the last we have seen in 1.4, for examples of the other cases see 3.2 below).

3. Non-absolutely convergent series

3.1. The alternating series. We have already seen that $\lim a_n$ generally does not suffice to make a series convergent. There is, however, an important case where it does.

Proposition. *Let* $a_n \geq a_{n+1}$ *for all* n*. Then the series*

$$a_1 - a_2 + a_3 - a_4 + \cdots$$

converges if and only if $\lim_n a_n = 0$.

Proof. Set $s_n = \sum_{k=0}^{n} (-1)^{n+1} a_k$. We have

$$s_{2n+2} = s_{2n} + a_{2n+1} - a_{2n+2} \leq s_{2n} \quad \text{and} \quad s_{2n+3} = s_{2n+1} - a_{2n+2} + a_{2n+3} \geq s_{2n+1}.$$

Thus we have two sequences,

$$\begin{aligned} &s_1 \geq s_3 \geq \cdots \geq s_{2n+1} \geq \cdots , \\ &s_2 \leq s_4 \leq \cdots \leq s_{2n} \leq \cdots , \end{aligned}$$

both of them convergent by II.2.6. Now we have $s_{2n+1} - s_{2n} = a_{2n+1}$ so that these two sequences converge to the same number (and hence to $\lim_n s_n$) if and only if $\lim_n a_n = 0$. $\square$

3.2. Notes. 1. In particular we have the convergent series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots . \tag{$*$}$$

By 1.4 it is not asolutely convergent. Note that here $\lim_n \left|\frac{a_{n+1}}{a_n}\right| = 1$ (cf. 2.5.3)

2. Take the series $(*)$ and transform it to

$$\left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \cdots ,$$

that is to

$$\frac{1}{1\cdot 2} + \frac{1}{3\cdot 4} + \frac{1}{5\cdot 6} + \cdots .$$

This is a series of positive numbers with the same sum as $(*)$. Hence it is absolutely convergent and also here we have $\lim_n \left|\frac{a_{n+1}}{a_n}\right| = \lim_n \left|\frac{(2n+1)(2n+2)}{(2n+3)(2n+4)}\right| = 1$ (cf. 2.5.3 again).

3.3. Finally we will show that a convergent but not absolutely convergent series is just the limit from the definition, and cannot be viewed as a "countable sum".

Thus let $\sum_{n=1}^{\infty} a_n$ be a convergent series that is not absolutely convergent. Divide the sequence $(a_n)_n$ into two sequences

$$\begin{aligned} B:&\quad b_1, b_2, b_3, \dots, \\ C:&\quad c_1, c_2, c_3, \dots, \end{aligned}$$

the first consisting of the non-negative a_n, the second consisting of the negative ones, in the order as they occur in $(a_n)_n$.

3.3.1. Lemma. *Neither of the sequences* $(\sum_{k=1}^{n} b_k)_n$, $(\sum_{k=1}^{n}(-c_k))_n$ *has an upper bound.*

Proof. 1. Suppose both of them have. Then $\sum_{n=1}^{\infty} b_n$ and $\sum_{n=1}^{\infty} c_n$ are absolutely convergent. For $\varepsilon > 0$ choose n_1 such that for every finite $K \subseteq \{n\,|\,n \geq n_1\}$ we have $\sum_{k\in K} |b_k| < \varepsilon$ and $\sum_{k\in K} |c_k| < \varepsilon$. Now if we choose n_0 such that $\{a_1, \dots, a_{n_0}\}$ contains both $\{b_1, \dots, b_{n_1}\}$ and $\{c_1. \dots, c_{n_1}\}$ then for every finite $K \subseteq \{n\,|\,n \geq n_0\}$ we have $\sum_{k\in K} |a_k| < 2\varepsilon$ and we see that $\sum_{n=1}^{\infty} a_n$ is absolutely convergent.

2. Let, say, $(\sum_{k=1}^{n}(-c_k))_n$ be bounded but $(\sum_{k=1}^{n} b_k)_n$ not. Then $\sum_{n=1}^{\infty} c_n$ is absolutely convergent; choose n_1 such that for every finite $K \subseteq \{n \,|\, n \geq n_1\}$ we have $\sum_{k\in K} |c_k| < 1$. If n_0 is such that $\{a_1, \dots, a_{n_0}\}$ contains the segment $\{c_1, \dots, c_{n_1}\}$ then for $n \geq n_0$ we have $\sum_{k=1}^{n} a_k > \sum_{k=1}^{n} b_k - \sum_{k=1}^{n_1} |c_k| - 1$ and hence $(\sum_{k=1}^{n} a_k)_n$ is not bounded and cannot converge. $\square$

3.3.2. Proposition. *Let $\sum_{n=1}^{\infty} a_n$ be a convergent but not absolutely convergent series and let r be an arbitrary real number. Then the series can be reshuffled to $\sum_{n=1}^{\infty} a_{p(n)}$ ($p : \mathbb{N} \to \mathbb{N}$ is a one-to-one onto mapping) equal to r.*

Proof. Let, say, $r \geq 0$. Let n_1 be the first natural number such that $\sum_{k=1}^{n_1} b_k > r$. Then take the least m_1 such that $\sum_{k=1}^{n_1} b_k + \sum_{k=1}^{m_1} b_k < r$. Further let n_2 be first such that

$$\sum_{k=1}^{n_1} b_k + \sum_{k=1}^{n_1} b_k + \sum_{k=n_1+1}^{n_2} b_k > r$$

and m_2 first such that

$$\sum_{k=1}^{n_1} b_k + \sum_{k=1}^{n_1} b_k + \sum_{k=n_1+1}^{n_2} b_k + \sum_{k=m_1+1}^{m_2} c_k < r.$$

Proceeding this way and taking into account that both $(b_n)_n$ and $(c_n)_n$ (subsequences of $(a_n)_n$) converge to zero we see that

$$\begin{aligned} &b_1 + \cdots + b_{n_1} + c_1 + \cdots + c_{m_1} + b_{n_1+1} + \cdots + b_{n_2} + c_{m_1+1} + \cdots + c_{m_2} + \cdots \\ &\quad \cdots + b_{n_k+1} + \cdots + b_{n_{k+1}} + c_{m_k+1} + \cdots + c_{m_{k+1}} + \cdots = r \qquad \square \end{aligned}$$

.

IV. Continuous real functions

1. Intervals

1.1. Notation and terminology. Recall the standard notation. For $a \le b$ set

$$\begin{aligned}
(a,b) &= \{x \,|\, a < x < b\} \\
\langle a,b) &= \{x \,|\, a \le x < b\} \\
(a,b\rangle &= \{x \,|\, a < x \le b\} \\
\langle a,b\rangle &= \{x \,|\, a \le x \le b\} \\
(a,+\infty) &= \{x \,|\, a < x\} \\
\langle a,+\infty) &= \{x \,|\, a \le x\} \\
(-\infty,b) &= \{x \,|\, x < b\} \\
(-\infty,b\rangle &= \{x \,|\, x \le b\}
\end{aligned}$$

These subsets of $\mathbb{R}$, and further $\emptyset$ and $\mathbb{R}$ itself, are referred to as (real) *intervals*. The first four and $\emptyset$ are said to be *bounded*.

Further, in the cases 1, 5, 7, $\emptyset$ and $\mathbb{R}$ one speaks of *open intervals*, and in the cases 4, 5, 8, $\emptyset$ and $\mathbb{R}$ one speaks of *closed intervals*. Note that $\emptyset$ and $\mathbb{R}$ are both open and closed, and they are the only such.

1.1.1. Caution. The symbol "(a,b)" has been alredy used for an ordered pair. We will keep this notation; the reader will certainly recognize from the context whether we speak of an ordered pair or of a bounded open interval.

1.2. General characteristics of intervals. A subset of $\mathbb{R}$ is said to be an *interval* if

$$\forall a,b \in J \ \ (a \le x \le b \ \Rightarrow \ x \in J). \qquad \text{(int)}$$

1.2.1. Proposition. *A subset $J \subseteq \mathbb{R}$ is an interval in the sense of* (int) *iff it is one of the subsets mentioned in 1.1, including $\emptyset$ and $\mathbb{R}$.*

Proof. Each of the subsets from 1.1 obviously satisfies (int).

Now let J satisfy (int) and let it be non-empty.

(a) Let J have both a lower and an upper bound. Then there are $a = \inf J$ and $b = \sup J$.

(a1) If $a, b \in J$ then obviously $J = \langle a,b\rangle$.

(a2) If $a \in J$ and if $b \notin J$ and $a \leq x < b$ then by the definition of infimum there is a $y \in J$ such that $x < y$ and hence by (int) $x \in J$ so that $J = \langle a, b)$.

(a3) Similarly if $a \notin J$ and $b \in J$ we infer that $J = (a, b\rangle$.

(a4) If neither a, b is in J and $a < x < b$ choose by the definitions of supremum and infimum $y, z \in J$ such that $a < y < x < z < b$ to infer that $J = (a, b)$.

(b) If J has a lower bound and no upper bound set $a = \inf J$.

(b1) If $a \in J$ then proceed like in (a2), with $y \in J$ such that $a \leq x < y$ obtained from the lack of upper bound to prove that $J = \langle a, +\infty)$.

(b2) If $a \notin J$ proceed like in (a4) with y from the definition of infimum and z from the lack of upper bound to obtain $J = (a, +\infty)$.

(c) If J has an upper bound and no lower bound set $b = \sup J$. Analogously like in (b) we learn that J is either $(+\infty, b\rangle$ or $(+\infty, b)$.

(d) Finally if J has no upper or lower bound, we easily see (similarly like in (a4)) that $J = \mathbb{R}$. □

1.3. Compact intervals. The bounded closed intervals $\langle a, b\rangle$ have particularly nice properties. They will be referred to as *compact intervals* (they are special cases of very important *compact spaces* we will meet later). In particular we will often use Theorem II.2.7 in the following reformulation.

1.3.1. Theorem. *Each sequence in a compact interval J contains a subsequence converging in J.*

2. Continuous real functions of one real variable

2.1. We will be interested in functions $f : D \to \mathbb{R}$ with the domain D typically an interval or a transparent union of intervals. Unless otherwise stated, we will speak of these *real functions of one real variable* briefly as of *functions*.

2.2. Continuity. A function $f : D \to \mathbb{R}$ is said to be *continuous at a point $x \in D$* if

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ \text{ such that } \ (y \in D \text{ and } |y - x| < \delta) \ \Rightarrow \ |f(y) - f(x)| < \varepsilon.$$

A function $f : D \to \mathbb{R}$ is *continuous* if it is continuous in all the $x \in D$, that is if

$$\forall x \in D \ \forall \varepsilon > 0 \ \exists \delta > 0 \quad ((y \in D \text{ and } |y - x| < \delta) \ \Rightarrow \ |f(y) - f(x)| < \varepsilon).$$

2.2.1. Constants and identity. For instance, the constant function $f : D \to \mathbb{R}$ defined by $f(x) = c$ for all $x \in D$, or the $f : D \to \mathbb{R}$ defined by $f(x) = x$ are continuous.

2.3. Arithmetic operations with functions. For $f, g : D \to \mathbb{R}$ and $\alpha \in \mathbb{R}$ define

$$f+g,\ \alpha f,\ fg \text{ and, if } g(x) \neq 0 \text{ for } x \in D,\ \frac{f}{g}$$

by setting

$$(f+g)(x) = f(x) + g(x), \quad (\alpha f)(x) = \alpha f(x),$$
$$(fg)(x) = f(x)g(x) \ \text{ and } \ \left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

2.3.1. Proposition. *Let $f, g : D \to \mathbb{R}$ be continuous in x and let α be a real number. Then $f + g$, αf, fg and, if $g(x) \neq 0$ for $x \in D$, also $\frac{f}{g}$, are continuous in x.*

Proof. The proof is quite analogous to that of II.2.3 - the only difference is in chosing δ's instead of n_0's. Just to illustrate it, let us prove it, this time with an extreme pedantery, for the product fg. Note that the pedantery, heading for a tidy ε instead of simply using the idea of "arbitrarily small" in fact obscures the matter. As an exercise do it again without the adjustments.

Let $\varepsilon > 0$. Choose

$$\begin{aligned}
&\delta_1 > 0 \text{ such that } |y - x| < \delta_1 \ \Rightarrow\ |f(y)| \le |f(x)|,\\
&\delta_2 > 0 \text{ such that } |y - x| < \delta_2 \ \Rightarrow\ |f(y) - f(x)| < \frac{\varepsilon}{2(|g(x)|+1)},\\
&\delta_3 > 0 \text{ such that } |y - x| < \delta_3 \ \Rightarrow\ |g(y) - g(x)| < \frac{\varepsilon}{2(|f(x)|+1)}
\end{aligned}$$

and set $\delta = \min(\delta_1, \delta_2, \delta_3)$. If $|y - x| < \delta$ we have

$$\begin{aligned}
|f(x)g(x) - f(y)g(y)| &= |f(x)g(x) - f(y)g(x) + f(y)g(x) - f(y)g(y)| =\\
&= |(f(x) - f(y))g(x) + f(y)(g(x) - g(y))| \le\\
&\le |g(x)||f(x) - f(y)| + |f(y)||g(x) - g(y)| <\\
&< (|g(x)|+1)\frac{\varepsilon}{2(|g(x)|+1)} + (|f(x)|+1)\frac{\varepsilon}{2(|f(x)|+1)} = \varepsilon. \quad \square
\end{aligned}$$

2.3.2. The following can be left to the reader as an easy exercise.

Proposition. *For $f, g : D \to \mathbb{R}$ define* $\max(f,g)$*,* $\min(f,g)$ *and* $|f|$ *by setting*

$$\max(f,g)(x) = \max(f(x), g(x)),\ \min(f,g) = \min(f(x), g(x))$$
$$\text{and}\ \ |f|(x) = |f(x)|.$$

Let f and g be continuous in x. Then $\max(f,g)$*,* $\min(f,g)$ *and* $|f|$ *are continuous in x.*

2.4. Compositions of real functions. Let $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be real functions and let $f[D] = \{f(x) \,|\, x \in D\} \subseteq E$. Then we define the composition of f and g, denoted

$$g \circ f,$$

by setting $(g \circ f)(x) = g(f(x))$.

2.4.1. Proposition. *Let $f : D \to \mathbb{R}$ be continuous in x and let $g : E \to \mathbb{R}$ be continuous in $f(x)$ Then $g \circ f$ is continuous in x.*

Proof. Let $\varepsilon > O$. Choose $\eta > 0$ such that $|z - f(x)| < \eta$ implies $|g(z) - g(f(x))| < \varepsilon$ and $\delta > 0$ such that $|y - x| < \delta$ implies $|f(y) - f(x)| < \eta$. Then $|y - x| < \delta$ implies $|g(f(y)) - g(f(x))| < \varepsilon$, □

3. Intermediate Value and Darboux Theorems

3.1. Theorem. (Intermediate Value Theorem) *Let $f : J \to \mathbb{R}$ be a continuous function defined on an interval J. Let $a, b \in J$, $a < b$, and let $f(a)f(b) < 0$. Then there exists a $c \in (a,b)$ such that $f(c) = 0$.*

Proof. Let, say, $f(a) < 0 < f(b)$ (else take $-f$ and use the fact that it is continuous iff f is).

Set

$$M = \{x \,|\, a \le x \le b,\ f(x) \le 0\}.$$

Since $a \in M$, $M \neq \emptyset$, and M has the upper bound b by definition. Hence there exists

$$c = \sup M$$

and we have $a \le c \le b$ and hence $c \in J$ and $f(c)$ is defined.

Suppose $f(c) < 0$. Set $\varepsilon = -f(c)$ and consider a $\delta > 0$ such that for x with $|c-x| \leq \delta$ one has $f(c)-\varepsilon < f(x) < f(c)+\varepsilon$. In particular one has for $c \leq x < c+\delta$ still $f(x) < f(c) + (-f(c)) = 0$ and c is not an upper bound od M.

Suppose $f(c) > 0$. Set $\varepsilon = f(c)$ and consider a $\delta > 0$ such that for x such that $|c-x| \leq \delta$ one has $f(c)-\varepsilon < f(x) < f(c)+\varepsilon$. Now one has in particular for $c-\delta < x$ already $0 = f(c)-f(c) < f(x)$ (for $x > c$, $0 < f(x)$ by definition of M) and there are upper bounds smaller than c, a contradiction again.

Thus, $f(c)$ is neither smaller nor greater than 0 and we are left with $f(c) = 0$. □

3.2. Theorem. (Darboux) *Let $f : D \to \mathbb{R}$ be a continuous function and let J be an interval, $J \subseteq D$. Then its image $f[J]$ is an interval.*

Proof. Let $a < b$ be in J and let $f(a) < y < f(b)$ or $f(a) > y > f(b)$. Define $g : D \to \mathbb{R}$ by setting $g(x) = f(x) - y$. By 2.2.1 and 2.3.1 g is continuous. We have $g(a)g(b) < 0$ and hence by 3.1 there is an x with $a < x < b$ (and hence $x \in J$) such that $g(x) = f(x) - y = 0$ and hence $f(x) = y$. □

3.3. Convention. A function $f : D \to \mathbb{R}$ is said to be *increasing, non-decreasing, non-increasing, decreasing*, respectively, if

$$x < y \quad \Rightarrow \quad f(x) < f(y),\ f(x) \leq f(y),\ f(x) \geq f(y),\ f(x) > f(y), \quad \text{resp..}$$

Unlike in general theory of partially ordered sets (where one distinguishes monotone and antitone maps), in analysis one uses the expression *monotone mapping* as a general term for all these cases.

If $x < y$ implies $f(x) < f(y)$ resp. $f(x) > f(y)$ we speek of a *strictly monotone mapping.*

3.4. Proposition. *Let J be an interval and let $f : J \to \mathbb{R}$ be a continuous one-to-one mapping. Then f is strictly monotone.*

Proof. If not there are $a < b < c$ such that $f(a) < f(b) > f(c)$ or $f(a) > f(b) < f(c)$. We will consider the first case, the other is quite analogous. Choose a y such that $\max(f(a), f(c)) < y < f(b)$. Using Theorem 3.2 for the interval $\langle a, b\rangle$ we obtain an x_1, $a < x_1 < b$, with $f(x_1) = y$, and using it for the interval $\langle b, c\rangle$ we obtain an x_2, $b < x_2 < c$ also with $f(x_1) = y$. Thus, f is not one-to-one. □

4. Continuity of monotone and inverse functions

4.1. Theorem. *Let J be an interval and let $f : J \to \mathbb{R}$ be monotone. Then it is continuous if and only if $f[J]$ is an interval.*

Proof. I. If $f[J]$ is not an interval then f is not continuous, by 3.2.

II. Now let $f[J]$ be an interval. Let $x \in J$; suppose it is not an extreme point of the interval so that there are $x_1 < x < x_2$ still in J. Let $\varepsilon > 0$.

If $f(x_1) = f(x) = f(x_2)$ it suffices to choose $0 < \delta \leq x - x_1, x_2 - x$ to have $|f(x) - f(y)| = 0$ for $x - \delta < y < x + \delta$.

If $f(x_1) < f(x) = f(x_2)$ choose a u such that $\min(f(x_1), f(x) - \varepsilon) < u < f(x)$ and, by 3.2, x_1' such that $f(x_1') = u$. If we choose $0 < \delta \leq x - x_1', x_2 - x$ we have, by monotonicity, $f(x) - \varepsilon < f(y) \leq f(x)$ for $x - \delta < y < x + \delta$.

If $f(x_1) = f(x) < f(x_2)$ choose a v such that $f(x) < v < \min(f(x_2), f(x) + \varepsilon)$ and and, by 3.2, x_2' such that $f(x_2') = v$. If we choose $0 < \delta \leq x - x_1, x_2' - x$ we have, by monotonicity, $f(x) \leq f(y) < f(x) + \varepsilon$ for $x - \delta < y < x + \delta$.

If $f(x_1) < f(x) < f(x_2)$ choose u, v such that $\max(f(x_2), f(x) - \varepsilon) < u < f(x) < v < \min(f(x_2), f(x) + \varepsilon)$ and, by 3.2, x_1', x_2' such that $f(x_1') = u$ and $f(x_2') = v$. If we choose $0 < \delta \leq x - x_1', x_2' - x$ we have, by monotonicity, $f(x) - \varepsilon < f(y) < f(x) + \varepsilon$ for $x - \delta < y < x + \delta$.

The cases of the extremal points of the interval are quite analogous, only easier because we have to take care only of one the sides of x. □

Note. The cases of $f(x_1) = f(x) = f(x_2)$, $f(x_1) < f(x) = f(x_2)$ and $f(x_1) = f(x) < f(x_2)$ had to be discussed because the mapping f is supposed to be just monotonous, not strictly monotonous. The reader, of course, sees that the gist is in the case $f(x_1) < f(x) < f(x_2)$; in the first reading the previous three cases may be skipped, and the proof will become (even) more transparent.

4.2. The inverse of a real function $f : D \to \mathbb{R}$. The inverse of $f : D \to E$ is a real function $g : E \to \mathbb{R}$ such that $g \circ f$ and $f \circ g$ exist and $f(g(x)) = x$ and $g(f(x)) = x$ for all $x \in E$ resp. all $x \in D$.

4.2.1. Observation. *If $g : E \to \mathbb{R}$ is inverse to $f : D \to \mathbb{R}$ then $f : D \to \mathbb{R}$ is inverse to $g : E \to \mathbb{R}$, we have $f[D] = E$ and $g[E] = D$, and the f, g restricted to D, E are mutually inverse mappings.*

(Indeed, the first statement is obvious. If $y \in E$ set $x = g(y)$ to obtain $f(x) = y$. Thus we have the restrictions $D \to E$ and $E \to D$ one-to-one onto.)

4.3. Proposition. *Let J be an interval, $f : J \to \mathbb{R}$. Then f has an inverse $g : J' \to \mathbb{R}$ if and only if it is strictly monotone, and this g is continuous.*

Proof. f has to be one-to-one and hence, by 3.4, it is strictly monotone. This makes the $J' = f[J]$ an interval, by 2.3, and the inverse $g : J' \to \mathbb{R}$ also strictly monotone. Now $g[J'] = J$, also an interval, and hence by 4.1 g is continuous. $\square$

4.4. Remark. Now we start to have a sizable stock of continuous functions. From 2.2.1 and 2.3.1 we immediately see that the functions given by polynomial formulas

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

and also the functions

$$f(x) = \frac{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n}{b_0 + b_1x + b_2x^2 + \cdots + b_mx^m}$$

(so called *rational functions*) provided the domain does not contain any x with $b_0 + b_1x + b_2x^2 + \cdots + b_mx^m = 0$.

Further, by 4.3 we have continuous functions given by formulas

$$f(x) = \sqrt{x}, \quad f(x) = \sqrt[n]{x}$$

(with the obvious provisos about the domains) and all the functions obtained from the mentioned ones in finitely many steps using compositions, arithmetic operations, and the operations from 2.3.2. We will have more in the next chapter.

5. Continuous functions on compact intervals

5.1. Theorem. *A function $f : D \to \mathbb{R}$ is continuous if and only if for every convergent sequence in D, $\lim_n f(x_n) = f(\lim_n x_n)$.*

Proof. I. Let f be continuous and let $\lim_n x_n = x$. For $\varepsilon > 0$ choose by continuity a $\delta > 0$ such that $|f(y) - f(x)| < \varepsilon$ for $|y - x| < \delta$. Now by the definition of the convergence of sequences there is an n_o such that for $n \geq n_0$, $|x_n - x| < \delta$. Thus, if $n \leq n_0$ we have $|f(x_n) - f(x)| < \varepsilon$ so that $\lim_n f(x_n) = f(\lim_n x_n)$.

II. Let f not be continuous. Then there is an $x \in D$ and an $\varepsilon_0 > 0$ such that for every $\delta > 0$ there is an $x(\delta)$ such that

$$|x - x(\delta)| < \delta \quad \text{but} \quad |f(x) - f(x(\delta))| \geq \varepsilon_0.$$

Set $x_n = x(\frac{1}{n})$. Then $\lim_n x_n = x$ but $(f(x_n))_n$ cannot converge to $f(x)$. □

5.2. Theorem. *A continuous function $f : \langle a, b\rangle \to \mathbb{R}$ on a compact interval attains a maximum and a minimum. That is, there are $x_0, x_1 \in \langle a, b\rangle$ such that for all $x \in \langle a, b\rangle$,*

$$f(x_0) \leq f(x) \leq f(x_1).$$

Proof. The proof will be done for the maximum. Set

$$M = \{f(x) \,|\, x \in \langle a, b\rangle\}$$

I. Suppose M is not bounded from above. Then for each n we can choose an $x_n \in \langle a, b\rangle$ such that $f(x_n) > n$. By 1.3.1 there is a subsequence x_{k_n} with $\lim_n x_{k_n} = x \in \langle a, b\rangle$. By 5.1, $\lim_n f(x_{k_n}) = f(x)$ in contradiction with $f(x_{k_n})$ being arbitrarily large.

II. Thus, M, obviously non-empty, has to be bounded from above and hence there is an $s = \sup M$. Thus, by the definition of supremum we have $x_n \in \langle a, b\rangle$ such that

$$s - \frac{1}{n} < f(x_n) \leq s. \tag{$*$}$$

Choose a subsequence x_{k_n} with $\lim_n x_{k_n} = x \in \langle a, b\rangle$. By 5.1, $\lim_n f(x_{k_n}) = f(x)$ and by $(*)$ this limit is s. Thus, $f(x) = \sup M = \max M$. □

5.4. Corollary. *Let all the values of a continuous function on a compact interval J be positive. Then there is a $c \geq 0$ such that all the values of f are $\geq c$.*

(Take $c = \min_M f(x)$.)

5.5. Corollary. *Let $f : J \to \mathbb{R}$ be continuous and let J be a compact interval. Then $f[J]$ is a compact interval.*

More generally, if $f : D \to \mathbb{R}$ is continuous and if $J \subseteq D$ is a compact interval then $f[J]$ is a compact interval.

5.5.1. Remark. Compact intervals and $\emptyset$ are the only intervals for which the type is preserved in arbitrary continuous images. For the other ones, $f[J]$ is an interval again, but not necessarily of the same type.

6. Limit of a function at a point

6.1. In the following, to avoid too many letters, we will omit specifying the domain in some of the formulas (as for instance, if we have already specified that our function is $f : D \to \mathbb{R}$ and speak of continuity we write just "$\forall\varepsilon > 0\ \exists\delta > 0$ s. t. $|y - x| < \delta \ \Rightarrow\ |f(y) - f(x)| < \varepsilon$".

We say that a function $f : D \to \mathbb{R}$ *has a limit b at a point a*, and write

$$\lim_{x\to a} f(x) = b$$

if

$$\forall\varepsilon > 0\ \exists\delta > 0 \text{ such that } (0 < |x - a| < \delta) \ \Rightarrow\ |f(x) - b| < \varepsilon.$$

Remark. Note the striking similarity with the definition of continuity, but also the fundamental difference:

> *In this definition there is no reference to a possible value of the function f in the point a. Indeed a does not have to be in the domain D, and even if it is, the value $f(a)$ does not play any role and has nothing to do with the value b.*

6.2. One-sided limits. We say that a function $f : D \to \mathbb{R}$ *has a limit b at a point a from the right*, and write

$$\lim_{x\to a+} f(x) = b$$

if

$$\forall\varepsilon > 0\ \exists\delta > 0 \text{ such that } (0 < x - a < \delta) \ \Rightarrow\ |f(x) - b| < \varepsilon.$$

It *has a limit b at a point a from the left*, written

$$\lim_{x\to a-} f(x) = b,$$

if

$$\forall\varepsilon > 0\ \exists\delta > 0 \text{ such that } (0 < a - x < \delta) \ \Rightarrow\ |f(x) - b| < \varepsilon.$$

6.2.1. Remark. The reader has certainly noted that formally we could obtain the one-sided limits by changing the domain: defining the f just for the $x > a$ for the limit from the right, and similarly for the other one. But it would be misleading. Whatever the domain, the intuitive sense of the concepts is the behaviour of the values when approaching the point a (without attaining it), in the one-sided limits approaching it from above or from below.

6.3. Observation.*A function $f : D \to \mathbb{R}$ is continuous at a point a if and only if* $\lim_{x\to a} f(x) = f(a)$.

(Just compare the definitions.)

6.3.1. One-sided continuity. A function $f : D \to \mathbb{R}$ is said to be *continuous at a point a from the right* (resp *from the left* if $\lim_{x\to a+} f(x) = f(a)$ (resp. $\lim_{x\to a-} f(x) = f(a)$).

6.4. Proposition. *Let* $\lim_{x\to a}(f)(x) = A$ *and* $\lim_{x\to a} g(x) = B$ *exist and let α be a real number, Then* $\lim_{x\to a}(f+g)(x)$, $\lim_{x\to a}(\alpha f)(x)$, $\lim_{x\to a}(fg)(x)$ *exist. and if $B \neq 0$ also* $\lim_{x\to a} \frac{f}{g}(x)$ *exists, and they are equal, in this order, to $A + B$, αA, AB and $\frac{A}{B}$.*

Proof. Use 6.3 and 2.3.1. Note that if $B \neq 0$ there is a $\delta_0 > 0$ such that for $|x - a| < \delta_0$ we have $g(x) \neq 0$. □

6.4.1. Note that obviously the same holds for one-sided limits.

6.5. Now one may expect that in analogy with 2.4.1 we will have that if $\lim_{x\to a} f(x) = b$ and $\lim_{x\to b} = c$ then $\lim_{x\to a}(g(f(x)) = c$. This is almost true, but not quite so; we have to be careful.

Consider the following example. Define $f, g : \mathbb{R} \to \mathbb{R}$ by setting

$$f(x) = \begin{cases} x \text{ for rational } x, \\ 0 \text{ for irrational } x \end{cases} \quad \text{and} \quad g(x) = \begin{cases} 0 \text{ for } x \neq 0, \\ 1 \text{ for } x = 0. \end{cases}$$

Here we have $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} = 0$ while $\lim_{x\to 0} g(f(x))$ does not exist at all.

We have, however, a very useful

6.5.1. Proposition. *Let* $\lim_{x\to a} f(x) = b$ *and* $\lim_{x\to b} g(x) = c$. *Let either*

(1) $g(b) = c$ *(that is, $g(b)$ is defined and g is continuous in b)*

or

(2) *for sufficiently small* $\delta_0 > 0$, $0 < |x-a| < \delta_0 \;\Rightarrow\; f(x) \neq b$.
Then $\lim_{x\to a} g(f(x))$ *exists and is equal to* c.

Proof. For $\varepsilon > 0$ choose an $\eta > 0$ such that

$$0 < |y-b| < \eta \quad\Rightarrow\quad |g(y)-c| < \varepsilon$$

and for this η choose a $\delta > 0$ (in the second case, $\delta \leq \delta_0$) such that

$$0 < |x-a| < \delta \quad\Rightarrow\quad |f(x)-b| < \eta.$$

Thus if $0 < |x-\delta| < \varepsilon$ we have in case (2) $|g(f(x))-c| < \varepsilon$ because $|f(x)-b| > 0$. In case (1), $|f(x)-b| = 0$ can occur, but no harm done: in such values we have $|g(f(x))-c| = 0$. □

6.6. Proposition. *Let* $\lim_{x\to a} f(x) = b = \lim_{x\to a} g(x)$ *and let* $f(x) \leq h(x) \leq g(x)$ *for* $|x-a|$ *smaller than some* $\delta_0 > 0$. *Then* $\lim_{x\to a} h(x)$ *exists and is equal to* b.

Proof. This is obvious: if $|f(x)-b| < \varepsilon$ and $|g(x)-b| < \varepsilon$ then $b-\varepsilon < f(x) \leq h(x) \leq g(x) < b+\varepsilon$. □

6.7. Discontinuities of the first and of the second kind. If a function defined at a point $a \in D$ is not continuous in this point we speak of a discontinuity of the first kind if the one-sided limits in this point exist, but either they are not equal, or the value $f(a)$ is not equal to the limit.

Otherwise we speak of a discontinuity of the second kind.

.

V. Elementary functions

In IV.4.4 we introduced some basic continuous real functions given by simple formulas (polynomials, rational functions, roots), and everything that one obtains from them by compositions, arithemtic operations, and inverses, applied repeatedly.

In this chapter we will extend this storage of functions by logarithms, exponentials, goniometric and cyclometric functions. The system of functions obtained from those mentioned above and the new ones by compositions, arithemtic operations, inverses, and also by restrictions, applied repeatedly, are called *elementary functions.*

The new functions will be introduced with a different degree of precision.

The logarithm will be defined axiomatically, and the reader will have to believe, for the time being, that a function with the required properties really exists. This will be mended after we will have the technique of Riemann integral.

Goniometric functions will be used in the form in which the student already knows them. We will need some very transparent facts about limits where we will use geometric intuition about the length of segments of a circle (hopefully persuasive enough, but lacking in rigour).

1. Logarithms

.

1.1. The function

$$\lg : (0, +\infty) \to \mathbb{R}$$

has the following properies[1].

(1) lg increases on the whole interval $(0, +\infty)$

(2) $\lg(xy) = \lg(x) + \lg(y)$

(3) $\lim_{x\to 1} \frac{\lg x}{x-1} = 1$.

[1] The existence of such a funcion will be proved in XII.4 below.

1.2. Two equalities. We have

$$\lg 1 = 0, \quad \lg \frac{x}{y} = \lg x - \lg y.$$

($\lg 1 = \lg(1 \cdot 1) = \lg 1 + \lg 1$. Further, $\lg \frac{x}{y} + \lg y = \lg(\frac{x}{y}y) = \lg x$.)

1.3. Three limits. We have

$$\lim_{x\to 0} \frac{\lg(1+x)}{x} = 1, \quad \lim_{x\to 1} \lg x = 0, \quad \lim_{x\to a} \lg \frac{x}{a} = 0.$$

(For the first use IV.6.5.1 and the obvious $\lim_{x\to 0}(x+1) = 1$. For the second, $\lim_{x\to 1} \lg x = \lim_{x\to 1} \frac{\lg x}{x-1} \lim_{x\to 1}(x-1) = 1 \cdot 0 = 0$; for the third use the second, IV.5.1 and the obvious $\lim_{x\to a} \frac{x}{a} = 1$.)

1.4. Proposition. *The function* lg *is continuous and* $\lg[(0,+\infty)] = \mathbb{R}$.

Proof. For an arbitrary $a > 0$ we have $\lim_{x\to a} \lg x = \lim_{x\to a} \lg(a\frac{x}{a}) = \lim_{x\to a}(\lg a + \lg \frac{x}{a}) = \lim_{x\to a} \lg a + \lim_{x\to a} \lg \frac{x}{a} = \lg a + 0 = \lg a$ so that lg is continuous by IV.6.3.

Now we know by IV.3.2 that $J = \lg[(0,+\infty)]$ is an interval. By 1.1(1), $K = \lg 2 > 0$ and we have, by 1.2, $-K = \lg \frac{1}{2}$. Hence we have in J arbitrarily large positive numbers, namely $nK = \lg(2^n)$ and arbitrarily large negative numbers, namely $-nK = \lg \frac{1}{2^n}$ so that by the definition of interval, $x \in J$ for all $x \in \mathbb{R}$. $\square$

1.5. Logarithm with general base. So far only a definition. The *logarithm with base a*, where $a > 0$ and $a \neq 1$, is

$$\log_a x = \frac{\lg x}{\lg a}.$$

2. Exponentials

.

2.1. By 1.4 (and IV.4.3), lg has a continuous inverse

$$\exp : \mathbb{R} \to \mathbb{R} \quad \text{with all values } \exp(x) \text{ positive.}$$

From the rules in 1.1 and 1.2 we immediately obtain that

$$\begin{aligned}&\exp 0 = 1,\\ &\exp(x+y) = \exp x \cdot \exp y, \text{ and}\\ &\exp(x-y) = \frac{\exp x}{\exp y}.\end{aligned}$$

2.1.1. From 1.1.(3) and IV.5.5.1 we obtain an important limit

$$\lim_{x\to 1} \frac{\exp(x)-1}{x} = 1.$$

2.2. The function exp **as exponentiation. Euler's number.** Set

$$e = \exp(1).$$

This number e is called *Euler's number* or *Euler's constant.*

We obviously have, for natural n,

$$\exp n = \exp(\overbrace{1+1+\cdots+1}^{n}) = e^n$$

and by 2.1,

$$\exp(-n) = \frac{1}{\exp(n)} = e^{-n}.$$

Further, recalling the standard rational exponents $a^{\frac{p}{q}}$ defined as $\sqrt[q]{a^p}$ we see that

$$\exp(\frac{p}{q}) = e^{\frac{p}{q}}$$

since $\exp(\frac{p}{q})^q = \exp(p) = e^p$ and it is the only positive real number with this property. Taking into account the continuity of exp we now see that it is natural to view

$$\exp(x) = e^x$$

as the x-th power of e.

2.2.1. The limit from 2.1.1 will be used in the form

$$\lim_{x\to 1} \frac{e^x-1}{x} = 1.$$

2.3. Since $e^{\lg a} = \exp \lg a = a$ we can define, for $a > 0$,

$$a^x = e^{x \lg a}$$

and easily check that this is a natural exponentiaion in the same sense as e^x is (coinciding with classical $a^n = \overbrace{aa \cdots a}^{n \text{ times}}$ etc.).

2.3.1. Now we can give more sense to the $\log_a x$ from 1.5: it is the inverse to the exponentiation a^x similarly like $\lg x$ is the inverse to e^x. Indeed we have $a^{\log_a x} = a^{\frac{\lg x}{\lg a}} = e^{\frac{\lg x}{\lg a} \lg a} = e^{\lg x} = x$ and $\log_a(a^x) = \frac{\lg(a^x)}{\lg a} = \frac{\lg(e^{x \lg a})}{\lg a} = \frac{x \lg a}{\lg a} = x$.

2.3.2. Finally we can use this general exponentiation (albeit only for $x > 0$) to define continuous

$$x \mapsto x^a = e^{a \lg x}.$$

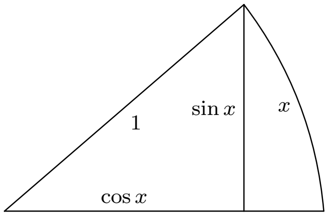
As an easy exercise check that that it coincides with classical x^n and $x^{\frac{p}{q}}$ (restricted to $x > 0$).

3. Goniometric and cyclometric functions

3.1. Recall the functions

$$\sin, \cos : \mathbb{R} \to \mathbb{R}$$

usually defined as the ratio of the opposite resp. adjacent side to the hypotenuse in a rectangular triangle. The argument in these functions is the angle (to which the side in question is opposite or adjacent). To measure the angle (and thus to obtain the argument x) one uses the length of a segment of the unit circle (see the picture below); we assume that we know what the length of such a curve is[2].



[2] Rigorous defintions can be found in XXIII.1 below

Both the functions are defined on the whole of $\mathbb{R}$ as periodic with the period 2π, see below ("the argument length is wound up around the circle").

3.1.1. Let us summarize some basic facts:

$$\begin{aligned}
&\sin^2 x + \cos^2 x = 1,\\
&|\sin x|, |\cos x| \leq 1.\\
&\sin(x + 2\pi) = \sin x, \quad \cos(x + 2\pi) = \cos x,\\
&\sin(x + \pi) = -\sin x, \quad \cos(x + \pi) = -\cos x,\\
&\cos x = \sin(\frac{\pi}{2} - x), \quad \sin x = \cos(\frac{\pi}{2} - x).\\
&\sin(-x) = -\sin x, \quad \cos(-x) = \cos x.
\end{aligned}$$

3.1.2. Further let us recall the very important formulas

$$\begin{aligned}
\sin(x + y) &= \sin x \cos y + \cos x \sin y,\\
\cos(x + y) &= \cos x \cos y - \sin x \sin y.
\end{aligned}$$

3.1.3. From 3.1.2 we easily deduce the following often used equalities.

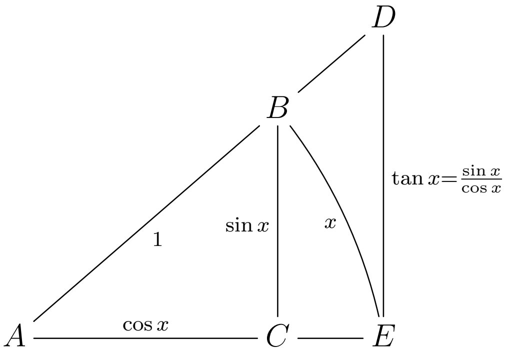
$$\begin{aligned}
\sin x \cos y &= \frac{1}{2}(\sin(x + y) - \sin(x - y)),\\
\sin x \sin y &= \frac{1}{2}(\cos(x - y) - \cos(x + y)),\\
\cos x \cos y &= \frac{1}{2}(\cos(x - y) + \cos(x + y)).
\end{aligned}$$

3.2. Four important limits.

1. $\lim_{x\to 0} \sin x = 0$,
2. $\lim_{x\to 0} \cos x = 1$,
3. $\lim_{x\to 0} \frac{\sin x}{x} = 1$,
4. $\lim_{x\to 0} \frac{\cos x - 1}{x} = 1$.

Explanation. I write, rather, "Explanation" than "Proof". The deduction will be based on the intuitive understanding of the length of a segment x of the unit circle.

Consider the following picture.



1. Since $|\sin(-x)| = |\sin x|$ it suffices to consider positive x. The curved segment x is longer than $\sin x$ (segment BC) (it is even longer than the straight segment BE), hence for small positive x we have $0 < \sin x < x$, and since $\lim_{x\to 0} x = 0$ the statement follows.

2. By 1 we have $\lim_{x\to 0} \cos^2 x = 1 - \lim_{x\to 0} \sin^2 x = 1$ and since $x \mapsto \sqrt{x}$ is continuous in 1 we have $\lim_{x\to 0} \cos x = 1$.

3. Comparing the areas of the triangles ABC, ADE and the intermediate triangle with the curved base x, ABE, we obtain

$$\frac{1}{2}\sin x \cos x \leq \frac{1}{2}x \leq \frac{1}{2}\frac{\sin x}{\cos x}$$

and from this, further,

$$\cos x \leq \frac{\sin x}{x} \leq \frac{1}{\cos x}.$$

Use 2 and IV.6.6.

4. Since $\sin^2 x = 1 - \cos^2 x = (1 + \cos x)(1 - \cos x)$ we have

$$\frac{1 - \cos x}{x} = \frac{1}{1 + \cos x} \cdot \sin x \cdot \frac{\sin x}{x}.$$

Use 2, 1, and 3. □

3.3. Proposition. *The functions* sin *and* cos *are continuous.*

Proof. Since $\cos x = \sin(\frac{\pi}{2} - x)$ it suffices to prove that sin is continuous. We have

$$\sin x = \sin(a + (x - a)) = \sin a \cdot \cos(x - a) + \cos a \cdot \sin(x - a)$$

and hence, by 3.2 and IV.6.5.1,

$$\lim_{x\to a}\sin x = \sin a\cdot 1 + \cos a\cdot 0 = \sin a.$$

Use IV.6.3. □

3.4. Tangens and cotangens. We have $\sin x = 0$ precisely when $x = k\pi$ for an integer k, and $\cos x = 0$ precisely when $x = k\pi + \frac{\pi}{2}$. Hence one can correctly define the function *tangens*,

$$\tan : D \to \mathbb{R} \ \text{ where } D = \bigcup_{-\infty}^{+\infty}((k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi)$$

by setting

$$\tan x = \frac{\sin x}{\cos x}.$$

We have

Fact. *The function* tan *is continuous and increases on each interval* $((k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi)$*, we have* $\tan(x+\pi) = \tan x$*, and* $\tan[((k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi)] = \mathbb{R}$.

Proof. We will start with the period π: the functions sin and cos have a period 2π but we have $\frac{\sin(x+\pi)}{\cos(x+\pi)} = \frac{-\sin x}{-\cos x} = \frac{\sin x}{\cos x}$.

Since sin obviously increases and cos decreases on the interval $\langle 0, \frac{\pi}{2}\rangle$, tan increases on this interval, and since $\tan(-x) = \frac{\sin(-x)}{\cos(-x)} = \frac{-\sin x}{\cos x} = -\tan x$ we infer that tan increases on the whole of $(-\frac{\pi}{2}, \frac{\pi}{2})$. Finally, because of the continuity there is a $\delta > 0$ such that for $\frac{\pi}{2}-\delta < x < \frac{\pi}{2}$ we have $\cos x < \frac{1}{2n}$ and $\sin x > \frac{1}{2}$ so that $\tan x > n$ and $\tan(-x) < -n$ so that $\tan[((k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi)]$ (since it is an interval) has to be $\mathbb{R}$. □

Similarly we have a function *cotangens*

$$\cot : D \to \mathbb{R} \ \text{ where } D = \bigcup_{-\infty}^{+\infty}(k\pi, (k+1)\pi)$$

defined by setting

$$\cot x = \frac{\cos x}{\sin x}$$

with period π, continuous and decreasing on each $(k\pi, (k+1)\pi)$, and mapping this interval onto $\mathbb{R}$.

3.5. Cyclometric functions. The function sin restricted to $\langle -\frac{\pi}{2}, \frac{\pi}{2}\rangle$ is strictly monotone and maps this interval onto $\langle -1, 1\rangle$. Its inverse

$$\arcsin : \langle -1, 1\rangle \to \mathbb{R}$$

is called *arcussinus.* Similarly we have the function *arcuscosinus*

$$\arccos : \langle -1, 1\rangle \to \mathbb{R}$$

inverse to cos restricted to $\langle 0, \pi\rangle$.

Of a particular interest is the inverse to tan restricted to $(-\frac{\pi}{2}, \frac{\pi}{2})$, the *arcustangens*

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

defined on the whole of $\mathbb{R}$.

VI. Derivative

1. Definition and a characteristic

1.1. Convention. When speaking of a derivative of a function $f : D \to \mathbb{R}$ at a point x we assume that the domain D contains an interval $(x-\delta, x+\delta)$ for sufficiently small $\delta > 0$ (we say that x is an *interior point* of the domain D).

When speaking of a derivative of a function $f : D \to \mathbb{R}$ at a point x from the right resp. left we assume that D contains $\langle x, x+\delta)$ resp. $(x-\delta, x\rangle$.

1.2. Derivative. The *derivative of a function* $f : D \to \mathbb{R}$ *at a point* x_0 is the limit

$$A = \lim_{h \to 0} \frac{f(x_0+h) - f(x)}{h}$$

if it exists. If it does we say that f *has a derivative in* x_0.

The derivative (the limit A) is usually denoted by

$$f'(x_0).$$

Other notation used is, e.g.,

$$\frac{\mathrm{d}f(x_0)}{\mathrm{d}x}, \quad \frac{\mathrm{d}f}{\mathrm{d}x}(x_0), \quad \text{or} \quad \left(\frac{\mathrm{d}}{\mathrm{d}x}f\right)(x_0).$$

(The second and the third comes from replacing the symbol f', without specifying the x_0, by $\frac{\mathrm{d}f}{\mathrm{d}x}$ or $\frac{\mathrm{d}}{\mathrm{d}x}f$.)

Note. The process of determining derivative is often called *differentiation.*

1.2.1. From IV.6.5.1 we immediately obtain another expression for the derivative,

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}. \qquad (*)$$

1.3. One-sided derivatives. The *derivative of* f *in* x_0 *from the right* resp. *from the left* is the one-sided limit

$$f'_+(x_0) = \lim_{h \to 0+} \frac{f(x_0+h) - f(x)}{h} \quad \text{resp.} \quad f'_-(x_0) = \lim_{h \to 0-} \frac{f(x_0+h) - f(x)}{h}.$$

Most rules for the one-sided derivatives will be the same as those for the plain derivative, and will not need any particular discussion. The exception will be the composition rule 2.2 – see 2.2.2 below.

1.4. Notes. There are (at least) three motivations resp. interpretations of a derivative.

1. *Geometry.* Think of the f as an equation of a curve

$$C = \{(x, f(x)) \,|\, x \in D\}$$

in the plane. Then $f'(x_0)$ is the slope of the tangent of C in the point $(0, f(x_0))$. More precisely, the tangent is given by the equation

$$y = f(x_0) + f'(x_0)(x - x_0).$$

2. *Physics.* Suppose $f(x)$ is the length of the trajectory achieved by a moving body after elapsing time x. Then

$$\frac{f(y) - f(x)}{y - x}$$

is the average velocity between times x and y, and $f'(x_0)$ is the actual velocity in the moment x_0.

Even more important in physics is the change of velocity, the *acceleration.* This is expressed by the *second derivative*, see Section 4 below.

3. *Approximation.* Linear functions $L(x) = C + Ax$ are easy to compute. The derivative provides an approximation of the given function in small neighbourhoods of a given argument by a linear function with an error considerably smaller than the change of the argument. See 1.5.1.

1.5. Theorem. *A function f has a derivative A at a point x if and only if there exists for a sufficiently small $\delta > 0$ a real function $\mu : (-\delta, +\delta) \smallsetminus \{0\} \to \mathbb{R}$ such that*

(1) $\lim_{h\to 0} \mu(h) = 0$, *and*

(2) *for* $0 < |h| < \delta$,

$$f(x + h) - f(x) = Ah + \mu(h)h.$$

Proof. Suppose $A = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$ exists. Set

$$\mu(h) = \frac{f(x+h)-f(x)}{h} - A.$$

Then μ obviously has the required properties.

On the other hand, let such μ exist. Then we have for small $|h|$,

$$\frac{f(x+h)-f(x)}{h} = A + \mu(h)$$

and $f'(x)$ exists and is equal to A by the rule of the limit of the sum. □

1.5.1. Recall 1.4.3. If we have $f(x+h) - f(x) = Ah + \mu(h)h$ as in (2) above then the linear function $L(y) = f(x) + A(y-x)$ approximates $f(y)$ in a small neighbourhood of x with the error $\mu(|y-x|)|y-x|$, hence $\mu(|y-x|)$-times smaller that $|y-x|$.

1.6. Corollary. *Let f have a derivative at a point x. Then it is continuous in x.*

(Indeed, set $h = y - x$. Then

$$|f(y) - f(x)| \le |A(y-x)| + |\mu(y-x)||(y-x)| < (|A|+1)|y-x|$$

for sufficiently small $|y-x|$.)

2. Basic differentiation rules

2.1. Arithmetic rules. In the following rules, $f, g : D \to \mathbb{R}$ are supposed to have a derivative in the point x and the statement includes the claim that the $f+g$, αf, fg and $\frac{f}{g}$ have the derivative.

Proposition.

(1) $(f+g)'(x) = f'(x) + g'(x)$,

(2) *for any real* α, $(\alpha f)'(x) = \alpha f'(x)$,

(3) $(fg)'(x) = f(x)g'(x) + f'(x)g(x)$, *and*

(4) *if* $g(x) \neq 0$ *then*

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}.$$

Proof. We will transform the formulas so that the rules will immediately follow by applying the limit rules IV.6.4 (and 1.6).

(1) We have

$$\begin{aligned}\frac{(f+g)(x+h)-(f+g)(x)}{h} &= \frac{f(x+h)+g(x+h)-f(x)-g(x)}{h} = \\ &= \frac{f(x+h)-f(x)}{h} + \frac{g(x+h)-g(x)}{h}.\end{aligned}$$

(2)

$$\frac{(\alpha f)(x+h)-(\alpha f)(x)}{h} = \frac{\alpha f(x+h)-\alpha f(x)}{h} = \alpha\frac{f(x+h)-f(x)}{h}.$$

(3)

$$\begin{aligned}&\frac{(fg)(x+h)-(fg)(x)}{h} = \frac{f(x+h)g(x+h)-f(x)g(x)}{h} = \\ &\frac{f(x+h)g(x+h)-f(x+h)g(x)+f(x+h)g(x)-f(x)g(x)}{h} = \\ &= f(x+h)\frac{g(x+h)-g(x)}{h} + g(x)\frac{f(x+h)-f(x)}{h}.\end{aligned}$$

(4) In view of (3) it suffices to derive the rule for $\frac{1}{g}$. We have

$$\begin{aligned}&\frac{(\frac{1}{g})(x+h)-(\frac{1}{g})(x)}{h} = \frac{\frac{1}{g(x+h)}-\frac{1}{g(x)}}{h} = \frac{\frac{g(x)-g(x+h)}{g(x+h)g(x)}}{h} = \\ &= \frac{g(x)-g(x+h)}{g(x+h)g(x)h} = \frac{1}{g(x+h)g(x)}\left(-\frac{g(x+h)-g(x)}{h}\right).\end{aligned}$$

□

2.2. The rule for composition. Let $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be such that $f[D] \subseteq E$. so that the composition $g \circ f$ is defined.

2.2.1. Theorem. *Let f have a derivative at a point x and let g have a derivative in $y = f(x)$. Then $g \circ f$ has a derivative in x and we have*

$$(g \circ f)'(x) = g'(f(x))f'(x).$$

Proof. By 1.5 we have μ and ν with $\lim_{h\to 0}\mu(h) = 0$ and $\lim_{k\to 0}\nu(k) = 0$ such that

$$\begin{aligned}f(x+h) - f(x) &= Ah + \mu(h)h \ \text{ and}\\ g(y+k) - g(y) &= Bk + \nu(k)k.\end{aligned}$$

To be able to use IV.6.5.1 we will define $\nu(0) = 0$ which does not change the limit of ν in 0.

Now we have

$$\begin{aligned}(g\circ f)(x+h) - (g\circ f)(x) &= g(f(x+h)) - g(f(x)) =\\ &= g(f(x) + (f(x+h) - f(x))) - g(f(x)) = g(y+k) - g(y)\end{aligned}$$

where $k = f(x+h) - f(x)$, and hence

$$\begin{aligned}(g\circ f)(x+h) - (g\circ f)(x) &= Bk + \nu(k)k =\\ &= B(f(x+h) - f(x)) + \nu(f(x+h) - f(x))(f(x+h) - f(x)) =\\ &= B(Ah + \mu(h)h) + \nu(Ah + \mu(h)h)(Ah + \mu(h)h) =\\ &= (BA)h + (A\mu(h) + \nu((A + \mu(h))h)(A + \mu(h)))h.\end{aligned}$$

Now if we define $\overline{\mu}(h) = A\mu(h) + (A + \mu(h))\nu((A + \mu(h))h)$ we obtain

$$(g\circ f)(x+h) - (g\circ f)(x) = (BA)h + \overline{\mu}(h)h$$

and since $\lim_{h\to 0}\overline{\mu}(h) = 0$ (indeed, we have trivially $\lim_{h\to 0} A\mu(h) = 0$, and $\lim_{h\to 0}\nu((A + \mu(h))h) = 0$ by IV.6.5.1 – recall augmenting ν by setting $\nu(0) = 0$ above) the statement follows from 1.5. □

2.2.2. Note on one-sided derivatives. Unlike the arithmetic rules 2.1, and also unlike the inverse rule 2.3 to follow, one has to be careful with the one-sided derivatives in composition. Even if x keeps to the right resp. left of x_0, the $f(x)$ can oscilate around the $f(x_0)$.

2.3. The rule for inverse.

Theorem. *Let $f : D \to \mathbb{R}$ be an inverse of $g : E \to \mathbb{R}$ and let g have a non-zero derivative in y_0. Then f has a derivative in $x_0 = g(y_0)$ and we have*

$$f'(x_0) = \frac{1}{g'(y_0)} = \frac{1}{g'(f(x_0))}.$$

Proof. We have $f(x_0) = f(g(y_0)) = y_0$. Thus, the function

$$F(y) = \frac{y - y_0}{g(y) - g(y_0)} = \frac{y - f(x_0)}{g(y) - x_0}$$

has a non-zero limit $\lim_{y\to y_0} f(y) = \frac{1}{g'(y_0)}$. The function f is continuous (recall IV.4.2) and since it has an inverse, it is one-to-one. Hence we can use IV.6.5.1 for $F \circ f$ to obtain

$$\lim_{x\to x_0} F(f(x)) = \frac{1}{g'(y_0)}.$$

Now since

$$F(f(x)) = \frac{f(x) - f(x_0)}{g(f(x)) - x_0} = \frac{f(x) - f(x_0)}{x - x_0},$$

the statement follows. □

2.3.1. Note. The point of the previous theorem is that

$$f'(x_0) \text{ \emph{exists.}}$$

The value follows from 2.2: obviously the derivative of the identical function $\mathrm{id}(y) = y$ is constant 1, and since $\mathrm{id}(y) = y = f(g(y))$ we have $1 = f'(g(y))g'(y)$. But, of course, to apply 2.2.1 we have to assume the existence of the derivative of f.

2.4. Summary. In the following section we will learn how to differentiate x, $\ln x$ and $\sin x$. Then, 2.1, 2.2 and 2.3 will provide an algorithm for differentiating general elementary functions.

3. Derivatives of elementary functions.

It would suffice to present derivatives of constants, of the identity (which we already know anyway, the first is constant 0 and the other is constant 1), and of sinus and logarithm: every elementary function can be obtained from these by repeatedly applying the arithmetic constructions, compositions and taking inverse, and for all this we have differentiation rules. For various reasons we will be sometimes more explicit.

3.1. Polynomials. We have

$$(x^n)' = nx^{n-1} \text{ for all natural } n.$$

This can be derived by induction using 2.1(3), but we can compute it directly.

For $n = 0$ the formula is trivial. Let $n > 0$. Then

$$\begin{aligned}\lim_{h\to 0}\frac{(x+h)^n - x^n}{h} &= \lim_{h\to 0}\frac{\sum_{k=0}^{n}\binom{n}{k}x^{n-k}h^k - x^n}{h} = \\ &= \lim_{h\to 0}\frac{\binom{n}{1}x^{n-1}h + h^2\sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^{k-2}}{h} = \\ &= nx^{n-1} + \lim_{h\to 0} h\sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^{k-2} = nx^{n-1}.\end{aligned}$$

Consequently we have

$$(\sum_{k=0}^{n} a_k x^k)' = \sum_{k=1}^{n} k a_k x^{k-1}.$$

3.1.1. Negative powers. Also for $-n$, n natural, we have, by 2.1.4

$$(x^{-n})' = \frac{1}{x^n} = \frac{-nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

3.1.2. Roots and rational powers. By 2.3 we obtain for $f(x) = \sqrt[q]{x}$ (since $g(y) = y^q$)

$$(\sqrt[q]{x})' = \frac{1}{q(\sqrt[q]{x})^{q-1}} = \frac{1}{q}(\sqrt[q]{x})^{1-q}.$$

Thus, using 2.2.1 we obtain (again)

$$(x^{\frac{p}{q}})' = \frac{1}{q}(\sqrt[q]{x^p})^{1-q}px^{p-1} = \frac{p}{q}x^{(\frac{p(1-q)}{q}+p-1)} = \frac{p}{q}x^{\frac{p-q}{q}} = \frac{p}{q}x^{\frac{p}{q}-1}.$$

3.2. Logarithm. We have

$$(\lg x)' = \frac{1}{x}.$$

Indeed, using V.1.2, V.1.3 and IV.6.5.1 we obtain

$$\lim_{h\to 0}\frac{\lg(x+h) - \lg x}{h} = \lim_{h\to 0}\frac{\lg\frac{x+h}{h}}{h} = \lim_{h\to 0}\frac{1}{x}\frac{\lg(1+\frac{h}{x})}{\frac{h}{x}} = \frac{1}{x}\lim_{h\to 0}\frac{\lg(1+\frac{h}{x})}{\frac{h}{x}} = \frac{1}{x}.$$

3.3. Exponentials, general powers. By 3.2 and 2.3 we have

$$(e^x)' = \frac{1}{\lg'(e^x)} = \frac{1}{\frac{1}{e^x}} = e^x.$$

Consequently, by 2.2,

$$(a^x)' = (e^{x \lg a})' = \lg a \cdot e^{x \lg a} = \lg a \cdot a^x.$$

For the general exponent a (albeit for positive x only) we obtain, not surprisingly,

$$(x^a)' = (e^{a \lg x})' = (e^{a \lg x})a\frac{1}{x} = ax^{a-1}.$$

3.4. Goniometric functions. We have

$$(\sin x)' = \cos x \quad \text{and} \quad (\cos x)' = -\sin x.$$

Indeed, by V.3.1.2 and V.3.2,

$$\lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h} = \lim_{h\to 0}\frac{\sin x\cos h + \sin h\cos x - \sin x}{h} =$$
$$= \lim_{h\to 0}\frac{\sin x(\cos h - 1) + \sin h\cos x}{h} = \sin x \cdot \lim_{h\to 0}\frac{\cos h - 1}{h} + \cos x \cdot \lim_{h\to 0}\frac{\sin h}{h} =$$
$$= \sin x \cdot 0 + \cos x \cdot 1 = \cos x,$$

and by V.3.1.1 and 2.2,

$$(\cos x)' = (\sin(\frac{\pi}{2} - x))' = \cos(\frac{\pi}{2} - x)\cdot(-1) = -\sin x.$$

Further we have, by 3.2.1(4),

$$(\tan x)' = \left(\frac{\sin x}{\cos x}\right)' = \frac{\cos x\cos x - \sin x(-\sin x)}{\cos^2 x} = \frac{\cos^x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

3.5. Cyclometric functions. By 2.3 we obtain

$$(\arcsin x)' = \frac{1}{\sin(\arcsin x)} = \frac{1}{\sqrt{1-\sin^2(\arcsin x)}} = \frac{1}{\sqrt{1-x^2}}.$$

The following formula is of a particular interest:

$$(\arctan x)' = \frac{1}{1+x^2}.$$

For this realize first (contemplating the rectangular triangle with sides 1 and $\tan x$) that

$$\cos^2 x = \frac{1}{1+\tan^2 x}$$

and using 2.3 compute

$$(\arctan\ x)' = \frac{1}{\tan'(\arctan x)} = \cos^2(\arctan x) = \frac{1}{1+\tan^2(\arctan x)} = \frac{1}{1+x^2}.$$

4. Derivative as a function. Higher order derivatives.

4.1. So far, strictly speaking, we have spoken just about derivatives of a function in that or other point. In fact, a function $f : D \to \mathbb{R}$ has often derivatives in all the points of D, or in its substantial part D'. We then have a function

$$f' : D' \to \mathbb{R}$$

and we speak of this *function* as of the derivative of f. As we have already indicated in 1.2, this function is often denoted by

$$\frac{\mathrm{d}f}{\mathrm{d}x} \quad \text{or} \quad \frac{\mathrm{d}}{\mathrm{d}x}f.$$

4.2. Derivatives of higher order. The function f' can have, again, a derivative f'', called the *second derivative* of f, and we also can have, further, the *third derivative* f''' and so on. We speak of the *derivatives of higher order*. Instead of n dashes one uses the symbol

$$f^{(n)}$$

and the symbols $\frac{\mathrm{d}f}{\mathrm{d}x}$, $\frac{\mathrm{d}}{\mathrm{d}x}f$ are in this sense extended to

$$\frac{\mathrm{d}^n f}{\mathrm{d}x^n} \quad \text{and} \quad \frac{\mathrm{d}^n}{\mathrm{d}x^n}f.$$

4.3. Note. The reader has observed that the derivative (in particular that of lg or of arctan) may be substantially simpler than the original function. This is not quite such a good news as it may appear. In fact it shows that when we will stand before the reverse task to differentiation, the integration, we can expect the results substantially more complex than the originals. And indeed the means for integration are very limited and integrals of elementary functions are often not elementary.

VII. Mean Value Theorems.

1. Local extremes.

1.1. Increasing and decreasing at a point. A function $f : D \to \mathbb{R}$ *increases* (resp. *decreases*) at a point x if there is an $\alpha > 0$ such that

$$x - \alpha < y < x \;\Rightarrow\; f(y) < f(x) \quad\text{and}\quad x < y < x + \alpha \;\Rightarrow\; f(x) < f(y)$$
$$(\text{resp. } x - \alpha < y < x \;\Rightarrow\; f(y) > f(x) \quad\text{and}\quad x < y < x + \alpha \;\Rightarrow\; f(x) > f(y).$$

1.1.1. Note. If a function increases resp. decreases in an interval then it obviously increases resp. decreases in each point of the interval. On the other hand, if a function (say) increases at a point x there may be no open interval $J \ni x$ in which the function would increase. For instance, the function

$$f(x) = \begin{cases} x + \frac{1}{2}x \sin \frac{1}{x} & \text{for } x \neq 0, \\ 0 \ \ \text{for } x = 0. \end{cases}$$

(draw a picture) increases in 0, but does not increase in any open interval J containing 0.

The question naturally arises whether a function that increases in every point of J increases in J. This is not straightforward, but see the easy 3.1 below.

1.1.2. Proposition. *Let $f'(x) > 0$ (resp. < 0). Then f increases (resp. decreases) in x.*

Proof. Recall VI.1.5 with $A = f'(x)$. Consider $\alpha > 0$ such that $|\mu(x)| < |A|$ for $-\alpha < x < \alpha$. Then in the expression

$$f(x + h) - f(x) = (A + \mu(h))h$$

the $A + \mu(h)$ is positive (resp. negative) iff A is, and hence $f(x + h) - f(x)$ has the same sign as h (resp. the opposite one). $\square$

1.2. Local extremes. A function $f : D \to \mathbb{R}$ has a *local maximum* (resp. *local minimum*) $M = f(x)$ at a point x if there is an $\alpha > 0$ such that for the points y in D

$$x - \alpha < y < x \;\Rightarrow\; f(y) \leq f(x) \quad\text{and}\quad x < y < x + \alpha \;\Rightarrow\; f(x) \geq f(y)$$
$$(\text{resp. } x - \alpha < y < x \;\Rightarrow\; f(y) \geq f(x) \quad\text{and}\quad x < y < x + \alpha \;\Rightarrow\; f(x) \leq f(y).$$

The common term for local maxima and local minima is

local extremes.

Note. We have emphasized that the condition is applied for the elements of D only (which we usually do not do, recall the convention in IV.5.1). For instance, the function $f : \langle 0, 1\rangle \to \mathbb{R}$ defined by $f(x) = x$ has a local minimum 0 in $x = 0$ and local maximum 1 in $x = 1$.

1.3. Comparing the definitions 1.1 and 1.2, and using Proposition 1.1.2 we immediately obtain

Proposition. *If f is increasing or decreasing at a point x. in particular if it has a non-zero derivative in x then it does not have a local extreme in x.*

2. Mean Value Theorems.

2.1. Theorem. (Rolle Theorem.) *Let f be continuous on a compact interval $J = \langle a, b\rangle$, $a < b$, let it have a derivative on the open interval (a, b) and let $f(a) = f(b)$. Then there is a point $c \in (a, b)$ such that $f'(c) = 0$.*

Proof. By Theorem IV.5.2 the function f achieves a maximum (and hence a local maximum) at a point $x \in J$ and a minimum (and hence a local minimum) at a point $y \in J$.

I. If $f(x) = f(y)$: then f is constant on J and hence has the derivative equal to 0 everywhere in (a, b).

II. If $f(x) \neq f(y)$ then at least one of x, y is neither a nor b. If we denote it by c we see by 1.3 that $f'(c) = 0$. □

2.2. Theorem. (Mean Value Theorem, Lagrange Theorem.) *Let f be continuous on a compact interval $J = \langle a, b\rangle$, $a < b$ and let it have a derivative on the open interval (a, b). Then there is a point $c \in (a, b)$ such that*

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

Proof. Define a function $F : \langle a, b\rangle \to \mathbb{R}$ by setting

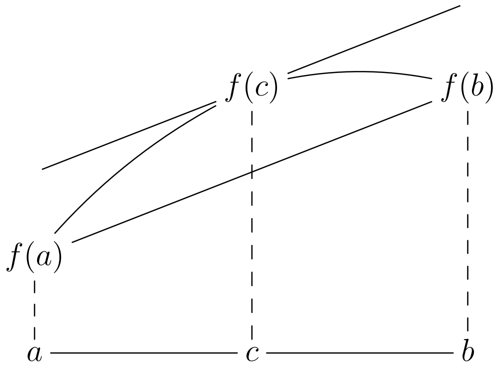
$$F(x) = (f(x) - f(a))(b - a) - (f(b) - f(a))(x - a).$$

Then F is continuous on $\langle a,b\rangle$, has (by the standard rules from the previous chapter) a derivative, namely

$$F'(x) = f'(x)(b-a) - (f(b)-f(a)), \tag{$*$}$$

and $F(a) = F(b) = 0$. Hence we can apply Rolle Theorem 2.1 and $(*)$ yields $0 = f'(c)(b-a) - (f(b)-f(a))$, that is. $f'(c)(b-a) = f(b) - f(a)$ and the statement follows by dividing both sides of this equation by $b-a$. □

2.2.1. Here is a geometric interpretation. The curve ("diagram of the function f") $\{(x, f(x)) \,|\, x \in J\}$ has a tangent parallel to the segment connecting the point $(a, f(a))$ with $(b, f(b))$. See the picture below.



2.2.2. Slight, but often expedient, reformulations. First note that the formula from 2.2 also holds if $b < a$ (than we of course speak about a c in (b,a). If the derivative makes sense between x and $x+h$ we can state that

$$f(x+h) - f(x) = f'(x+\theta h)h \quad \text{with } 0 < \theta < 1$$

(compare with the formula in V.1.5). This is often written in the form

$$f(y) - f(x) = f'(x + \theta(y-x))(y-x) \quad \text{with } 0 < \theta < 1.$$

2.3. Theorem. Generalized Mean Value Theorem, Generalized Lagrange Theorem.) *Let f, g be continuous on a compact interval $J = \langle a,b\rangle$, $a < b$, and let them have derivatives on the open interval (a,b). Let g' be non-zero on (a,b). Then there is a point $c \in (a,b)$ such that*

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

Proof. Is practically the same as in 2.2. Define a function $F : \langle a, b\rangle \to \mathbb{R}$ by setting

$$F(x) = (f(x) - f(a))(g(b) - g(a)) - (f(b) - f(a))(g(x) - g(a)).$$

Then F has a derivative, namely

$$F'(x) = f'(x)(g(b) - g(a)) - (f(b) - f(a))g'(x), \tag{$*$}$$

and $F(a) = F(b) = 0$. Hence we can apply Rolle Theorem again and $(*)$ yields $0 = f'(c)(g(b)-g(a)) - (f(b)-f(a))g'(c)$, that is, $f'(c)(g(b)-g(a)) = (f(b)-f(a))g'(c)$. Now by 2.2, $g(b)-g(a) = g'(\xi)(b-a) \neq 0$ and our formula immediately follows dividing both sides by $(g(b) - g(a))g'(c)$. □

3. Three simple consequences.

3.1. Proposition. *Let $f : D \to \mathbb{R}$ be continuous on $\langle a, b\rangle$ and let it have a positive (resp. negative) derivative on $(a, b) \smallsetminus \{a_1, \dots, a_n\}$ for some finite sequence $a < a_1 < a_2 < \dots < a_n < b$. Then f increases (resp. decreases) on $\langle a, b\rangle$.*

Proof. Since the statement obviously holds if it holds for the restrictions to $\langle a, a_1\rangle$, $\langle a_i, a_{i+1}\rangle$ and $\langle a_n, b\rangle$, it suffices to prove it disregarding the a_i. Let $a \leq x < y \leq b$. Then we have a c such that $f(y) - f(x) = f'(c)(y - x)$. If $f'(c)$ is positive, $f(y) > f(x)$. □

3.2. Discontinuities of derivatives. Let a derivative of a function $f : J \to \mathbb{R}$ where J is an open interval exist on the whole of J. The function f has to be continuous (recall VI.1.6), but f' may not be so. Consider the $f : \mathbb{R} \to \mathbb{R}$ defined by setting

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

For $x \neq 0$ we obtain, using the rules from VI.2 and VI3,

$$f'(x) = 2x \sin \frac{1}{x} + x^2 \cdot \cos \frac{1}{x} \cdot \left(-\frac{1}{x^2}\right) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$$

and hence $\lim_{x\to 0} f'(x)$ does not exist: the value of f' in $\frac{1}{2k\pi + \frac{\pi}{2}}$ is $-1 + \frac{2}{2k\pi + \frac{\pi}{2}}$ while it is 0 in $\frac{1}{2k\pi}$.

However, $f'(0)$ does exist and it is equal to 0, since $\left|\frac{f(h)-f(0)}{h}\right| = \left|h \sin \frac{1}{h}\right| \leq |h|$.

3.2.1. The discontinuity of the f' above was of the second kind (recall IV.6.7.). This is all that can happen: a derivative cannot have a discontinuity of the first kind. We have

Proposition. *Let* $\lim_{y\to x} f'(y)$ *(or* $\lim_{y\to x+} f'(y)$, $\lim_{y\to x-} f'(y)$, *resp.) exist. Then* $f'(x)$ *(*$f'_+(x)$, $f'_-(x)$, *resp.) exists and is equal to the respective limit.*

Proof will be done for f'_+. We have, by 2.2.2,

$$\frac{f(x+h)-f(x)}{h} = f'(x+\theta_h h), \quad 0 < \theta_h < 1,$$

and $\lim_{h\to 0+} f'(x+\theta_h h) = \lim_{h\to 0+} f'(x+h) = \lim_{y\to x+} f'(y)$. □

3.3. Unicity of a primitive function. Later on we will be interested in the task reverse to differentiation (recall VI.4.3), in determining the *primitive function* F of f, that is, an F such that $F' = f$. Such an F cannot be uniquely determined (for instance $(x+1)' = x' = 1$), but the situation is fairly transparent. We have

Proposition. *Let* J *be an interval and let* $F.G : J \to \mathbb{R}$ *be functions such that* $F' = G' = f$. *Then there is a constant* C *such that* $F = G + C$.

Proof. Set $H = F - G$. Then $H' = \mathrm{const}_0$ and since H is defined on an interval we have by 2.2 for any x, y,

$$H(x) - H(y) = H'(c)(x-y) = 0. \qquad \square$$

3.3.1. Note. The assumption that the domain is an interval is, of course, essential.

.

VIII. Several applications of differentiation.

1. First and second derivatives in physics.

Recall VI.1.4. One of the first motivations (and applications) came in physics.

1.1. Represent a moving body in the Euclidean space $\mathbb{E}_3$ by its position in time

$$(x(t), y(t), z(t))$$

(here, the coordinates x, y, z are the real functions to be analyzed, and the real argument, representing time, will be denoted by t). The velocity is then represented by the vector function (that is, a function $D \to \mathbb{R}^3$ with coordinates real functions)

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}(t), \frac{\mathrm{d}y}{\mathrm{d}t}(t), \frac{\mathrm{d}z}{\mathrm{d}t}(t)\right). \qquad (*)$$

1.2. Acceleration. One of the most important concepts of Newtonian physics (and of physics in general), the *force*, is connected with the *acceleration*, the second derivative of (x, y, z),

$$\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2}(t), \frac{\mathrm{d}^2y}{\mathrm{d}t^2}(t), \frac{\mathrm{d}^2z}{\mathrm{d}t^2}(t)\right).$$

The reader certainly knows that the force is given as $M(\frac{\mathrm{d}^2x}{\mathrm{d}t^2}, \frac{\mathrm{d}^2y}{\mathrm{d}t^2}, \frac{\mathrm{d}^2z}{\mathrm{d}t^2})$ where M is the mass.

1.3. Tangent of a curve. The same way as in 1.1 we can express the tangent of a curve given parametrically as (f_1, f_2, f_3) with $f_i : J \to \mathbb{R}$ real functions. We have then $(f_1'(x_0), f_2'(x_0), f_3'(x_0))$ the vector determining the direction of the tangent in the point $(f_1(x_0), f_2(x_0), f_3(x_0))$, and the tangent is expressed parametrically as

$$(f_1(x_0), f_2(x_0), f_3(x_0)) + x(f_1'(x_0), f_2'(x_0), f_3'(x_0)), \quad x \in \mathbb{R}.$$

1.3. Note. In VI.1.4 we have also mentioned another aspect of the derivative, the approximation. More on that will come later, in particular in the Section about Taylor's formula.

2. Determining local extremes.

2.1. Proposition. *For a function $f : D \to \mathbb{R}$ consider the set $E(f)$ consisting of all the $x \in D$ such that*

- *x is not an interior point of D, or*
- *$f'(x)$ does not exist, or*
- *$f'(x) = 0$.*

Then $E(f)$ contains all the points in which there is a local extreme.

Proof. In all the points that are not in $E(f)$ there is a non-zero derivative. Use VII.1.2. □

2.2. Notes. 1. When looking for local extremes one should not forget about the non-interior points and the points without a derivative. Determining the x such that $f'(x) = 0$ does not finish the task.

2. Proposition 1.3.1 provides a list of all possible candidates of a local extreme. This does not say that all the elements of $E(f)$ are local extremes. See the following examples.

(a) Define $f : \langle 0, \infty) \to \mathbb{R}$ by setting

$$f(x) = \begin{cases} x \sin \frac{1}{x} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

There is no local extreme in the non-interior point 0.

(b) Define $f : (0, 2) \to \mathbb{R}$ by setting

$$f(x) = \begin{cases} x & \text{for } 0 < x \leq 1, \\ 2x - 1 & \text{for } 1 \leq x < 2. \end{cases}$$

f has not a derivative in $x = 1$, but there is no extreme there.

(c) $f(x) = x^3$ defined on all the $\mathbb{R}$ has no extreme in $x = 0$ although $f'(0) = 0$.

3. Convex and concave functions

From VII.3.1 w know that the sign of the (first) derivative determines whether a function increases or decreases. The second derivative determines whether a function is *convex* (“rounded downwards”) or *concave* (“rounded upwards).

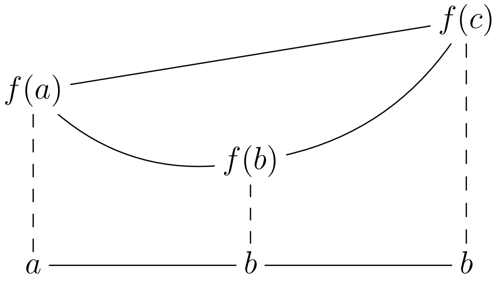
3.1. We say that a function $f : D \to \mathbb{R}$ is *convex* (resp. *strictly convex*) on an interval $J \subseteq D$ if for any a, b, c in J such that $a < b < c$ we have

$$\frac{f(c) - f(b)}{c - b} - \frac{f(b) - f(a)}{b - a} \geq 0 \quad (\text{resp. } > 0). \tag{$*$}$$

We say that it is *concave* (resp. *strictly concave*) on the J if for any for $a < b < c$ in J we have

$$\frac{f(c) - f(b)}{c - b} - \frac{f(b) - f(a)}{b - a} \leq 0 \quad (\text{resp. } < 0).$$

3.2. The formula for convexity expresses the fact that the values $f(b)$ of f in the intermediate points between a, c lie below the segment connecting the points $(a, f(a))$ and $(c, f(c))$ in the plane $\mathbb{R}^2$. See the following picture.



The connecting segment is given by

$$y = f(a) + \frac{f(c) - f(a)}{c - a}(x - a), \quad a \leq x \leq b,$$

and if we set $x = b$ we obtain $y(b) = f(a) + \frac{f(c)-f(a)}{c-a}(b - a)$ so that, say, requiring that the value $f(b)$ is below the segment, that is, $f(b) < y(b)$, yields

$$\frac{f(c) - f(a)}{c - a} - \frac{f(b) - f(a)}{b - a} > 0. \tag{$**$}$$

For $x, y > 0$ we have $\frac{X}{x} > \frac{Y}{y}$ iff $\frac{X+Y}{x+y} > \frac{Y}{y}$ (the first says that $Xy > Yx$, the second that $Xy + Yy > Yx + Yy$) so that the formula $(**)$ is equivalent with the $(*)$ for the strict convexity.

3.3. Proposition. *Let $f : D \to \mathbb{R}$ be continuous on $\langle a, b\rangle$ and let it have a second derivative on $(a, b) \smallsetminus \{a_1, \dots, a_n\}$ for some finite sequence $a < a_1 < a_2 < \cdots < a_n < b$. Let $f''(x) > 0$ (≥ 0, ≤ 0, < 0, resp.) in $(a, b) \smallsetminus \{a_1, \dots, a_n\}$. Then f is strictly convex (convex, concave, strictly concave, resp.) on $\langle a, b\rangle$.*

Proof. Similarly like in VII.3.1 we can disregard the exceptional points a_i and prove the theorem for f continuous on $\langle a, b\rangle$ with the specified second derivative in (a, b). We will consider, say, $f''(x) > 0$ in this open interval.

By Mean Value Theorem we have have for $x < y < z$ in $\langle a, b\rangle$

$$V = \frac{f(z) - f(y)}{z - y} - \frac{f(y) - f(x)}{y - x} = f'(v) - f'(u)$$

for some $x < u < y < v < z$. Using the same theorem again we obtain

$$V = f''(w)(v - u)$$

with $u < w < v$, hence $v - u > 0$ and $w \in (a, b)$ so that also $f''(w) > 0$ and $V > 0$. □

3.4. Inflection. An *inflection point* of a function $f : D \to \mathbb{R}$ is an element $x \in D$ such that there is a $\delta > 0$ with $(x - \delta, x + \delta) \subseteq D$ such that

– either f is convex on $(x - \delta, x\rangle$ and concave on $\langle x, x + \delta)$,
– or f is concave on $(x - \delta, x\rangle$ and convex on $\langle x, x + \delta)$.

From 3.3 we immediately obtain

3.4.1. Corollary. *Let J be an interval and let $f : J \to \mathbb{R}$ have a continuous second derivative in J. Then we have $f''(x) = 0$ in every inflection point of f.*

3.4.2. Note. Thus, for a function on an interval with continuous second derivative, we have a list $\{x \,|\, f''(x) = 0\}$ containing all inflection points. But not all the x with $f''(x) = 0$ are necessarily inflection ones. Consider the functions $f(x) = x^{2n}$: they are convex on the whole of $\mathbb{R}$ while $f''(0) = 0$.

4. Newton's Method

(Also known as Newton-Raphson Method.) This is a method of finding a sucession of approximative solutions of an equation $f(x) = 0$. It can be very effective – see 4.3 below.

4.1. Suppose you wish to solve an equation

$$f(x) = 0 \tag{$*$}$$

where f is a real function such that f' exists. Suppose that the values of f and of f' are not hard to compute. Then the following procedure often yields a very fast convergence to the solution.

For a $b \in D$ consider the point $(b, f(b))$ on the graph $\Gamma = \{(x, f(x)) \mid x \in D\}$ of the function f. Then take the tangent of Γ in this point. This tangent is the graph of the linear function

$$L(x) = f(b) + f'(b)(x - b).$$

In a reasonably small neighbourhood of b the function $L(x)$ is a good approximation of the function f and hence we can conjecture that the solution of

$$L(x) = 0 \tag{$**$}$$

approximates a solution of the equation $(*)$ above. The solution of $(**)$ is easy to compute: it is

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

Draw a picture!

The point $\widetilde{b}$ is much closer to the solution of $(*)$ then b, and if we repeat the procedure, the resulting $\widetilde{\widetilde{b}}$ is much closer again.

4.2. This leads to the following procedure called *Newton's method.* To solve approximatively the equation $(*)$ above,

- first, choose an approximation a_0 (not necessarily a good one, just something to start with), and
- second, define

$$a_{n+1} = \widetilde{a_n} = a_n - \frac{f(a_n)}{f'(a_n)}.$$

The resulting sequence

$$a_0, a_1, a_2, \ldots$$

(if certain conditions are satisfied) converges to a solution, and often very fast – see 4.3.

4.2.1. Example. Let us compute the square root of 3, that is the solution of the equation

$$x^2 - 3 = 0.$$

We get

$$a_{n+1} = a_n - \frac{a_n^2 - 3}{2a_n} = \frac{a_n^2 + 3}{2a_n}.$$

If we start, say, with $a_0 = 2$, we get

$$\begin{aligned} a_1 &= 1.75, \\ a_2 &= 1.732142657, \\ a_3 &= 1.73205081 \end{aligned}$$

Thus, a_1 agrees with with the $\sqrt{3}$ (given in the tables as 1.7320508075) in two digits. a_2 in four digits, and a_3 already in eight digits!

4.3. In the example we have seen that (under favourable circumstances) the error may diminish very rapidly. Let us present an easy estimate under the condition that the second derivative exists.

Denote by a the solution, that is, an a with $f(a) = 0$. We have

$$a_{n+1} - a = a_n - a - \frac{f(a_n)}{f'(a_n)} = a_n - a - \frac{f(a_n) - f(a)}{f'(a_n)},$$

and hence by Mean Value Theorem there is an α between a_n and a such that

$$a_{n+1} - a = (a_n - a) - (a_n - a)\frac{f(\alpha)}{f'(a_n)} = (a_n - a)\left(1 - \frac{f(\alpha)}{f'(a_n)}\right),$$

and, further, using Mean Value Theorem again, this time for the first derivative f', we obtain a β between a and α such that

$$a_{n+1} - a = (a_n - a)\left(\frac{f'(a_n) - f'(\alpha)}{f'(a_n)}\right) = (a_n - a)(a_n - \alpha)\frac{f''(\beta)}{f'(a_n)}$$

so that, since α is between a_n and a we obtain, taking an upper estimate K of $|\frac{f''(\beta)}{f'(a_n)}|$ (which does not have to be very large),

$$|a_{n+1} - a| \leq |a_n - a|^2 K.$$

Thus if we start with an error less than 10^{-1} we have in the next step less than $K \cdot 10^{-2}$, then $K^2 \cdot 10^{-4}$, $K^3 \cdot 10^{-8}$, $K^4 \cdot 10^{-16}$, etc., which may be a very satisfactory convergence indeed, as we have seen at the $\sqrt{3}$ above.

4.4. Note. Needless to say, the choice of a_0 is essential. Sometimes the adjustment comes automatically: in the example 4.2.1 we started with $a_0 = 2$ "on the right side of the convexity". If we started "on the wrong side", say in 1, we obtain $a_1 = 2$ so that the first step just get us to the "right side", and we proceed just with one step delay (draw a picture).

On the other hand, one can start very badly. Consider $f(x) = -\frac{7}{4}x^4 + \frac{15}{4}x^2 - 1$. Then $f(1) = f(-1) = 1$, $f'(1) = -f'(-1) = \frac{1}{2}$ and if we start with $a_0 = 1$ we obtain

$$a_1 = -1, a_2 = 1, a_3 = -1, a_4 = 1, \text{ etc.}$$

5. L'Hôpital Rule

(Also L'Hospital Rule; believed to be discovered by Johann Bernoulli.)

5.1. The simple L'Hôpital Rule. We will have a harder one later; this one is very easy.

Proposition. *Let $\eta > 0$. let f, g have deivatives in all the x such that $0 < |x - a| < \eta$, and let $\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = 0$. Let $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ exist. Then also $\lim_{x\to a} \frac{f(x)}{g(x)}$ exists and we have*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Proof. We can define $f(a) = g(a) = 0$ to obtain continuous functions on $\langle a, x\rangle$ resp. $\langle x, a\rangle$ for $|x - a|$ sufficiently small. Furthermore, because $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ exists, if $|x-a|$ is sufficiently small there are derivatives on (a, x) resp. (x, a), and the derivative g' is there non-zero. Therefore we can apply VII.2.3 and obtain

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(c)}{g'(c)}$$

for some c between a and x. Thus, if $0 < |x-a| < \delta$ we also have $0 < |c-a| < \delta$ and hence if we choose a $\delta > 0$ such that for $0 < |c-a| < \delta$ we have $|\frac{f'(c)}{g'(c)} - L| < \varepsilon$, we also have for $0 < |x-a| < \delta$, $|\frac{f'(x)}{g'(x)} - L| < \varepsilon$. $\square$

5.1.1. Note. In the previous proof, if $x > a$ then $c > a$ and if $x < a$ then $c < a$. Thus we have in fact also proved that, under the corresponding conditions,

$$\lim_{x\to a+} \frac{f(x)}{g(x)} = \lim_{x\to a+} \frac{f'(x)}{g'(x)} \quad \text{and} \quad \lim_{x\to a-} \frac{f(x)}{g(x)} = \lim_{x\to a-} \frac{f'(x)}{g'(x)}.$$

5.1.2. Examples. Let us recall the limits from V.1.3 and V.3.2

$$\lim_{x\to 0} \frac{\lg(1+x)}{x} = \lim_{x\to 0} \frac{\frac{1}{1+x}}{1} = 1, \quad \lim_{x\to 0} \frac{\sin x}{x} = \lim_{x\to 0} \frac{\cos x}{1} = 1$$

(of course, we would not be able to compute the derivatives of lg or sin in VI.3 without knowing these limits beforehand; this is just an illustration). Or we can compute

$$\lim_{x\to 0} \frac{\cos x - 1}{x^2} = \lim_{x\to 0} \frac{-\sin x}{2x} = \lim_{x\to 0} \frac{-\cos x}{2} = \frac{-1}{2}.$$

5.2. Infinite limits and limits in infinity. To be able to extend L'Hôpital rule to its full generality we will have to extend our concept of limit of a function.

We say that a function $f : D \to \mathbb{R}$ *has a limit* $+\infty$ (resp. $-\infty$) at a point a, and write

$$\lim_{x\to a} f(x) = +\infty \quad (\text{resp. } -\infty)$$

if $\forall K\ \exists\delta > 0$ such that $(0 < |x-a| < \delta) \Rightarrow f(x) > K$ (resp. $< K$).

A function $f : D \to \mathbb{R}$ *has a limit b in* $+\infty$ (resp. $-\infty$), written

$$\lim_{x\to +\infty} f(x) = b \quad (\text{resp. } \lim_{x\to -\infty} f(x) = b)$$

if $\forall\varepsilon > 0\ \exists K$ such that $x > K$ (resp.$x < K$) $\Rightarrow |f(x) - b| < \varepsilon$.

A function $f : D \to \mathbb{R}$ *has a limit* $+\infty$ *in* $+\infty$, written

$$\lim_{x\to +\infty} f(x) = +\infty$$

if $\forall K\ \exists K'$ such that $x > K' \ \Rightarrow\ f(x) > K$ (similarly for limits $+\infty$ in $-\infty$, $-\infty$ in $-\infty$ and $-\infty$ in $+\infty$).

5.2.1. Remark. The one-sided variants of the previous definitions are obvious. Note that the limits in $+\infty$ and in $-\infty$ are one-sided as they are.

5.3. Scrutinizing the proof of 5.1 we see that this Proposition also holds for infinite limits in finite points, and also for the one-sided ones.

5.4. Proposition.*Let $\eta > 0$, let f, g have deivatives in all the x such that $0 < |x-a| < \eta$ and let $\lim_{x\to a} |g(x)| = +\infty$. Let $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ exist (finite or infinite). Then also $\lim_{x\to a} \frac{f(x)}{g(x)}$ exists and we have*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Proof. This proof will not be quite so transparent as that of 5.1, although the principle is similar. We cannot, of course, use the trick with defining the values in a as zero.

We can write

$$\frac{f(x)}{g(x)} = \left(\frac{f(x)-f(y)}{g(x)-g(y)} + \frac{f(y)}{g(x)-g(y)}\right)\frac{g(x)-g(y)}{g(x)}.$$

Thus, for a suitable ξ between x and y we have

$$\frac{f(x)}{g(x)} = \left(\frac{f'(\xi)}{g'(\xi)} + \frac{f(y)}{g(x)-g(y)}\right)\frac{g(x)-g(y)}{g(x)}. \qquad (*)$$

For technical reasons we will proceed in three alternative cases.

I. $\lim_{x\to a} \frac{f'(x)}{g'(x)} = 0$:

Choose a $\delta_1 > 0$ such that for $0 < |x-a| < \delta_1$ we have $|\frac{f'(x)}{g'(x)}| < \varepsilon$. Fix a y with $0 < |y-a| < \delta_1$. Further, choose a δ with $0 < \delta < \delta_1$ such that

$$0 < |x-a| < \delta \quad \Rightarrow \quad \left|\frac{f(y)}{g(x)-g(y)}\right| < \varepsilon \ \text{ and } \ \left|\frac{g(y)}{g(x)}\right| < 1.$$

Then by $(*)$ we have for $0 < |x-a| < \delta$

$$\left|\frac{f(x)}{g(x)}\right| < (\varepsilon+\varepsilon)2 = 4\varepsilon$$

and hence $\lim_{x\to a}\frac{f(x)}{g(x)}=0$

II. $\lim_{x\to a}\frac{f'(x)}{g'(x)}=L$ finite.

Set $h(x)=f(x)-Lg(x)$. Then $h'(x)=f'(x)-Lg'(x)$ and we have $\frac{h(x)}{g(x)}=\frac{f(x)}{g(x)}-L$ and $\frac{h'(x)}{g'(x)}=\frac{f'(x)}{g'(x)}-L$. Apply the previous step for $\frac{h(x)}{g(x)}$.

III, $\lim_{x\to a}\frac{f'(x)}{g'(x)}=+\infty$ ($-\infty$ is quite analogous):

For K choose a $\delta_1>0$ such that for $0<|x-a|<\delta_1$ we have $\frac{f'(x)}{g'(x)}>2K$. Fix a y with $0<|y-a|<\delta_1$. Choose a δ with $0<\delta<\delta_1$ such that

$$0<|x-a|<\delta \quad\Rightarrow\quad \left|\frac{f(y)}{g(x)-g(y)}\right|<K \text{ and } \left|\frac{g(y)}{g(x)}\right|<\frac{1}{2}.$$

Then by $(*)$ we have for $0<|x-a|<\delta$

$$\frac{f(x)}{g(x)}>(2K-K)(1-\frac{1}{2})>\frac{1}{2}K$$

and the statement follows. $\square$

5.5. In the following, "$\square$" stands for $a, a+, a-, +\infty$ or $-\infty$. To have a derivative "close to $\square$“ means that the function in question has a derivative in $(a-\delta,a+\delta)\smallsetminus\{a\}$ for some $\delta>0$, $(a,a+\delta)$ for some $\delta>0$, $(a-\delta,a)$ for some $\delta>0$, $(K,+\infty)$ for some K, or $(-\infty,K)$ for some K, in this order.

Theorem. (L'Hôpital Rule) *Let* $\lim_{x\to\square}f(x)=\lim_{x\to\square}g(x)=0$ *or* $\lim_{x\to\square}|g(x)|=+\infty$. *Let f,g have derivative close to $\square$ and let* $\lim_{x\to\square}\frac{f'(x)}{g'(x)}=L$ *(finite or infinite) exist. Then* $\lim_{x\to\square}\frac{f(x)}{g(x)}$ *exists and is equal to L.*

Proof. The cases of $\square=a, a+$ or $a-$ are contained in 5.1, 5.3 and 5.4. Thus we are left with $+\infty$ and $-\infty$. They are quite analogous and hence we will discuss just the former.

By IV.6.5.1 adapted for limits in $+\infty$,

$$\lim_{x\to+\infty}H(x)=\lim_{x\to0+}H(\frac{1}{x}).$$

So if we set $F(x)=f(\frac{1}{x})$ and $G(x)=g(\frac{1}{x})$ we have $F'(x)=f(\frac{1}{x})\cdot\frac{1}{x^2}$ and $G'(x)=g(\frac{1}{x})\cdot\frac{1}{x^2}$, and

$$\lim_{x\to0+}\frac{F'(x)}{G'(x)}=\lim_{x\to0+}\frac{f'(\frac{1}{x})\cdot\frac{1}{x^2}}{g'(\frac{1}{x})\cdot\frac{1}{x^2}}=\lim_{x\to0+}\frac{f'(\frac{1}{x^2})}{g'(\frac{1}{x^2})}=\lim_{x\to+\infty}\frac{f'(x)}{g'(x)}=L.$$

Hence by the previous facts,

$$\lim_{x\to+\infty}\frac{f(x)}{g(x)} = \lim_{x\to 0+}\frac{F(x)}{G(x)} = L. \qquad \square$$

5.5.1. Example. Let $a > 1$. By 5.5,

$$\lim_{x\to+\infty}\frac{a^x}{x^n} = \lim\frac{\lg a \cdot a^x}{nx^{n-1}} = \lim\frac{(\lg a)^2\cdot a^x}{n(n-1)x^{n-2}} = \cdots = \lim_{x\to+\infty}\frac{(\lg a)^n\cdot a^x}{n!} = +\infty.$$

Thus, for arbitrarily small $\varepsilon > 0$ the exponential function $(1+\varepsilon)^x$ grows to infinity faster than any polynomial.

Or, for any $b > 0$,

$$\lim_{x\to+\infty}\frac{x^b}{\lg x} = \lim\frac{bx^{b-1}}{\frac{1}{x}} = \lim_{x\to+\infty} bx^b = +\infty.$$

Thus, for arbitrarily small positive b the function x^b (for instance, any root $\sqrt[n]{x}$) grows to infinity faster than logarithm.

5.6. Indeterminate expressions. This is a common name of limits of functions obtained by simple expressions from functions f, g where we know $\lim f$ and $\lim g$ but the arithmetic rules or similar operations fail. They are indicated by expressions pointing out the trouble. Often we are helped by using the L'Hôpital rule.

5.6.1. The types $\frac{0}{0}$ and $\frac{\infty}{\infty}$. Here we are often helped by using Theorem 5.5: the task in f, g may be indeterminate while the corresponding one in f', g' may be not.

Note. Needless to say, differentiating is a task of type $\frac{0}{0}$.

5.6.2. The type $0\cdot\infty$. This can be made to the type $\frac{0}{0}$ or $\frac{\infty}{\infty}$ rewriting $f(x)g(x)$ as

$$\frac{f(x)}{\frac{1}{g(x)}} \quad\text{or}\quad \frac{g(x)}{\frac{1}{f(x)}},$$

whichever is more expedient.

5.6.3. The type $\infty - \infty$. This is slightly harder. Often the following rewriting helps:

$$f(x) - g(x) = \frac{1}{\frac{1}{f(x)}} - \frac{1}{\frac{1}{g(x)}} = \frac{\frac{1}{g(x)} - \frac{1}{f(x)}}{\frac{1}{f(x)g(x)}}.$$

5.6.4. The types 0^0, 1^∞ and ∞^0. We use the fact that $f(x)^{g(x)} = e^{g(x)\cdot \lg f(x)}$ and that e^x is continuous. Thus it suffices to be able to compute $\lim(g(x)\cdot \lg f(x))$; in the first case we have the type $0\cdot(-\infty)$, in the second, $\infty\cdot 0$, and in the last one, $0\cdot(+\infty)$.

6. Drawing graphs of functions

Suppose we would like to get an idea of the behaviour of a function f presented by a formula. It becomes apparent viewing the graph of f,

$$\Gamma = \{(x, f(x))\,|\, x \in D\},$$

if we can draw it.

For drawing Γ the facts we have learned can be of a great help.

6.1. First, the formula can give us an information about continuity and discontinuity. L'Hôpital rule can help with the limits (also with the one-sided ones) in the critical points, and with the asymptotic behaviour if the domain is not bounded.

6.2. Then try to find the points

$$\cdots < a_i < a_{i+1} < \cdots$$

in which $f(a_i) = 0$. In the intervals (a_i, a_{i+1}) note whether the function is positive or negative.

6.3. Next, consider the first derivative and try to find the points

$$\cdots < b_i < b_{i+1} < \cdots$$

in which $f'(b_i) = 0$ or in which the derivative does not exist. In the intervals (b_i, b_{i+1}) note the sign to learn whether the function increases or decreases. At the b_i where the sign changes we have local extremes.

Determine $f(b_i)$ and if $f'(b_i) = 0$ draw the tangent in $(b_i, f(b_i))$ (parallel with the x-axis). Whether the $f(b_i)$ is a local extreme or not, it will be handy for the curve Γ to lean against. If $f'(b_i)$ does not exist but there are (distinct) one-sided derivatives, draw the "half-tangents".

It may also help to draw the tangents in $(a_i, 0)$ – the more tangents one has to lean against the easier is the final (approximate) drawing the curve.

6.4. Now consider the second derivative and try to find the

$$\cdots < c_i < c_{i+1} < \cdots$$

in which $f''(c_i) = 0$ or in which the second derivative does not exist. In the intervals (c_i, c_{i+1}) note the sign to learn whether the function is convex (that is, rounded dowwards) or concave (rounded upwards). In $(c_i, f(c_i))$ where $f''(c_i) = 0$ draw tangents (these are usually very helpful: they often approximate the curve very closely).

6.5. Now it is usually very easy to draw a curve between the tangents (following the convexity and concavity).

6.5. Note. 1. We may not be able to determine all the values above. But even a part of them may present quite a good image.

2. Needless to say, for solving the equations $f(x) = 0$, $f'(x) = 0$ and $f''(x) = 0$ we can use Newton's Method. But often just a good estimate suffices. For determining useful limits and asymptotics, L'Hôpital rule is often of a help.

6.7. Exercises. 1. Draw the graph the function f from 4.4 and see why the Newton's method with the badly chosen a_0 failed.

2. Draw the graph of $f(x) = \frac{4x}{1+x^2}$ (with domain the whole of $\mathbb{R}$).

2. Draw the graph of $f(x) = e^{\frac{1}{x}}$ (with domain $\mathbb{R} \smallsetminus \{0\}$).

7. Taylor Polynomial and Remainder

7.1. By VI.1.5, a function with a derivative at a point a can be approximated by the linear function (first degree polynomial)

$$p(x) = f(a) + f'(a)(x - a).$$

This polynomial p is characterized by the fact that it agrees with f in $p^{(0)}(a) = f^{(0)}(a) = f(a)$ and $p^{(1)}(a) = f^{(1)}(a)$.

It is natural to conjecture that if we consider a polynomial p of degree n such that

$$p^{(0)}(a) = f^{(0)}(a),\ p^{(1)}(a) = f^{(1)}(a),\ \dots\ ,\ p^{(n)}(a) = f^{(n)}(a) \qquad (*)$$

(we think of the f itself as of its own 0-th derivative) we will get, with the growing n, better and better fit, that is, the remainder $R(x)$ in

$$f(x) = p(x) + R(x)$$

will be getting smaller. This is (with exceptions) really the case as we will shortly.

7.2. Taylor polynomial. First we will see that the conditions $(*)$ uniquely determine a polynomial p of degree n. If $p(x) = \sum_{k=0}^{n} b_k(x-a)^k$ we have

$$p'(x) = \sum_{k=1}^{n} kb_k(x-a)^{k-1},\ p''(x) = \sum_{k=2}^{n} k(k-1)b_k(x-a)^{k-2}, \dots, p^{(n)}(x) = n!b_n,$$

that is,

$$\begin{aligned}
&p^{(1)}(x) = 1 \cdot b_1 + (x-a)\sum_{k=2}^{n} kb_k(x-a)^{k-2},\\
&p^{(2)}(x) = 1 \cdot 2 \cdot b_2 + (x-a)\sum_{k=3}^{n} k(k-1)b_k(x-a)^{k-3},\\
&p^{(3)}(x) = 1 \cdot 2 \cdot 3 \cdot b_3 + (x-a)x\sum_{k=4}^{n} k(k-1)(k-2)b_k(x-a)^{k-4},\\
&\dots,\\
&p^{(n)}(x) = n! \cdot b_n
\end{aligned}$$

so that if $p^{(k)}(a) = f^{(k)}(a)$ for $k = 0, \dots, n$ we have

$$b_k = \frac{1}{k!}p^{(k)}(a) = \frac{f^{(k)}(a)}{k!}, \quad k = 0, \dots, n.$$

The resulting polynomial

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

is called the *Taylor polynomial of degree n* of the function f (in a).

7.3. Theorem. *Let a function f have derivatives $f^{(k)}$, $k = 0, \dots, n+1$ in an interval $J = (a - \Delta, a + \Delta)$. Then we have for all $x \in J$*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}$$

where ξ is a real number between x and a.

Proof. Consider the function of real variable t (x is viewed as a constant)

$$R(t) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!}(x-t)^k.$$

Thus, $R(x) = 0$ and is $R(a) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$ is the remainder, the error when replacing f by its Taylor polynomial.

For the derivative of R we obtain, using the rules for differentiating sums and products (and also the rule for composition taking into account that $\frac{\mathrm{d}}{\mathrm{d}t}(x-t) = -1$),

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = -\sum_{k=0}^{n} \frac{f^{(k+1)}(t)}{k!}(x-t)^k + \sum_{k=1}^{n} \frac{f^{(k)}(t)}{(k-1)!}(x-t)^{k-1}.$$

Replacing the k in the first summand and the $k-1$ in the second summand by r we obtain

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = -\sum_{r=0}^{n} \frac{f^{(r+1)}(t)}{r!}(x-t)^r + \sum_{r=0}^{n-1} \frac{f^{(r+1)}(t)}{r!}(x-t)^r = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n.$$

Now take any g such that g' is non-zero between a and x. Since $R(x) = 0$, we obtain from VII.2.3 that

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

for a ξ between a and x.

If we now set $g(t) = (x-t)^{n+1}$ we have $g'(t) = -(n+1)(x-t)^n$ and $g(x) = 0$ so that

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

the remainder from the statement. □

7.4. Notes. 1. Choosing $g(t) = (x-t)^{n+1}$ belongs to Lagrange, and one often speaks of the remainder in our formulation as of the *remainder in Lagrange form.* Note that it is very easy to remember: one just takes one more summand with $f^{(n+1)}(\xi)$ replacing $f^{(n+1)}(a)$.

One can take, of course, simpler g, but the results are not quite so satisfactory. If we set $g(t) = t$ we obtain

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

the so called *Cauchy remainder formula*, not quite so transparent.

2. For $n = 0$ we obtain

$$f(x) = f(a) + f'(\xi)(x-a),$$

The Mean Value Theorem.

3. The remainder often diminishes quickly (see the examples below), sometimes not quite so (for instance if we try to compute the logarithm lg with the center in $a = 1$).

It can also happen, though, that the whole of the function is in the remainder. Consider

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

Then f has derivatives of all orders, and $f^{(k)}(0) = 0$ for all k.

7.5. Examples. For instance for the exponential we obtain

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + e^{\xi}\frac{x^{n+1}}{(n+1)!},$$

or for the sinus,

$$\sin x = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \pm \frac{x^{2n+1}}{(2n+1)!} \pm \cos\xi\frac{x^{2n+2}}{(2n+2)!}.$$

In both cases the remainder rapidly decreases.

8. Osculating circle. Curvature.

8.1. The value $f'(x)$ of the first derivative determines how fast the function increases or decreases in x, regardless other data concerning f or x.

Since the second derivative f'' determines whether the function f is convex or concave one might, just for a moment, conjecture that it should determine the curvature, that is, that the value of $f''(x)$ should tell us how much the graph is bent in the vicinity of x.

Even the most primitive examples, however, show that it cannot be quite so simple. Consider $f(x) = x^2$. The second derivative is constantly 2, while the bending is not constant at all: the curve is rounded close to $x = 0$ but gets very flat with increasing x.

8.2. Osculating circle. Similarly like the slope of the function was apparent from the tangent (and therefore, from the first derivative), that is the straight line approximating f, the problem of curvature is naturally aproached by trying to find, instead od a straight line, a *circle* that is a good approximation of the graph. It will be a circle touching the graph of f and agreeing with f in the first derivative (that is, having a common tangent in the point in question) and, moreover, also agreeing in the value of the second derivative. Such circle is called

the osculating circle.

8.2.1. So consider a point x_0 and suppose that

- f has in x_0 a second derivative, and
- $f''(x_0) \neq 0$ (“f is convex or concave in the vicinity of x_0”).

To simplify the notation we will write

$$y_0 = f(x_0), \quad y_0' = f'(x_0) \quad \text{and} \quad y_0'' = f''(x_0).$$

The equation of the circle with the center in (a, b) and radius r is

$$(x - a)^2 + (y - b)^2 = r^2 \tag{$*$}$$

and hence if k is a function defined in the vicinity of x_0 whose graph is a part of the circle $(*)$ we have

$$(x - a)^2 + (k(x) - b)^2 = r^2 \tag{1}$$

and taking the first and second derivatives of both sides of the equation (1) (and in the first case also dividing by 2) we obtain

$$(x-a)+(k(x)-b)k'(x)=0 \tag{2}$$
$$1+(k'(x))^2+(k(x)-b)k''(x)=0. \tag{3}$$

Now if k agrees with f as desired we have $k(x_0)=y_0$, $k'(x_0)=y_0'$ and $k''(x_0)=y_0''$ and obtain from (1), (2) and (3) the following system of equations.

$$(x_0-a)^2+(y_0-b)^2=r^2 \tag{1y}$$
$$(x_0-a)+(y_0-b)y_0'=0 \tag{2y}$$
$$1+(y_0')^2+(y_0-b)y_0''=0. \tag{3y}$$

From (2y) we obtain

$$(x_0-a)=-(y_0-b)y_0'$$

so that, by (1y),

$$(y_0-b)^2(1+(y_0')^2)=r^2$$

and since, by (3y), $(y_0-b)=-\frac{1+(y_0')^2}{y_0''}$, we can conclude the following.

8.2.2. Proposition. *The radius of the osculation circle of f in the point x_0 is*

$$r=\frac{(1+(f'(x_0))^2)^{\frac{3}{2}}}{|f''(x_0)|}.$$

□

Note. Now we can also easily compute the coordinates of the a,b of the center. This can be left to the reader as an easy exercise.

8.3. Curvature. the *curvature* of the (graph of) the function f is the inverse $\frac{1}{r}$ of the radius r of the osculating circle. Thus we have

8.3.1. Proposition. *The curvature of f in the point x is*

$$r=\frac{|f''(x)|}{(1+(f'(x))^2)^{\frac{3}{2}}}.$$

□

Note. We see that the conjecture about $f''(x)$ determining the curvature was not so bad, after all. The curvature indeed linearly depends on the second derivative; only, its value has to be adjusted by $\frac{1}{(1+(f'(x))^2)^{\frac{3}{2}}}$.

2nd semester

IX. Polynomials and their roots

1. Polynomials

1.1. We are interested in real analysis but we will still need some basic facts about polynomials with coefficients and variables in the field

$$\mathbb{C}$$

of complex numbers.

From Chapter I, 3.4, recall the abso;ute value $|a| = \sqrt{a_1^2 + a_2^2}$ of the complex number $a = a_1 + a_2 i$ and the triangle inequality

$$|a + b| \le |a| + |b|.$$

Further recall the complex conjugate $\overline{a} = a_1 - a_2 i$ of $a = a_1 + a_2 i$ and the facts that

$$\overline{a+b} = \overline{a} + \overline{b}, \quad \overline{ab} = \overline{a}\overline{b} \quad \text{and} \quad |a| = \sqrt{a\overline{a}}.$$

1.1.1. Note that

$a + \overline{a}$ and $a\overline{a}$ are always real numbers.

1.2. Degree of a polynomial. If the coefficient a_n in the polynomial

$$p \;\equiv\; a_n x^n + \cdots + a_1 x + a_0$$

is not 0 we say that the degree of p is n and write

$$\deg(p) = n.$$

This leaves out $p = \mathsf{const}_0$ which is usually not given a degree.

1.2.1. We immediately see that

$$\deg(pq) = \deg(p) + \deg(q).$$

1.3. Dividing polynomials. Consider polynomials p, q with degrees $n = \deg(p) \geq k = \deg(q)$,

$$\begin{aligned} p &\equiv a_n x^n + \cdots + a_1 x + a_0, \\ q &\equiv b_k x^k + \cdots + b_1 x + b_0. \end{aligned}$$

Subtracting $\frac{a_n}{b_k} x^{n-k} q(x)$ from $p(x)$ we obtain zero or a polynomial p_1 with $\deg(p_1) < n$, and

$$p(x) = c_1 x^{n_1} q(x) + p_1(x).$$

If $\deg(p_1) \geq \deg(q)$ we similarly obtain $p_1(x) = c_2 x^{n_2} q(x) + p_2(x)$ and repeating this procedure we finish with

$$p(x) = s(x)q(x) + r(x)$$

with $r = \mathsf{const}_0$ or $\deg(r) < \deg(q)$. One speaks of the r as of the *remainder* when dividing p by q.

1.3.1. An important observation. *If the coefficients of p and q are real then also the coefficients of s and r are real.*

2. Fundamental Theorem of Algebra. Roots and decomposition.

2.1. A *root* of a polynomial p is a number x such that $p(x) = 0$. A polynom with real coefficients does not have to have a real root (consider for example $p \equiv x^2 + 1$) but in the field of complex numbers we have

Theorem. (Fundamental Theorem of Algebra) *Each polynomial p of $\deg(p) > 0$ with complex coefficients has a complex root.*[3]

2.2. Decomposition of complex polynomials. Recall the obvious formula

$$x^k - \alpha^k = (x - a)(x^{k-1} + x^{k-2}\alpha + \cdots + x\alpha^{k-2}x + \alpha^{k-1})$$

[3] This theorem, which is, rather, a theorem of analysis or geometry, has several proofs based on different principles. One of them is in XXIII.3 below.

and denote the polynomial $x^{k-1}+x^{k-2}\alpha+\cdots+x\alpha^{k-2}x+\alpha^{k-1}$ (in x) of degree $k-1$ by $s_k(x,\alpha)$. If α_1 is a root of $p(x)=\sum_{k=0}^n a_kx^k$ of degree n we have

$$\begin{aligned}p(x)=p(x)-p(\alpha_1)&=\sum_{k=0}^n a_kx^k-\sum_{k=0}^n a_k\alpha_1^k=\\&=\sum_{k=0}^n a_k(x^k-\alpha_1^k)=(x-\alpha_1)\sum_{k=0}^n a_ks_k(x,\alpha_1)\end{aligned}$$

where the polynomial $p_1(x)=\sum_{k=0}^n a_ks_k(x,\alpha)$ has by 1.2.1 degree precisely $n-1$. Repeating the procedure we obtain

$$p_1(x)=(x-\alpha_2)p_2(x),\quad p_2(x)=(x-\alpha_3)p_3(x),\quad \text{etc.}$$

with $\deg(p_k)=n-k$, and ultimately

$$p(x)=a(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n) \qquad (*)$$

with $a\neq 0$.

2.3. Proposition. *A polynomial of degree n has at most n roots.*

Proof. Let x be a root of $p(x)=a(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)$. Then $(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)=0$ and hence some of the $x-\alpha_k$ has to be zero, that is, $x=\alpha_k$. □

2.3.1. The unicity of the coefficients. So far we have worked with a polynomial as with the expression $p(x)=a_nx^n+\cdots+a_1x+a_0$. Now we can prove that it is determined by the function p. We have

Proposition. *The coefficients a_k in the expression $p(x)=a_nx^n+\cdots+a_1x+a_0$ are uniquely determined by the function $(x\mapsto p(x))$. Consequently, this function also determines* $\deg(p)$.

Proof. Let $p(x)=a_nx^n+\cdots+a_1x+a_0=b_nx^n+\cdots+b_1x+b_0$ (any of a_k,b_k may be zero). Then $a_nx^n+\cdots+a_1x+a_0-b_nx^n-\cdots-b_1x-b_0=(a_n-b_n)x^n+\cdots+(a_1-b_1)x+(a_0-b_0)$ has infinitely many roots and hence cannot have a degree. Thus, $a_k=b_k$ for all k. □

2.3.2. Proposition. *The polynomials s,r obtained when dividing polynomial p by a polynomial q as in 1.3 are uniquely determined.*

Proof. Let $p(x)=s_1(x)q(x)+r_1(x)=s_2(x)q(x)+r_2(x)$. Then $q(x)(s_1(x)-s_2(x))+(r_1(x)-r_2(x))$ is a zero polynomial and since $\deg(q)>\deg(r_1-r_2)$

(if the last is determined at all), $s_1 = s_2$. Then $r_1 - r_2 \equiv 0$ and hence also $r_1 = r_2$. $\square$

2.4. Multiple roots. On the other hand, $p(x)$ does not have to have $\deg(p)$ many distinct roots: see for instance $p(x) = x^n$ with only one root, namely zero. The roots α_k in the decomposition $(*)$ can appear several times, and after suitable permutation of the factors, $(*)$ can be rewritten as

$$p(x) = a(x - \beta_1)^{k_1}(x - \beta_2)^{k_2} \cdots (x - \beta_r)^{k_r} \text{ with } \beta_k \text{ distinct.} \qquad (**)$$

The power k_j is called *multiplicity* of the root β_j and we have $\sum_{j=1}^{r} k_j = n$.

2.4.1. Proposition. *The multiplicity of a root is uniquely defined. Consequently, the decomposition* $(**)$ *is determined up to the permutation of the factors.*

Proof. Suppose we have $p(x) = (x-\beta)^k q(x) = (x-\beta)^\ell r(x)$ such that β is not a root of neither q nor r. Suppose $k < \ell$. Dividing $p(x)$ by $(x - \beta)^k$ we obtain (using the unicity of division, see 2.3.2 above) that $q(x) = (x = \beta)^{\ell-k} r(x)$ so that β is a root of p, a contradiction. $\square$

2.5. Note. The set of all complex polynomials forms an integral domain (similarly like the set of integers. Now $q|p$ (q divides p) if $p(x) = s(x)q(x)$ and both $q|p$ and $q|p$ iff there is a number $c \neq 0$ such that $p(x) = c \cdot q(x)$. The primes in this division are the (equivalence classes of) binoms $x - \alpha$. In the propositions above we have seen that in the integral domain of complex polynomials we have unique prime decomposition.

3. Decomposition of polynomials with real coefficients.

3.1. Proposition. *Let the coefficients a_n of a polynomial $p(x) = a_n x^n + \cdots + a_1 x + a_0$ be real. Let α be a root of p. Then the complex conjugate $\overline{\alpha}$ is also a root of p.*

Proof, We have (recall 1.1) $p(\overline{\alpha}) = a_n\overline{\alpha}^n + \cdots + a_1\overline{\alpha} + a_0 = \overline{a}_n\overline{\alpha}^n + \cdots + \overline{a}_1\overline{\alpha} + \overline{a}_0 = \overline{a_n\alpha^n} + \cdots + \overline{a_1\alpha} + \overline{a}_0 = \overline{a_n\alpha^n + \cdots + a_1\alpha + a_0} = \overline{0} = 0.$ $\square$

3.2. Proposition. *Let α be a root of multiplicity k of a polynomial p with real coefficients. Then the multiplicity of the root $\overline{\alpha}$ is also k.*

Proof. If α is real there is nothing to prove. Now let α not be real. Then we have

$$p(x) = (x-\alpha)(x-\overline{\alpha})q(x) = (x^2 - (\alpha+\overline{\alpha})x + \alpha\overline{\alpha})q(x)$$

and since $x^2 - (\alpha+\overline{\alpha})x + \alpha\overline{\alpha}$ has real coefficients (recall 1.1.1), q also has realc coefficients (recall 1.3.1). Now if α is a root of q again we have another root $\overline{\alpha}$ of q, and the statement follows inductively. □

3.3. The trinoms $x^2 + \beta x + \gamma = x^2 - (\alpha+\overline{\alpha})x + \alpha\overline{\alpha}$ have no real roots: they already have roots α and $\overline{\alpha}$, and cannot have more by 2.3. They are called *irreducible* trinoms,

3.4. From 2.4, 3.1 and 3.2 we now obtain

3.4.1. Corollary. *Let p be a polynomial of degree n with real coefficients. Then*

$$p(x) = a(x-\beta_1)^{k_1}(x-\beta_2)^{k_2}\cdots(x-\beta_r)^{k_r}(x^2+\gamma_1 x+\delta_1)^{\ell_1}\cdots(x^2+\gamma_s x+\delta_s)^{\ell_s}$$

with $\beta_j, \gamma_j, \delta_j$ real, $x^2+\gamma_j x+\delta_j$ irreducinle and $\sum_{j=1}^r k_j + 2\sum_{j=1}^s \ell_j = n$ (s can be equal to 0).

3.4.1. Note. Thus, in the integrity domain of real polynomials we have a greater variety of primes. Besides the $x-\beta$ we also have the ireeducible $x^2+\gamma x+\delta$.

4. Sum decomposition of rational functions.

4.1. We have already used the term *integral domain* in Notes 2.5 and 3.4.1. To be more specific, an integral domain is a commutative ring J with unit 1 and such that for $a, b \in J$, $a, b \neq 0$ implies $ab \neq 0$

As in the domain $\mathbb{Z}$ of integers, in a general integral domain (and in particular in the domain of polynomials with coefficients in $\mathbb{C}$ resp. $\mathbb{R}$) we say that a divides b and write $a|b$ if there is an x such that $b = xa$. a and b are equivalent if $a|b$ and $b|a$; we write $a \sim b$.

The *greatest common divisor* a, b is a d such that $d|a$ and $d|b$ and such that whenever $x|a$ and $x|b$ then $x|d$. The unit divides every a; elements a and b are *coprime* (or *relatively prime*) if they have (up to equivalence) no other common divisor.

4.2. Theorem. *Let J be an integrity domain and let us have a function $\nu : J \to \mathbb{N}$ and a rule of division with remainder for $a, b \neq 0$ and b not dividing a,*

$$a = sb + r \quad \textit{with} \quad \nu(r) > \nu(b).$$

Then for any $a, b \neq 0$ there exist x, y such that $xa+yb$ is the greatest common divisor of a, b.

Proof. Let $d = xa + yb$ with the least possible $\nu(d)$. Suppose d does not divide a. Then

$$a = sd + r \quad \text{with} \quad \nu(r) < \nu(d).$$

But then $(1 - sx)a - syb = r$ and $\nu((1 - sx)a - syb) = \nu(r) < \nu(d)$, a contradiction. Thus, $d|a$ and for the same reason $d|b$. On the other hand, if $c|a$ and $c|b$ then obviously $c|(xa + yb)$. Thus. d is the greatest common divisor. □

4.2.1. Note. For the integrity domain of integers (with $\nu(n) = |n|$) this was proved by Bachet (16.-17. century), in the more general form – in particular for our polynomials – this is by Bézout (18.century). One usually speaks of Bézout lemma; Bachet-Bézout Theorem should be appropriate.

4.3. A *rational function (in one variable)* is a complex or real function of one (complex resp. real variable) that can be written as

$$P(x) = \frac{p(x)}{q(x)}$$

where p, q are polynomials.

medskip

4.3.1. Theorem. *A complex rational function $P(x) = \frac{p(x)}{q(x)}$ can be written as*

$$P_1(x) + \sum_j V_j(x)$$

where each of the expression is of the form

$$\frac{A}{(x - \alpha)^k}$$

where A is a number and α is a root of the polynomial q with multiplicity at least k.

Proof by induction on $\deg(q)$. The statement is trivial for $\deg(q) = 0$. For $\deg(q) = 1$ (and hence $q(x) = C(x - \alpha)$) we obtain from 1.3 that

$$p(x) = s(x)q(x) + B$$

and

$$\frac{p(x)}{q(x)} = s(x) + \frac{B'}{x - \alpha} \quad \text{where} \quad B' = \frac{B}{C}.$$

Now let the theorem hold for $\deg(q) < n$. It suffices to prove it for $\frac{p(x)}{(x-\alpha)q(x)}$ with $\deg q < n$. This can be written, by the induction hypothesis as

$$\frac{P_1(x)}{x - \alpha} + \sum_j \frac{V_j(x)}{x - \alpha}.$$

If $V_j = \frac{A}{(x-\alpha)^k}$ the corresponding summand will be $\frac{A}{(x-\alpha)^{k+1}}$. If it is $\frac{A}{(x-\beta)^k}$ with $\beta \neq \alpha$ we realize first that the greatest common divisor of $(x - \alpha)$ and $(x - \beta)^k$ is 1 and hence by 4.2 we have polynomials u, v such that

$$u(x)(x - \alpha) + v(x)(x - \beta)^k = 1$$

so that

$$\frac{A}{(x - \alpha)(x - \beta)^k} = \frac{A(u(x)(x - \alpha) + v(x)(x - \beta)^k)}{(x - \alpha)(x - \beta)^k} = \frac{Au(x))}{(x - \beta)^k} + \frac{Av(x)}{(x - \alpha)}$$

and by induction hypothesis both of the last summands can be written as desired. □

4.3.2. Theorem. *A real rational function $P(x) = \frac{p(x)}{q(x)}$ can be written as*

$$P_1(x) + \sum_j V_j(x)$$

where each of the expression is of the form

$$\frac{A}{(x - \alpha)^k}$$

where A is a number and α is a root of the polynomial q with multiplicity at least k or of the form

$$\frac{Ax + B}{(x^2 + ax + b)^k}$$

where $x^2 + ax + b$ is some of the ireeducible trinoms from 3.4.1 and k is less or equal to the corresponding ℓ.

Proof can be done following the lines of the proof of 4.3.1, only distinguishing more cases of the relative primeness of the $x - \alpha$ and $x^2 + ax + b$.

With careful checking it can be also deduced from 4.3.1: namely, if a root α is not real we have to have with each

$$\frac{A}{(x-\alpha)^k}$$

a summand

$$\frac{B}{(x-\overline{\alpha})^k}$$

with the same power k: else, the sum would not be real. Now we have

$$\frac{A}{(x-\alpha)^k} + \frac{B}{(x-\overline{\alpha})^k} = \frac{A(x-\overline{\alpha}) + B(x-\alpha)}{(x^2 - (\alpha+\overline{\alpha})x + \alpha\overline{\alpha})^k} = \frac{A_1 x + B_1}{(x^2 + ax + b)^k}$$

and again we have to check that the A_1, B_1 have to be real.

In fact, the variation of the proof of 4.3.1 may be less laborious then the latter, but in the latter we perhaps (even if we do not do the details) see better what is happening. □

4.3.3. Note. In practical computing one simply takes into account that the expression as in 4.3.1 or 4.3.2 is possible and obtains the coefficients A resp. A and B as solutions of a system of linear equations.

X. Primitive function (indefinite integral).

1. Reversing differentiation

1.1. In Chapter VI we defined a derivative of a function and learned how to compute the derivatives of elementary functions.

Now we will reverse the task. Given a function f we will be interested in a function F such that $F' = f$. Such a function F will be called the *primitive function*, or *indefinite integral* of f (in the next chapter we will discuss a basic definite one, the Riemann integral).

In differentiation we had, first, a derivative of a function at a point, which was a number, and then we defined a derivative of a function f as a function $f' : D \to \mathbb{R}$, provided f had a derivative $f'(x)$ in every point x of a domain D. In taking the primitive function we have nothing like the former. It will be always a search of a function (the F above) associated with a given one.

1.2. Unlike a derivative f' that is uniquely determined by the function f, the primitive function is not, for obvious reasons: the derivative of a constant C is zero so that if $F(x)$ is a primitive function of $f(x)$ then so is any $F(x) + C$. But the situation is not much worse than that, as we have already proved in VIII.3.3. We have

1.2.1. Fact. *If F and G are primitive functions of F on an interval J then there is a constant C such that*

$$F(x) = G(x) + C$$

for all $x \in J$.

1.3. Notation. Primitive function of a function f is often denoted by

$$\int f$$

Instead of this concise symbol we equally often use a more explicit

$$\int f(x)\mathrm{d}x.$$

This latter is not just an elaborate indication what the variable in question is (as in $\int f(x, y)\mathrm{d}x$). In Section 4 it will be of a great advantage in computing an integral by means of the substitution method. But its natural meaning will be even more apparent in connection with the definite integral in the next chapter. See XI.2.5, XI.2.6 and XI.5.5.1.

Since a primitive function is not uniquely determined, the expression "$F = \int f$ " should be understood as "F is a primitive function of f ", not as an equality of two entities (we have $\frac{1}{2}x^2 = \int x\mathrm{d}x$ and $\frac{1}{2}x^2 + 5 = \int x\mathrm{d}x$ but we cannot conclude from these two "equalities" that $\frac{1}{2}x^2 = \frac{1}{2}x^2 + 5$). To be safer one usually writes

$$\int f(x)\mathrm{d}x = F(x) + C \quad \text{or} \quad \int f = F(x) + C,$$

but even this can be misleading: the statement 1.2.1 holds for an interval only and the domains of very natural functions are not always intervals intervals; see 2.2.2.2 below. One has to be careful.

2. A few simple formulas.

2.1. Reversing the basic rule of differentiation we immediately obtain

Proposition. *Let f, g be functions with the same domain D and let a, b be numbers. Let $\int f$ and $\int g$ exist on D. Then $\int(af + bg)$ exists and we have*

$$\int(af + bg) = a\int f + b\int g.$$

2.1.1. Note. This is the only arithmetic rule for integration. For principial reasons there cannot be a formula for $\int f(x)g(x)\mathrm{d}x$ or for $\int \frac{f(x)}{g(x)}\mathrm{d}x$, see 2.2.2.1 and 2.3.1.

2.2. Reversing the rule for differentiating x^n with $n \neq -1$ we obtain

$$\int x^n\mathrm{d}x = \frac{1}{n+1}x^{n+1}.$$

(In fact, this does not hold for integers n only. If D is $\{x \in \mathbb{R} \,|\, x > 0\}$ then we have by VI.3.3 the formula

$$\int x^a\mathrm{d}x = \frac{1}{a+1}x^{a+1} \quad \text{for any real } a \neq -1.)$$

Hence, using 2.1 we have for a polynomial $p(x) = \sum_{k=0}^{n} a_k x^k$,

$$\int p(x)\mathrm{d}x = \sum_{k=0}^{n} \frac{a_k}{k+1} x^{k+1}.$$

2.2.1. For $n = -1$ (and domain $\mathbb{R} \smallsetminus \{0\}$) we have the formula

$$\int \frac{1}{x}\mathrm{d}x = \lg|x|.$$

(Indeed, for $x > 0$ we have $|x| = x$ and hence $(\lg|x|)' = \frac{1}{x}$. For $x < 0$ we have $|x| = -x$ and hence $(\lg|x|)' = (\lg(-x))' = \frac{1}{-x} \cdot (-1) = \frac{1}{x}$ again.)

2.2.2. Notes. 1. This last formula indicates that there can hardly be a simple rule for integration $\frac{f(x)}{g(x)}$ in terms on $\int f$ and $\int g$: this would mean an arithmetic formula producing $\lg x$ from $x = \int 1$ and $\frac{1}{2}x^2 = \int x$.

2. The domain of function $\frac{1}{x}$ is not an interval. Note that we have, a.o.,

$$\int \frac{1}{x}\mathrm{d}x = \begin{cases} \lg|x| + 2 \text{ for } x < 0, \\ \lg|x| + 5 \text{ for } x > 0. \end{cases}$$

which shows that using the expression $\int f(x)\mathrm{d}x = F(x) + C$ is not without danger.

2.3. For goniometric function we immediately obtain

$$\int \sin x = -\cos x \quad \text{and} \quad \int \cos x = \sin x.$$

2.3.1. Note. In general, a primitive function of an elementary function (although it always exists as we will see in the next chapter) may not be elementary. One such is

$$\int \frac{\sin x}{x}$$

(proving this is far beyond our means, you have to believe it). Now we have an easy $\int \frac{1}{x}$ and $\int \sin x$; thus there cannot be a rule for computing $\int f(x)g(x)\mathrm{d}x$ in terms of $\int f$ and $\int g$.

2.4. For the exponential we have, trivially,

$$\int e^x \mathrm{d}x = e^x \quad \text{and by VI.3.3 more generally} \quad \int a^x \mathrm{d}x = \frac{1}{\lg a} a^x.$$

2.5. Let us add two more obvious formulas:

$$\int \frac{\mathrm{d}x}{1+x^2} = \arctan x \quad \text{and} \quad \int \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \arcsin x.$$

In the following two sections we will learn two useful methods for finding primitive functions in more involved cases.

3. Integration per partes.

3.1. Let f, g have derivatives. From the rule of differentiating products we immediately obtain

$$\int f' \cdot g = f \cdot g - \int f \cdot g'. \tag{$*$}$$

At the first sight we have not achieved much: we wish to integrate the product $f' \cdot g$ and we are left with integrating a similar one, $f \cdot g'$. But

(1) $\int f \cdot g'$ can be much simpler than $\int f' \cdot g$, or

(2) the formula can result in an equation from which the desired integral can be easily computed, or

(3) the formula may yield a recursive one that leads to our goal.

Using the formula $(*)$ is called *integration per partes.*

3.2. Example: Illustration of 3.1.(1). Let us compute

$$J = \int x^a \lg x \quad \text{with } x > 0 \text{ and } a \neq -1.$$

If we set $f(x) = \frac{1}{a+1}x^{a+1}$ and $g(x) = \lg x$ we obtain $f'(x) = x^a$ and $g'(x) = \frac{1}{x}$ so that

$$J = \frac{1}{a+1}x^{a+1}\lg x - \frac{1}{a+1}\int x^{a+1}\cdot\frac{1}{x} = \frac{1}{a+1}(x^{a+1}\lg x - \int x^a) =$$
$$= \frac{1}{a+1}(a^{a+1}\lg x - \frac{1}{a+1}x^{a+1}) = \frac{x^{a+1}}{a+1}(\lg x - \frac{1}{a+1})$$

and hence for instance for $a = 1$ we obtain

$$\int \lg x \mathrm{d}x = x(\lg x - 1).$$

3.3. Example: Illustration of 3.1.(2). Let us compute

$$J = \int e^x \sin x \mathrm{d}x.$$

Setting $f(x) - f'(x) = e^x$ and $g(x) = \sin x$ we obtain

$$J = e^x \sin x - \int e^x \cos x \mathrm{d}x.$$

Now the new integral on the left hand side is about as complex as the one we have started with. But let us repeat the procedure, this time with $g(x) = \cos x$. We obtain

$$\int e^x \cos x \mathrm{d}x = e^x \cos x - \int e^x(-\sin x)\mathrm{d}x$$

and hence

$$J = e^x \sin x - (e^x \cos x - \int e^x(-\sin x)\mathrm{d}x) = e^x \sin x - e^x \cos x - J$$

and conclude that

$$J = \frac{e^x}{2}(\sin x - \cos x).$$

3.4. Example: Illustration of 3.1.(3). Let us compute

$$J_n = \int x^n e^x \mathrm{d}x \quad \text{for integers } n \geq 0.$$

Setting $f(x) = x^n$ and $g(x) = g'(x) = e^x$ we obtain

$$J_n = x^n e^x - \int n x^{n-1} e^x = x^n e^x - nJ_{n-1}.$$

Iterating the procedure we get

$$\begin{aligned} J_n &= x^n e^x - n x^{n-1} e^x + n(n-1)J_{n-2} = \cdots = \\ &= x^n e^x - n x^{n-1} + n(n-1)x^{n-2}e^x + \cdots \pm n! J_0 \end{aligned}$$

and since $J_0 = \int e^x = e^x$ this makes

$$J_n = e^x \cdot \sum_{k=0}^{n} \frac{n!}{(n-k)!}(-1)^k \cdot x^{n-k}.$$

4. Substitution method.

4.1. The rule of differentiating composed function VI.2.2 can be for our purposes reinterpreted as follows.

Fact. *Let $\int f = F$, let a function ϕ have derivative ϕ', and let the composition $F \circ \phi$ make sense. Then*

$$\int f(\phi(x)) \cdot \phi'(x)\,dx = F(\phi(x)).$$

4.1.1. Thus, to obtain $\int f(\phi(x)) \cdot \phi'(x)\mathrm{d}x$ we compute $\int f(y)\mathrm{d}y$ and in the result substitute $\phi(x)$ for all the occurences of y. Using this trick is called the *substitution method.*

Here the notation

$$\int f(x)\mathrm{d}x$$

instead of the plain $\int f$ is of a great help. Recall the notation

$$\frac{\mathrm{d}\phi(x)}{\mathrm{d}x} \quad \text{for the derivative} \quad \phi'(x).$$

Now the expression $\frac{\mathrm{d}\phi(x)}{\mathrm{d}x}$ is not really a fraction with numerator $\mathrm{d}\phi(x)$ and denominator $\mathrm{d}x$, but let us pretend for a moment it is. Thus,

$$\mathrm{d}\phi(x) = \phi'(x)\mathrm{d}x \quad \text{or} \quad \text{“ } \mathrm{d}y = \phi'(x)\mathrm{d}x \text{ where } \phi(x) \text{ is substituted for } y \text{ ”.}$$

Hence, using the substitution method (substituting $\phi(x)$ for y) consists of computing

$$\int f(y)\mathrm{d}y$$

as an integral in variable y, and when substituting $\phi(x)$ for y writing

$$\mathrm{d}y = \phi'(x)\mathrm{d}x \ \text{ as obtained from } \ \frac{\mathrm{d}y}{\mathrm{d}x} = \phi'(x).$$

This is very easy to remember.

4.2. Example. To determine $\int \frac{\lg x}{x}\mathrm{d}x$ substitute $y = \lg x$. Then $\mathrm{d}y = \frac{\mathrm{d}x}{x}$ and we obtain

$$\int \frac{\lg x}{x}\mathrm{d}x = \int y\mathrm{d}y = \frac{1}{2}y^2 = \frac{1}{2}(\lg x)^2.$$

4.3. Example. To compute $\int \tan x\mathrm{d}x$ recall that $\tan x = \frac{\sin x}{\cos x}$ and that $(-\cos x)' = \sin x$. Hence, substituting $y = -\cos x$ we obtain

$$\int \tan x\mathrm{d}x = \int \frac{\sin x}{\cos x}\mathrm{d}x = \int \frac{\mathrm{d}y}{-y} = -\lg|y| = -\lg|\cos x|.$$

We will meet many more examples in the following two sections.

5. Integrals of rational functions.

5.1. In view of 2.1 and IX.4.3.2 it suffices to find the integrals

$$\int \frac{1}{(x-a)^k}\mathrm{d}x \tag{5.1.1}$$

and

$$\int \frac{Ax+B}{(x^2+ax+b)^k}\mathrm{d}x \quad \text{with} \ \ x^2+ax+b \ \ \text{irreducible} \tag{5.1.2}$$

for natural numbers k.

5.2. The first, (5.1.1) is very easy. If we substitute $y = x - a$ then $\mathrm{d}y = \mathrm{d}x$ and we compute our integral as $\int \frac{1}{y^k}$ and by 2.2 and 2.2.1 (substituting back $x - a$ for y)

$$\int \frac{1}{(x-a)^k}\mathrm{d}x = \begin{cases} \frac{1}{1-k} \cdot \frac{1}{(x-a)^{k-1}} & \text{for } k \neq 1, \\ \lg|x-a| & \text{for } k = 1. \end{cases}$$

5.3. Lemma. *Set*

$$J(a,b,x,k) = \int \frac{1}{(x^2+ax+b)^k} dx.$$

Then we have

$$\int \frac{Ax+B}{(x^2+ax+b)^k} dx = \begin{cases} \frac{A}{2(1-k)} \cdot \frac{1}{(x^2+ax+b)^{k-1}} + (B - \frac{Aa}{2})J(a,b,x,k) & \textit{for } k \neq 1, \\ \frac{A}{2} \lg|x^2+ax+b| + (B - \frac{Aa}{2})J(a,b,x,k) & \textit{for } k = 1. \end{cases}$$

Proof. We have

$$\frac{Ax+B}{x^2+ax+b} = \frac{A}{2}\frac{2x+a}{x^2+ax+b} + (B - \frac{Aa}{2})\frac{1}{x^2+ax+b}$$

Now in the first we can compute

$$\int \frac{2x+a}{x^2+ax+b}\mathrm{d}x$$

substituting $y = x^2 + ax + b$; then we have $\mathrm{d}y = (2x+a)\mathrm{d}x$ and the task, as in 5.2, reduces to determining $\int \frac{1}{y^k}\mathrm{d}y$. □

5.4. Hence, (5.1.2) will be solved by computing

$$\int \frac{1}{(x^2+ax+b)^k}\mathrm{d}x$$

with irreducible $x^2 + ax + b$.

5.4.1. First observe that because of the irreducibility we have $b-\frac{a^2}{4}>0$ (otherwise, x^2+ax+b would have real roots). Therefore we have a real c with

$$c^2 = b-\frac{a^2}{4}$$

and

$$x^2+ax+b = c^2\left(\left(\frac{x+\frac{1}{2}a}{c}\right)^2+1\right).$$

Thus, if we substitute $y=\frac{x+\frac{1}{2}a}{c}$ (hence, $\mathrm{d}y=\frac{1}{c}\mathrm{d}x$) in $\int\frac{1}{(x^2+ax+b)^k}\mathrm{d}x$ we obtain

$$\frac{1}{c^{2k-1}}\int\frac{1}{(y^2+1)^k}\mathrm{d}y$$

and we have further reduced our task to finding $\int\frac{1}{(x^2+1)^k}\mathrm{d}x$.

5.4.2. Proposition. *The integral*

$$J_k=\int\frac{1}{(x^2+1)^k}dx$$

can be computed recursively from the formula

$$J_{k+1}=\frac{1}{2k}\cdot\frac{x}{x^2+1}+\frac{2k-1}{2k}J_k \qquad (*)$$

with $J_1=\operatorname{arctg}x$.

Proof. First set

$$f(x)=\frac{1}{(x^2+1)^k}\quad\text{and}\quad g(x)=x.$$

Then

$$f'(x)=-k\frac{2x}{(x^2+1)^{k+1}}\quad\text{and}\quad g'(x)=1$$

and from the per partes formula we obtain

$$J_k=\frac{x}{(x^2+1)^k}+2k\int\frac{x^k}{(x^2+1)^{k+1}}=$$

$$=\frac{x}{(x^2+1)^k}+2k\left(\int\frac{x^k+1}{(x^2+1)^{k+1}}-\int\frac{1}{(x^2+1)^{k+1}}\right)=$$

$$=\frac{x}{(x^2+1)^k}+2kJ_k-2kJ_{k+1}$$

and the formula $(*)$ follows; the $J_1 = \arctan x$ was already mentioned in 2.5 □

6. A few standard substitutions.

6.1. First let us extend the terminology from Chapter IX. An expression

$$\sum_{r,s\leq n} a_{rs}x^r y^s$$

will be called a *polynomial in two variables* x, y. If $p(x,y)$, $q(x,y)$ are polynomials in two variables we speak of

$$R(x,y) = \frac{p(x,y)}{q(x,y)}$$

as of *rational function in two variables.*

6.1.1. Convention. In the rest of this section, $R(x,y)$ will always be a rational function in two variables.

6.1.2. Observation. *Let $P(x), Q(x)$ be rational function as in Chapter IX. Then $S(x) = R(P(x), Q(x))$ is a rational function.*

6.2. The integral $\int R\big(x, \sqrt{\frac{ax+b}{cx+d}}\big)\mathbf{d}x$. Substitute $y = \sqrt{\frac{ax+b}{cx+d}}$. Then $y^2 = \frac{ax+b}{cx+d}$ from which we obtain

$$x = \frac{b - dy^2}{ay^2 + a}$$

and hence

$$\frac{\mathrm{d}x}{\mathrm{d}y} = S(y)$$

where $S(y)$ is a rational function (the explicit formula can be easily obtained). Hence, the substitution transforms

$$\int R\left(x, \sqrt{\frac{ax+b}{cx+d}}\right)\mathrm{d}x \quad \text{to} \quad \int R\left(\frac{b-dy^2}{ay^2+a}, y\right)S(y)\mathrm{d}y$$

and this we can compute using the procedures from the previous section.

6.3. Euler substitution: the integral $\int R(x, \sqrt{ax^2+bx+c})\mathrm{d}x$. First let us dismiss the case of $a \le 0$. Since we assume that the function makes sense, we have to have $ax^2+bx+c \ge 0$ on its domain which implies (in case of $a \le 0$) real roots α, β and

$$R(x, \sqrt{ax^2+bx+c}) = R(x, \sqrt{-a}\sqrt{(x-\alpha)(x-\beta)}) =$$
$$= R\big(x, \sqrt{-a}(x-\alpha)\sqrt{\frac{x-\beta}{x-\alpha}}\big)$$

and this is a case already dealt with in 5.2.

But if $a > 0$ the situation is new. Then let us substitute the t from the equation

$$\sqrt{ax^2+bx+c} = \sqrt{a}x + t$$

(this is the *Euler substitution*). The squares of both sides yield

$$ax^2+bx+c = ax^2 + 2\sqrt{a}xt + t^2$$

and we obtain

$$x = \frac{t^2-c}{b-2t\sqrt{a}} \quad \text{and hence} \quad \frac{dx}{\mathrm{d}t} = S(t)$$

where $S(t)$ is a rational function. Thus we can compute our integral as

$$\int R\left(\frac{t^2-c}{b-2t\sqrt{a}}, \sqrt{a}\frac{t^2-c}{b-2t\sqrt{a}} + t\right) S(t)\mathrm{d}t.$$

6.4. Goniometric functions in a rational one: $\int R(\sin x, \cos x)\mathbf{d}x$. To compute

$$\int R(\sin x, \cos x)\mathrm{d}x$$

we will be helped by the substitution

$$y = \tan\frac{x}{2}.$$

Recall the standard formula

$$\cos^2 x = \frac{1}{1+\tan^2 x}$$

from which we obtain

$$\sin x = 2\sin\frac{x}{2}\cos\frac{x}{2} = 2\tan\frac{x}{2}\cos^2\frac{x}{2} = \frac{2\tan\frac{x}{2}}{1+\tan\frac{x}{2}^2} = \frac{2y}{1+y^2},$$

$$\cos x = \cos^2\frac{x}{2} - \sin^2\frac{x}{2} = 2\cos^2\frac{x}{2} - 1 = \frac{2}{1+y^2} - 1 = \frac{1-y^2}{1+y^2}.$$

Further we have

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{2}\cdot\frac{1}{\cos^2\frac{x}{2}} = \frac{1}{2}\cdot(1+\tan^2\frac{x}{2}) = \frac{1}{2}(1+y^2)$$

and hence

$$\mathrm{d}x - \frac{2}{1+y^2}\mathrm{d}y$$

so that we can solve our task by computing

$$\int R\left(\frac{2y}{1+y^2}, \frac{1-y^2}{1+y^2}\right)\frac{2}{1+y^2}\mathrm{d}y.$$

6.5. Note. The procedures in Section 4 and Section 5 are admittedly very laborious and time consuming. This is because they should cover fairly general cases. In a concrete case we sometimes can find a combination of the per partes and substitution methods leading to our goal in a much shorter procedure. Compare for instance $\int \tan x \mathrm{d}x$ as computed in 4.3 with 5.4.

XI. Riemann integral

1. The area of a planar figure.

1.1. Let us denote by $\mathsf{vol}(M)$ the area of a planar figure $M \subseteq \mathbb{R}^2$. A figure may be too exotic to be assigned an area, but we will not work with such here. Using the symbol vol includes the claim that the area in question makes sense.

The reader may wonder why we use the abreviation vol and not something like "ar". This is because later we will work in higher dimensions and referring to $M \subseteq \mathbb{R}^n$ with general n, "volume" is used rather than "area".

1.2. The following are rules we can certainly easily agree upon.

(1) $\mathsf{vol}(M) \geq 0$ whenever it makes sense,

(2) if $M \subseteq N$ then $\mathsf{vol}(M) \leq \mathsf{vol}(N)$,

(3) if M and N are disjoint then $\mathsf{vol}(M \cup N) = \mathsf{vol}(M) + \mathsf{vol}(N)$, and

(4) if M is a rectangle with sides a, b then $\mathsf{vol}(M) = a \cdot b$.

1.3. Observation. 1. $\mathsf{vol}(\emptyset) = 0$.

2. *Let M be a segment. Then* $\mathsf{vol}(M) = 0$.

Proof. 1: $\emptyset$ is a subset of any rectangle, hence the statement follows from (1),(2) and (4)

2 follows similarly: a segment of length a is a subset of a rectangle with sides a, b with arbitrarily small positive b. $\square$

1.3.1. Note. Thus we see that it was not necessary to specify whether we included in 1.2(4) the border segments, or just parts of them.

1.4. Proposition. *If the areas make sense we have*

$$\mathsf{vol}(M \cup N) = \mathsf{vol}(M) + \mathsf{vol}(N) - \mathsf{vol}(M \cap N).$$

In particular we have

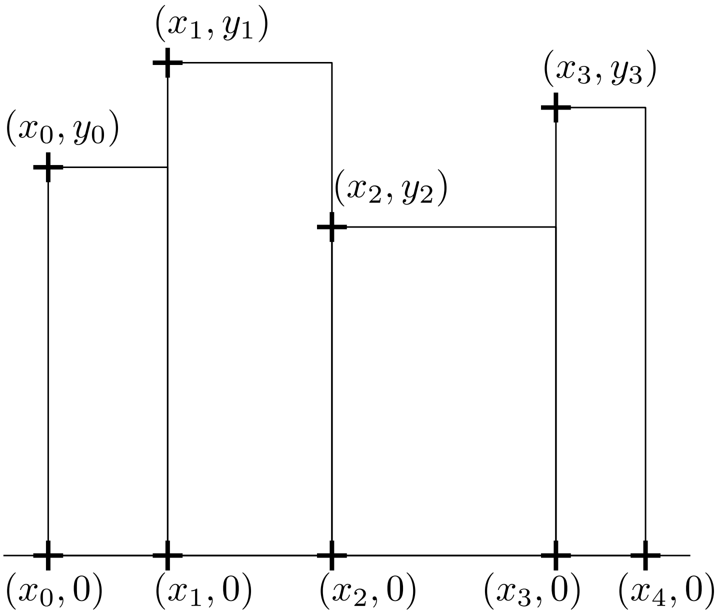
$$\mathsf{vol}(M \cup N) = \mathsf{vol}(M) + \mathsf{vol}(N) \quad \textit{whenever} \quad \mathsf{vol}(M \cap N) = 0.$$

Proof. Follows from 1.2(4) taking into account the disjoint unions

$$M \cup N = M \cup (N \smallsetminus M) \quad \text{and} \quad N = (N \smallsetminus M) \cup (N \cap M).$$

□

1.5. In the sequel the areas of figures of the following type



will play a fundamental role. By the previous trivial statements, their areas are simply the sums of the areas of the rectangles involved. In particular, the area of the figure in the picture is

$$y_0(x_1 - x_0) + y_1(x_2 - x_1) + y_2(x_3 - x_2) + y_3(x_4 - x_3).$$

2. Definition of the Riemann integral.

2.1. Convention. In this chapter we will be interested in *bounded* real functions $f : J \to \mathbb{R}$ defined on compact intervals J, that is, functions such that there are constants m, M such that for all $x \in J$, $m \leq f(x) \leq M$. Recall that (because of the compactness) a continuous function on J is always bounded. But our functions will not be always necessarily continuous.

2.2. A *partition* of a compact interval $\langle a, b\rangle$ is a sequence

$$P: \ a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b.$$

Another partition

$$P': \ a = t'_0 < t'_1 < \cdots < t'_{n-1} < t_m = b$$

is said to *refine* P (or to *be a refinement* of P) if the set $\{t_j \,|\, j = 1, \dots, n-1\}$ is contained in $\{t'_j \,|\, j = 1, \dots, m-1\}$.

The *mesh* of P, denoted $\mu(P)$, is defined as the maximum of the differences $t_j - t_{j-1}$.

2.3. For a bounded function $f : J = \langle a, b\rangle \to \mathbb{R}$ and a partition $P : a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$ define the *lower* resp. *upper sum* of f in P by setting

$$s(f, P) = \sum_{j=1}^{n} m_j(t_j - t_{j-1}) \quad \text{resp.} \quad S(f, P) = \sum_{j=1}^{n} M_j(t_j - t_{j-1})$$

where $m_j = \inf\{f(x) \,|\, t_{j-1} \le x \le t_j\}$ and $M_j = \sup\{f(x) \,|\, t_{j-1} \le x \le t_j\}$.

2.3.1. Proposition. *Let P' refine P. Then*

$$s(f, P) \le s(f, P') \quad \textit{and} \quad S(f, P) \ge S(f, P')$$

Proof will be done for the upper sum. Let $t_{k-1} = t'_l < t'_{l+1} < \cdots < t'_{l+r} = t_k$. For $M'_{l+j} = \sup\{f(x) \,|\, t'_{l+j-1} \le x \le t'_{l+j}\}$ and $M_k = \sup\{f(x) \,|\, t_{k-1} \le x \le t_k\}$ we have $\sum_j M'_j(t'_{l+j} - t'_{l+j-1}) \le \sum_j M_k(t'_{l+j} - t'_{l+j-1}) = M_k(t_k - t_{k-1})$ and the statement follows. □

2.3.2. Proposition. *For any two partitions P_1, P_2 we have*

$$s(f, P_1) \le S(f, P_2).$$

Proof. Obviously, $s(f, P) \le S(f, P)$ for any partition. Further, for any two partitions P_1, P_2 there is a common refinement P: it suffices to take the union of the dividing points of the two partitions. Thus, by 2.3.1,

$$s(f, P_1) \le s(f, P) \le S(f, P) \le S(f, P_2).$$

□

2.4. By 2.3.2 we have the set of real numbers $\{s(f,P)\,|\,P \text{ a partition}\}$ bounded from above and $\{S(f,P)\,|\,P \text{ a partition}\}$ bounded from below. Hence there are finite

$$\underline{\int}_a^b f(x)\mathrm{d}x = \sup\{s(f,P)\,|\,P \text{ a partition}\} \quad \text{and}$$
$$\overline{\int}_a^b f(x)\mathrm{d}x = \inf\{S(f,P)\,|\,P \text{ a partition}\}.$$

The first is called the *lower Riemann integral* of f over $\langle a,b\rangle$, the second is the *upper Riemann integral* of f.

From 2.3.2 again we see that

$$\underline{\int}_a^b f(x)\mathrm{d}x \le \overline{\int}_a^b f(x)\mathrm{d}x;$$

If $\underline{\int}_a^b f(x)\mathrm{d}x = \overline{\int}_a^b f(x)\mathrm{d}x$ the common value is denoted by

$$\int_a^b f(x)\mathrm{d}x$$

and called the *Riemann integral* of f over $\langle a,b\rangle$.

2.4.1. Observation. *Set $m = \inf\{f(x)\,|\,a \le x \le b\}$ and $M = \sup\{f(x)\,|\,a \le x \le b\}$. We have*

$$m(b-a) \le \underline{\int}_a^b f(x)dx \le \overline{\int}_a^b f(x)dx \le M(b-a).$$

2.4.2. Proposition. *The Riemann integral $\int_a^b f(x)dx$ exists if and only if for every $\varepsilon > 0$ there is a partition P such that*

$$S(f,P) - s(f,P) < \varepsilon.$$

Proof. I. Let $\int_a^b f(x)\mathrm{d}x$ exist and let $\varepsilon > 0$. Then there are partitions P_1 and P_2 such that

$$S(f,P_1) < \int_a^b f(x)\mathrm{d}x + \frac{\varepsilon}{2} \quad \text{and} \quad s(f,P_2) > \int_a^b f(x)\mathrm{d}x + \frac{\varepsilon}{2}.$$

Then we have, by 2.3.1, for the common refinement P of P_1, P_2,

$$S(f,P) - s(f,P) < \int_a^b f(x)\mathrm{d}x + \frac{\varepsilon}{2} - \int_a^b f(x)\mathrm{d}x + \frac{\varepsilon}{2} = \varepsilon.$$

II. Let the statement hold. Choose an $\varepsilon > 0$ such that $S(f,P) - s(f,P) < \varepsilon$. Then

$$\overline{\int}_a^b f(x)\mathrm{d}x \le S(f,P) < s(f,P) + \varepsilon \le \underline{\int}_a^b f(x)\mathrm{d}x + \varepsilon,$$

and since ε was arbitrary we conclude that $\overline{\int}_a^b f(x)\mathrm{d}x = \underline{\int}_a^b f(x)\mathrm{d}x$. $\square$

2.5. Notes. 1. We will see best what is happening if we analyse the case of a non-negative function f. Consider $F = \{(x,y) \,|\, x \in \langle a,b\rangle,\ 0 \le f(x)\}$. that is, the figure bordered by the x-axis, the graph of f and the vertical lines passing through $(a,0)$ and $(b,0)$. Take the largest union $F_l(P)$ of rectangles with the lower horizontal sides $\langle t_{j-1}, t_j\rangle$ (recall the picture in 1.5) that is contained in F; obviously $\mathsf{vol}(F_l(P)) = s(f,P)$. The similar smallest union of rectangles $F_u(P)$ that contains F has $\mathsf{vol}(F_u(P)) = S(f,P)$. Thus, if the area of F makes sense we have to have

$$s(f,P) = \mathsf{vol}(F_l(P)) \le \mathsf{vol}(F) \le \mathsf{vol}(F_u(P)) = S(f,P),$$

and if $\int_a^b f(x)\mathrm{d}x$ exists then this number is the only candidate for $\mathsf{vol}(F)$ and it is only natural to take it for the definition of the area.

2. The notation $\int_a^b f(x)\mathrm{d}x$ comes from not quite correct but useful intuition. Think of $\mathrm{d}x$ as of a very small interval (one would like to say “infinitely small, but with non-zero length”, which is not quite such a nonsense as it sounds); anyway, the $\mathrm{d}x$ are disjoint and cover the segment $\langle a,b\rangle$, and $\int$ stands for “sum” of the areas of the “very thin rectangles” with the horizontal side $\mathrm{d}x$ and height $f(x)$. Note how close this intuition is to the more correct view from 1 if we take P with a very small mesh.

2.6. Notation. If there is no danger of confusion we abbreviate (in analogy with the notation in Chapter X) the expressions

$$\underline{\int}_a^b f(x)\mathrm{d}x,\ \overline{\int}_a^b f(x)\mathrm{d}x.\ \int_a^b f(x)\mathrm{d}x \quad \text{to} \quad \underline{\int}_a^b f,\ \overline{\int}_a^b f.\ \int_a^b f.$$

3. Continuous functions.

3.1. Uniformm continuity. A real function $f : D \to \mathbb{R}$ is said to be *uniformly continuous* if

$$\forall \varepsilon > 0\ \exists \delta > 0 \ \text{ such that } \ \forall x, y \in D, \ \ |x - y| < \delta \ \Rightarrow \ |f(x) - f(y)| < \varepsilon.$$

3.1.1. Remark. Note the subtle difference between continuity and uniform continuity. In the former the δ depends not only on the ε but also on the x, while in the latter it does not. A uniformly continuous function is obviously continuous, but the reverse implication does not hold even in very simple cases. Take for instance

$$f(x) = (x \mapsto x^2) : \mathbb{R} \to \mathbb{R}.$$

we have $|x^2 - y^2| = |x - y| \cdot |x + y|$; thus, if we wish to have $|x^2 - y^2| < \varepsilon$ in the neighbourhood of $x = 1$ it suffices to take δ close to ε itself, in the neighbourhood of $x = 100$ one needs something like $\delta = \frac{\varepsilon}{100}$.

3.1.2. Perhaps somewhat surprisingly, in a compact domain these concepts coincide. We have

Theorem. *A function $f : \langle a, b\rangle \to \mathbb{R}$ is continuous if and only if it is uniformly continuous.*

Proof. Let f not be uniformly continuous. We will prove it is not continuous either.

Since the formula for uniform continuity does not hold we have an $\varepsilon_0 > 0$ such that for every $\delta > 0$ there are $x(\delta), y(\delta)$ such that $|x(\delta) - y(\delta)| < \delta$ while $|f(x(\delta)) - f(y(\delta))| \geq \varepsilon_0$. Set $x_n = x(\frac{1}{n})$ and $y_n = y(\frac{1}{n})$. By IV.1.3.1 we can choose convergent subsequences $(\widetilde{x}_n)_n$, $(\widetilde{y}_n)_n$ (first choose a convergent subsequence $(x_{k_n})_n$ of $(x_n)_n$ then a convergent subsequence $(y_{k_{l_n}})_n$ of $(y_{n_k})_k$ and finally set $\widetilde{x}_n = x_{k_{l_n}}$ and $\widetilde{y}_n = y_{k_{l_n}}$). Then $|\widetilde{x}_n - \widetilde{y}_n| < \frac{1}{n}$ and hence $\lim \widetilde{x}_n = \lim \widetilde{y}_n$ Because of $|f(\widetilde{x}_n) - f(\widetilde{y}_n)| \geq \varepsilon_0$, however, we cannot have $\lim f(\widetilde{x}_n) = \lim f(\widetilde{y}_n)$ so that by IV.5.1 f is not continuous. □

3.2. Theorem. *For every continuous function $f : \langle a, b\rangle \to \mathbb{R}$ the Riemann integral $\int_a^b f$ exists.*

Proof. Since f is by 3.1.2 uniformly continuous we can choose, for $\varepsilon > 0$ a $\delta > 0$ such that

$$|x - y| < \delta \quad \Rightarrow \quad |f(x) - f(y)| < \frac{\varepsilon}{b - a}.$$

Recall the mesh $\mu(P) = \max_j(t_j - t_{j-1})$ of $P : t_0 < t_1 < \cdots < t_k$. If $\mu(P) < \delta$ we have $t_j - t_{j-1} < \delta$ for all j, and hence

$$\begin{aligned} M_j - m_j = \sup\{f(x) \,|\, t_{j-1} \le x \le t_j\} - \inf\{f(x) \,|\, t_{j-1} \le x \le t_j\} \le \\ \le \sup\{|f(x) - f(y)| \,|\, t_{j-1} \le x, y \le t_j\} \le \frac{\varepsilon}{b-a} \end{aligned}$$

so that

$$\begin{aligned} S(f,P) - s(f,P) = \sum (M_j - m_j)(t_j - t_{j-1}) \le \\ \le \frac{\varepsilon}{b-a} \sum (t_j - t_{j-1}) = \frac{\varepsilon}{b-a}(b-a) = \varepsilon. \end{aligned}$$

Now use 2.4.2 $\square$

3.2.1. Scrutinizing the proof above we obtain a somewhat stronger

Theorem. *Let $f : \langle a, b\rangle \to \mathbb{R}$ be a continuous function and let $P_1, P_2, \dots$ be a sequence of partitions such that $\lim_n \mu(P_n) = 0$. Then*

$$\lim_n s(f, P_n) = \lim_n S(f, P_n) = \int_a^b f.$$

(Indeed, with ε and δ as above choose an n_0 such that for $n \ge n_0$ we have $\mu(P_n) < \delta$.)

3.3. Theorem. (The Integral Mean Value Theorem) *Let $f : \langle a, b\rangle \to \mathbb{R}$ be continuous. Then there exists a $c \in \langle a, b\rangle$ such that*

$$\int_a^b f(x)dx = f(c)(b-a).$$

Proof. Set $m = \min\{f(x) \,|\, a \le x \le b\}$ and $M = \max\{f(x) \,|\, a \le x \le b\}$ (recall IV.5.2). Then

$$m(b-a) \le \int_a^b f(x)\mathrm{d}x \le M(b-a).$$

Hence there is a K with $m \le K \le M$ such that $\int_a^b f(x)\mathrm{d}x = K(b-a)$. By IV.3.2 there exists a $c \in \langle a.b\rangle$ such that $K = f(c)$. $\square$

4. Fundamental Theorem of Calculus.

4.1. Proposition. *Let $a < b < c$ and let f be bounded on $\langle a, c\rangle$. Then*

$$\underline{\int}_a^b f + \underline{\int}_b^c f = \underline{\int}_a^c f \quad \textit{and} \quad \overline{\int}_a^b f + \overline{\int}_b^c f = \overline{\int}_a^c f.$$

Proof for the lower integral. Denote by $\mathcal{P}(u, v)$ the set of all partitions of $\langle u, v\rangle$. For $P_1 \in \mathcal{P}(a, b)$ and $P_2 \in \mathcal{P}(b, c)$ define $P_1 + P_2 \in \mathcal{P}(a, c)$ as the union of the two sequences. Then obviously

$$s(f, P_1 + P_2) = s(f, P_1) + s(f, P_2)$$

and hence

$$\begin{aligned}\underline{\int}_a^b f + \underline{\int}_b^c f &= \sup_{P_1 \in \mathcal{P}(a,b)} s(f, P_1) + \sup_{P_2 \in \mathcal{P}(b,c)} s(f, P_2) = \\ &= \sup\{s(f, P_1) + s(f, P_2) \mid P_1 \in \mathcal{P}(a, b), P_2 \in \mathcal{P}(b, c)\} = \\ &= \sup\{s(f, P_1 + P_2) \mid P_1 \in \mathcal{P}(a, b), P_2 \in \mathcal{P}(b, c)\}.\end{aligned}$$

Now every $P \in \mathcal{P}(a, c)$ can be refined to a $P_1 + P_2$: it suffices to add b into the sequence. Thus, by 2.3.1 this last supremum is equal to

$$\sup\{s(f, P) \mid P \in \mathcal{P}(a, c)\} = \underline{\int}_a^c f.$$

□

4.2. Convention. For $a = b$ we set $\int_a^a f = 0$ and for $a > b$ we set $\int_a^b f = \int_b^a f$. Then by straightforward checking we obtain

4.2.1. Observation. *For any a, b, c,*

$$\int_a^b f + \int_b^c f = \int_a^c f.$$

4.3. Theorem. (Fundamental Theorem of Calculus) *Let $f : \langle a, b\rangle \to \mathbb{R}$ be continuous. For $x \in \langle a, b\rangle$ set*

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

Then $F'(x) = f(x)$ (to be precise, the derivative in a is from the right and the one in b is from the left).

Proof. By 4.2.1 and 3.3 we have for $h \neq 0$

$$\frac{1}{h}(F(x+h)-f(x)) = \frac{1}{h}(\int_a^{x+h} f - \int_a^{x} f) = \frac{1}{h}\int_x^{x+h} f = \frac{1}{h}f(x+\theta h)h = f(x+\theta h)$$

where $0 < \theta < 1$ and as f is continuous, $\lim_{h\to 0}\frac{1}{h}(F(x+h) - f(x)) = \lim_{h\to 0} f(x+\theta h) = f(x)$. □

4.3.1. Corollary. *Let $f : \langle a,b\rangle \to \mathbb{R}$ be continuous. Then it has a primitive function on (a,b) continuous on $\langle a,b\rangle$. If G is any primitive function of f on (a,b) continuous on $\langle a,b\rangle$ then*

$$\int_a^b f(t)\,dt = G(b) - G(a).$$

.

(By 4.3 we have $\int_a^b f(t)\mathrm{d}t = F(b) - F(a)$. Recall IX.1.2.)

4.3.2. Remark. Note the contrast between derivatives and primitive functions. Having a derivative is a very strong property of a continuous function, but differentiating of elementary functions – that is, the functions we typically encounter – is very easy. On the other hand, each continuous function has a primitive one, but it is hard to compute.

4.4. Recall the Integral Mean Value Theorem (3.3). The fundamental theorem of calculus puts it in a close connection with the Mean Value Theorem of differential calculus. Indeed if we denote by F the primitive function of f, the formula in 3.3 reads

$$F(b) - F(a) = F'(c)(b-a).$$

5. A few simple facts.

5.1. Proposition. *Let g and f differ in finitely many points. Then*

$$\underline{\int}_a^b f = \underline{\int}_a^b g \quad \textit{and} \quad \overline{\int}_a^b f = \overline{\int}_a^b g.$$

In particular, if $\int_a^b f$ exists then also $\int_a^b g$ exists and $\int_a^b f = \int_a^b g$.

Proof for the lower integral. Recall the mesh $\mu(P)$ from 2.2. If $|f(x)|$ and $|g(x)|$ are $\le A$ for all x and if f and g differ in n points then

$$|s(f,P) - s(g,P)| \le n \cdot A \cdot \mu(P),$$

and $\mu(P)$ can be arbitrarily small. $\square$

5.2. Proposition. *Let f have only finitely many points of discontinuity in $\langle a,b\rangle$, all of them of the first kind. Then the Riemann integral $\int_a^b f$ exists.*

Proof. Let the discontinuity points be $c_1 < c_2 < \cdots < c_n$. Then we have

$$\int_a^b f = \int_a^{c_1} f + \int_{c_1}^{c_2} f + \cdots + \int_{c_n}^b f.$$

$\square$

5.3. Proposition. *Let $\int_a^b f$ and $\int_a^b g$ exist and let α, β be real numbers. Then $\int_a^b(\alpha f + \beta g)$ exists and we have*

$$\int_a^b (\alpha f + \beta g) = \alpha \int_a^b f + \beta \int_a^b g.$$

Proof. I. First we easily see that $\int_a^b \alpha f = \alpha \int_a^b f$. Indeed, for $\alpha \ge 0$ we obviously have $s(\alpha f, P) = \alpha s(f,P)$ and $S(\alpha f, P) = \alpha S(f,P)$, and for $\alpha \le 0$ we have $s(\alpha f, P) = \alpha S(f,P)$ and $S(\alpha f, P) = \alpha s(f,P)$.

II. Thus, it suffices to prove the statement for the sum $f+g$. Set $m_i = \inf\{f(x)+g(x) \,|\, x \in \langle t_{i-1}, t_i\rangle\}$, $m'_i = \inf\{f(x) \,|\, x \in \langle t_{i-1}, t_i\rangle\}$ and $m''_i = \inf\{g(x) \,|\, x \in \langle t_{i-1}, t_i\rangle\}$. Obviously $m'_1 + m''_i \le m_i$ and consequently

$$s(f,P)+s(g,P) \le s(f+g,P), \quad \text{and similarly} \quad S(f+g,P) \le S(f,P)+S(g,P)$$

and we easily conclude that

$$\underline{\int}_a^b f + \underline{\int}_a^b g \le \underline{\int}_a^b (f+g) \quad \text{and} \quad \overline{\int}_a^b (f+g) \le \overline{\int}_a^b f + \overline{\int}_a^b g$$

and hence

$$\int_a^b f + \int_a^b g \le \underline{\int}_a^b (f+g) \le \overline{\int}_a^b \le \int_a^b f + \int_a^b g.$$

□

5.4. Per partes. Set

$$[h]_a^b = h(b) - h(a).$$

Then we trivially obtain from 4.3 and X.3.1

$$\int_a^b f \cdot g' = [f \cdot g]_a^b - \int_a^b f' \cdot g.$$

5.5. Theorem. (Substitution theorem for Riemann integral) *Let $f : \langle a, b\rangle \to \mathbb{R}$ be continuous and let $\phi : \langle a, b\rangle \to \mathbb{R}$ be a one-to-one map with derivative. Then*

$$\int_a^b f(\phi(x))\phi'(x)\,dx = \int_{\phi(a)}^{\phi(b)} f(x)\,dx.$$

Proof. Recall 4.4 including the definition of F. We immediately have

$$\int_{\phi(a)}^{\phi(b)} f(x)\mathrm{d}x = F(\phi(b)) - F(\phi(a)).$$

But from X.4.1 and 4.4 we also have

$$F(\phi(b)) - F(\phi(a)) = \int_a^b f(\phi(x))\phi'(x)\mathrm{d}x,$$

and the statement follows.

5.5.1. There is a strong geometric intuition behind the substitution formula.

Recall 2.5 and 2.6. Think of ϕ as of a deformation of the interval $\langle a, b\rangle$ to obtain $\langle \phi(a), \phi(b)\rangle$. The derivative $\phi'(x)$ is a measure of how a very small interval around x is stretched resp. compressed. Thus, if we compute the integral $\int_{\phi(a)}^{\phi(b)} f$ as an integral over the original $\langle a, b\rangle$ we have to adjust the "small element" of length $\mathrm{d}x$ by the stretch or compression and obtain a corrected "small element" of length $\phi'(x)\mathrm{d}x$.

XII. A few applications of Riemann integral.

In this short chapter we will present a few applications of Riemann integral. Some of them wil concern computing volumes and similar, but there will be also two theoretical ones.

1. The area of a planar figure again.

1.1. We motivated the definition of Riemann integral by the idea of the area of the planar figure

$$F = \{(x, y) \,|\, x \in \langle a, b\rangle, 0 \le y \le f(x)\}$$

where f was a non-negative continuous function. Given a partition $P : a = t_0 < t_1 \cdots < t_n = b$ of $\langle a, b\rangle$ this F was minorized by the union of rectangles

$$\bigcup_{j=1}^{n} \langle t_{j-1}, t_j\rangle \times \langle 0, m_j\rangle \quad \text{with} \quad m_j = \inf\{f(x) \,|\, t_{j-1} \le x \le t_j\},$$

with the area

$$s(f, D) = \sum_{j=1}^{n} m_j(t_j - t_{j-1}),$$

and majorized by the union of rectangles

$$\bigcup_{j=1}^{n} \langle t_{j-1}, t_j\rangle \times \langle 0, M_j\rangle \quad \text{with} \quad M_j = \sup\{f(x) \,|\, t_{j-1} \le x \le t_j\},$$

with the area

$$S(f, D) = \sum_{j=1}^{n} M_j(t_j - t_{j-1}).$$

Thus (recall XI.2.5), the only candidate for the area of F is

$$\mathsf{vol}(F) = \int_a^b f(x)\mathrm{d}x,$$

the common value of the supremum of the former and the infimum of the latter.

1.2. Thus for instance the area of the section of parabola

$$F = \{(x,y)\,|\, -1 \leq x \leq 1, 0 \leq y \leq 1 - x^2\}$$

is

$$\int_{-1}^{1} (1 - x^2)\mathrm{d}x = [x - \frac{1}{3}x^3]_{-1}^{1} = 1 - \frac{1}{3} + 1 - \frac{1}{3} = \frac{4}{3}.$$

1.3. Let us compute the area of the circle with radius r. A half of it is given by

$$J = \int_{-r}^{r} \sqrt{r^2 - x^2}\mathrm{d}x.$$

Substitute $x = r\sin y$. Then $\mathrm{d}x = r\cos y\mathrm{d}y$ and $\sqrt{r^2 - x^2} = r\cos y$ so that we have J transformed to

$$J = r^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2 y \,\mathrm{d}y.$$

Now $\cos^2 y = \frac{1}{2}(\cos 2y + 1)$, and we proceed

$$\frac{J}{r^2} = \frac{1}{2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos 2y \,\mathrm{d}y + \frac{1}{2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}y = \frac{1}{2}\left(\left[\frac{1}{2}\sin 2y\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + [y]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\right) = \frac{1}{2}(0 + \pi)$$

and hence the area in question is $2J = \pi r^2$.

2. Volume of a rotating body.

2.1. Consider again a non-negative continuous function f and the curve

$$C = \{(x, f(x), 0)\,|\, a \leq x \leq b\}$$

in the three-dimensional Euclidean space. Now rotate C around the x-axis $\{x, 0, 0)\,|\, x \in \mathbb{R}\}$ and consider the set F surrounded by the result.

It is easy to compute the volume of F. Instead of the union of rectangles $\bigcup_{j=1}^{n} \langle t_{j-1}, t_j\rangle \times \langle 0, m_j\rangle$ as in 1.1, we will now minorize the set F by the union of discs (cylinders)

$$\bigcup_{j=1}^{n} \langle t_{j-1}, t_j\rangle \times \{(y,z)\,|\, y^2 + z^2 \leq m_i^2\} \quad \text{with} \quad m_j = \inf\{f(x)\,|\, t_{j-1} \leq x \leq t_j\}$$

with the volume

$$\sum_{j=1}^{n} \pi m_j^2 (t_j - t_{j-1})$$

and similarly we obtain the upper estimate of the volume by

$$\sum_{j=1}^{n} \pi M_j^2 (t_j - t_{j-1}) \quad \text{with} \quad M_j = \sup\{f(x) \,|\, t_{j-1} \le x \le t_j\}.$$

Thus, we compute the volume of F as

$$\mathsf{vol}(F) = \pi \int_a^b f^2(x)\mathrm{d}x.$$

2.2. For istance we obtain the three-dimensional ball B_3 as bounded by the rotating curve $\{(x, \sqrt{r^2 - x^2}) \,|\, -r \le x \le r\}$ and hence obtain

$$\mathsf{vol}(B_3) = \pi \int_{-r}^{r} (r^2 - x^2)\mathrm{d}x = \pi \left[r^2 x - \frac{1}{3}x^3 \right]_{-r}^{r} = 2\pi \left(r^3 - \frac{1}{3}r^3 \right) = \frac{4}{3}\pi r^3.$$

3. Length of a planar curve and surface of a rotating body.

3.1. Let f be a continuous function on $\langle a, b \rangle$ (later, we will assume it to have a derivative) and the curve

$$C = \{(x, f(x)) \,|\, a \le x \le b\}.$$

Take a partition

$$P: \ a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$$

of the interval $\langle a, b \rangle$, and approximate C by the system of segments $S(P)$ connecting

$$(t_{j-1}, f(t_{j-1})) \quad \text{with} \quad (t_j, f(t_j)).$$

The length $L(P)$ of this approximation, the overall sum of the lengths of these segments, is

$$L(P) = \sum_{j=1}^{n} \sqrt{(t_j - t_{j-1})^2 + (f(t_j) - f(t_{j-1}))^2}.$$

Now suppose f has a derivative. Then we can use the Mean Value Theorem (VII.2.2) to obtain

$$L(P) = \sum_{j=1}^{n} \sqrt{(t_j - t_{j-1})^2 + f'(\theta_i)^2(t_j) - t_{j-1})^2} = \sum_{j=1}^{n} \sqrt{1 + f'(\theta_i)^2}(t_j - t_{j-1}).$$

Obviously if P_1 refines P we have from the triangle inequality

$$L(P_1) \geq L(P)$$

so that

$$L(C) = \sup\{L(P) \mid P \text{ partition of } \langle a, b \rangle\}$$

can be naturally viewed as the length of the curve C. By XI 3.2.1 the sums converge to

$$L(C) = \int_a^b \sqrt{1 + f'(x)^2}\mathrm{d}x.$$

3.2. Similarly, approximating the surface of a rotating body by the relevant parts of truncated cones with heights $(t_j - t_{j-1})$ and radii $f(t_i)$ and $f(t_{j-1})$ of the bases, we obtain the formula

$$2\pi \int_a^b f(x)\sqrt{1 + f'(x)^2}\mathrm{d}x.$$

4. Logarithm.

4.1. In V.1.1 we introduced logarithm axiomatically as a function L that

(1) increases in $\langle 0, +\infty)$,

(2) satisfies $L(xy) = L(x) + L(y)$,

(3) and such that $\lim_{x \to 0} \frac{L(x)}{x-1} = 1$.

The existence of such a function (which we had to believe in in V.1.1) will be now proven by a simple construction.

4.2. Set

$$L(x) = \int_1^x \frac{1}{t}\mathrm{d}t$$

If $x > 0$ this is correct: the function $\frac{1}{t}$ is well defined and continuous on the closed interval between 1 and x.

4.2.1. If $x < y$ then $L(y) - L(x) = \int_x^y \frac{1}{t}\mathrm{d}t$ is an integral of a positive function over $\langle x, y\rangle$ and hence a positive number. Hence $L(x)$ increases.

4.2.2. We have

$$L(xy) = \int_1^{xy} \frac{1}{t}\mathrm{d}t = \int_1^x \frac{1}{t}\mathrm{d}t + \int_x^{xy} \frac{1}{t}\mathrm{d}t. \qquad (*)$$

In the last summand substitute $z = \phi(t) = xt$ to obtain

$$\int_x^{xy} \frac{1}{z}\mathrm{d}z = \int_1^y \frac{1}{xt}\phi'(t)\mathrm{d}t = \int_1^y \frac{x}{xt}\mathrm{d}t = \int_1^y \frac{1}{t}\mathrm{d}t$$

so that $(*)$ yields

$$L(x, y) = \int_1^x \frac{1}{t}\mathrm{d}t + \int_1^y \frac{1}{t}\mathrm{d}t = L(x) + L(y).$$

4.2.3. Finally we have

$$\lim_{x\to 0} \frac{L(x)}{x-1} = \lim_{x\to 0} \frac{L(x) - L(1)}{x-1} = L'(1) = \frac{1}{1} = 1$$

by XI.4.3.

5. Integral criterion of convergence of a series.

5.1. Consider a series $\sum a_n$ with $a_1 \geq a_2 \geq a_3 \geq \cdots \geq 0$. Let f be a non-increasing continuous function defined on the interval $\langle 1, +\infty)$ such that

$$a_n = f(n).$$

5.2. Theorem. (Integral Criterion of Convergence) *The series $\sum a_n$ converges if and only if the limit*

$$\lim_{n\to\infty} \int_1^n f(x)dx$$

is finite.

Proof. The trivial estimate of Riemann integral yields

$$a_{n+1} = f(n+1) \le \int_n^{n+1} f(x)\mathrm{d}x \le f(n) = a_n.$$

Thus,

$$a_2 + a_3 + \cdots + a_n \le \int_1^n f(x)\mathrm{d}x \le a_1 + a_2 + \cdots + a_{n-1}.$$

Hence, if $L = \lim_{n\to\infty} \int_1^n f(x)\mathrm{d}x$ is finite then

$$\sum_1^n a_k \le a_1 + L$$

and the series converges. On the other hand, if the sequence $(\int_1^n f(x)\mathrm{d}x)_n$ is not bounded then also $(\sum_1^n a_n)_n$ is not bounded. □

5.3. Remark. Note that unlike the criteria in III.2.5, the Integral Criterion is a necessary and sufficient condition. Hence, of course, it is much finer. This will be illustrated by the following example.

5.4. Proposition. *Let $\alpha > 1$ be a real number. Then the series*

$$\frac{1}{1^\alpha} + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \cdots + \frac{1}{n^\alpha} + \cdots \tag{$*$}$$

converges.

Proof. We have

$$\int_1^n x^{-\alpha}\mathrm{d}x = \left[\frac{1}{1-\alpha} \cdot x^{1-\alpha}\right] = \frac{1}{1-\alpha}\left(\frac{1}{n^{\alpha-1}} - 1\right) \le \frac{1}{\alpha - 1}.$$

□

Note that the convergence of the series $(*)$ does not follow from the criteria III.2.5 even for big α.

XIII. Metric spaces: basics

1. An example.

1.1. In the following chapters we will study real functions of several real variables. Hence, domains of such functions will be subsets of Euclidean spaces. We will need to understand better the basic notions like convergence or continuity: as we will see in the following example they cannot be reduced to the behaviour of functions in the individual variables. In this chapter we will discuss some concepts to be used in the general context of metric spaces.

1.2. Define a function of two real variables $f : \mathbb{E}_2 \to \mathbb{R}$ by setting

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

For any fixed y_0 the function $\phi : \mathbb{R} \to \mathbb{R}$ defined by $\phi(x) = f(x, y_0)$ is evidently a continuous one (if $y_0 \neq 0$ it is defined by an arithmetic expression, and for $y_0 = 0$ it is the constant 0) and similarly for any fixed x_0 the formula $\psi(y) = f(x_0)$ defines a continuous function $\psi : \mathbb{R} \to \mathbb{R}$. But the function f as a whole behaves wierdly: if we approach $(0,0)$ in the arguments (x,x) with $x \neq 0$ the values of f are constantly $\frac{1}{2}$ and at $x = 0$ we jump to 0, an evident discontinuity in any reasonable intuitive meaning of the word.

2. Metric spaces, subspaces, continuity.

2.1. A *metric* (or *distance function*, or briefly *distance*) on a set X is a function

$$d : X \times X \to \mathbb{R}$$

such that

(1) $\forall x, y,\ d(x,y) \geq 0$ and $d(x,y) = 0$ iff $x = y$,

(2) $\forall x, y,\ d(x,y) = d(y,x)$ and

(3) $\forall x, y, z,\ d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).

A *metric space* (X, d) is a set X endowed by a metric d.

Note. The assumptions (1) and (3) are rather intuitive: (1) requires that the distance of two distinct points is not zero, (3) says that the shortest path between x and z cannot be longer than the one subjected to the condition that we visit a point y on the way. The symmetry condition (2) is somewhat less satisfactory (consider the distances between two places in town one has to cover by car), but for our purposes is is quite acceptable.

2.2. Examples. 1. The **real line**, that is, $\mathbb{R}$ with the distance $d(x, y) = |x - y|$.

2. The **Gauss plane**, that is, the set of complex numbers $\mathbb{C}$ with the distance $d(x, y) = |x - y|$. Note that the fact that this formula is a distance in $\mathbb{C}$ is less trivial than the fact about the $|x - y|$ in $\mathbb{R}$.

3. The ***n*-dimensional Euclidean space** $\mathbb{E}_n$: The set

$$\{(x_1, \dots, x_n) \mid x_i \in \mathbb{R}\}$$

with the metric

$$d((x_1, \dots, x_n), (y_1, \dots, y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (*)$$

4. Let J be an interval. Consider the set

$$F(J) = \{f \mid f : J \to \mathbb{R} \text{ bounded}\}$$

endowed with the distance

$$d(f, g) = \sup\{|f(x) - g(x)| \mid x \in J\}.$$

2.2.1. More about $\mathbb{E}_n$. The Euclidean space $\mathbb{E}_n$ (and its subsets) will play a fundamental role in the sequel. It deserves a few comments.

(a) The reader knows from linear algebra the n-dimensional vector space V_n, the scalar product $x \cdot y = (x_1, \dots, x_n) \cdot (y_1, \dots, y_n) = \sum_{i=1}^{n} x_i y_i$, the norm $\|x\| = \sqrt{x \cdot x}$, and the Cauchy Schwarz inequality

$$|x \cdot y| \leq \|x\| \cdot \|y\|.$$

From this inequality one easily infers that $d(x,y) = \|x-y\|$ is a distance on V_n (do it as a simple exercise). Now $\mathbb{E}_n$ is nothing else than (V_n, d) with the structure of vector space neglected.

(b) The Gauss plane is the Euclidean plane $\mathbb{E}_2$. Only, similarly as V_n as compared with $\mathbb{E}_n$, it has more structure.

(c) The (Pythagorean) metric $(*)$ in $\mathbb{E}_n$ is in acordance with the standard Euclidean geometry. It can be, however, somewhat inconvenient to work with. More expedient distances (equivalent with $(*)$ for our purposes) will be introduced in 4.3 below.

2.3. Continuous and uniformly continuous maps. Let (X_1, d_1) and (X_2, d_2) be metric spaces. A mapping $f : X_1 \to X_2$ is said to be *continuous* if

$$\forall x \in X_1\ \forall \varepsilon > 0\ \exists \delta > 0 \text{ such that } \forall y \in X_1,\ d_1(x,y) < \delta\ \Rightarrow d_2(f(x), f(y)).$$

It is said to be *uniformly continuous* if

$$\forall \varepsilon > 0\ \exists \delta > 0 \text{ such that } \forall x \in X_1\ \forall y \in X_1,\ d_1(x,y) < \delta\ \Rightarrow d_2(f(x), f(y)).$$

Note that obviously each uniformly continuous mapping is continuous.

2.3.1. Observations. (1) *The identity mapping* id : $(X, d) \to (X.d)$ *is continuous.*

(2) *The composition* $g \circ f : (X_1, d_1) \to (x_3, d_3)$ *of (uniformly) continuous maps* $f : (X_1, d_1) \to (x_2, d_2)$ *and* $g : (X_2, d_2) \to (x_3, d_3)$ *is (uniformly) continuous.*

2.4. Subspaces. Let (X, d) be a metric space and let $Y \subseteq X$ be a subset. Defining $d_Y(x,y) = d(x,y)$ for $x, y \in Y$ we obtain a metric on Y; the resulting metric space (Y, d_Y) is said to be a *subspace* of (X, d).

2.4.1. Observation. *Let* $f : (X_1, d_1) \to (X_2, d_2)$ *be a (uniformly) continuous maping. Let* $Y_i \subseteq X_i$ *be such that* $f[Y_1] \subseteq Y_2$. *Then the mapping* $g : (Y_1, d_{1Y_1}) \to (Y_2, d_{2Y_2})$ *defined by* $g(x) = f(x)$ *is (uniformly) continuous.*

2.5. Conventions. 1. Often, if there is no danger of confusion, we use the same symbol for distinct metrics. In particular we will mostly omit the subscript Y in the subsapce metric d_Y.

2. Unless stated otherwise, we will endow a subset of a metric space automatically with the subspace metric. We will speak of subspaces as of

the corresponding subsets, and of subsets as of the corresponding subspaces. Thus we will speak of a “ finite subspace”, an “open subspace” (see 3.4 below) or, on the other hand. of a “compact subset” (see Section 7), etc.

3. Several topological concepts.

3.1. Convergence. A sequence $(x_n)_n$ in a metric space (X, d) *converges* to $x \in X$ if

$$\forall \varepsilon > 0 \ \exists n_0 \text{ such that } \forall n \geq n_0,\ d(x_n, x) < \varepsilon.$$

We then speak of a *convergent sequence* and the x is called its *limit*, and we write

$$x = \lim_n x_n.$$

3.1.1. Observation. *Let $(x_n)_n$ be a convergent sequence and let x be its limit. Then each subsequence $(x_{k_n})_n$ of $(x_n)_n$ converges and we have $\lim_n x_{k_n} = x$.*

3.1.2. Theorem. *A mapping $f : (X_1, d_1) \to (X_2, d_2)$ is continuous if and only if for each convergent sequence $(x_n)_n$ in (X_1, d_1) the sequence $(f(x_n))_n$ converges in (X_2, d_2) and $\lim_n f(x_n) = f(\lim_n x_n)$.*

Proof. I. Let f be continuous and let $\lim_n x_n = x$. For $\varepsilon > 0$ choose by continuity a $\delta > 0$ such that $d_2(f(y), f(x)) < \varepsilon$ for $d_1(x, y) < \delta$. Now by the definition of the convergence of sequences there is an n_o such that for $n \geq n_0$, $d_1(x_n, x) < \delta$. Thus, if $n \leq n_0$ we have $d_2(f(x_n), f(x)) < \varepsilon$ so that $\lim_n f(x_n) = f(\lim_n x_n)$.

II. Let f not be continuous. Then there is an $x \in X_1$ and an $\varepsilon_0 > 0$ such that for every $\delta > 0$ there is an $x(\delta)$ such that

$$d_1(x, x(\delta)) < \delta \quad \text{but} \quad d_2(f(x), f(x(\delta))) \geq \varepsilon_0.$$

Set $x_n = x(\frac{1}{n})$. Then $\lim_n x_n = x$ but $(f(x_n))_n$ cannot converge to $f(x)$. □

Note that the proof is the same as that in IV.5.1, only with the $|u - v|$ substituted by the distances in the two spaces. In this respect there is nothing specific about the real functions of one variable.

3.2. Neighbourhoods. For a point x in a metric space (X, d) and $\varepsilon > 0$ set

$$\Omega_{(X,d)}(x, \varepsilon) = \{y \mid d(x, y) < \varepsilon\}$$

(if there is no danger of confusion, the subscript "(X, d)" is often omitted, or replaced just by "X".

A *neighbourhood* of a point x in (X, d) is any $U \subseteq X$ such that there is an $\varepsilon > 0$ with $\Omega(x, \varepsilon) \subseteq U$.

3.3.1. Proposition. 1. *If U is a neighbourhood of x and $U \subseteq V$ then V is a neighbourhood of x.*

2. *If U and V are neighbourhoods of x then the intersection $U \cap V$ is a neighbourhood of x.*

Proof. 1 is trivial.

2: If $\Omega(x, \varepsilon_1) \subseteq U$ and $\Omega(x, \varepsilon_2) \subseteq V$ then $\Omega(x, \min(\varepsilon_1, \varepsilon_2)) \subseteq U \cap V$. □

3.3.2. Proposition. *Let Y be a subspace of a metric space (X, d). Then $\Omega_Y(x, \varepsilon) = \Omega_X(x, \varepsilon) \cap Y$ and $U \subseteq Y$ is a neighbourhood of $x \in Y$ iff there is a neighbourhood V of x in (X, d) such that $U = V \cap Y$.*

Proof is straightforward. □

3.4. Open sets. A subset $U \subseteq (X, d)$ is *open* if it is a neighbourhood of each of its points.

3.4.1. Proposition. *Each $\Omega_X(x, \varepsilon)$ is open in (X, d).*

Proof. Let $y \in \Omega_X(x, \varepsilon)$. Then $d(x, y) < \varepsilon$. Set $\delta = \varepsilon - d(x, y)$. By triangle inequality, $\Omega(y, \delta) \subseteq \Omega(x, \varepsilon)$. □

3.4.2. Observation. *$\emptyset$ and X are open. If U_i, $i \in J$, are open then $\bigcup_{i \in J} U_i$ is open, and if U and V are open then $U \cap V$ is open.*

Proof. The first three statements are obvious and the third one immediately follows from 2.3.1. □

3.4.3. Proposition. *Let Y be a subspace of a metric space (X, d). Then U is open in Y iff there is a V open in X such that $U = V \cap Y$.*

Proof. For every V open in X, $U \cap Y$ is open in Y by 3.3.2. On the other hand, if U is open in Y choose for each $x \in U$ an $\Omega_Y(x, \varepsilon_x) \subseteq U$ and set $V = \bigcup_{x \in U} \Omega_X(x, \varepsilon_x)$. □

3.5. Closed sets. A subset $A \subseteq (X, d)$ is *closed* in (X, d) if for every sequence $(x_n)_n \subseteq A$ convergent in X the limit $\lim_n x_n$ is in A.

3.5.1. Proposition. *A subset $A \subseteq (X, d)$ is* closed *in (X, d) iff the complement $X \smallsetminus A$ is open.*

Proof. I. Let $X \smallsetminus A$ not be open. Then there is a point $x \in X \smallsetminus A$ such that for every n, $\Omega(x, \frac{1}{n}) \nsubseteq X \smallsetminus A$, that is, $\Omega(x, \frac{1}{n}) \cap A \neq \emptyset$. Choose

$x_n \in \Omega(x, \frac{1}{n}) \cap A$. Then $(x_n)_n \subseteq A$ and the sequence converges to $x \notin A$ and hence A is not closed.

II. Let $X \smallsetminus A$ be open and let $(x_n)_n \subseteq A$ converge to $x \in X \smallsetminus A$. Then for some $\varepsilon > 0$, $\Omega(x, \varepsilon) \subseteq X \smallsetminus A$ and hence for sufficiently large n, $x_n \in \Omega(x, \varepsilon) \subseteq X \smallsetminus A$, a contradiction. □

From 3.5.1, 3.4.2 and DeMorgan formulas we immediately obtain

3.5.2. Corollary. *$\emptyset$ and X are closed. If A_i, $i \in J$, are closed then $\bigcap_{i \in J} A_i$ is closed, and if A and B are closed then $A \cup B$ is closed.*

3.5.3. Corollary. *Let Y be a subspace of a metric space (X, d). Then A is closed in Y iff there is a B closed in X such that $A = B \cap Y$.*

3.6. Distance of a point from a subset. Closure. Let x be a point and $A \subseteq X$ be a subset of a metric space (X, d). Define the distance of x from A as

$$d(x, A) = \inf\{d(x, a) \mid a \in A\}.$$

The *closure* of a set A is

$$\overline{A} = \{x \mid d(x, A) = 0\}.$$

3.6.1. Proposition. (1) $\overline{\emptyset} = \emptyset$.
(2) $A \subseteq \overline{A}$,
(3) $A \subseteq B \quad \Rightarrow \quad \overline{A} \subseteq \overline{B}$,
(4) $\overline{A \cup B} = \overline{A} \cup \overline{B}$, *and*
(5) $\overline{\overline{A}} = \overline{A}$.

Proof. (1): $d(x, \emptyset) = +\infty$.

(2) and (3) are trivial.

(4): By (3) we have $\overline{A \cup B} \supseteq \overline{A} \cup \overline{B}$. Now let $x \in \overline{A \cup B}$ but not $x \in \overline{A}$. Then $\alpha = d(x, A) > 0$ and hence all the $y \in A \cup B$ such that $d(x, y) < \alpha$ are in B; hence $x \in \overline{B}$.

(5): Let $d(x, \overline{A})$ be 0. Choose $\varepsilon > 0$. There is a $z \in \overline{A}$ such that $d(x, z) < \frac{\varepsilon}{2}$ and for this z we can choose a $y \in A$ such that $d(z, y) < \frac{\varepsilon}{2}$. Thus, by triangle inequality, $d(x, y) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ and we see that $x \in \overline{A}$. □

3.6.2. Proposition. *$\overline{A}$ is the set of all the limits of convergent sequences $(x_n)_n \subseteq A$*

Proof. A limit of a convergenr $(x_n)_n \subseteq A$ is obviously in $\overline{A}$.

Now let $x \in \overline{A}$. If $x \in A$ then it is the limit of the constant sequence $x, x, x, \ldots$. If $x \in \overline{A} \smallsetminus A$ then for each n there is an $x_n \in A$ such that $d(x, x_n) < \frac{1}{n}$. Obviously $x = \lim_n x_n$. □

3.6.3. Proposition. *$\overline{A}$ is closed and it is the least closed set containing A. That is,*

$$\overline{A} = \bigcap \{B \mid A \subseteq B,\ B \textit{ closed}\}.$$

Proof. Let $(x_n)_n \subseteq \overline{A}$ converge to x. For each n choose $y_n \in A$ such that $d(x_n, y_n) < \frac{1}{n}$. Then $\lim_n y_n = x$ and x is in $\overline{A}$ by 3.5.1.

Now let B be closed and let $A \subseteq B$. If $x \in \overline{A}$ we can choose, by 3.5.1, a convergent sequence $(x_n)_n$ in A, and hence in B, such that $\lim x_n = x$. Thus, $x \in B$. □

3.6.4. Corollary. *Let Y be a subspace of a metric space (X, d). Then the closure of A in Y is equal to $\overline{A} \cap Y$ (where $\overline{A}$ is the closure in X).*

3.7. Theorem. *Let $(X_1, d_1), (X_2.d_2)$ be metric spaces and let $f : X_1 \to X_2$ be a mapping. Then the following statements are equivalent.*

(1) *f is continuous.*

(2) *for every $x \in X_1$ and for every neighbourhood V of $f(x)$ there is a neighbourhood U of x such that $f[U] \subseteq V$.*

(3) *for every open U in X_2 the preimage $f^{-1}[U]$ is open in X_1.*

(4) *for every closed A in X_2 the preimage $f^{-1}[A]$ is closed in X_1.*

(5) *for every $A \subseteq X_1$, $f[\overline{A}] \subseteq \overline{f[A]}$.*

Proof. (1)⇒(2): There is an $\varepsilon > 0$ such that $\Omega(f(x), \varepsilon) \subseteq V$. Take the δ from the definition of continuity and set $U = \Omega(x, \delta)$. Then $f[U] \subseteq \Omega(f(x), \varepsilon) \subseteq V$.

(2)⇒(3): Let U be open and $x \in f^{-1}[U]$. Thus, $f(x) \in U$ and U is a neighbourhood of $f(x)$. There is a neighbourhood V of x such that $f[V] \subseteq U$. Consequently $x \in V \subseteq f^{-1}[U]$ and $f^{-1}[U]$ is a neighbourhood of x. Since $x \in f^{-1}[U]$ was arbitrary, the preimage is open.

(3)⇔(4) by 3.5.1 since preimage preserves complements.

(4)⇒(5): We have $A \subseteq f^{-1}[f[A]] \subseteq f^{-1}[\overline{f[A]}]$. By (4), $f^{-1}[\overline{f[A]}]$ is closed and hence by 3.5.3, $\overline{A} \subseteq f^{-1}[\overline{f[A]}]$ and finally $f[\overline{A}] \subseteq \overline{f[A]}$.

(5)⇒(1): Let $\varepsilon > 0$. Set $B = X_2 \smallsetminus \Omega(f(x), \varepsilon)$ and $A = f^{-1}[B]$. Then $f[\overline{A}] \subseteq \overline{f[f^{-1}[B]]} \subseteq \overline{B}$. Hence $x \notin \overline{A}$ (the distance $d(f(x), B)$ is at least ε) and hence there is a $\delta > 0$ such that $\Omega(x, \delta) \cap A = \emptyset$ and we easily conclude that $f[\Omega(x, \delta)] \subseteq \Omega(f(x), \varepsilon)$. □

3.8. Homeomorfism. Topological concepts. A continuous mapping $f : (X, d) \to (Y, d')$ is called *homeomorphism* if there is a continuous $g : (Y, d') \to (X, d)$ such that $f \circ g = \mathrm{id}_Y$ and $g \circ f = \mathrm{id}_X$. If there exists a homeomorphism $f : (X, d) \to (Y, d')$ we say that the spaces (X, d) and (Y, d') are *homeomorphic.*

A property or definition is said to be *topological* if it is preserved by homeomorphisms. Thus we have the following topological properties:

- convergence (see 3.1.2),
- openness (see 3.7),
- closedness (see 3.7).
- closure (although $d(x, A)$ is not topological; see, however, 3.6.3),
- neighbourhood (although $\Omega(x, \varepsilon)$ is not topological; but realize that A is a neighbourhood of x if there is an open U such that $x \in U \subseteq A$),
- or continuity itself.

On the other hand, for instance uniform continuity is not a topological property.

3.9. Isometry. An onto mapping $f : (X, d) \to (Y, d')$ is called *isometry* if $d'(f(x), f(y)) = d(x, y)$ for all $x, y \in X$. Then, trivially,

- f is one-to-one and continuous, and
- its inverse is also an isometry; thus, f is a homeomorphism.

If there is an isometry $f : (X, d) \to (Y, d')$ the spaces (X, d) and (Y, d') are said to be isometric. Of course, an isometry preserves all topological concepts, but much more, indeed everything that can be defined in terms of distance.

4. Equivalent and strongly equivalent metrics.

4.1. Two metrics d_1, d_2 on a set are said to be *equivalent* if $\mathrm{id}_X : (X, d_1) \to (X.d_2)$ is a homeomorphism. Thus, replacing a metric by an equivalent one we obtain a space in which all topological notions from the original space are preserved.

4.2. A much stronger concept is that of a strong equivalence. We say that d_1, d_2 on a set are *strongly equivalent* if there are positive constants α and β such that for all $x, y \in X$.

$$\alpha \cdot d_1(x,y) \leq d_2(x,y) \leq \beta \cdot d_1(x,y)$$

(this relation is of course symmetric: consider $\frac{1}{\alpha}$ and $\frac{1}{\beta}$).

Note that

> *replacing a metric by a strongly equivalent one preserves not only topological properties but also for instance the uniform convergence.*

4.3. The concept of strong equivalence will help us to reason much more easily in Euclidean spaces $\{(x_1, \dots, x_n) \mid x_i \in \mathbb{R}\}$ where we so far had the distance

$$d((x_1, \dots, x_n), (y_1, \dots, y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Set

$$\lambda((x_1, \dots, x_n), (y_1, \dots, y_n)) = \sum_{i=1}^{n} |x_i - y_i|, \quad \text{and}$$
$$\sigma((x_1, \dots, x_n), (y_1, \dots, y_n)) = \max_i |x_i - y_i|.$$

4.3.1. Proposition. *d, λ and σ are strongly equivalent metrics on $\mathbb{E}_n$.*

Proof. It is easy to see that λ and σ are metrics.

Now we have

$$\lambda((x_i)_i, (y_i)_i) = \sum_{i=1}^{n} |x_i - y_i| \leq n\sigma((x_j)_j, (y_j)_j)$$

since for each i, $|x_i - y_i| \le \sigma((x_j)_j, (y_j)_j)$, and for the same reason

$$d((x_i)_i, (y_i)_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \le \sqrt{n}\sigma((x_j)_i, (y_j)_j).$$

On the other hand obviously

$$\sigma((x_i)_i, (y_i)_i) \le \lambda((x_i)_i, (y_i)_i) \quad \text{and} \quad \sigma((x_i)_i, (y_i)_i) \le d((x_i)_i, (y_i)_i).$$

□

In the sequel we will mostly work with the Euclidean space as with $(\mathbb{E}_n, \sigma)$.

5. Products.

5.1. Let (X_1, d_i), $i = 1, \dots, n$ be metric spaces. On the cartesian product

$$\prod_{i=1}^{n} X_i$$

define a metric

$$d((x_1, \dots, x_n), (y_1, \dots, y_n)) = \max_i d_i(x_i, y_i).$$

The resulting metric space will be denoted by $\prod_{i=1}^{n}(X_i, d_i)$.

5.1.1. Notation. We will also write

$$(X_1, d_1) \times (X_2, d_2) \quad \text{or} \quad (X_1, d_1) \times (X_2, d_2) \times (X_3, d_3)$$

for $\prod_{i=1}^{2}(X_i, d_i)$ or $\prod_{i=1}^{3}(X_i, d_i)$, and sometimes also

$$(X_1, d_1) \times \cdots \times (X_n, d_n)$$

for the general $\prod_{i=1}^{n}(X_i, d_i)$.

Further, if $(X_i, d_i) = (X, d)$ for all i we write

$$\prod_{i=1}^{n}(X_i, d_i) = (X, d)^n.$$

5.1.2. Remarks. 1. Thus, $(\mathbb{E}_n, \sigma)$ is the product $\overbrace{\mathbb{R} \times \cdots \times \mathbb{R}}^{n \text{ times}} = \mathbb{R}^n$.

2. For all purposes we could have defined the metric in the product by

$$d((x_i)_i, (y_i)_i) = \sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)^2} \quad \text{or} \quad d((x_i)_i, (y_i)_i) = \sum_{i=1}^{n} d_i(x_i, y_i),$$

but working with the d above is much easier.

5.2. Lemma. *A sequence*

$$(x_1^1, \dots, x_n^1), (x_1^2, \dots, x_n^2), \dots, (x_1^k, \dots, x_n^k), \dots$$

converges to $(x_1, \dots, x_n)$ *in* $\prod(X_i, d_i)$ *if and only if each of the sequences* $(x_i^k)_k$ *converges to* x_i *in* (X_i, d_i).

(Caution: the superscripts k are indices, not powers.)

Proof. $\Rightarrow$ immediately follows from the fact that $d_i(u_i, v_i) \leq d((u_j)_j, (v_j)_j)$.

$\Leftarrow$: Let each of the $(x_i^k)_k$ converge to x_i. For an $\varepsilon > 0$ and i we have k_i such that for $k \geq k_i$, $d_i(x_i^k, x_i) < \varepsilon$. Then for $k \geq \max_i k_i$ we have

$$d((x_1^k, \dots, x_n^k), (x_1, \dots, x_n)) < \varepsilon.$$

□

5.3. Theorem. 1. *The projection mappings* $p_j = ((x_i)_i \mapsto x_j) : \prod_{i=1}^n (X_i, d_i) \to (X_j, d_j)$ *are continuous.*

2. *Let* $f_{:}(Y, d') \to (X_j, d_j)$ *be arbitrary continuous mapings. Then the unique mapping* $f : (Y, d') \to \prod_{i=1}^n (X_i, d_i)$ *such that* $p_j \circ f = f_j$, *namely that defined by* $f(y) = (f_1(y), \dots, f_n(y))$, *is continuous.*

Proof. 1 immediately follows from the fact that $d_j(x_j, y_j) \leq d((x_i)_i, (y_i)_i)$.

2: Follows from 3.1.2 and 5.2. If $\lim_k y_k = y$ in (Y, d') then $\lim_k f_j(y_k) = f_j(y)$ in (X_j, d_j) for all j and hence $(f(y_k))_k$, that is,

$$(f_1(y_1),\ \dots, f_n(y_1)),\ (f_1(y_2), \dots, f_n(y_2)),\ \dots,\ (f_1(y_k), \dots, f_n(y_k)),\ \dots$$

converges to $(f_1(y), \dots, f_n(y))$. □

5.4. Observation. *Obviously* $\prod_{i=1}^{n+1}(X_i, d_i)$ *is isometric (recall 3.9) with* $\prod_{i=1}^n (X_i, d_i) \times (X_{n+1}, d_{n+1})$. *Consequenly, it usually suffices to prove a statement on finite products for products of two spaces only.*

6. Cauchy sequences. Completeness.

6.1. A sequence $(x_n)_n$ in a metric space (X,d) is said to be *Cauchy* if

$$\forall \varepsilon > 0 \ \exists n_0 \text{ such that } \ m,n \geq n_0 \ \Rightarrow \ d(x_m, x_n) < \varepsilon.$$

6.1.1. Observation. *Each convergent sequence is Cauchy.*
(Just like in $\mathbb{R}$: if $d(x_n, x) < \varepsilon$ for $n \geq n_0$ then for $m,n \geq n_0$,

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < 2\varepsilon.)$$

6.2. Proposition. *Let a Cauchy sequence have a convergent subsequence. Then it converges (to the limit of the subsequence).*

Proof. Let $(x_n)_n$ be Cauchy and let $\lim_n x_{k_n} = x$. Let $d(x_m, x_n) < \varepsilon$ for $m,n \geq n_1$ and $d(x_{k_n}, x) \leq \varepsilon$ for $n \geq n_2$. If we set $n_0 = \max(n_1, n_2)$ we have for $n \geq n_0$ (since $k_n \geq n$)

$$d(x_n, x) \leq d(x_n, x_{k_n}) + d(x_{k_n}, x) < 2\varepsilon.$$

□

6.3. A metric space (X,d) is *complete* if each Cauchy sequence in (X,d) converges.

6.3.1. Thus, by Bolzano-Cauchy Theorem (II.3.4) the real line $\mathbb{R}$ with the standard metric is complete.

6.4. Proposition. *A subspace of a complete space is complete if and only if it is closed.*

Proof. I. Let $Y \subseteq (X,d)$ be closed. Let $(y_n)_n$ be Cauchy in Y. Then it is Cauchy and hence convergent in X, and the limit, by closedness, is in Y.

II. Let Y not be closed. Then there is a sequence $(y_n)_n$ in Y convergent in X such that $\lim_n y_n \notin Y$. Then $(y_n)_n$ is Cauchy in X, but since the distance is the same, also in Y. But in Y it does not converge. □

6.5. Lemma. *A sequence*

$$(x_1^1, \dots, x_n^1), (x_1^2, \dots, x_n^2), \dots, (x_1^k, \dots, x_n^k), \dots$$

is Cauchy in $\prod_{i=1}^n (X_i, d_i)$ *if and only if each of the sequences* $(x_i^k)_k$ *is Cauchy in* (X_i, d_i).

Proof. $\Rightarrow$ immediately follows from the fact that $d_i(u_i, v_i) \leq d((u_j)_j, (v_j)_j)$.

$\Leftarrow$: Let each of the $(x_i^k)_k$ be Cauchy. For an $\varepsilon > 0$ and i we have k_i such that for $k, l \geq k_i$, $d_i(x_i^k, x_i^l) < \varepsilon$. Then for $k, l \geq \max_i k_i$ we have

$$d((x_1^k, \dots, x_n^k), (x_1^l, \dots, x_n^l)) < \varepsilon.$$

□

Combining 5.2 and 6.5 (and, of course, 6.3.1) we immediately obtain

6.6. Proposition. *A product of complete spaces is complete. In particular, the Euclidean space $\mathbb{E}_n$ is complete.*

From 6.6 and 6.4 we imediately infer

6.7. Corollary. *A subspace Y of the Euclidean space $\mathbb{E}_n$ is complete if and only if it is closed.*

6.8. Note. Neither the Cauchy property nor completeness is a topological property. Consider $\mathbb{R}$ and any bounded open interval J in $\mathbb{R}$. They are homeomorphic (if for instance $J = (-\frac{\pi}{2}, +\frac{\pi}{2})$ we have the mutually inverse homeomorphisms $\tan : J \to \mathbb{R}$ and $\mathrm{arctg} : \mathbb{R} \to J$). But $\mathbb{R}$ is complete and J is not.

But it is easy to see that the properties are preserved when replacing a metric by a strongly equivalent one. This concerns, of course, in particular the metrics in $\mathbb{E}_n$ mentioned in Section 4.

7. Compact metric spaces.

7.1. A metric space (X, d) is said to be *compact* if each sequence in (X, d) contains a convergent subsequence.

7.1.1. Note. Thus the compact intervals, that is the bounded closed intervals $\langle a, b\rangle$ are compact in this definition, and they are the only compact ones among the various types of intervals.

7.2. Proposition. *A subspace of a compact space is compact if and only if it is closed.*

Proof. I. Let $Y \subseteq (X, d)$ be closed. Let $(y_n)_n$ be a sequence in Y. In X it has a convergent subsequence $(y_{k_n})_n$ convergent in X, and the limit, by closedness, is in Y.

II. Let Y not be closed. Then there is a sequence $(y_n)_n$ in Y convergent in X such that $y = \lim_n y_n \notin Y$. Then $(y_n)_n$ cannot have a subsequence convergent in Y since each subsequence converges to y. $\square$

7.3. Proposition. *Let (X,d) be arbitrary and let a subspace Y of X be compact. Then Y is closed in (X,d).*

Proof. Let $(y_n)_n$ be a sequence in Y convergent in X to a limit y. Then each subsequence of $(y_n)_n$ converges to y and hence $y \in Y$. $\square$

7.4. A metric space (X,d) is said to be *bounded* if there is a constant K such that

$$\forall x, y \in X, \quad d(x,y) < K.$$

7.4.1. Proposition. *Each compact metric space is bounded.*

Proof. Suppose not. Choose x_1 arbitrarily and then x_n so that $d(x_1, x_n) > n$. The sequence $(x_n)_n$ has no convergent subsequence: if x were a limit of such a subsequence we would have infinitely many members of this subsequence closer to x_1 than $d(x_1, n) + 1$, a contradiction. $\square$

7.5. Theorem. *A product of finitely many compact metric spaces is compact.*

Proof. By 5.4 it suffices to prove the statement for two spaces.

Let (X, d_1), (Y, d_2) be compact and let $((x_n, y_n))_n$ be a sequence in $X \times Y$. Choose a convergent subsequence $(x_{k_n})_n$ of $(x_n)_n$ and a convergent subsequence $(y_{k_{l_n}})_n$ of $(y_{k_n})_n$. Then by 5.2.

$$((x_{k_{l_n}}, y_{k_{l_n}}))_n$$

is a convergent subsequence of $((x_n, y_n))_n$. $\square$

7.6. Theorem. *A subspace of the Euclidean space $\mathbb{E}_n$ is compact if and only if it is bounded and closed.*

Proof. I. A compact subspace of any metric space is closed by 7.3 and bounded by 7.4.1.

II. Now let $Y \subseteq \mathbb{E}_n$ be bounded and closed. Since it is bounded we have for a sufficiently large compact interval

$$Y \subseteq J^n \subseteq \mathbb{E}_n.$$

Now by 7.5 J^n is compact and since Y is closed in $\mathbb{E}_n$ it is also closed in J^n and hence compact by 7.2. $\square$

7.7. Proposition. *Each compact space is complete.*

Proof. A Cauchy sequence has by compactness a convergent subsequence and hence it converges, by 6.2. □

7.8. Proposition. *Let $f : (X, d) \to (Y, d')$ be a continuous mapping and let $A \subseteq X$ be compact. Then $f[A]$ is compact.*

Proof. Let $(y_n)_n$ be a sequence in $f[A]$. Choose $x_n \in A$ such that $y_n = f(x_n)$. Let $(x_{k_n})_n$ be a convergent subsequence of $(x_n)_n$. Then $(y_{k_n})_n = (f(x_{k_n}))_n$ is by 3.1.2 a convergent subsequence of $(x_n)_n$. □

7.9. Proposition. *Let (X, d) be compact. Then a continuous function $f : (X, d) \to \mathbb{R}$ attains a maximum and a minimum.*

Proof. By 7.8, $Y = f[X] \subseteq \mathbb{R}$ is compact. Hence it is bounded by 7.4.1 and it has to have a supremum M and an infimum m. We have obviously $d(m, Y) = d(M, Y) = 0$ and since Y is closed, $m, M \in Y$. □

7.9.1. Corollary. *Let all the values of a continuous function on a compact space be positive. Then there is a $c > 0$ such that all the values of f are greater or equal c.*

We already know that a continuous mapping f is characterized by the property that *preimages* of closed sets are closed. Now by 7.2 and 7.8 we see that if the domain is compact we also have that *images* of closed sets are closed. This results (a.o.) in the following theorem.

7.10. Theorem. *Let (X, d) be compact and let $f : (X, d) \to (Y, d')$ be a one-to-one and onto continuous map. Then f is a homeomorphism.*

More generally, let $f : (X, d) \to (Y, d')$ be an onto continuous map let $g : (X, d) \to (Z, d'')$ be a continuous map, and let $h : (Y, d') \to (Z, d'')$ be such that $h \circ f = g$. Then h is continuous.

Proof. We will prove the second statement: the first one follows by setting $g = \mathrm{id}_Y$.

Let B be closed in Z. Then $A = g^{-1}[B]$ is closed and hence compact in X and hence $f[A]$ is compact and hence closed in Y. Since f is onto we have $f[f^{-1}[C]] = C$ for any C. Thus,

$$h^{-1}[B] = f[f^{-1}[h^{-1}[B]]] = f[(h \circ f)^{-1}[B]] = f[g^{-1}[B]] = f[A]$$

is closed. □

7.11. Theorem. *Let (X, d) be a compact space. Then a mapping $f : (X, d) \to (Y, d')$ is continuous if and only if it is uniformly continuous.*

Note. Similarly as in 3.1.2 we can repeat practically verbatim the proof of the corresponding statement on real functions on compact intervals.

Proof. Let f not be uniformly continuous. We will prove it is not continuous either.

Since the formula for uniform continuity does not hold we have an $\varepsilon_0 > 0$ such that for every $\delta > 0$ there are $x(\delta), y(\delta)$ such that $d(x(\delta), y(\delta)) < \delta$ while $d'(f(x(\delta)), f(y(\delta))) \geq \varepsilon_0$. Set $x_n = x(\frac{1}{n})$ and $y_n = y(\frac{1}{n})$. Choose convergent subsequences $(\widetilde{x}_n)_n$, $(\widetilde{y}_n)_n$ (first choose a convergent subsequence $(x_{k_n})_n$ of $(x_n)_n$ then a convergent subsequence $(y_{k_{l_n}})_n$ of $(y_{n_k})_k$ and finally set $\widetilde{x}_n = x_{k_{l_n}}$ and $\widetilde{y}_n = y_{k_{l_n}}$). Then $d(\widetilde{x}_n, \widetilde{y}_n) < \frac{1}{n}$ and hence $\lim \widetilde{x}_n = \lim \widetilde{y}_n$. Because of $d'(f(\widetilde{x}_n), f(\widetilde{y}_n)) \geq \varepsilon_0$, however, we cannot have $\lim f(\widetilde{x}_n) = \lim f(\widetilde{y}_n)$ so that by 3.1.2, f is not continuous. □

XIV. Partial derivatives and total differential. Chain rule

1. Conventions.

1.1. We will work with real functions of several real variables, that is, with mappings $f : D \to \mathbb{R}$ where the domain D is a subset of $\mathbb{E}_n$. When taking derivatives, D will be typically open. Sometimes we will also have closed domains, usually closures of open sets with transparent boundaries.

We already know (recall XIII.1) that the behaviour of such functions cannot be reduced to that of functions of one variable obtained by fixing all the variables but one. But this will not prevent us from such fixings in some constructions (for instance already in the definition of partial derivative in the next section).

1.2. Convention. To simplify notation, we will often use bold-face letters to indicate points of the Euclidean space $\mathbb{E}_n$ (that is, n-tuples of real numbers, real arithmetic vectors). For example, we will write

$$\mathbf{x} \text{ for } (x_1, \dots, x_n) \quad \text{or} \quad \mathbf{A} \text{ for } (A_1, \dots, A_n).$$

We will also write

$$\mathbf{o} \quad \text{for} \quad (0, 0, \dots, 0).$$

In the rare cases when we will use subscripts with bold-face letters, e.g. $\mathbf{a}_1, \mathbf{a}_2, \dots$ we will always have in mind several points, never coordinates of a single point $\mathbf{a}$.

The scalar product of vectors $\mathbf{x}$, $\mathbf{y}$, that is, $\sum_{j=1}^{n} x_j y_j$, can be written as

$$\mathbf{xy}.$$

1.3. Extending the convention. The "bold face" convention will be also used for *vector functions*, that is,

$$\mathbf{f} = (f_1, \dots, f_m) : D \to \mathbb{E}_m, \quad f_j : D \to \mathbb{R}.$$

Note that here there is no problem with continuity: $\mathbf{f}$ is continuous iff all the f_i are continuous (recall XIII.5.3).

1.4. Composition. Vector functions $\mathbf{f} : D \to \mathbb{E}_m$, $D \subseteq \mathbb{E}_n$, and $\mathbf{g} : D' \to \mathbb{E}_k$, $D \subseteq \mathbb{E}_n$ can be composed if $\mathbf{f}[D] \subseteq D'$, and we shall write

$$\mathbf{g} \circ \mathbf{f} : D \to \mathbb{E}_k, \quad \text{(if there is no danger of confusion, just} \quad \mathbf{gf} : D \to \mathbb{E}_k\text{)},$$

Note that, similarly like with real functions of one real variable, we refrain from pedantic renaming the $\mathbf{f}$ when restricted to a map $D \to D'$.

2. Partial derivatives

2.1. Let $f : D \to \mathbb{R}$ be a real function of n variables. Consider the functions

$$\phi_k(t) = f(x_1, \dots, x_{k-1}, t, x_{k+1}, \dots, x_n), \quad \text{all } x_j \text{ with } j \neq k \text{ fixed.}$$

The *partial derivative* of f by x_k (at the point $(x_1, \dots, x_n)$) is the (ordinary) derivative of the function ϕ_k, that is, the limit

$$\lim_{h \to 0} \frac{f(x_1, \dots x_{k-1}, x_k + h, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_n)}{h}.$$

One sometimes speaks of the *k-th partial derivative* of f but one has to be careful not to confuse this expression with a derivative of higher order.

The standard notation is

$$\frac{\partial f(x_1, \dots, x_n)}{\partial x_k} \quad \text{or} \quad \frac{\partial f}{\partial x_k}(x_1, \dots, x_n),$$

in case of denoting variables by different letters, say $f(x, y)$, we write, of course,

$$\frac{\partial f(x, y)}{\partial x} \quad \text{and} \quad \frac{\partial f(x, y)}{\partial y}, \quad \text{etc.}$$

This notation is not quite consistent: the x_k in the "denominator" ∂x_k just indicates focusing to the k-th variable while the x_n in the $f(x_1, \dots, x_n)$ in the "numerator" may refer to an actual value of the argument. This usually does not create any misunderstanding. If there is a danger of confusion we can write e.g.

$$\left.\frac{\partial f(x_1, \dots, x_n)}{\partial x_k}\right|_{(x_1, \dots, x_n) = (a_1, \dots, a_n)}.$$

However, one rarely needs such a specification.

2.2. Similarly as with the standard derivative it can happen (and typically it does) that a partial derivative $\frac{\partial f(x_1,\dots,x_n)}{\partial x_k}$ exists for all $(x_1,\dots,x_n)$ in some domain D'. In such case, we have a function

$$\frac{\partial f}{\partial x_k} : D' \to \mathbb{R}.$$

It is usually obvious from the context whether, speaking of a partial derivative, we have in mind a function or just a number (the value of the limit above).

2.3. The function f from XIII.1.2 has both partial derivatives in every point (x, y). Thus we see that unlike the standard derivative of a real function with one real variable, the existence of partial derivatives does not imply continuity. For calculus in several variables we will need a stronger concept. It will be discussed in the next section.

3. Total differential.

3.1. Recall VI.1.5. The formula $f(x + h) - f(x) = Ah$ (we are neglecting the "small part" $|h| \cdot \mu(h)$) expresses the line tangent to the curve $\{(t, f(t)) \,|\, t \in D\}$ at the point $(x, f(x))$. Or, it can be viewed as a linear approximation of the function in the vicinity of this point.

Now think of a function $f(x, y)$ in this vein (the problem with more than two variable is the same) and consider the surface

$$S = \{(t, u, f(t, u)) \,|\, (t, u) \in D\}.$$

The two partial derivatives express the directions of two tangent lines to S in the point $(x, y, f(x, y))$,

- but not the tangent plane (and only that would be a desirable extension of the fact in VI.1.5),
- and do not provide any linear approximation of the function.

This will be mended by the concept of total differential.

3.2. The norm. For a point $\mathbf{x} \in \mathbb{E}_n$ we define the norm $\|\mathbf{x}\|$ as the distance of $\mathbf{x}$ from $\mathbf{o}$. Thus, we will typically use the formula

$$\|\mathbf{x}\| = \max_i |x_i|$$

(but $\|\mathbf{x}\| = \sum_{i=1}^n |x_i|$ or the standard Pythagorean $\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$ would yield the same results, recall XIII.4).

3.3. Total differential. We say that $f(x_1, \dots, x_n)$ has a *total differential* at a point $\mathbf{a} = (a_1, \dots, a_n)$ if there exists a function μ continuous in a neighborhood U of $\mathbf{o}$ which satisfies $\mu(\mathbf{o}) = 0$ (in another, equivalent, formulation, one requires μ to be defined in $U \smallsetminus \{\mathbf{o}\}$ and satisfy $\lim_{\mathbf{h}\to\mathbf{o}} \mu(\mathbf{h}) = 0$), and numbers $A_1, \dots, A_n$ such that

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \sum_{k=1}^n A_k h_k + \|\mathbf{h}\|\mu(\mathbf{h}).$$

3.3.1. Notes. 1. Using the scalar product we may write $f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \mathbf{A}\mathbf{a} + \|\mathbf{h}\|\mu(\mathbf{h}))$.

2. Note that we have not defined a total differential as an entity, only the property of a function "to have a total differential". We will leave it at that.

3.4. Proposition. *Let a function f have a total differential at a point $\mathbf{a}$. Then*

1. *f is continuous in $\mathbf{a}$.*

2. *f has all the partial derivatives in $\mathbf{a}$, with values*

$$\frac{\partial f(\mathbf{a})}{\partial x_k} = A_k.$$

Proof. 1. We have

$$|f(\mathbf{x}-\mathbf{y})| \le |\mathbf{A}(\mathbf{x}-\mathbf{y})| + |\mu(\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|$$

and the limit of the right hand side for $\mathbf{y} \to \mathbf{x}$ is obviously 0.

2. We have

$$\frac{1}{h}(f(x_1, \dots x_{k-1}, x_k + h, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_n)) =$$
$$= A_k + \mu((0, \dots, 0, h, 0, \dots, 0))\frac{\|(0, \dots, h, \dots, 0)\|}{h},$$

and the limit of the right hand side is clearly A_k. $\square$

3.5. Now we have a linear approximation: the formula

$$f(x_1+h_1\dots,x_n+h_n)-f(x_1,\dots x_n)=f(\mathbf{a}+\mathbf{h})-f(\mathbf{a})=\sum_{k=1}^{n}A_kh_k+\|\mathbf{h}\|\mu(\mathbf{h})$$

can be interpreted as saying that in a small neighborhood of $\mathbf{a}$, the function f is well approximated by the linear function

$$L(x_1,\dots,x_n)=f(a_1,\dots,a_n)+\sum A_k(x_k-a_k).$$

By the required properties of μ, the error is much smaller than the difference $\mathbf{x}-\mathbf{a}$.

In case of just one variable, there is no difference between having a derivative at a point a and having a total differential at the same point (recall VI.1.5). In case of more than one variable, however, the difference between having all partial derivatives and having a total differential at a point is tremendous.

What is happening geometrically is this: If we think of a function f as represented by its "graph", the hypersurface

$$S=\{(x_1,\dots,x_n,f(x_1,\dots,x_n))\,|\,(x_1,\dots,x_n)\in D\}\subseteq\mathbb{E}_{n+1},$$

the partial derivatives describe just the tangent lines in the directions of the coordinate axes (recall 3.1), while the total differential describes the entire tangent hyperplane.

3.6. It may be slightly surprising that, while the plain existence of partial derivative does not amount to much, possessing *continuous* partial derivatives is quite another matter. We have

Theorem. *Let f have continuous partial derivatives in a neighborhood of a point $\mathbf{a}$. Then f has a total differential at $\mathbf{a}$.*

Proof. Let

$$\mathbf{h}^{(0)}=\mathbf{h},\ \mathbf{h}^{(1)}=(0,h_2,\dots,h_n),\ \mathbf{h}^{(2)}=(0,0,h_3,\dots,h_n)\quad\text{etc.}$$

(so that $\mathbf{h}^{(n)}=\mathbf{o}$). Then we have

$$f(\mathbf{a}+\mathbf{h})-f(\mathbf{a})=\sum_{k=1}^{n}(f(\mathbf{a}+\mathbf{h}^{(k-1)})-f(\mathbf{a}+\mathbf{h}^{(k)}))=M.$$

By Lagrange Theorem (VII.2.2), there are $0 \le \theta_k \le 1$ such that

$$f(\mathbf{a}+\mathbf{h}^{(k-1)}) - f(\mathbf{a}+\mathbf{h}^{(k)}) = \frac{\partial f(a_1,\dots,a_{k-1},a_k+\theta_k h_k, a_{k+1},\dots,a_n)}{\partial x_k} h_k$$

and hence we can proceed

$$\begin{aligned} M &= \sum \frac{\partial f(a_1,\dots,a_k+\theta_k h_k,\dots,a_n)}{\partial x_k} h_k = \\ &= \sum \frac{\partial f(\mathbf{a})}{\partial x_k} h_k + \sum \left(\frac{\partial f(a_1,\dots,a_k+\theta_k h_k,\dots,a_n)}{\partial x_k} - \frac{\partial f(\mathbf{a})}{\partial x_k} \right) h_k = \\ &= \sum \frac{\partial f(\mathbf{a})}{\partial x_k} h_k + \|\mathbf{h}\| \sum \left(\frac{\partial f(a_1,\dots,a_k+\theta_k h_k,\dots,a_n)}{\partial x_k} - \frac{\partial f(\mathbf{a})}{\partial x_k} \right) \frac{h_k}{\|\mathbf{h}\|}. \end{aligned}$$

Set

$$\mu(\mathbf{h}) = \sum \left(\frac{\partial f(a_1,\dots,a_k+\theta_k h_k,\dots,a_n)}{\partial x_k} - \frac{\partial f(\mathbf{a})}{\partial x_k} \right) \frac{h_k}{\|\mathbf{h}\|}.$$

Since $\left| \frac{h_k}{\|\mathbf{h}\|} \right| \le 1$ and since the functions $\frac{\partial f}{\partial x_k}$ are continuous, $\lim_{\mathbf{h}\to\mathbf{o}} \mu(\mathbf{h}) = 0$.
□

3.7. Thus, we can write schematically

$$\text{continuous PD} \;\Rightarrow\; \text{TD} \;\Rightarrow\; \text{PD}$$

(where PD stands for all partial derivatives and TD for total differential). Note that neither of the implication can be reversed. We have already discussed the second one; for the first one, recall that for functions of one variable the existence of a derivative at a point coincides with the existence of a total differential, while a derivative is not necessarily a continuous function even when it exists at every point of an open set.

In the rest of this chapter, simply assuming that partial derivatives exist will almost never be enough. Sometimes the existence of the total differential will suffice, but more often than not we will assume the stronger existence of continuous partial derivatives.

4. Higher order partial derivatives. Interchangeability

4.1. Recall 2.2. When we have a function $g(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_k}$ then similarly as taking the second derivative of a function of one variable, we may consider partial derivatives of $g(\mathbf{x})$, that is,

$$\frac{\partial g(\mathbf{x})}{\partial x_l}.$$

The result, if it exists, is then denoted by

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_l}.$$

More generally, iterating this procedure we may obtain

$$\frac{\partial^r f(\mathbf{x})}{\partial x_{k_1} \partial x_{k_2} \dots \partial x_{k_r}},$$

the *partial derivatives of order* r.

Note that the order is given by the number of taking derivatives and does not depend on repeated individual variables. Thus for example,

$$\frac{\partial^3 f(x,y,x)}{\partial x \partial y \partial z} \quad \text{and} \quad \frac{\partial^3 f(x,y,x)}{\partial x \partial x \partial x}$$

are derivatives of third order (even though in the former case we have taken a partial derivative by each variable only once).

To simplify notation, taking a partial derivatives by the same variable more than once consecutively may be indicated by an exponent, e.g.

$$\frac{\partial^5 f(x,y)}{\partial x^2 \partial y^3} = \frac{\partial^5 f(x,y)}{\partial x \partial x \partial x \partial y \partial y},$$

$$\frac{\partial^5 f(x,y)}{\partial x^2 \partial y^2 \partial x} = \frac{\partial^5 f(x,y)}{\partial x \partial x \partial y \partial y \partial x}.$$

4.2. Example. Compute the “mixed” second order derivatives of the function $f(x,y) = x \sin(y^2 + x)$. We obtain, first,

$$\frac{\partial f(x,y)}{\partial x} = \sin(y^2 + x) + x \cos(y^2 + x) \quad \text{and} \quad \frac{\partial f(x,y)}{\partial y} = 2xy \cos(y^2 + x).$$

Now for the second order derivatives we get

$$\frac{\partial^2 f}{\partial x \partial y} = 2y\cos(y^2+x) - 2xy\sin(y^2+x) = \frac{\partial^2 f}{\partial y \partial x}.$$

Whether it is surprising or not, it suggests that higher order partial derivatives may not depend on the order of differentiation. In effect this is true – provided all the derivatives in question are continuous (it should be noted, though, that without this assumption the equality does not necessarily hold).

4.2.1. Proposition. *Let $f(x,y)$ be a function such that the partial derivatives $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ are defined and continuous in a neighborhood of a point (x,y). Then we have*

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x}.$$

Proof. The idea of the proof is easy: we compute the second derivative in one step. This leads, as one easily sees, to computing the limit $\lim_{h\to 0} F(h)$ of the function

$$F(h) = \frac{f(x+h,y+h) - f(x,y+h) - f(x+h,y) + f(x,y)}{h^2}$$

and this is what we are going to do.

Setting

$$\begin{aligned}\varphi_h(y) &= f(x+h,y) - f(x,y) \quad \text{and}\\ \psi_k(x) &= f(x,y+k) - f(x,y),\end{aligned}$$

we obtain two expressions for $F(h)$:

$$F(h) = \frac{1}{h^2}(\varphi_h(y+h) - \varphi_h(y)) \quad \text{and} \quad F(h) = \frac{1}{h^2}(\psi_h(x+h) - \psi_h(x)).$$

Let us compute the first one. The function φ_h, which is a function of one variable y, has the derivative

$$\varphi_h'(y) = \frac{\partial f(x+h,y)}{\partial y} - \frac{\partial f(x,y)}{\partial y}$$

and hence by Lagrange Formula VI.2.2, we have

$$\begin{aligned}F(h) = \frac{1}{h^2}(\varphi_h(y+h) - \varphi_h(y)) = \frac{1}{h}\varphi_h'(y+\theta_1 h) =\\ = \frac{\partial f(x+h,y+\theta_1 h)}{\partial y} - \frac{\partial f(x,y+\theta_1 h)}{\partial y}.\end{aligned}$$

Then, using VI.2.2 again, we obtain

$$F(h) = \frac{\partial}{\partial x}\left(\frac{\partial f(x+\theta_2 h, y+\theta_1 h)}{\partial y}\right) \qquad (*)$$

for some θ_1, θ_2 between 0 and 1.

Similarly, computing $\frac{1}{h^2}(\psi_h(x+h) - \psi_h(x))$ we obtain

$$F(h) = \frac{\partial}{\partial y}\left(\frac{\partial f(x+\theta_4 h, y+\theta_2 h)}{\partial x}\right). \qquad (**)$$

Now since both $\frac{\partial}{\partial y}(\frac{\partial f}{\partial x})$ and $\frac{\partial}{\partial x}(\frac{\partial f}{\partial y})$ are continuous at the point (x, y), we can compute $\lim_{h\to 0} F(h)$ from either of the formulas $(*)$ or $(**)$ and obtain

$$\lim_{h\to 0} F(h) = \frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x}.$$

□

4.3. Iterating the interchanges allowed by 4.2.1 we obtain by an easy induction

Corollary. *Let a function f of n variables possess continuous partial derivatives up to the order k. Then the values of these drivatives depend only on the number of times a partial derivative is taken in each of the individual variables $x_1, \dots, x_n$.*

4.3.1. Thus, under the assumption of Theorem 4.3, we can write a general partial derivative of the order $r \le k$ as

$$\frac{\partial^r f}{\partial x_1^{r_1} \partial x_2^{r_2} \dots \partial x_n^{r_n}} \quad \text{with} \quad r_1 + r_2 + \dots + r_n = r$$

where, of course, $r_j = 0$ is allowed and indicates the absence of the symbol ∂x_j.

5. Composed functions and the Chain Rule.

Recall the proof of the rule of the derivative for composed functions in VI.2.2.1. It was based on the "total differential formula for one variable". By an analogous procedure we will obtain the following

5.1. Theorem. (Chain Rule in its simplest form) *Let $f(\mathbf{x})$ have a total differential at a point $\mathbf{a}$. Let real functions $g_k(t)$ have derivatives at a point b and let $g_k(b) = a_k$ for all $k = 1, \dots, n$. Put*

$$F(t) = f(\mathbf{g}(t)) = f(g_1(t), \dots, g_n(t)).$$

Then F has a derivative at b, and

$$F'(b) = \sum_{k=1}^{n} \frac{\partial f(\mathbf{a})}{\partial x_k} \cdot g_k'(b).$$

Proof. Applying the formula from 3.3 we get

$$\frac{1}{h}(F(b+h) - F(b)) = \frac{1}{h}(f(\mathbf{g}(b+h)) - f(\mathbf{g}(b)) =$$
$$= \frac{1}{h}(f(\mathbf{g}(b) + (\mathbf{g}(b+h) - \mathbf{g}(b))) - f(\mathbf{g}(b)) =$$
$$= \sum_{k=1}^{n} A_k \frac{g_k(b+h) - g_k(b)}{h} + \mu(\mathbf{g}(b+h) - \mathbf{g}(b)) \max_k \frac{|g_k(b+h) - g_k(b)|}{h}.$$

We have $\lim_{h\to 0} \mu(\mathbf{g}(b+h) - \mathbf{g}(b)) = 0$ since the functions g_k are continuous at b. Since the functions g_k have derivatives, the values $\max_k \frac{|g_k(b+h)-g_k(b)|}{h}$ are bounded in a sufficiently small neighborhood of 0. Thus, the limit of the last summand is zero and we have

$$\lim_{h\to 0} \frac{1}{h}(F(b+h) - F(b)) = \lim_{h\to 0} \sum_{k=1}^{n} A_k \frac{g_k(b+h) - g_k(b)}{h} =$$
$$= \sum_{k=1}^{n} A_k \lim_{h\to 0} \frac{g_k(b+h) - g_k(b)}{h} = \sum_{k=1}^{n} \frac{\partial f(\mathbf{a})}{\partial x_k} g_k'(b).$$

□

5.1.1. Corollary. (The Chain Rule) *Let $f(\mathbf{x})$ have a total differential at a point $\mathbf{a}$. Let real functions $g_k(t_1, \dots, t_r)$ have partial derivatives at $\mathbf{b} = (b_1, \dots, b_r)$ and let $g_k(\mathbf{b}) = a_k$ for all $k = 1, \dots, n$. Then the function*

$$(f \circ \mathbf{g})(t_1, \dots, t_r) = f(\mathbf{g}(t)) = f(g_1(t), \dots, g_n(t))$$

has all the partial derivatives at b, and

$$\frac{\partial(f\circ\mathbf{g})(\mathbf{b})}{\partial t_j}=\sum_{k=1}^{n}\frac{\partial f(\mathbf{a})}{\partial x_k}\cdot\frac{\partial g_k(\mathbf{b})}{\partial t_j}.$$

5.1.2. Note. Just possessing partial derivatives would not suffice. The assumption of the existence of total differential in 5.1 is essential and it is easy to see why. Recall the geometric intuition from 3.1 and the last paragraph of 3.5. The n-tuple of functions $\mathbf{g}=(g_1,\dots,g_n)$ represents a parametrized curve in D, and $f\circ\mathbf{g}$ is then a curve on the hypersurface S. The partial derivatives of f (or the tangent lines of S in the directions of the coordinate axes) have in general nothing to do with the behaviour on this curve.

5.2. The rules for multiplication and division as a consequence of the chain rule. As we have already mentioned, the Chain Rule (including its proof) is a more or less immediate extension of the composition rule in one variable. It may come as a surprise that it includes the rules for multiplication and division.

Consider $f(x,y)=xy$. Then $\frac{\partial f}{\partial x}=y$ and $\frac{\partial f}{\partial y}=x$ and hence

$$\begin{aligned}(u(t)v(t))'=f(u(t),v(t))'=&\frac{\partial f(u(t),v(t))}{\partial x}v'(t)+\frac{\partial f(u(t),v(t))}{\partial y}u'(t)=\\ &=v(t)\cdot u'(t)+u(t)\cdot v'(t).\end{aligned}$$

Similarly for $f(x,y)=\frac{x}{y}$ we have $\frac{\partial f}{\partial x}=\frac{1}{y}$ and $\frac{\partial f}{\partial y}=-\frac{x}{y^2}$ and consequently

$$\frac{u(t)}{v(t)}'=\frac{1}{v(t)}u'(t)-\frac{u(t)}{v^2(t)}=\frac{v(t)u'(t)-u(t)v'(t)}{v^2(t)}.$$

5.3. Chain rule for vector functions. Let us make one more step and consider in 5.1.1 a mapping $\mathbf{f}=(f_1,\dots,f_s):D\to\mathbb{E}_s$. Take its composition $\mathbf{f}\circ\mathbf{g}$ with a mapping $\mathbf{g}:D'\to\mathbb{E}_n$ (recall the convention in 1.4). Then we have

$$\frac{\partial(\mathbf{f}\circ\mathbf{g})}{\partial t_j}=\sum_k\frac{\partial f_i}{\partial x_k}\cdot\frac{\partial g_k}{\partial x_j}. \qquad (*)$$

It certainly has not escaped the reader's attention that the right hand side is the product of matrices

$$\left(\frac{\partial f_i}{\partial x_k}\right)_{i,k}\left(\frac{\partial g_k}{\partial x_j}\right)_{k,j}. \qquad (**)$$

Recall from linear algebra the role of matrices in describing linear functions $L : V_n \to V_m$. In particular recall that a composition of linear mappings results in the product of the associated matrices. Then the formulas $(*)$ resp. $(**)$ should not be surprising: they represent a fact to be expected, namely that the linear approximation of a composition $\mathbf{f}\circ\mathbf{g}$ is the composition of the linear approximations of $\mathbf{f}$ and $\mathbf{g}$.

5.3.1. Following the above comment, we may express the chain rule in matrix form as follows. For an $\mathbf{f} = (f_1, \dots, f_s) : U \to \mathbb{E}_s$, $D \subseteq \mathbb{E}_n$, define $D\mathbf{f}$ as the matrix

$$D\mathbf{f} = \left(\tfrac{\partial f_i}{\partial x_k}\right)_{i,k}.$$

Then we have

$$D(\mathbf{f} \circ \mathbf{g}) = D\mathbf{f} \cdot D\mathbf{g}.$$

More explicitly, in a concrete argument $\mathbf{t}$ we have

$$D(\mathbf{f} \circ \mathbf{g})(\mathbf{t}) = D(\mathbf{f}(\mathbf{g}))(\mathbf{t}) \cdot D\mathbf{g}(\mathbf{t}).$$

Compare it with the one variable rule

$$(f \circ g)'(t) = f'(g(t)) \cdot g'(t);$$

for 1×1 matrices we of course have $(a)(b) = (ab)$.

5.4. Lagrange Formula in several variables. Recall that a subset $U \subseteq \mathbb{E}_n$ is said to be *convex* if

$$\mathbf{x}, \mathbf{y} \in U \quad \Rightarrow \quad \forall t,\ 0 \le t \le 1,\ \ (1-t)\mathbf{x} + t\mathbf{y} = \mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in U.$$

5.4.1. Proposition. *Let f have continuous partial derivatives in a convex open set $U \subseteq \mathbb{E}_n$. Then for any two points $x, y \in D$, there exists a θ with $0 \le \theta \le 1$ such that*

$$f(\mathbf{y}) - f(\mathbf{x}) = \sum_{j=1}^{n} \frac{\partial f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x}))}{\partial x_j}(y_j - x_j).$$

Proof. Set $F(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. Then $F = f \circ \mathbf{g}$ where $\mathbf{g}$ is defined by $g_j(t) = x_j + t(y_j - x_j)$, and

$$F'(t) = \sum_{j=1}^{n} \frac{\partial f(\mathbf{g}(t))}{\partial x_j} g_j'(t) = \sum_{j=1}^{n} \frac{\partial f(\mathbf{g}(t))}{\partial x_j}(y_j - x_j).$$

Hence by VII.2.2,

$$f(\mathbf{y}) - f(\mathbf{x}) = F(1) - F(0) = F'(\theta)$$

which yields the statement. □

Note. The formula is often used in the form

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \sum_{j=1}^{n} \frac{\partial f(\mathbf{x}+\theta\mathbf{h})}{\partial x_j} h_j.$$

Compare this with the formula for total differential.

XV. Implicit Function Theorems

1. The task.

1.1. Suppose we have m real functions $F_k(x_1,\dots,x_n,y_1,\dots y_m)$, $k=1,\dots,m$, of $n+m$ variables each. Consider the system of equations

$$\begin{aligned}&F_1(x_1,\dots,x_n,y_1,\dots y_m)=0\\&\dots\quad\dots\quad\dots\\&F_m(x_1,\dots,x_n,y_1,\dots y_m)=0\end{aligned}$$

We would like to find a solution $y_1,\dots,y_m$. Better, using the convention of XIV.1, we have a system of m equations of m unknowns (the number n of the variablex x_j is inessential)

$$F_k(\mathbf{x},y_1,\dots y_m)=0,\quad k=1,\dots,m \tag{$*$}$$

and we are looking for solutions $y_k=f_k(\mathbf{x})$ $(=f(x_1,\dots,x_n))$.

1.2. Even in simplest cases we cannot expect to have necessarily a solution, not to speak of a unique one. Take for example the following single equation

$$F(x,y)=x^2+y^2-1=0.$$

For $|x|>1$ there is no y with $f(x,y)=0$. For $|x_0|<1$, we have in a sufficiently small open interval containing x_0 two solutions

$$f(x)=\sqrt{1-x^2}\quad\text{and}\quad g(x)=-\sqrt{1-x^2}.$$

This is better, but we have *two* values in each point, contradicting the definition of a function. To achieve uniqueness, we have to restrict not only the values of x, but *also the values of* y to an interval $(y_0-\Delta,y_0+\Delta)$ (where $F(x_0,y_0)=0$). That is, if we have a particular solution (x_0,y_0) we have a "window"

$$(x_0-\delta,x_0+\delta)\times(y_0-\Delta,y_0+\Delta)$$

through which we see a unique solution.

But in our example there is also the case $(x_0,y_0)=(1,0)$, where there is a unique solution, but no suitable window as above, since in every neighborhood of $(1,0)$, there are no solutions on the right hand side of $(1,0)$, and two solutions on the left hand side.

Note that in the critical points $(1,0)$ and $(-1,0)$ we have

$$\frac{\partial F}{\partial y}(1,0) = \frac{\partial F}{\partial y}(-1,0) = 0. \tag{$**$}$$

1.3. In this chapter we will show that for functions F_k with continuous partial derivatives the situation is not worse than in the example above:

- we will have to have some points $\mathbf{x}^0$, $\mathbf{y}^0$ such that $F_k(\mathbf{x}^0, \mathbf{y}^0) = 0$ to start with;
- with certain exceptions we then have "windows" $U \times V$ such that for $\mathbf{x} \in U$ there is precisely one $\mathbf{y} \in V$, that is, $y_k = f(x_1, \dots, x_n)$, satisfying the system of equations;
- and the exceptions are natural extensions of the condition associated with the $(**)$ above: instead of $\frac{\partial F}{\partial y}(x^0, y^0) \neq 0$ we will have $\frac{\mathsf{D}(\mathbf{F})}{\mathsf{D}(\mathbf{y})}(\mathbf{x}^0, \mathbf{y}^0) \neq 0$ for something related, called Jacobian.

Furthermore, the solutions will have continuous partial derivatives as long as the F_j have them.

2. One equation.

2.1. Theorem. *Let $F(\mathbf{x}, y)$ be a function of $n+1$ variables defined in a neighbourhood of a point $(\mathbf{x}^0, y_0)$. Let F have continuous partial derivatives up to the order $k \geq 1$ and let*

$$F(\mathbf{x}^0, y_0) = 0 \quad \textit{and} \quad \left| \frac{\partial F(\mathbf{x}^0, y_0)}{\partial y} \right| \neq 0.$$

Then there exist $\delta > 0$ and $\Delta > 0$ such that for every $\mathbf{x}$ with $||\mathbf{x} - \mathbf{x}^0|| < \delta$ there exists precisely one y with $|y - y_0| < \Delta$ such that

$$F(\mathbf{x}, y) = 0.$$

Furthermore, if we write $y = f(\mathbf{x})$ for this unique solution y, then the function

$$f : (x_1^0 - \delta, x_1^0 + \delta) \times \cdots \times (x_n^0 - \delta, x_n^0 + \delta) \to \mathbb{R}$$

has continuous partial derivatives up to the order k.

Before the proof. The reader is advised to reproduce the following proof as if there were just one real variable x. This simplification will make the procedure more transparent without losing anything of the ideas. The general $\mathbf{x}$ just needs more complicated notation which might slightly obscure some of the steps.

Proof. The norm $\|\mathbf{x}\|$ will be as in IV.3.2, that is $\max_i |x_i|$. Set

$$U(\gamma) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^0\| < \gamma\} \quad \text{and} \quad A(\gamma) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^0\| \leq \gamma\}$$

(the “window” we are seeking will turn out to be $U(\delta) \times (y_0 - \Delta, y_0 + \delta)$).

Without loss of generality let, say,

$$\frac{\partial F(\mathbf{x}^0, y_0)}{\partial y} > 0.$$

The first partial derivatives of F are continuous and $A(\delta)$ is closed and bounded and hence compact by XIII.7.6. Hence, by XIII.7.9 there exist $a > 0$, K, $\delta_1 > 0$ and $\Delta > 0$ such that for all $(\mathbf{x}, y) \in U(\delta_1) \times \langle y_0 - \Delta, y_0 + \Delta\rangle$ we have

$$\frac{\partial F(\mathbf{x}, y)}{\partial y} \geq a \quad \text{and} \quad \left|\frac{\partial F(\mathbf{x}, y)}{\partial x_i}\right| \leq K. \qquad (*)$$

I. *The function f* : Fix an $\mathbf{x} \in U(\delta_1)$, and define a function of one variable $y \in (y_0 - \Delta, y_0 + \Delta)$ by

$$\varphi_{\mathbf{x}}(y) = F(\mathbf{x}, y).$$

Then $\varphi'_{\mathbf{x}}(y) = \frac{\partial F(\mathbf{x},y)}{\partial y} > 0$ and hence

all $\varphi_{\mathbf{x}}(y)$ are increasing functions of y, and
$\varphi_{\mathbf{x}_0}(y_0 - \Delta) < \varphi_{\mathbf{x}_0}(y_0) = 0 < \varphi_{\mathbf{x}_0}(y_0 + \Delta)$.

By XIV.2.5 and XIV.3.4, F is continuous, and hence there is a δ, $0 < \delta \leq \delta_1$, such that

$$\forall \mathbf{x} \in U(\delta), \quad \varphi_{\mathbf{x}}(y_0 - \Delta) < 0 < \varphi_{\mathbf{x}}(y_0 + \Delta).$$

Now $\varphi_{\mathbf{x}}$ is increasing and hence one-to-one. Thus, by IV.3 there is precisely one $y \in (y_0 - \Delta, y_0 + \Delta)$ such that $\varphi_{\mathbf{x}}(y) = 0$ – that is, $F(\mathbf{x}, y) = 0$. Denote this y by $f(\mathbf{x})$.

Note this f is so far just a function; we know nothing about its properties, in particular, we do not know whether it is continuous or not.

II. *The first derivatives.* Fix an index j, abbreviate the sequence $x_1, \dots, x_{j-1}$ by $\mathbf{x}_b$ and the the sequence $x_{j+1}, \dots, x_n$ by $\mathbf{x}_a$; thus, we have

$$\mathbf{x} = (\mathbf{x}_b, x_j, \mathbf{x}_a).$$

We will compute $\frac{\partial f}{\partial x_j}$ as the derivative of $\psi(t) = f(\mathbf{x}_b, t, \mathbf{x}_a)$.

By XIV.5.4.1 we have

$$\begin{aligned} 0 &= F(\mathbf{x}_b, t+h, \mathbf{x}_a, \psi(t+h)) - F(\mathbf{x}_b, t, \mathbf{x}_a, \psi(t)) = \\ &= F(\mathbf{x}_b, t+h, \mathbf{x}_a, \psi(t) + (\psi(t+h) - \psi(t))) - F(\mathbf{x}_b, t, \mathbf{x}_a, \psi(t)) = \\ &= \frac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial x_j} h \\ &+ \frac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial y} (\psi(t+h) - \psi(t)) \end{aligned}$$

and hence

$$\psi(t+h) - \psi(t) = -h \cdot \frac{\dfrac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial x_j}}{\dfrac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial y}} \qquad (**)$$

for some θ between 0 and 1.

Now we can infer that f is continuous. From $(*)$ we obtain

$$|\psi(t+h) - \psi(t)| \le |h| \cdot \left| \frac{K}{a} \right|$$

Using this fact we can compute from $(**)$ further

$$\begin{aligned} &\lim_{h\to 0} \frac{\psi(t+h) - \psi(t)}{h} = \\ &= -\lim_{h\to 0} \frac{\dfrac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial x_j}}{\dfrac{\partial F(\mathbf{x}_b, t+\theta h, \mathbf{x}_a, \psi(t) + \theta(\psi(t+h) - \psi(t)))}{\partial y}} = -\frac{\dfrac{\partial F(\mathbf{x}_b, t, \mathbf{x}_a, \psi(t))}{\partial x_j}}{\dfrac{\partial F(\mathbf{x}_b, t, \mathbf{x}_a, \psi(t))}{\partial y}} \end{aligned}$$

III. *The higher derivatives.* Note that we have not only proved the *existence* of the first derivative of f, but also the formula

$$\frac{\partial f(\mathbf{x})}{\partial x_j} = -\frac{\partial F(\mathbf{x}, f(\mathbf{x}))}{\partial x_j} \cdot \left(\frac{\partial F(\mathbf{x}, f(\mathbf{x}))}{\partial y} \right)^{-1}. \qquad (***)$$

From this we can inductively compute the higher derivatives of f (using the standard rules of differentiation) as long as the derivatives

$$\frac{\partial^r F}{\partial x_1^{r_1} \cdots \partial x_n^{r_n} \partial y^{r_{n+1}}}$$

exist and are continuous. $\square$

2.2. We have obtained the formula $(***)$ as a by-product of the proof that f has a derivative (it was useful further on, but this is not the point). Note that if we knew beforehand that f had one we could deduce (5.2.3) immediately from the Chain Rule. In effect, we have

$$0 \equiv F(\mathbf{x}, f(\mathbf{x}));$$

taking a derivative of both sides we obtain

$$0 = \frac{\partial F\mathbf{x}, f(\mathbf{x}))}{\partial x_j} + \frac{\partial F\mathbf{x}, f(\mathbf{x}))}{\partial y} \cdot \frac{\partial f(\mathbf{x})}{\partial x_j}.$$

Differentiating further, we obtain inductively linear equations from which we can compute the values od all the derivatives guaranteed by the theorem.

2.3. Note. The solution f in 2.1 has as many derivatives as the initial F – provided F has at least the first ones. One sometimes thinks of the function itself as of its 0-th derivative. The theorem, however, *does not guarantee a continuous solution f of an equation $F(x, f(x)) = 0$ with continuous F*. We had to use the first derivatives already for the existence of the f.

3. A warm-up: two equations.

3.1. Consider a pair of equations

$$F_1(\mathbf{x}, y_1, y_2) = 0,$$
$$F_2(\mathbf{x}, y_1, y_2) = 0$$

and try to find a solution $y_i = f_i(\mathbf{x})$, $i = 1, 2$, in a neighborhood of a point $(\mathbf{x}^0, y_1^0, y_2^0)$ (at which the equalities hold). We will apply the "substitution method" based on Theorem 2.1. First think of the second equation as an equation for the y_2; in a neighborhood of $(\mathbf{x}^0, y_1^0, y_2^0)$ we then obtain y_2 as a function $\psi(\mathbf{x}, y_1)$. Substitute this into the first equation to obtain

$$G(\mathbf{x}, y_1) = F_1(\mathbf{x}, y_1, \psi(\mathbf{x}, y_1));$$

if we find a solution $y_1 = f_1(\mathbf{x})$ in a neighborhood of $(\mathbf{x}^0, y_1^0)$ we can substitute it into ψ and obtain $y_2 = f_2(\mathbf{x}) = \psi(\mathbf{x}, f_1(\mathbf{x}))$.

3.2. Now we have a solution let us summarize what exactly we have assumed:

– First we had to have the continuous partial derivatives of the functions F_i.

– Then, to be able to obtain ψ by 2.1 the way we did, we needed to have

$$\frac{\partial F_2}{\partial y_2}(\mathbf{x}^0, y_1^0, y_2^0) \neq 0. \tag{$*$}$$

– Finally, we also need to have (use the Chain Rule)

$$0 \neq \frac{\partial G}{\partial y_1}(\mathbf{x}^0, x^0) = \frac{\partial F_1}{\partial y_1} + \frac{\partial F_1}{\partial y_2}\frac{\partial \psi}{\partial y_1} \neq 0. \tag{$**$}$$

Use the formula for the first derivative

$$\frac{\partial \psi}{\partial y_1} = -\left(\frac{\partial F_1}{\partial y_2}\right)^{-1} \frac{\partial F_2}{\partial y_1}$$

from the proof of 2.1 and transform $(**)$ to

$$\left(\frac{\partial F_1}{\partial y_2}\right)^{-1} \left(\frac{\partial F_1}{\partial y_1}\frac{\partial F_2}{\partial y_2} - \frac{\partial F_1}{\partial y_2}\frac{\partial F_2}{\partial y_1}\right) \neq 0,$$

that is,

$$\frac{\partial F_1}{\partial y_1}\frac{\partial F_2}{\partial y_2} - \frac{\partial F_1}{\partial y_2}\frac{\partial F_2}{\partial y_1} \neq 0.$$

This is a familiar formula, namely that for a determinant. Thus we have in fact assumed that

$$\begin{vmatrix} \dfrac{\partial F_1}{\partial y_1}, & \dfrac{\partial F_1}{\partial y_2} \\ \dfrac{\partial F_2}{\partial y_1}, & \dfrac{\partial F_2}{\partial y_2} \end{vmatrix} = \det\left(\frac{\partial F_i}{\partial y_j}\right)_{i,j} \neq 0.$$

And this condition suffices: if we assume that this determinant is non-zero we have *either*

$$\frac{\partial F_2}{\partial y_2}(\mathbf{x}^0, y_1^0, y_2^0) \neq 0$$

and/or

$$\frac{\partial F_2}{\partial y_1}(\mathbf{x}^0, y_1^0, y_2^0) \neq 0,$$

so if the latter holds, we can start by solving $F_2(\mathbf{x}, y_1, y_2) = 0$ for y_1 instead of y_2.

4. The general system.

4.1. Jacobi determinant. Let $\mathbf{F}$ be a sequence of functions

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = (F_1(\mathbf{x}, y_1, \dots, y_m), \dots, F_m(\mathbf{x}, y_1, \dots, y_m)).$$

For this $\mathbf{F}$ and the sequence $\mathbf{y} = (y_1, \dots, y_m)$ define the *Jacobi determinant* (briefly, the *Jacobian*)

$$\frac{\mathsf{D}(\mathbf{F})}{\mathsf{D}(\mathbf{y})} = \det \left(\frac{\partial F_i}{\partial y_j} \right)_{i,j=1,\dots,m}$$

Note that if $m = 1$, that is if we have one function F and one y, we have

$$\frac{\mathsf{D}(F)}{\mathsf{D}(y)} = \frac{\partial F}{\partial y}.$$

4.2. Theorem. *Let $F_i(\mathbf{x}, y_1, \dots, y_m)$, $i = 1, \dots, m$, be functions of $n+m$ variables with continuous partial derivatives up to an order $k \geq 1$. Let*

$$\mathbf{F}(\mathbf{x}^0, \mathbf{y}^0) = \mathbf{o}$$

and let

$$\frac{\mathsf{D}(\mathbf{F})}{\mathsf{D}(\mathbf{y})}(\mathbf{x}^0, \mathbf{y}^0) \neq 0.$$

Then there exist $\delta > 0$ and $\Delta > 0$ such that for every

$$\mathbf{x} \in (x_1^0 - \delta, x_1^0 + \delta) \times \cdots \times (x_n^0 - \delta, x_n^0 + \delta)$$

there exists precisely one

$$\mathbf{y} \in (y_1^0 - \Delta, y_1^0 + \Delta) \times \cdots \times (y_m^0 - \Delta, x_m^0 + \Delta)$$

such that

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = 0.$$

Furthermore, if we write this $\mathbf{y}$ *as a vector function* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$*, then the functions* f_i *have continuous partial derivatives up to the order* k*.*

Before the proof. The procedure will follow the idea of the substitution method from Section 3. Only, we will have to do something more with determinants (but this is linear algebra, well known to the reader) and at the end we will have to tidy up the Δ and δ (which we have so far neglected).

Proof will be done by induction. The statement holds for $m = 1$ (see 2.1). Now let it hold for m, and let us have a system of equations

$$F_i(\mathbf{x}, \mathbf{y}),\ i = 1, \dots, m + 1$$

satisfying the assumptions (note that the unknown vector $\mathbf{y}$ is $m+1$-dimensional, too). Then, in particular, in the Jacobian determinant we cannot have a column consisting entirely of zeros, and hence, after possibly reshufling the F_i's, we can assume that

$$\frac{\partial F_{m+1}}{\partial y_{m+1}}(\mathbf{x}^0, \mathbf{y}^0) \neq 0.$$

Write $\widetilde{\mathbf{y}} = (y_1, \dots, y_m)$; then, by the induction hypothesis, we have $\delta_1 > 0$ and $\Delta_1 > 0$ such that for

$$(\mathbf{x}, \widetilde{\mathbf{y}}) \in (x_1^0 - \delta_1, x_1^0 + \delta_1) \times \cdot \times (x_n^0 - \delta_1, x_1^n + \delta_1) \times \cdots \times (y_m^0 - \delta_1, y_m^0 + \delta_1),$$

there exists precisely one $y_{m+1} = \psi(\mathbf{x}, \widetilde{\mathbf{y}})$ satisfying

$$F_{m+1}(\mathbf{x}, \widetilde{\mathbf{y}}, y_{m+1}) = 0 \quad \text{and} \quad |y_{m+1} - y_{m+1}^0] < \Delta_1.$$

This ψ has continuous partial derivatives up to the order k and hence so have the functions

$$G_i(\mathbf{x}, \widetilde{\mathbf{y}}) = F_i(\mathbf{x}, \widetilde{\mathbf{y}}, \psi(\mathbf{x}, \widetilde{\mathbf{y}})),\ i = 1, \dots m + 1$$

(the last G_{m+1} is constant 0). By the Chain Rule we obtain

$$\frac{\partial G_j}{\partial y_i} = \frac{\partial F_j}{\partial y_i} + \frac{\partial F_j}{\partial y_{m+1}} \frac{\partial \psi}{\partial y_i}. \tag{$*$}$$

Now consider the determinant

$$\frac{\mathsf{D}(\mathbf{F})}{\mathsf{D}(\mathbf{y})} = \begin{vmatrix} \dfrac{\partial F_1}{\partial y_1}, & \ldots, & \dfrac{\partial F_1}{\partial y_m}, & \dfrac{\partial F_1}{\partial y_{m+1}} \\ \ldots, & \ldots, & \ldots, & \ldots \\ \dfrac{\partial F_m}{\partial y_1}, & \ldots, & \dfrac{\partial F_m}{\partial y_m}, & \dfrac{\partial F_m}{\partial y_{m+1}} \\ \dfrac{\partial F_{m+1}}{\partial y_1}, & \ldots, & \dfrac{\partial F_{m+1}}{\partial y_m}, & \dfrac{\partial F_{m+1}}{\partial y_{m+1}} \end{vmatrix}.$$

Multiply the last column by $\frac{\partial \psi}{\partial y_i}$ and add it to the ith one. By ($*$), taking into account that $G_{m+1} \equiv 0$ and hence

$$\frac{\partial G_{m+1}}{\partial y_i} = \frac{\partial F_{m+1}}{\partial y_i} + \frac{\partial F_{m+1}}{\partial y_{m+1}} \frac{\partial \psi}{\partial y_i} = 0,$$

we obtain

$$\frac{\mathsf{D}(\mathbf{F})}{\mathsf{D}(\mathbf{y})} = \begin{vmatrix} \dfrac{\partial G_1}{\partial y_1}, & \ldots, & \dfrac{\partial G_1}{\partial y_m}, & \dfrac{\partial F_1}{\partial y_{m+1}} \\ \ldots, & \ldots, & \ldots, & \ldots \\ \dfrac{\partial G_m}{\partial y_1}, & \ldots, & \dfrac{\partial G_m}{\partial y_m}, & \dfrac{\partial F_m}{\partial y_{m+1}} \\ 0, & \ldots, & 0, & \dfrac{\partial F_{m+1}}{\partial y_{m+1}} \end{vmatrix} = \frac{\partial F_{m+1}}{\partial y_{m+1}} \cdot \frac{D(G_1, \ldots, G_m)}{D(y_1, \ldots, y_m)}.$$

Thus,

$$\frac{\mathsf{D}(G_1, \ldots, G_m)}{\mathsf{D}(y_1, \ldots, y_m)} \neq 0$$

and hence by the induction hypthesis there are $\delta_2 > 0$, $\Delta_2 > 0$ such that for $|x_i - x_i^0| < \delta_2$ there is a uniquely determined $\widetilde{\mathbf{y}}$ with $|y_i - y_i^0| < \Delta_2$ such that

$$G_i(\mathbf{x}, \widetilde{\mathbf{y}}) = 0 \quad \text{for} \quad i = 1, \dots, m$$

and that the resulting $f_i(\mathbf{x})$ have continuous partial derivatives up to the order k. Finally defining

$$f_{i+1}(\mathbf{x}) = \psi(\mathbf{x}, f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$$

we obtain a solution $\mathbf{f}$ of the original system of equations $\mathbf{F}(\mathbf{x}, \mathbf{y}) = 0$.

To finish the proof we need the constraints $\|\mathbf{x} - \mathbf{x}^0\| < \delta$ and $\|\mathbf{y} - \mathbf{y}^0\| < \Delta$ within which the solution is correct (that is, unique).

Choose $0 < \Delta \le \delta_1, \Delta_1, \Delta_2$ and then $0 < \delta < \delta_1, \delta_2$ and sufficiently small so that for $|x_1 - x_i^0| < \delta$ one has $|f_j(\mathbf{x}) - f_j(\mathbf{x}^0)| < \Delta$ (the last condition makes sure to have in the Δ-interval *at least one* solution). Now let

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{o}, \quad \text{and} \quad \|\mathbf{x} - \mathbf{x}^0\| < \delta \text{ and } \|\mathbf{y} - \mathbf{y}^0\| < \Delta. \qquad (**)$$

We have to prove that then necessarily $y_i = f_i(\mathbf{x})$ for all i. Since $|x_i - x_i^0| < \delta \le \delta_1$ for $i = 1, \dots, n$, $|y_i - y_i^0| < \Delta \le \delta_1$ for $i = 1, \dots, m$ and $|y_{m+1} - y_{m+1}^0| < \Delta \le \Delta_1$ we have necessarily $y_{m+1} = \psi(\mathbf{x}, \widetilde{\mathbf{y}})$. Thus, by $(**)$,

$$\mathbf{G}(\mathbf{x}, \widetilde{\mathbf{y}}) = \mathbf{o}$$

and since $|x_i - x_i^0| < \delta \le \delta_2$ and $|y_i - y_i^0| < \Delta \le \Delta_2$ we have indeed $y_i = f_i(\mathbf{x})$. □

5. Two simple applications: regular mappings

5.1. Let $U \subseteq \mathbb{E}_n$ be an open set. Let f_i, $i = 1, \dots, n$, be mappings with continuous partial derivatives (and hence continuous themselves). The resulting (continuous) mapping $\mathbf{f} = (f_1, \dots, f_n) : U \to \mathbb{E}_n$ is said to be *regular* if

$$\frac{\mathsf{D}(\mathbf{f})}{\mathsf{D}(\mathbf{x})}(\mathbf{x}) \neq 0$$

for all $\mathbf{x} \in U$.

5.2. Recall that continuous mappings are characterized by preserving openness (or closedness) by *preimage* (recall XIII.3.7). Also recall the very special fact II.7.10) that if the domain is compact, also *images* of closed sets are closed. For regular maps we have something similar.

Proposition. *If* $\mathsf{f} : U \to \mathbb{E}_n$ *is regular then the image* $\mathsf{f}[V]$ *of every open* $V \subseteq U$ *is open.*

Proof. Let $f(\mathsf{x}^0) = \mathsf{y}^0$. Define $\mathsf{F} : V \times \mathbb{E}_n \to \mathbb{E}_n$ by setting

$$F_i(\mathsf{x}, \mathsf{y}) = f_i(\mathsf{x}) - y_i. \qquad (*)$$

then $\mathsf{F}(\mathsf{x}^0, \mathsf{y}^0) = \mathsf{o}$ and $\frac{\mathsf{D(F)}}{\mathsf{D(x)}} \neq 0$, and hence we can apply 4.2 to obtain $\delta > 0$ and $\Delta > 0$ such that for every y with $\|\mathsf{y} - \mathsf{y}^0\| < \delta$, there exists a x such that $\|\mathsf{x} - \mathsf{x}^0\| < \Delta$ and $F_i(\mathsf{x}, \mathsf{y}) = f_i(\mathsf{x}) - y_i = 0$. This means that we have $\mathsf{f}(\mathsf{x}) = \mathsf{y}$ (do not get confused by the reversed roles of the x_i and the y_i: the y_i are here the independent variables), and

$$\Omega(\mathsf{y}^0, \delta) = \{\mathsf{y} \mid \|\mathsf{y} - \mathsf{y}^0\| < \delta\} \subseteq \mathsf{f}[V]. \qquad \square$$

5.3. Proposition. *Let* $\mathsf{f} : U \to \mathbb{E}_n$ *be a regular mapping. Then for each* $\mathsf{x}^0 \in U$ *there exists an open neighborhood* V *such that the restriction* $\mathsf{f}|V$ *is one-to-one. Moreover, the mapping* $\mathsf{g} : f[V] \to \mathbb{E}_n$ *inverse to* $\mathsf{f}|V$ *is regular.*

Proof. We will use again the mapping $\mathsf{F} = (F_1, \dots, F_n)$ from $(*)$. For a sufficiently small $\Delta > 0$ we have precisely one $\mathsf{x} = \mathsf{g}(\mathsf{y})$ such that $\mathsf{F}(\mathsf{x}, \mathsf{y}) = 0$ and $\|\mathsf{x} - \mathsf{x}^0\| < \Delta$. This g has, furthermore, continuous partial derivatives. By XIV.5.3 we have

$$D(\mathrm{id}) = D(\mathsf{f} \circ \mathsf{g}) = D(\mathsf{f}) \cdot D(\mathsf{g}).$$

By the Chain Rule (and the theorem on product of determinants)

$$\frac{\mathsf{D(f)}}{\mathsf{D(x)}} \cdot \frac{D(\mathsf{g})}{D(\mathsf{y})} = \det D(\mathsf{f}) \cdot \det D(\mathsf{g}) = 1$$

and hence for each $\mathsf{y} \in \mathsf{f}[V]$, $\frac{\mathsf{D(g)}}{\mathsf{D(y)}}(\mathsf{y}) \neq 0$. $\square$

5.3.1. Corollary. *A one-to-one regular mapping* $\mathsf{f} : U \to \mathbb{E}_n$ *has a regular inverse* $\mathsf{g} : \mathsf{f}[U] \to \mathbb{E}_n$.

6. Local extremes and extremes with constraints.

6.1. Recall looking for local extremes of a real-valued function of one real variable f in VII.1. If f was defined on an interval $\langle a, b\rangle$ and had a derivative in (a, b) we learned by an easy application of the formula VI.1.5 that in the local extremes the derivative had to be zero. Then it sufficed to check the values in the boundary points a and b and we had a complete list of candidates.

Now consider the local extremes of a function of several real variables. Pinpointing possible local extremes *in the interior of its domain* is equally easy: similarly as in the function of one variable we deduce from the total differential formula (but we really do not even need that, partial derivatives would suffice) that at the points of local extreme $\mathbf{a}$, we must have

$$\frac{\partial f}{\partial x_i}(\mathbf{a}) = 0, \quad i = 1, \dots, n. \tag{$*$}$$

But the boundary is now another matter. Typically it does not consist of finitely many isolated points to be checked one at a time.

6.1.1. Example. Suppose we want to find the local extremes of the function $f(x, y) = x + 2y$ on the ball $B = \{(x, y) \,|\, x^2 + y^2 \leq 1\}$. The domain B is compact, and hence the function f certainly attains a minimum and a maximum on B. They cannot be in the interior of B: we have constantly, $\frac{\partial f}{\partial x} = 1$ and $\frac{\partial f}{\partial y} = 2$; thus, the extremes must be located somewhere in the infinte set $\{(x, y) \,|\, x^2 + y^2 = 1\}$, and the rule $(*)$ is of no use.

6.2. Hence we will try to find local extremes of a function $f(x_1, \dots, x_n)$ *subject to certain constraints* $g_i(x_1, \dots, x_n) = 0$, $i = 1, \dots, k$. We have the following

Theorem. *Let $f, g_1, \dots, g_k$ be real functions defined in an open set $D \subseteq \mathbb{E}_n$, and let them have continuous partial derivatives. Suppose that the rank of the matrix*

$$M = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1}, & \dots, & \dfrac{\partial g_1}{\partial x_n} \\ \dots, & \dots, & \dots \\ \dfrac{\partial g_k}{\partial x_1}, & \dots, & \dfrac{\partial g_k}{\partial x_n} \end{pmatrix}$$

is the largest possible, that is k, at each point of D.

If the function f achieves at a point $\mathbf{a} = (a_1, \dots, a_n)$ a local extreme subject to the constraints

$$g_i(x_1, \dots, x_n) = 0, \quad i = 1, \dots, k$$

then there exist numbers $\lambda_1, \dots, \lambda_k$ such that for each $i = 1, \dots, n$ we have

$$\frac{\partial f(\mathbf{a})}{\partial x_i} + \sum_{j=1}^{k} \lambda_j \cdot \frac{\partial g_j(\mathbf{a})}{\partial x_i} = 0.$$

Notes. 1. The functions f, g_i were assumed to be defined in an open D so that we can take derivatives whenever we need them. In typical applications one works with functions that can be extended to an open set containing the area in question.

2. The force of the statement is in asserting the existence of $\lambda_1, \dots, \lambda_k$ that satisfy *more than* k equations. See the solution of 6.1.1 in 6.3 below.

3. The numbers λ_i are known as *Lagrange multipliers*.

Proof. From linear algebra we know that a matrix M has rank k iff at least one of the $k \times k$ submatrices of M is regular (and hence has a non-zero determinant). Without loss of generality we can assume that at the extremal point we have

$$\begin{vmatrix} \frac{\partial g_1}{\partial x_1}, & \dots, & \frac{\partial g_1}{\partial x_k} \\ \dots, & \dots, & \dots \\ \frac{\partial g_k}{\partial x_1}, & \dots, & \frac{\partial g_k}{\partial x_k} \end{vmatrix} \neq 0. \tag{1}$$

If this holds, we have by the Implicit Function Theorem in a neighborhood of the point $\mathbf{a}$ functions $\phi_i(x_{k+1}, \dots, x_n)$ with continuous partial derivatives such that (we write $\widetilde{\mathbf{x}}$ for $(x_{k+1}, \dots, x_n)$)

$$g_i(\phi_1(\widetilde{\mathbf{x}}), \dots, \phi_k(\widetilde{\mathbf{x}}), \widetilde{\mathbf{x}}) = 0 \quad \text{for} \quad i = 1, \dots, k,$$

Thus, a local maximum or a local minimum of $f(\mathbf{x})$ at $\mathbf{a}$, subject to the given constraints, implies the corresponding extreme property (without constraints) of the function

$$F(\widetilde{\mathbf{x}}) = f(\phi_1(\widetilde{\mathbf{x}}), \dots, \phi_k(\widetilde{\mathbf{x}}), \widetilde{\mathbf{x}}),$$

at $\widetilde{\mathbf{a}}$, and hence by 5.1

$$\frac{\partial F(\widetilde{\mathbf{a}})}{\partial x_i} = 0 \quad \text{for} \quad i = k+1, \dots, n,$$

that is, by the Chain Rule,

$$\sum_{r=1}^{k} \frac{\partial f(\mathbf{a})}{\partial x_r} \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} + \frac{\partial f(\mathbf{a})}{\partial x_i} \quad \text{for} \quad i = k+1, \dots, n. \tag{2}$$

Taking derivatives of the constant functions $g_i(\phi_1(\widetilde{\mathbf{x}}), \dots, \phi(\widetilde{\mathbf{x}}), \widetilde{\mathbf{x}}) = 0$ we obtain for $j = 1, \dots, k$,

$$\sum_{r=1}^{k} \frac{\partial g_j(\mathbf{a})}{\partial x_r} \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} + \frac{\partial g_j(\mathbf{a})}{\partial x_i} \quad \text{for} \quad i = k+1, \dots, n. \tag{3}$$

Now we will use (1) again, for another purpose. Because of the rank of the matrix, the system of linear equations

$$\frac{\partial f(\mathbf{a})}{\partial x_i} + \sum_{j=1}^{n} \lambda_j \cdot \frac{\partial g_j(\mathbf{a})}{\partial x_i} = 0, \quad i = 1, \dots, k,$$

has a unique solution $\lambda_1, \dots, \lambda_k$. These are the equalities from the statement, but so far for $i \leq k$ only. It remains to be shown that the same equalities hold also for $i > k$. In effect, by (2) and (3), for $i > k$ we obtain

$$\frac{\partial f(\mathbf{a})}{\partial x_i} + \sum_{j=1}^{n} \lambda_j \cdot \frac{\partial g_j(\mathbf{a})}{\partial x_i} = -\sum_{r=1}^{k} \frac{\partial f(\mathbf{a})}{\partial x_r} \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} - \sum_{j=1}^{k} \lambda_j \sum_{r=1}^{k} \frac{\partial g_j(\mathbf{a})}{\partial x_r} \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} =$$
$$-\sum_{r=1}^{n} \left(\frac{\partial f(\mathbf{a})}{\partial x_i} + \sum_{j=1}^{n} \lambda_j \cdot \frac{\partial g_j(\mathbf{a})}{\partial x_i} \right) \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} = -\sum_{r=1}^{n} 0 \cdot \frac{\partial \phi_r(\widetilde{\mathbf{a}})}{\partial x_i} = 0. \quad \square$$

6.3. Solution of 6.1.1. We have $\frac{\partial f}{\partial x} = 1$ and $\frac{\partial f}{\partial y} = 2$, $g(x, y) = x^2 + y^2 - 1$ and hence $\frac{\partial g}{\partial x} = 2x$ and $\frac{\partial g}{\partial y} = 2y$. There is *one* λ that satisfies *two* equations

$$1 + \lambda \cdot 2x = 0 \quad \text{and} \quad 2 + \lambda \cdot 2y = 0.$$

This is possible only if $y = 2x$. Thus, as $x^2 + y^2 = 1$ we obtain $5x^2 = 1$ and hence $x = \pm\frac{1}{\sqrt{5}}$; this localizes the extremes to $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$ and $(\frac{-1}{\sqrt{5}} \frac{-2}{\sqrt{5}})$.

6.4. The comstraints g_i do not necessarily come from describing boundaries. Here is an example of another nature.

Let us ask the question which rectangular parallelepiped of a given surface area has the largest volume. Denoting the lengths of the edges by $x_1, \dots, x_n$, the surface area is

$$S(x_1, \dots, x_n) = 2x_1 \cdots x_n \left(\frac{1}{x_1} + \cdots + \frac{1}{x_n} \right)$$

and the volume is

$$V(x_1, \dots, x_n) = x_1 \cdots x_n.$$

Hence

$$\begin{aligned} \frac{\partial V}{\partial x_i} &= \frac{1}{x_i} \cdot x_1 \cdots x_n \quad \text{and} \\ \frac{\partial S}{\partial x_i} &= \frac{2}{x_i}(x_1 \cdots x_n) \left(\frac{1}{x_1} + \cdots + \frac{1}{x_n} \right) - 2x_1 \cdots x_n \frac{1}{x_i^2}. \end{aligned}$$

If we denote $y_i = \frac{1}{x_i}$ and $s = y_1 + \cdots + y_n$, and divide the equation from the theorem by $x_1 \cdots x_n$, we obtain

$$2y_i(s - y_i) + \lambda y_i = 0 \quad \text{resulting in} \quad y_i = s + \frac{\lambda}{2}.$$

Thus, all the x_i are equal and the solution is the cube.

.

XVI. Multivariable Riemann integral

The idea of Riemann integral in several variables is the same as that in one variable. The only difference is that we will have n-dimensional intervals instead of the standard ones, and that the partitions will have to divide such intervals in all dimensions so that the resulting intervals of the partition will not be so tidily ordered as the small intervals $\langle t_0, t_1\rangle$,$\langle t_1, t_2\rangle$,$\dots$. But a finite sum is a finite sum and we will see that the ordering is not important.

What is new is the Fubini theorem (Section 4) allowing to compute multivariable integrals using integrals of one variable. All what will be done before that will be modifications of facts from Chapter XI.

1. Intervals and partitions.

1.1. In this chapter, an *n-dimensional compact interval* is a product

$$J = \langle a_1, b_1\rangle \times \dots \times \langle a_n, b_n\rangle$$

(such a J is indeed compact, recall XIII.7.6); if there will be no danger of confusion we will simply speak of an *interval*. We will also speak of *bricks*, in particular when they will be parts of bigger intervals.

A *partition* of J is a sequence $P = (P^1, \dots, P^n)$ of partitions

$$P^j: \ a_j = t_{j0} < t_{j1} < \dots < t_{j,n_j-1} < t_{j,n_j} = b_j, \quad j = 1, \dots n.$$

The intervals

$$\langle t_{1,i_1}, t_{1,i_1+1}\rangle \times \dots \times \langle t_{n,i_n}, t_{n,i_n+1}\rangle$$

will be called the *bricks* of P and the set of all the bricks of P will be denoted by

$$\mathcal{B}(P).$$

1.2. The volume of an interval $J = \langle a_1, b_1\rangle \times \dots \times \langle a_n, b_n\rangle$ is the number

$$\mathsf{vol}(J) = (b_1 - a_1)(b_2 - a_2)\cdots(b_n - a_n).$$

Since distinct bricks in $\mathcal{B}(P)$ obviously meet in a set of volume 0 (recall XI.1 applied for not necessaruly planar figures) we have an

1.2.1. Observation. $\mathsf{vol}(J) = \sum\{\mathsf{vol}(B) \mid B \in \mathcal{B}(J)\}$.

1.3. Mesh of a partition. The diameter of $J = \langle a_1, b_1\rangle \times \cdots \times \langle a_n, b_n\rangle$ is

$$\mathsf{diam}(J) = \max_i (b_i - a_i)$$

and the *mesh* of a partition P is

$$\mu(P) = \max\{\mathsf{diam}(B) \mid B \in \mathcal{B}(P)\}.$$

1.4. Refinement. Recall XI.2.2, A partition $Q = (Q^1, \dots, Q^n)$ *refines* a partition $P = (P^1, \dots, P^n)$ if every Q^j refines P^j.

Considering the segments $t_{j,k-1} = t'_{j,l} < t'_{j,l+1} < \cdots < t'_{j,l+r} = t_{j,k}$ of the finer partition Q we obtain

1.4.1. Observation. *A refinement Q of a partition P induces partitions*

$$Q_B \quad \textit{of the bricks} \quad B \in \mathcal{B}(P)$$

and we have a disjoint union

$$\mathcal{B}(Q) = \bigcup\{\mathcal{B}(Q_B) \mid B \in \mathcal{B}(P)\}.$$

1.5. Observation. *For any two partitions P, Q of an n-dimensional compact interval J there is a common refinement.*

(Indeed, recall the proof of XI.2.3.2. If $P = (P^1, \dots, P^n)$ and $Q = (Q^1, \dots, Q^n)$ are partitions of J consider the partition $R = (R^1, \dots, R^n)$ with R^j common refinements of P^j and Q^j.)

2. Lower and upper sums. Definition of Riemann integral.

2.1. Let f be a bounded real function on an an n-dimensional compact interval J and let $B \subseteq J$ be an an n-dimensional compact subinterval of J (a brick). Set

$$m(f, B) = \inf\{f(\mathbf{x}) \mid \mathbf{x} \in B\} \quad \text{and} \quad M(f, B) = \sup\{f(\mathbf{x}) \mid \mathbf{x} \in B\}.$$

We have

2.1.1. Fact. *$m(f,B) \le M(f,B)$ and if $C \subseteq B$ then*

$$m(f,C) \ge m(f,B) \quad \text{and} \quad M(f,C) \le M(f,B).$$

($\{f(\mathbf{x}) \,|\, \mathbf{x} \in C\}$ is a subset of $\{f(\mathbf{x}) \,|\, \mathbf{x} \in B\}$ and hence each lower (upper) bound of the latter is a lower (upper) bound of the former.)

2.2. Let P be a partition of an interval J and let $f : J \to \mathbb{R}$ be a bounded function. Set

$$\begin{aligned} s_J(f,P) &= \sum\{m(f,B) \cdot \mathsf{vol}(B) \,|\, B \in \mathcal{B}(P)\} \quad \text{and} \\ S_J(f,P) &= \sum\{M(f,B) \cdot \mathsf{vol}(B) \,|\, B \in \mathcal{B}(P)\}. \end{aligned}$$

The subscript J will be usually omitted

2.2.1. Proposition. *Let a partition Q refine P. Then*

$$s(f,Q) \ge s(f,P) \quad \text{and} \quad S(f,Q) \le S(f,P).$$

Proof. We have (the statement used is indicated over $=$ or $\le$)

$$\begin{aligned} S(f,Q) = &\sum\{M(f,C) \cdot \mathsf{vol}(C) \,|\, C \in \mathcal{B}(Q)\} \overset{1.4.1}{=} \\ &\overset{1.4.1}{=} \sum\{M(f,C) \cdot \mathsf{vol}(C) \,|\, C \in \text{(disjoint)} \bigcup\{\mathcal{B}(Q_B) \,|\, B \in \mathcal{B}(P)\}\} = \\ &= \sum\{\sum\{M(f,C) \cdot \mathsf{vol}(C) \,|\, C \in \mathcal{B}(Q_B)\} \,|\, B \in \mathcal{B}(P)\} \overset{2.1.1}{\le} \\ &\overset{2.1.1}{\le} \sum\{\sum\{M(f,B) \cdot \mathsf{vol}(C) \,|\, C \in \mathcal{B}(Q_B)\} \,|\, B \in \mathcal{B}(P)\} = \\ &= \sum\{M(f,B) \sum\{\mathsf{vol}(C) \,|\, C \in \mathcal{B}(Q_B)\} \,|\, B \in \mathcal{B}(P)\} \overset{1.2.1}{=} \\ &\overset{1.2.1}{=} \sum\{M(f,B) \cdot \mathsf{vol}(B) \,|\, B \in \mathcal{B}(P)\} = S(f,P). \end{aligned}$$

Similarly for $s(f,Q)$. □

2.2.2. Proposition. *Let P,Q be partitions of J. We have $s(f,P) \le S(f,Q)$.*

Proof. For a common partition R of P,Q (recall 1.5) we have by 2.2.1

$$s(f,P) \le s(f,R) \le S(f,R) \le S(fQ).$$

□

2.3. By 2.2.2 the set $\{s(f,P)\,|\,P \text{ a partition}\}$ is bounded from above and we can define the *lower Riemann integral* of f over J by

$$\underline{\int}_J f(\mathsf{x})\mathrm{d}\mathsf{x} = \sup\{s(f,P)\,|\,P \text{ a partition}\};$$

similarly, the set $\{S(f,P)\,|\,P \text{ a partition}\}$ is bounded from below and we can define the *upper Riemann integral* of f over J by

$$\overline{\int}_J f(\mathsf{x})\mathrm{d}\mathsf{x} = \inf\{S(f,P)\,|\,P \text{ a partition}\}.$$

If the lower and upper integrals are equal we call the common value the *Riemann integral of f over J* and denote it by

$$\int_J f(\mathsf{x})\mathrm{d}\mathsf{x} \quad \text{or simply} \quad \int_J f$$

2.3.1. Remark. The integral can be also denoted e.g. by

$$\int_J f(x_1,\dots,x_n)\mathrm{d}x_1,\dots x_n$$

which certainly does not surprise. The reader may encounter also symbols like

$$\int_J f(x_1,\dots,x_n)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n.$$

This may look peculiar, but it makes more sense than meets the eyes. See 4.2 below.

2.4. Obviously we have the simple estimate

$$\inf\{f(\mathsf{x})\,|\,\mathsf{x}\in J\}\cdot\mathsf{vol}(J) \le \underline{\int}_J f \le \overline{\int}_J f \le \sup\{f(\mathsf{x})\,|\,\mathsf{x}\in J\}\cdot\mathsf{vol}(J).$$

3. Continuous mappings.

3.1. Proposition. *The Riemann integral $\int_J f(\mathbf{x})d\mathbf{x}$ exists if and only if for every $\varepsilon > 0$ there is a partition P such that*

$$S_J(f,P) - s_J(f,P) < \varepsilon.$$

Note instead of a proof. The statement can be proved by repeating the proof of XI.2.4.2. But the reader may realize that rather than having here an easy generalization of IX.2.4.2, the statements are both special cases of a general simple statement on suprema and infima. Suppose you have a set $(X, \leq)$ partially ordered by $\leq$ such that for any two $x, y \in X$ there is a $z \leq x, y$. If we have $\alpha : X \to \mathbb{R}$ such that $x \leq y$ implies $\alpha(x) \geq \alpha(y)$ and $\beta : X \to \mathbb{R}$ such that $x \leq y$ implies $\beta(x) \leq \beta(y)$, and if $\alpha(x) \leq \beta(y)$ for all x, y then $\sup_x \alpha(x) = \inf_x \beta(x)$ iff for every $\varepsilon > 0$ there is an x such that $\beta(x) < \alpha(x) + \varepsilon$. This is a trivial fact that has nothing to do with sums and such. But of course the criterion is very useful.

3.2. For the proof of the following theorem we will use again the uniform continuity of a continuous function on a compact space (now in the more general version XIII.7.11).

Theorem. *For every continuous function $f : J \to \mathbb{R}$ on an n-dimensional compact interval the Riemann integral $\int_J f$ exists.*

Proof. We will use the distance σ in in $\mathbb{E}_n$ defined by

$$\sigma(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|.$$

Since f is uniformly continuous we can choose for $\varepsilon > 0$ a $\delta > 0$ such that

$$\sigma(\mathbf{x}, \mathbf{y}) < \delta \quad \Rightarrow \quad |f(\mathbf{x}) - f(\mathbf{y})| < \frac{\varepsilon}{\mathsf{vol}(J)}.$$

Recall the mesh $\mu(P)$ from 1.3. If $\mu(P) < \delta$ then $\mathsf{diam}(B) < \delta$ for all $B \in \mathcal{B}(P)$ and hence

$$\begin{aligned} M(f,B) - m(f,B) &= \sup\{f(\mathbf{x}) \,|\, \mathbf{x} \in B\} - \inf\{f(\mathbf{x}) \,|\, \mathbf{x} \in B\} \leq \\ &\leq \sup\{|f(\mathbf{x}) - f(\mathbf{y})| \,|\, \mathbf{x}, \mathbf{y} \in B\} < \frac{\varepsilon}{\mathsf{vol}(J)} \end{aligned}$$

so that

$$S(f,P)-s(f,P)=\sum\{(M(f,B)-m(f,B))\cdot\mathsf{vol}(B)\,|\,B\in\mathcal{B}(P)\}\le$$
$$\le\frac{\varepsilon}{\mathsf{vol}(J)}\sum\{\mathsf{vol}(B)\,|\,B\in\mathcal{B}(P)\}=\frac{\varepsilon}{\mathsf{vol}J}\mathsf{vol}(J)=\varepsilon$$

by 1.2.1. Now use 3.1. □

3.2.1. Similarly like in XI.3.2.1 the previous proof yields the following

Theorem. *Let $f:J\to\mathbb{R}$ be a continuous function and let $P_1,P_2,\dots$ be a sequence of partitions such that $\lim_n\mu(P_n)=0$. Then*

$$\lim_n s(f,P_n)=\lim_n S(f,P_n)=\int_J f.$$

(Indeed, with ε and δ as above choose an n_0 such that for $n\ge n_0$ we have $\mu(P_n)<\delta$.)

3.2.2. Corollary. *Let $f:J\to\mathbb{R}$ be a continuous function on an n-dimensional compact interval J. For every brick $B\subseteq J$ choose an element $\mathbf{x}_B\in B$ and define for a partition P of J*

$$\Sigma(f,P)=\sum\{f(\mathbf{x}_B)\cdot\mathsf{vol}(B)\,|\,B\in\mathcal{B}(P)\}.$$

Let $P_1,P_2,\dots$ be a sequence of partitions such that $\lim_n\mu(P_n)=0$. Then

$$\lim_n\Sigma(f,P_n)=\int_J f.$$

4. Fubini Theorem.

4.1. Theorem. *Consider the product $J=J'\times J''\subseteq\mathbb{E}_{m+n}$ of intervals $J'\subseteq\mathbb{E}_m$, $J''\subseteq\mathbb{E}_n$. Let $f:J\to\mathbb{R}$ be such that $\int_J f(\mathbf{x},\mathbf{y})d\mathbf{xy}$ exists and that for every $\mathbf{x}\in J'$ (resp. every $\mathbf{y}\in J''$) the integral $\int_{J''}f(\mathbf{x},\mathbf{y})d\mathbf{y}$ (resp. $\int_{J'}f(\mathbf{x},\mathbf{y})d\mathbf{x}$) exists (this holds in particular for every continuous function). Then*

$$\int_J f(\mathbf{x},\mathbf{y})d\mathbf{xy}=\int_{J'}(\int_{J''}f(\mathbf{x},\mathbf{y})d\mathbf{y})d\mathbf{x}=\int_{J''}(\int_{J'}f(\mathbf{x},\mathbf{y})d\mathbf{x})d\mathbf{y}.$$

Proof. We will discuss the first equality, the second one is analogous. Set

$$F(\mathbf{x}) = \int_{J''} f(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y}.$$

We will prove that $\int_{J'} F$ exists and that

$$\int_J f = \int_{J'} F.$$

Choose a partition P of J such that

$$\int f - \varepsilon \le s(f,P) \le S(f,P) \le \int f + \varepsilon.$$

This partition P is obviously constituted of a partition P' of J' and a partition P'' of J''. We have

$$\mathcal{B}(P) = \{B' \times B'' \mid B' \in \mathcal{B}(P'), B'' \in \mathcal{B}(P'')\},$$

and each brick of P appears as precisely one $B' \times B''$. By 2.4

$$F(\mathbf{x}) \le \sum_{B'' \in \mathcal{B}(P'')} \max_{\mathbf{y} \in B''} f(\mathbf{x},\mathbf{y}) \cdot \mathsf{vol} B''$$

and hence

$$\begin{aligned} S(F,P') &\le \sum_{B' \in \mathcal{B}(P')} \max_{\mathbf{x} \in B'} \Big(\sum_{B'' \in \mathcal{B}(P'')} \max_{\mathbf{y} \in B''} f(\mathbf{x},\mathbf{y}) \cdot \mathsf{vol}(B'')\Big) \cdot \mathsf{vol}(B') \le \\ &\le \sum_{B' \in \mathcal{B}(P')} \sum_{B'' \in \mathcal{B}(P'')} \max_{(\mathbf{x},\mathbf{y}) \in B' \times B''} f(\mathbf{x},\mathbf{y}) \cdot \mathsf{vol}(B'') \cdot \mathsf{vol}(B') \le \\ &\le \sum_{B' \times B'' \in \mathcal{B}(P)} \max_{\mathbf{z} \in B' \times B''} f(\mathbf{z}) \cdot \mathsf{vol}(B' \times B'') = S(f,P) \end{aligned}$$

and similarly

$$s(f,P) \le s(F,P').$$

Hence we have

$$\int_j f - \varepsilon \le s(F,P') \le \int_{J'} F \le S(F,P) \le \int_J f + \varepsilon$$

and therefore $\int_{J'} F$ is equal to $\int_J f$. $\square$

4.2. Corollary. *Let* $f : J = \langle a_1, b_1\rangle \times \cdots \times \langle a_n, b_n\rangle \to \mathbb{R}$ *be a continuous function. Then*

$$\int_J f(\mathbf{x})\,d\mathbf{x} = \int_{a_n}^{b_n}(\cdots(\int_{a_2}^{b_2}(\int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n)\,dx_1)\,dx_2)\cdots)\,dx_n.$$

Note. The notation mentioned in 2.3 comes, of course, from omitting the brackets.

3rd semester

XVII. More about metric spaces

1. Separability and countable bases.

1.1. Density. Recall the closure from XIII.3.6. A subset M of a metric space (X,d) is *dense* if $\overline{M}=X$. In other words, M is dense if for each $x\in X$ and each $\varepsilon>0$ there is an $m\in M$ such that $d(x,m)<\varepsilon$.

1.2. Separable spaces. A metric space (X,d) is said to be *separable* if there exists a countable dense subset $M\subseteq X$.

1.3. Bases of open sets. A subset $\mathcal{B}$ of the set $\mathsf{Open}(X,d)$ of all open sets in (X,d) is said to be a *basis* (of open sets) if every open set is a union of sets from $\mathcal{B}$, that is, if

$$\forall U\in\mathsf{Open}(X)\ \exists\mathcal{B}_U\subseteq\mathcal{B}\ \text{ such that }\ U=\bigcup\{B\,|\,B\in\mathcal{B}_U\}.$$

In other words,

$$\forall U\in\mathsf{Open}(X)\ \ U=\bigcup\{B\,|\,B\in\mathcal{B}_U,\ B\subseteq U\}.$$

1.3.1. Notes. 1. Thus the set of all open intervals (a,b), or already the set of all the (a,b) with rational a,b is a basis (of open sets) of the real line $\mathbb{R}$.

2. In every metric space the set

$$\{\Omega(x,\frac{1}{n})\,|\,x\in X,\ n=1,2,\dots\}$$

(recall XIII.3.2) is a basis.

3. The term “basis” is in a certain clash with the homonymous term from linear algebra. There is no minimality or independence in the concept of a basis of open sets. Rather, we have here a generating set.

1.4. Covers. A *cover* of a space (X,d) is a subset $\mathcal{U}\subseteq\mathsf{Open}(X,d)$ such that $\bigcup\{U\,|\,U\in\mathcal{U}\}=X$. A *subcover* $\mathcal{V}$ of a cover $\mathcal{U}$ is a subset $\mathcal{V}\subseteq\mathcal{U}$ such that (still) $\bigcup\{U\,|\,U\in\mathcal{V}\}=X$.

Note. More precisely we should speak of *open covers*. But we will not encounter other covers than covers by open sets.

1.5. Lindelöf property, Lindelöf spaces. A space $X = (X, d)$ is said to be *Lindelöf* or to have the *Lindelöf property* if every cover of X has a countable subcover.

1.6. Theorem. *The following statements about a metric space X are equivalent.*

(1) *X is separable.*

(2) *X has a countable basis.*

(3) *X has the Lindelöf property.*

Proof. (1)⇒(2): Let X be separable and let M be a countable dense subset. Set

$$\mathcal{B} = \{\Omega(m, r) \,|\, m \in M,\ r \text{ rational}\}.$$

$\mathcal{B}$ is obviously countable; we will prove it is a basis.

Let U be open and let $x \in U$. Then there is an $\varepsilon > 0$ such that $\Omega(x, \varepsilon) \subseteq U$. Choose an $m_x \in M$ and a rational r_m such that $d(x, m_x) < \frac{1}{3}\varepsilon$ and $\frac{1}{3}\varepsilon < r_x < \frac{2}{3}\varepsilon$. Then

$$x \in \Omega(m_x, r_x) \subseteq \Omega(x, \varepsilon) \subseteq U.$$

Indeed, $x \in \Omega(m_x, r_x)$ trivially and if $y \in \Omega(m_x, r_x)$ then $d(x, y) \leq d(x, m_x) + d(m_x, y) < \frac{1}{3}\varepsilon + \frac{2}{3}\varepsilon = \varepsilon$. Thus, $U = \bigcup\{\Omega(m_x, r_x) \,|\, x \in U\}$.

(2)⇒(3): Let $\mathcal{B}$ be a countable basis and let $\mathcal{U}$ be a cover of X. Since $U = \bigcup\{B \,|\, B \in \mathcal{B}, B \subseteq U\}$ for each $U \in \mathcal{U}$ we have

$$X = \bigcup\{B \in \mathcal{B} \,|\, \exists U_B \supseteq B,\ U_B \in U\}.$$

The cover $\mathcal{A} = \{B \in \mathcal{B} \,|\, \exists U_B \supseteq B,\ U_B \in U\}$ is countable and hence so is also the cover $\mathcal{V} = \{U_B \,|\, B \in \mathcal{A}\}$.

(3)⇒(1): Let X be Lindelöf. For covers

$$\mathcal{U}_n = \{\Omega(x, \frac{1}{n}) \,|\, x \in X\}$$

choose countable subcovers

$$\Omega(x_{n1},\frac{1}{n}),\Omega(x_{n2},\frac{1}{n}),\dots,\Omega(x_{nk},\frac{1}{n}),\dots\ .$$

Then $\{x_{nk}\,|\,n=1,2,\dots,k=1,2,\dots\}$ is dense. □

1.7. Remarks. 1. One often works with spaces that are more general than the metric ones. In the most stanard ones, the *topological spaces*, one gets the information about what are open sets, closed sets, neighbourhoods, etc., without having them constructed from a previously given distance (often, in fact, such spaces cannnot be based on a distance at all). All the concepts above make sense in this generalized context, but their relations are not necessarily the same. Thus for instance, (2) (the existence of countable basis) implies in general both the separability and the Lindelöf property, but none of the other implications holds generally.

2. Note that, trivially, a countable basis is inherited by every subspace (recall XIII.3.4.3), so that we also have that (for metric spaces)

- *every subspace of a separable space is separable*, and
- *every subspace of a Lindelöf space is Lindelöf.*

In particular the latter statement is somewhat surprising (see Section 3 and a similar characteristic of compactness that is inherited by closed subspaces only).

2. Totally bounded metric spaces.

2.1. A metric space (X,d) is *totally bounded* if

$$\forall\varepsilon>0\ \exists\text{ finite }M(\varepsilon)\ \text{ such that }\ \forall\,x\in X,\ d(x,M(\varepsilon))<\varepsilon.$$

Obviously

every totally bounded space is bounded (recall XIII.7.4)

(for any two $x,y\in X$, $d(x,y)\le\max\{d(a,b)\,|\,a,b\in M(1)\}+2$) but not every bounded space is totally bounded: take an infinite set X with $d(x,y)=1$ for $x\neq y$).

2.1.1. Observation. *Total boundedness (and the plain boundedness as well) is preserved when replacing a metric by a strongly equivalent one (recall XIII.4) but it is not a topological property.*

(For the second statement consider the bounded open interval (a, b) and the real line $\mathbb{R}$; recall XIII.6.8.)

2.2. Proposition. *A subspace of a totally bounded (X, d) is totally bounded.*

Proof. Let $Y \subseteq X$. For $\varepsilon > 0$ take the $M(\frac{\varepsilon}{2}) \subseteq X$ from the definition and set

$$M_Y = \{a \in M(\frac{\varepsilon}{2}) \mid \exists y \in Y,\ d(a, y) < \frac{\varepsilon}{2}\}.$$

Now for each $a \in M_Y$ choose an $a_Y \in Y$ such that $d(a, a_Y) < \frac{\varepsilon}{2}$ and set

$$N(\varepsilon) = \{a_Y \mid a \in M_Y\}.$$

Then for every $y \in Y$ we have $d(y, N(\varepsilon)) < \varepsilon$. □

2.3. Proposition. *A product $X = \prod_{j=1}^{n}(X_j, d_j)$ of totally bounded spaces is totally bounded.*

Proof. For the product take the distance d from XIII.5. Then if we take for X_i the $M_i(\varepsilon)$ from the definition, the set $M(\varepsilon) = \prod M_i(\varepsilon)$ has the property needed for X. □

2.4. Proposition. *A subspace of $\mathbb{E}_n$ is totally bounded if and only if it is bounded.*

Proof. In view of 2.2. and 2.3 it suffices to prove that the interval $\langle a, b\rangle$ is totally bounded. But this is easy: for $\varepsilon > 0$ take an n such that $\frac{b-a}{n} < \varepsilon$ and set

$$M(\varepsilon) = \{a + k\frac{b-a}{n} \mid k = 0, 1, 2, \dots\}.$$

□

2.5. A characteristics of total boundedness reminiscent of compactness.

2.5.1. Lemma. *If (X, d) is not totally bounded then there is a sequence that contains no Cauchy subsequence.*

Proof. If (X, d) is not locally bounded then there is an $\varepsilon_0 > 0$ such that for every finite $M \subseteq X$ there is an $x_M \in X$ such that $d(x_M, M) \geq \varepsilon_0$. Choose

x_1 arbitrarily and if $x_1, \dots, x_n$ are already chosen set $x_{n+1} = x_{\{x_1,\dots,x_n\}}$. Then any two elements of the resulting sequence have the distance at least ε_0 and hence there is no Cauchy subsequence. □

2.5.2. Theorem. *A metric space X is totally bounded if and only if every sequence in X contains a Cauchy subsequence.*

Proof. Let $(x_n)_n$ be a sequence in a totally bounded (X, d). Consider the

$$M(\frac{1}{n}) = \{y_{n1}, \dots, y_{nm_n}\}$$

from the definition. If $A = \{x_n \,|\, n = 1, 2, \dots\}$ is finite then $(x_n)_n$ contains a constant subsequence. Thus, suppose A is not finite. There is an r_1 such that $A_1 = A \cap \Omega(y_{1r_1}, 1)$ is infinite; choose $x_{k_1} \in A_1$. If we already have infinite

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_s, \quad A_j \subseteq \Omega(y_{jr_j}, \frac{1}{j})$$

and

$$k_1 < \cdots < k_s \quad \text{such that} \quad x_{k_j} \in A_j$$

choose r_{s+1} such that $A_{s+1} = A_s \cap \Omega(y_{s+1,r_{s+1}}, \frac{1}{s+1})$ is infinite and an $x_{k_{s+1}} \in A_{s+1}$ such that $k_{s+1} > k_s$. Then the subsequence $(x_{k_n})_n$ is Cauchy. □

The converse is in 2.5.1. □

2.6. Theorem. *A metric space is compact if and only if it is totally bounded and complete.*

Proof. Let X be compact. Then it is complete by XIII.7.7 and totally bounded by 2.5.1.

On the other hand let X be totally bounded and let $(x_n)_n$ be a sequence in X. Then it contains a Cauchy subsequence and if it is, moreover, complete, this subsequence converges. □

2.6.1. Remark. 1. We already know the characteristics of the compact subspaces of $\mathbb{E}_n$ as the closed bounded ones (XIII.7.6). Realize that it is a special case of 2.6: a subset of $\mathbb{E}_n$ is complete iff it is closed (see XIII.6.6 and XIII.6.4), and it is totally bounded iff it is bounded (see 2.4).

2. Note that neither completeness nor total boundedness are topological properties, while their conjunction is.

2.7. Proposition. *Every totally bounded metric space is separable.*

Proof. Take the sets $M(\varepsilon)$ from the definition again. The set

$$\bigcup_{n=1}^{\infty} M(\frac{1}{n})$$

is countable and evidently dense. □

2.7.1. Corollary. *Every compact space is separable and hence Lindelöf.*

3. Heine-Borel Theorem.

3.1. Accumulation points. A point is an *accumulation point* of a set A in a space X if every neighbourhood of x contains infinitely many points of A. The following is a straightforward but expedient modification of the definition of compactness by means of convergent subsequences.

Proposition. *A metric space X is compact iff every infinite M in X has an accumulation point.*

Proof. Let X be compact and let A be infinite. Choose an arbitrary sequence $x_1, x_2, \dots, x_n, \dots$ in A such that $x_i \neq x_j$ for $i \neq j$. Then every neighbourhood of a limit x of a subsequence $(x_{k_n})_n$ contains infinitely many x_j's and hence x is an accumulation point of A.

On the other hand let the second statement hold and let $(x_n)_n$ be a sequence in X. Then either $A = \{x_n \,|\, n = 1, 2, \dots\}$ is finite and then $(x_n)_n$ contains a constant subsequence, or A has an accumulation point x. Then we can proceed as follows. Choose x_{k_1} in $A \cap \Omega(x, 1)$ and if $x_{k_1}, \dots, x_{k_n}$ have been already chosen pick $x_{k_{n+1}}$ in $A \cap \Omega(x, \frac{1}{n+1})$ so that $k_{n+1} > k_n$ (this disqualifies only finitely many of infinite number of choices); then $\lim_n x_{k_n} = x$. □

3.2. Theorem. (Heine-Borel Theorem)*A metric space is compact if and only if each cover of X contains a finite subcover.*

Proof. I. Let X be compact but let there be a cover that has no finite subcover. By 2.7.1 X is Lindelöf and hence there is a *countable* cover

$$U_1, U_2, , \dots, U_n, \dots \qquad (*)$$

with no finite subcover. Define

$$V_1, V_2, , \dots, V_n, \dots$$

as follows:

- take for V_1 the first non-empty U_k, and
- if $V_1, V_2, , \dots, V_n$ have been already chosen take for V_{n+1} the first U_k such that $U_k \nsubseteq \bigcup_{j=1}^n V_j$.This way we have rejected precisely the U_j that were redundant for covering the space in the order of $(*)$ (that is, the sequence $(\bigcup_{k=1}^n V_n)_n$ of the already covered parts of X is the same as $(\bigcup_{k=1}^n U_n)_n$, only without repetition).

Hence

(1) $\{V_n \,|\, n = 1, 2, \dots\}$ is a subcover of $\{U_n \,|\, n = 1, 2, \dots\}$,

(2) the procedure cannot stop, else we had a finite subcover, and

(3) we can choose $x_n \in V_n \smallsetminus \bigcup_{k=1}^{n-1} V_k$.

Now all the x_n are distinct (if $k < n$ then $x_n \in V_n \smallsetminus V_k$ while $x_k \in V_k$) and hence we have an infinite set

$$A = \{x_1, x_2, \dots, x_n, \dots\}$$

and this set has to have an accumulation point x. Since $\{V_n \,|\, n = 1, 2, \dots\}$ is a cover, there is an n such that $x \in V_n$. This is a contradiction since V_n contains none of the x_k with $k > n$ and hence $V_n \cap A$ is not infinite.

II. Let the statement about covers hold and let there be an infinite A without an accumulation point. That is, no $x \in X$ is an accumulation point of A and hence we have open $U_x \ni x$ such that $U_x \cap A$ is finite. Choose a finite subcover

$$U_{x_1}, U_{x_2}, \dots, U_{x_n}$$

of the cover $\{U_x \,|\, x \in X\}$. Now we have

$$A = A \cap X = A \cap \bigcup_{k=1}^n U_{x_k} = \bigcup_{k=1}^n (A \cap U_{x_k})$$

which is a contradiction since the rightmost union is finite. □

3.3. Corollary. (Finite Intersection Property) *Let $\mathcal{A}$ be a system of closed subsets of a compact space. If $\bigcap\{A \,|\, A \in \mathcal{A}\} = \emptyset$ then there is a finite $\mathcal{A}_0 \subseteq \mathcal{A}$ such that $\bigcap\{A \,|\, A \in \mathcal{A}_0\} = \emptyset$. Consequently, if*

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$$

is a decreasing sequence of non-empty closed subsets of X *then* $\bigcap_{n=1}^{\infty} A_n \neq \emptyset$.

Proof. By De Morgan formula, $\{X \smallsetminus A \,|\, A \in \mathcal{A}\}$ is a cover. □

4. Baire Category Theorem.

4.1. Diameter. Generalizing the *diameter* from XVI.1.3 we define in a general metric space (X, d) for a subset $A \subseteq X$

$$\mathsf{diam}(A) = \sup\{d(x,y) \,|\, x, y \in A\}$$

Note that $\mathsf{diam}(A)$ can be infinite: in fact $\mathsf{diam}(X)$ is finite only if the space is bounded.

From the triangle inequality we immediately obtain

4.1.1. Observations. 1. $\mathsf{diam}(\Omega(x,\varepsilon)) \leq 2\varepsilon$, *and*

2. $\mathsf{diam}(\overline{A}) = \mathsf{diam}(A)$.

4.2. Lemma. *Let* (X, d) *be a complete metric space. Let*

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$$

be a decreasing sequence of non-empty closed subsets of X *with* $\lim_n \mathsf{diam}(A_n) = 0$. *Then*

$$\bigcap_{n=1}^{\infty} A_n \neq \emptyset.$$

Proof. Choose $a_n \in A_n$. Then, by the assumption on diameters, $(a_n)_n$ is a Cauchy sequence and hence, by completeness, it has a limit a. Now the subsequence

$$a_n, a_{n+1}, a_{n+2}, \ldots$$

is in the *closed* A_n and hence its limit a is in A_n. As n was arbitrary, $a \in \bigcap_{n=1}^{\infty} A_n$. □

4.2.1. Notes. 1. The assumption on diminishing diameter is essential: take e.g. the closed $A_n = \langle n, +\infty)$ in the complete $\mathbb{R}$. It may look at the first sight slightly paradoxical that an intersection of small sets is non-void but an intersection of large ones is not necessarily so. But the principle is, hopefully, obvious.

2. The reader may wonder whether it is not, on the other hand, essential that the diameters in the example above are infinite. In general it is easy to

give an example with $\mathsf{diam}(A_n) = 1$, but not in $\mathbb{R}$ or, more generally, not in $\mathbb{E}_n$: see 3.3. But this has to do with compactness, not with completeness.

3. Needless to say, the intersection in 4.2 consists necessarily of a single point.

4.3. Lemma. *If* $0 < \varepsilon < \eta$ *then* $\overline{\Omega(x, \varepsilon)} \subseteq \Omega(x, \eta)$

Proof. This is an immediate consequence of the triangle inequality: if $d(y, \Omega(x, \varepsilon)) = 0$ choose a $z \in \Omega(x, \varepsilon)$ with $d(y, z) < \eta - \varepsilon$; then $d(x, y) \leq d(x, z) + d(z, y) < \eta$. □

4.4. Nowhere dense sets. A subset A of a metric space X is said to be *nowhere dense* if $X \smallsetminus \overline{A}$ is dense, that is, if $\overline{X \smallsetminus \overline{A}} = X$. Note that

A is nowhere dense iff $\overline{A}$ is nowhere dense.

4.4.1. Reformulation. *$A \subseteq X$ is nowhere dense iff for every non-empty open U the intersection $U \cap (X \smallsetminus \overline{A})$ is non-empty.*

(Indeed, this amounts to stating that for every x and every $\varepsilon > 0$ the intersection $\Omega(x, \varepsilon) \cap (X \smallsetminus \overline{A})$ is non-empty.)

4.4.2. Proposition. *A union of finitely many nowhere dense sets is nowhere dense.*

Proof. It suffices to prove the statement for two sets. Let A, B be nowhere dense and let U be non-empty open. We have $U \cap (X \smallsetminus \overline{(A \cup B)}) = U \cap (X \smallsetminus (\overline{A} \cup \overline{B}) = U \cap (X \smallsetminus \overline{A}) \cap (X \smallsetminus \overline{B})$. Now the open set $V = U \cap (X \smallsetminus \overline{A})$ is non-empty, and hence $V \cap (X \smallsetminus \overline{B}))$ is non-empty as well. □

4.5. Sets of first category (meagre sets). A countable union of nowhere dense sets can be already very far from being nowhere dense. Take for instance the one-point subsets in the space X of all rational numbers: their union is the whole of X. But in complete spaces such countable unions can form only very small parts.

A subset of a metric space is said to be a *set of first category* (or a *meagre set*) if it is a countable union $\bigcup_{n=1}^{\infty} A_n$ of nowhere dense sets A_n.

4.5.1. Theorem. (Baire Category Theorem) *No complete metric space X is of the first category in itself.*

Proof. Suppose it is, that is,

$$X = \bigcup_{n=1}^{\infty} A_n \quad \text{with} \quad X \smallsetminus \overline{A_n} \text{ dense.}$$

We can assume all the A_n closed; hence we have $X \smallsetminus A_n$ dense open. Choose $U_1 = \Omega(x, \varepsilon)$ such that $\Omega(x, 2\varepsilon) \subseteq X \smallsetminus A_1$ and $2\varepsilon < 1$. Thus, by 4.1.1 and 4.3

$$B_1 = \overline{U}_1 \subseteq X \smallsetminus A_1 \quad \text{and} \quad \mathsf{diam}(B_1) < 1.$$

Let us have for $k \leq n$ non-empty open $U_1, \dots, U_n$ with

$$U_{k-1} \supseteq B_k = \overline{U}_k \text{ for } k \leq n,\ B_k \subseteq X \smallsetminus A_k, \text{ and } \mathsf{diam}(B_k) < \frac{1}{k}. \qquad (*)$$

Since $U_n \cap (X \smallsetminus A_{n+1})$ is non-empty open we can choose $U_{n+1} = \Omega(y, \eta)$ for some $y \in U_n \cap (X \smallsetminus A_{n+1})$ and η sufficiently small to have $\Omega(y, 2\eta) \subseteq U_n \cap (X \smallsetminus A_{n+1})$ and $2\eta < \frac{1}{n+1}$. Then we have, by 4.1.1 and 4.3, the system $(*)$ extended from n to $n+1$ and we inductively obtain a sequence of non-empty closed sets B_n such that

(1) $B_1 \supseteq B_2 \supseteq \cdots \supseteq B_n \supseteq \cdots$,

(2) $\mathsf{diam}(B_n) < \frac{1}{n}$, and

(3) $B_n \subseteq X \smallsetminus A_n$.

By (1),(2) and 4.2, $B = \bigcap_{n=1}^{\infty} B_n \neq \emptyset$, and by (3)

$$B \subseteq \bigcap_{n=1}^{\infty} (X \smallsetminus A_n) = X \smallsetminus \bigcup_{n=1}^{\infty} A_n = X \smallsetminus X = \emptyset,$$

a contradiction. $\square$

4.5.2. Note. Realize how small part of a complete space X a set of first category constitutes. A countable union of such sets is obviously still of first category, hence it not only smaller than X, but it is in effect so small that infinitely many disjoint copies cannot cover X.

5. Completion.

5.1. For various reasons, for applying metric spaces in analysis or geometry it is preferable to have the spaces complete. We have already seen the advantages of the real line $\mathbb{R}$ as compared with the rational one, $\mathbb{Q}$. Note that the extension of the rationals to reals is very satisfactory. We do not lose anything of the calculating power, in fact everything is in this respect only better, and $\mathbb{Q}$ is dense in $\mathbb{R}$ so that everything to be computed in $\mathbb{R}$ can be well approximated by computing with rationals.

In this section we will show that we can analogously extend every metric space. That is, for every (X, d) we have a space $(\widetilde{X}, \widetilde{d})$ such that

- (X, d) is dense in $(\widetilde{X}, \widetilde{d})$ (in our construction we will have an isometric embedding $\iota : (X, d) \to (\widetilde{X}, \widetilde{d})$ such that $\iota[X]$ is dense in $\widetilde{X}$), and
- $(\widetilde{X}, \widetilde{d})$ is complete.

5.2. The construction. The idea of the following construction is very natural. In the original space there can be Cauchy sequences without limits; thus, let us add the limits. This will be done by representing the limits by the so far limitless Cauchy sequences; only, we will have to identify the sequences that represent the same limit – see the equivalence $\sim$ below.

Denote by

$$\mathcal{C}(X, d), \quad \text{in short} \quad \mathcal{C}(X),$$

the set of all Cauchy sequences in X. For $(x_n)_n, (y_n)_n \in \mathcal{C}(X)$ define

$$d'((x_n)_n, (y_n)_n) = \lim_n d(x_n, y_n).$$

5.2.1. Lemma. *The limit in the definition of d' always exists and we have*

(1) $d'((x_n)_n, (x_n)_n) = 0$,

(2) $d'((x_n)_n, (y_n)_n) = d'((y_n)_n, (x_n)_n)$, *and*

(3) $d'((x_n)_n, (z_n)_n) \leq d'((x_n)_n, (y_n)_n) + d'((y_n)_n, (z_n)_n)$.

Proof. The first statement will be proved by showing that the sequence $(d(x_n, y_n))_n$ is Cauchy in $\mathbb{R}$. Indeed, $(x_n)_n$ and $(y_n)_n$ are Cauchy and hence for an $\varepsilon > 0$ we have an n_0 such that for $m, n > n_0$, $d(x_n, x_m) < \frac{\varepsilon}{2}$ and $d(y_n, y_m) < \frac{\varepsilon}{2}$. Then $d(x_n, y_n) \leq d(x_n, x_m) + d(x_m, y_m) + d(y_m, y_n) < \varepsilon + d(x_m, y_m)$, hence $d(x_n, y_n) - d(x_m, y_m) < \varepsilon$ and by symmetry also $d(x_m, y_m) - d(x_n, y_n) < \varepsilon$, and we conclude that $|d(x_n, y_n) - d(x_m, y_m)| < \varepsilon$.

(1) and (2) are trivial and (3) is very easy: choose k such that

$$|d'((x_n)_n,(z_n)_n) - d(x_k,z_k)| < \varepsilon, \quad |d'((x_n)_n,(y_n)_n) - d(x_k,y_k)| < \varepsilon$$
$$\text{and} \ \ |d'((y_n)_n,(z_n)_n) - d(y_k,z_k)| < \varepsilon.$$

Then we obtain from the triangle inequality of d that

$$d'((x_n)_n,(z_n)_n) \le d'((x_n)_n,(y_n)_n) + d'((y_n)_n,(z_n)_n) + 3\varepsilon$$

and since $\varepsilon > 0$ was arbitrary, (3) follows. $\square$

5.2.2. Define an equivalence relation $\sim$ on $\mathcal{C}(X)$ by setting

$$(x_n)_n \sim (y_n)_n \quad \text{iff} \quad d'((x_n)_n,(y_n)_n) = 0$$

(from 5.2.1 it immediately follows that $\sim$ is an equivalence relation), denote

$$\widetilde{X} = \mathcal{C}(X)/\sim,$$

and for classes $p = [(x_n)_n]$ and $q = [(y_n)_n]$ of this equivalence relation set

$$\widetilde{d}(p,q) = d'((x_n)_n,(y_n)_n).$$

5.2.3. Lemma. *The value of $\widetilde{d}(p,q)$ does not depend on the choice of representatives of p and q, and $(\widetilde{X},\widetilde{d})$ is a metric space.*

Proof. If $(x_n)_n \sim (x'_n)_n$ and $(y_n)_n \sim (y'_n)_n$ we have

$$\begin{aligned} d'((x_n)_n,(y_n)_n) &\le d'((x_n)_n,(x'_n)_n) + d'((x'_n)_n,(y'_n)_n) + d'((y'_n)_n,(y_n)_n) = \\ &= 0 + d'((x'_n)_n,(y'_n)_n) + 0 = d'((x'_n)_n,(y'_n)_n) \end{aligned}$$

and by symmetry also $d'((x'_n)_n,(y'_n)_n) \le d'((x_n)_n,(y_n)_n)$.

Now by 5.2.1, $\widetilde{d}$ satisfies the requirements XIII.2.1(2),(3) and the missing $\widetilde{d}(p,q) = 0 \ \Rightarrow \ p = q$ immediately follows from the definition of $\sim$: if $d(p,q) = d'((x_n)_n,(y_n)_n) = 0$ then $(x_n)_n \sim (y_n)_n$ and the sequences represent the same element of $\widetilde{X}$. $\square$

5.3. Set

$$\widetilde{x} = (x,x,\dots,x,\dots)$$

and define a mapping

$$\iota = \iota_{(X,d)} : (X,d) \to (\widetilde{X},\widetilde{d})$$

by

$$\iota(x) = [\widetilde{x}].$$

We have

$$d'(\widetilde{x}, \widetilde{y}) = d(x, y)$$

and hence ι is an isometric embedding.

Theorem. *The image of the isometric embedding $\iota_{(X,d)}$ is dense in $(\widetilde{X}, \widetilde{d})$, and the space $(\widetilde{X}, \widetilde{d})$ is complete.*

Proof. Take a $p = [(x_n)_n] \in \widetilde{X}$ and an $\varepsilon > 0$. Since $(x_n)_n$ is Cauchy there is an n_0 such that for $m, k > n_0$, $d(x_m, x_k) \leq \varepsilon$. But then $\widetilde{d}(\iota(x_{n_0}), p) = d'(\widetilde{x_{n_0}}, (x_k)_k) \leq d(x_{n_0}, x_k) < \varepsilon$.

Now let

$$p_1 = [(x_{1n})_n],\ p_2 = [(x_{2n})_n],\ \dots,\ p_k = [(x_{kn})_n],\ \dots \qquad (*)$$

be a Cauchy sequence in $(\widetilde{X}, \widetilde{d})$. For each p_n choose, by the already proved density, an $x_n \in X$ such that $\widetilde{d}(p_n, \iota(x_n)) < \varepsilon$. For $\varepsilon > 0$ choose $n_0 > \frac{3}{\varepsilon}$ such that for $m, n \geq n_0$, $\widetilde{d}(p_m, p_n) < \frac{\varepsilon}{3}$. Then for $m, n \geq n_0$,

$$d(x_m, x_n) = \widetilde{d}(\iota(x_m), \iota(x_n)) \leq \widetilde{d}(\iota(x_m), p_m) + \widetilde{d}(p_m, p_n) + \widetilde{d}(p_n, \iota(x_n)) < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

and we see that $(x_n)_n$ is Cauchy. We will prove that the sequence $(*)$ converges to $p = [(x_n)_n]$.

We know that $\widetilde{d}(p_n, \iota(x_n)) = \lim_k d(x_{nk}, x_n) < \frac{1}{n}$. Choose $n_0 > \frac{2}{\varepsilon}$ such that for $k, n \geq n_0$ we have $d(x_k, x_m) < \frac{\varepsilon}{2}$. Then

$$d(x_{nk}, x_k) \leq d(x_{nk}, x_n) + d(x_n, x_k) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

and hence $\widetilde{d}(p_n, p) = \lim_k d(x_{nk}, x_k) \leq \varepsilon$. □

5.4. Remark. The question naturally arises whether the completion extending the rational line $\mathbb{Q}$ to the real one, $\mathbb{R}$, can be constructed in the vein of the procedure just presented. The ansver is a cautious YES; one has to keep in mind that we will have some troubles formulating precisely what we are doing. The construction above already works with metric spaces and the distances already have *real* values. But we can speak of Cauchy sequences, define equivalence $\sim$ of Cauchy sequences (but not by means of limits the existence of which is based on the properties of reals), and obtain the desired. But many readers would view the usually used method of Dedekind cuts as somewhat simpler.

.

XVIII. Sequences and series of functions

1. Pointwise and uniform convergence.

1.1. Pointwise convergence. Let $X = (X, d)$ and $Y = (Y, d')$ be metric spaces and let $f_n : X \to Y$ be a sequence of continuous mappings. If for each $x \in X$ there is a $\lim_n f(x) = f(x)$ (in Y) we say that the sequence $(f_n)_n$ *converges pointwise* to the mapping f and usually write

$$f_n \to f.$$

1.1.1. Example. Pointwise convergence does not preserve nice properties of the functions f_n, not even continuity, not to speak of possessing derivatives. Consider the following extremely simple example. Let $X = Y = \langle 0, 1\rangle$ and let the functions f_n be defined by

$$f_n(x) = x^n.$$

Then $f(x) = \lim_n f_n(x)$ is 0 for $x < 1$ while $f(1) = 1$.

1.2. Uniform convergence. A sequence $(f_n : (X, d) \to (Y, d'))_n$ converges *uniformly* to $f : X \to Y$ if

$$\forall \varepsilon > 0 \; \exists n_0 \text{ such that } \forall x \in X \;\; (n \geq n_0 \;\Rightarrow\; d'(f_n(x), f(x)) < \varepsilon).$$

We speak of a *uniformly convergent sequence* of mappings and write

$$f_n \rightrightarrows f.$$

1.3. Theorem. *Let $f_n : X \to Y$ be continuous mappings and let $f_n \rightrightarrows f$. Then f is continuous.*

Proof. Choose $x \in X$ and $\varepsilon > 0$. Fix an n such that

$$\forall y \in X, \; d'(f_n(y), f(y)) < \frac{\varepsilon}{3}.$$

Since f_n is continuous there is a $\delta > 0$ such that

$$d(x, z) < \delta \quad \Rightarrow \quad d'(f_n(x), f_n(z)) < \frac{\varepsilon}{3}.$$

Hence for $d(x,z) < \delta$,

$$d'(f(x), f(z)) \le d'(f(x), f_n(x)) + d'(f_n(x), f_n(z)) + d'(f_n(z), f(z)) < \\ < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

□

1.4. Notes. 1. The adjective "uniform" refers, similarly as in "uniform continuity", to the indpendence of the property in question on the location in the domain space. One might for a moment expect that, similarly as in the uniform continuity, we will obtain something for free in case of a compact domain. But it is not so: the sequence in the example 1.1.1 has a very simple compact domain and range and it is not uniformly convergent.

2. Theorem 1.3 holds for uniform continuity as well, that is, we have that

> *if $f_n : X \to Y$ are uniformly continuous mappings and $f_n \rightrightarrows f$. then f is uniformly continuous.*

To prove this it suffices to adapt the proof of 1.3 by not fixing the x at the start. The reader may write down the details as a simple exercise.

1.5. We say that a sequence $(f_n)_n$ converges to f *locally uniformly* if for every $x \in X$ there exists a neighbourhood U such that $f_n|U \rightrightarrows f|U$ for the restictions on U. Since the continuity at a point is a local property (that is, f is continuous in x iff $f|U$ is continuous in x for a neighbourhood U of x) we immediately obtain from 1.3

1.5.1 Corollary. *Let $f_n : X \to Y$ be continuous mappings and let the sequence f_n converge to f locally uniformly. Then f is continuous.*

2. More about uniform convergence: derivatives, Riemann integral.

2.1. Example. Alhough uniform convergence preserves continuity, it does not preserve smoothness (existence of derivatives). Consider the functions

$$f_n : \langle -1, 1\rangle \to \langle 0, 1\rangle \quad \text{defined by} \quad f_n(x) = \sqrt{(1 - \frac{1}{n})x^2 + \frac{1}{n}}.$$

These smooth functions uniformly converge to $f(x) = |x|$ which has no derivative in $x = 0$: we have

$$\left|\sqrt{(1-\frac{1}{n})x^2+\frac{1}{n}}-|x|\right| = \frac{\frac{1}{n}(1-x^2)}{\left|\sqrt{(1-\frac{1}{n})x^2+\frac{1}{n}}+|x|\right|} \le \sqrt{\frac{1}{n}}.$$

However, smoothness is preserved if the uniform convergence concerns the derivatives.

2.2. Theorem. *Let f_n be continuous real functions defined on an open interval J and let them have continuous derivatives f_n'. Let $f_n \to f$ and $f_n' \rightrightarrows g$ on J. Then f has a derivative on J and $f' = g$.*

Proof. We have

$$\begin{aligned} A(h) &= \left|\frac{f(x+h)-f(x)}{h} - g(x)\right| = \\ &= \left|\frac{f(x+h)-f_n(x+h)}{h} - \frac{f(x)-f_n(x)}{h} + \frac{f_n(x+h)-f_n(x)}{h} - g(x)\right| \end{aligned}$$

and since by Lagrange theorem, $\frac{f_n(x+h)-f_n(x)}{h} = f_n'(x+\theta h)$ for some θ with $0 < \theta < 1$, we further obtain

$$\begin{aligned} A(h) = \Bigg|&\frac{f(x+h)-f_n(x+h)}{h} - \frac{f(x)-f_n(x)}{h} + f_n'(x+\theta h) - \\ &- g(x+\theta h) + g(x+\theta h) - g(x)\Bigg| \le \\ \le &\frac{1}{|h|}|f(x+h)-f_n(x+h)| + \frac{1}{|h|}|f(x)-f_n(x)| + \\ &+ |f_n'(x+\theta h) - g(x+\theta h)| + |g(x+\theta h) - g(x)|. \end{aligned}$$

Since $f_n' \rightrightarrows g$, the function g is continuous by 1.3. Choose $\delta > 0$ such that for $|x-y| < \delta$ we have $|g(x)-g(y)| < \varepsilon$; thus if $|h| < \delta$ the last summand is smaller than ε.

Now fix an h with $|h| < \delta$ and choose an n sufficiently large so that

$$\begin{aligned} &|f_n'(y) - g(y)| < \varepsilon, \\ &|f(x+h) - f_n(x+h)| < \varepsilon|h|, \text{ and} \\ &|f(x) - f_n(x)| < \varepsilon|h| \end{aligned}$$

(note that for the first we have to use the uniform convergence – we do not know precisely where $y = x + \theta h$ is; not so in the other two inequalities, where one uses just convergence in two fixed arguments x and $x + h$). Then we obtain

$$A(h) = \left| \frac{f(x+h) - f(x)}{h} - g(x) \right| < 4\varepsilon$$

and the statement follows. □

2.3. Integral of a limit of functions. For Riemann integral we do not generally have $\int_a^b \lim_n f_n = \lim_n \int_a^b f_n$ even if all the $\int_a^b f_n$ exist and all the functions f_n are bounded by the same constant. Here is an example.

Order all the rational numbers between 0 and 1 in a sequence

$$r_1, r_2, \dots, r_n, \dots \ .$$

Set

$$f_n(x) = \begin{cases} 1 & \text{if } x = r_k \text{ with } k \le n, \\ 0 & \text{otherwise.} \end{cases}$$

Then obviously $\int_0^1 f_n = 0$ for every n. But the limit f of the sequence f_n is the well-known Dirichlet function for which (obviously again) the lower integral is 0 and the upper integral is 1.

For uniform convergence we have, however

2.3.1. Theorem. *Let $f_n \rightrightarrows f$ on $\langle a, b\rangle$ and let the Riemann integrals $\int_a^b f_n$ exist. Then also $\int_a^b f$ exists and we have*

$$\int_a^b f = \lim_n \int_a^b f_n.$$

Proof. For $\varepsilon > 0$ choose an n_0 such that for $n \ge n_0$,

$$|f_n(x) - f(x)| < \frac{\varepsilon}{b-a} \tag{$*$}$$

for all $x \in \langle a, b\rangle$. Recall the notation from XI.2. For a partition $P : a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$ (which will be further specified) consider

$$m_j = \inf\{f(x) \,|\, t_{j-1} \le x \le t_j\}, \quad M_j = \sup\{f(x) \,|\, t_{j-1} \le x \le t_j\} \text{ and}$$
$$m_j^n = \inf\{f_n(x) \,|\, t_{j-1} \le x \le t_j\}, \quad M_j^n = \sup\{f_n(x) \,|\, t_{j-1} \le x \le t_j\}.$$

By $(*)$ we have for $n, k \geq n_0$

$$|m_j - m_j^n|,\ |M_j - M_j^n| \leq \frac{\varepsilon}{b-a} \ \text{ and hence also } \ |M_j^k - M_j^n| \leq \frac{2\varepsilon}{b-a}$$

and we obtain for the lower sums

$$\begin{aligned} |s(f,P) - s(f_n,P)| &= \Big|\sum (m_i - m_i^n)(t_i - t_{i-1})\Big| \leq \\ &\leq \sum |m_i - m_i^n|(t_i - t_{i-1}) \leq \varepsilon \end{aligned}$$

and similarly for the upper sums

$$|S(f,P) - S(f_n,P)| \leq \varepsilon \quad \text{and} \quad |S(f_k,P) - S(f_n,P)| \leq 2\varepsilon.$$

Now, first take a P such that $|\int f_n - S(f_n,P)| < \varepsilon$ and $|\int f_k - S(f_k,P)| < \varepsilon$; then we infer from the triangle inequality that $|\int f_k - \int f_n| < 4\varepsilon$ and see that $(\int f_n)_n$ is a Cauchy sequence. Hence there exists a limit $L = \lim_n \int f_n$. Choose $n \geq n_0$ sufficiently large to have $|\int f_n - L| < \varepsilon$.

Now if the partition P is chosen so as to have

$$S(f_n,P) - \varepsilon < \int f_n < s(f_n,P) + \varepsilon$$

we obtain

$$\begin{aligned} L - 3\varepsilon \leq \int f_n - 2\varepsilon < s(f_n,P) - \varepsilon \leq s(f,P) \leq \\ \leq S(f,P) \leq S(f_n,P) + \varepsilon \leq \int f_n + 2\varepsilon \leq L + 3\varepsilon \end{aligned}$$

and since $\varepsilon > 0$ was arbitrary we conclude that $L = \underline{\int} f = \overline{\int} f$. $\square$

2.3.2. Note. The example in 2.3 where the Riemann integrable functions pointwise converged to the Dirichlet function suggested that the trouble might be rather in the not-integrable limit function then in the value of the integral being different from the limit. This is only partly true. Indeed, if we take the more powerful Lebesgue integral (roughly speaking, based on the idea of sums of *countable* disjoint systems, while our Riemann integral is based on *finite* disjoint systems) the integral of the Dirichlet function is 0 (as the intuition suggests: the part of the interval in which the function is not 0 is infinitelly smaller than the one with values 0).

But whatever the strength of the integral might be, the formula

$$\int_a^b \lim_n f_n = \lim_n \int_a^b f_n$$

cannot hold generally. Consider the functions $f_n, g_n : \langle -1, 1\rangle \to \mathbb{R} \cup \{+\infty\}$ defined by setting

$$f_n(x) = \begin{cases} 0 \ \text{ for } \ x \leq -\frac{1}{n} \ \text{ and } \ x \geq \frac{1}{n}, \\ n + n^2 x \ \text{ for } \ -\frac{1}{n} \leq x \leq 0, \\ n - n^2 x \ \text{ for } \ 0 \leq x \leq \frac{1}{n}, \end{cases} \qquad g_n(x) = \begin{cases} 0 \ \text{ for } \ x \neq 0, \\ n \ \text{ for } \ x = 0 \end{cases}$$

(draw a picture of f_n). Then for each n, $\int_a^b f_n = 1$ and $\int_a^b g_n = 0$ while $\lim_n f_n = \lim_n g_n$.

In actual fact, for Lebesgue integral the formula $\int_a^b \lim_n f_n = \lim_n \int_a^b f_n$ holds for instance if the limit is monotone or if the functions are equally bounded by an integrable function. Thus, in the example above the formula $\int_a^b \lim_n g_n = \lim_n \int_a^b g_n$ is correct, the one with f_n is not.

2.4. Lemma. *Let* $\lim_{n\to\infty} g(x_n) = A$ *for each sequence* $(x_n)_n$ *such that* $\lim_n x_n = a$. *Then* $\lim_{x\to a} g(x) = A$.

Proof. Suppose $\lim_{x\to a} g(x)$ does not exist or is not equal to A. Then there is an $\varepsilon > 0$ such that for every $\delta > 0$ there is an $x(\delta)$, with $0 < |a - x(\delta)| < \delta$ and $|A - g(x(\delta))| \geq \varepsilon$. Set $x_n = x(\frac{1}{n})$. Then $\lim_n x_n = a$ while $\lim_{n\to\infty} g(x_n)$ is not A. □

2.4.1. Proposition. *Let* $f : \langle a, b\rangle \times \langle c, d\rangle \to \mathbb{R}$ *be a continuous function. Then*

$$\lim_{y\to y_0} \int_a^b f(x, y)\, dx = \int_a^b f(x, y_0)\, dx.$$

Proof. Since $\langle a, b\rangle \times \langle c, d\rangle$ is compact, f is uniformly continuous. Thus, for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\max\{|x_1 - x_2|, |y_1 - y_2|\} < \delta$ implies $|f(x_1, y_1) - f(x_2, y_2)| < \varepsilon$.

Let $\lim_n y_n = y_0$. Set $g(x) = f(x, y_0)$ and $g_n(x) = f(x, y_n)$. If $|y_n - y_0| < \delta$ as above, we have $|g_n(x) - g(x)| < \varepsilon$ independently of x, hence $g_n \rightrightarrows g$ so that by 2.3, $\lim_n \int_a^b g_n(x)\mathrm{d}x = \int_a^b g(x)\mathrm{d}x$, that is, $\lim_n \int_a^b f(x, y_n)\mathrm{d}x = \int_a^b f(x, y_0)\mathrm{d}x$, and the statement follows from Lemma 2.4. □

2.4.2. Proposition. *Let $f : \langle a,b\rangle \times \langle c,d\rangle \to \mathbb{R}$ be continuous and let it have a continuous partial derivative $\frac{\partial f(x,y)}{\partial y}$ in $\langle a,b\rangle \times (c,d)$. Then $F(y) = \int_a^b f(x,y)dx$ has a derivative in (c,d) and we have*

$$\frac{d}{dy}\int_a^b f(x,y)dx = \int_a^b \frac{\partial f(x,y)}{\partial y}dx.$$

Proof. Fix $y \in (c,d)$ and choose an $\alpha > 0$ such that $c < y-\alpha < y+\alpha < d$. Set $F(y) = \int_a^b f(x,y)\mathrm{d}x$ and define

$$g(x,t) = \begin{cases} \frac{1}{t}(f(x,y+t) - f(x,y)) & \text{for } \ t \neq 0, \\ \frac{\partial f(x,y)}{\partial y} \ \text{ for } \ t = 0. \end{cases}$$

This function g is continuous on the compact $\langle a,b\rangle \times \langle -\alpha, +\alpha\rangle$. This is obvious in the points (x,t) with $t \neq 0$, and since by Lagrange theorem

$$g(x,t) - g(x,0) = \frac{1}{t}(f(x,y+t) - f(x,y)) - \frac{\partial f(x,y)}{\partial y} = \frac{\partial f(x,y+\theta t)}{\partial y} - \frac{\partial f(x,y)}{\partial y},$$

the continuity in $(x,0)$ follows from the continuity of the partial derivative.

Hence we can apply 2.4.1 to obtain

$$\lim_{t\to 0}\int_a^b g(x,t)\mathrm{d}x = \int_a^b \frac{\partial f(x,y)}{\partial y}\mathrm{d}x.$$

and since for $t \neq 0$

$$\int_a^b g(x,t) = \frac{1}{t}\left(\int_a^b f(x,y+t) - \int_a^b f(x,y)\right) = \frac{1}{t}(F(y+t) - F(y))$$

the statement follows. □

3. The space of continuous functions.

3.1. Let $X = (X,d)$ be a metric space. Denote by

$$C(X)$$

the set of all bounded continuous real functions endowed by the metric

$$d(f,g) = \sup\{|f(x) - g(x)| \mid x \in X\}$$

(checking that thus defined d is indeed a metric is straightforward).

3.1.1. Note. There is no harm in allowing infinite distances; in effect, it has advantages. However, we have worked so far with finite distances and we will keep doing so. This is why we assume our functions bounded. But

- most of what we will do in this section holds without the boundedness, and
- if X is compact we have the functions bounded anyway.

3.2. Proposition. *A sequence $(f_n)_n$ converges to f in $C(X)$ if and only if $f_n \rightrightarrows f$.*

Proof. We have $\lim_n f_n = f_n$ in $C(X)$ if for every $\varepsilon > 0$ there is an n_0 such that $d(f_n, f) = \sup\{|f_n(x) - f(x)| \,|\, x \in X\} \le \varepsilon$ for $n \ge n_0$. This is to say that for every $\varepsilon > 0$ there is an n_0 such that for all $n \ge n_0$ and for all $x \in X$ it holds that $|f_n(x) - f(x)| \le \varepsilon$, which is the definition of uniform convergence. □

3.3. Observation. *Let a be a real number. Then the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = |a - x|$ is continuous.*

(Indeed, we have $|a-y| \le |a-x| + |x-y|$, hence $|a-y| - |a-x| \le |x-y|$ and by symmetry $||a-y| - |a-x|| \le |x-y|$.)

3.3.1. Theorem. *$C(X)$ is a complete metric space.*

Proof. Let $(f_n)_n$ be a Cauchy sequence in $C(X)$. Thus, for every $\varepsilon > 0$ there is an n_0 such that

$$\forall m, n \ge n_0, \quad \forall x \in X \quad |f_m(x) - f_n(x)| < \varepsilon. \qquad (*)$$

Thus in particular each of the sequences $(f_n(x))_n$ is Cauchy in $\mathbb{R}$ and we have a limit $f(x) = \lim_n f_n(x)$.

Fix an $m \ge n_0$. Taking a limit in $(*)$ and using Observation 3.3 we obtain

$$\forall m \ge n_0, \quad |f_m(x) - \lim_n f_n(x)| = |f_m(x) - f(x)| \le \varepsilon,$$

independently on x.

Thus $f_n \rightrightarrows f$ and hence

- by 1.3, f is continuous; it is also bounded since if we fix an $m \ge n_0$ obviously $|f(x)| \le |f_m(x)| + \varepsilon$ (and f_m is bounded) and hence $f \in C(X)$,

- and by 3.2 $\lim_n f_n = f$ in $C(X)$.

□

4. Series of continuous functions.

4.1. Series of continuous functions

$$\sum_{n=0}^{\infty} f_n(x) = f_0(x) + f_1(x) + \cdots + f_n(x) + \cdots$$

are treated as limits

$$\lim_n \sum_{k=0}^{n} f_k(x)$$

of the partial finite sums. However, as with series of numbers, for obvious reasons, the really important ones are the *absolutely convergent series of functions*, namely those for which $\sum_{n=0}^{\infty} f_n(x)$ is absolutely convergent for each x in the domain. In particular (recall III.2.4)

> *if $\sum_{n=0}^{\infty} f_n(x)$ is absolutely convergent then the sum does not depend on the order of the summands.*

4.2. A series $\sum_{n=0}^{\infty} f_n(x)$ is said to *coverge uniformly* (resp. *converge locally uniformly*) if

$$(\sum_{k=0}^{n} f_k(x))_n$$

is a uniformly convergent (resp. locally uniformly convergent) sequence of functions.

In the first case we will sometimes use the symbol

$$\sum_{n=0}^{\infty} f_n(x) \rightrightarrows f(x) \quad \text{or} \quad f_0(x) + f_1(x) + \cdots + f_n(x) + \cdots \rightrightarrows f(x).$$

From 1.3 we immediately obtain

4.3. Proposition. *Let $\sum_{n=0}^{\infty} f_n(x)$ be a uniformly convergent series of functions. Then the sum is continuous.*

From 2.2 we obtain, using the fact that the derivative of finite sums are sums of derivatives,

4.4. Proposition. *Let the series $\sum_{n=0}^{\infty} f_n(x)$ converge to $f(x)$, let the functions $f_n(x)$ have derivatives $f'_n(x)$ and let the series $\sum_{n=0}^{\infty} f'_n(x)$ uniformly converge. Then $f(x)$ has a derivative*

$$\left(\sum_{n=0}^{\infty} f_n(x)\right)' = \sum_{n=0}^{\infty} f'_n(x).$$

4.5. The following extension of the criterion III.2.2 will be very useful.

Theorem. *Let $b_n \geq 0$ and let $\sum_{n=0}^{\infty} b_n$ converge. Let $f_n(x)$ be real functions on a domain D such that $|f_n(x)| \leq b_n$ for all $x \in D$. Then $\sum_{n=0}^{\infty} f_n(x)$ converges on D absolutely and uniformly.*

Proof. The absolute convergence is in the definition. Now let $\varepsilon > 0$. The sequence $(\sum_{k=0}^{n} b_k)_n$ is Cauchy and hence there is an n_0 such that for $m, n+1 \geq n_0$, $\sum_{n}^{m} b_k < \varepsilon$. Then we have for $x \in D$

$$\left|\sum_{n+1}^{m} f_k(x)\right| \leq \sum_{n+1}^{m} |f_k(x)| \leq \sum_{n+1}^{m} b_k < \varepsilon$$

and hence in $C(D)$

$$d(\sum_{k=0}^{m} f_k, \sum_{k=0}^{n} f_k) = \sup\{\left|\sum_{n+1}^{m} f_k(x)\right| \mid x \in D\} \leq \varepsilon.$$

Thus, the sequence $(\sum_{k=0}^{n} f_k)_n$ is Cauchy in $C(D)$ and by 3.2 (and the definition 2.2) $\sum_{k=0}^{\infty} f_k(x)$ uniformly converges. □

4.5.1. Corollary. *Let $f(x) = \sum_{n=0}^{\infty} f_n(x)$ converge and let $f_n(x)$ have derivatives. Let there be a convergent series $\sum_{n=0}^{\infty} b_n$ such that $|f'_n(x)| \leq b_n$ for all n and x. Then the derivative of f exists and we have*

$$\left(\sum_{n=0}^{\infty} f_n(x)\right)' = \sum_{n=0}^{\infty} f'_n(x).$$

XIX. Power series

1. Limes superior.

1.1. We will allow infinite limits of sequenced of real numbers, that is,

$$\lim_n a_n = +\infty \quad \text{if} \quad \forall K\ \exists n_0\ (n \geq n_0\ \Rightarrow\ a_n \geq K),$$
$$\lim_n a_n = -\infty \quad \text{if} \quad \forall K\ \exists n_0\ (n \geq n_0\ \Rightarrow\ a_n \leq K),$$

and infinite suprema for $M \subseteq \mathbb{R}$,

$$\sup M = +\infty \quad \text{if } M \text{ has no finite upper bound.}$$

We will set

$$(+\infty)\cdot a = a\cdot(+\infty) = +\infty \ \text{ for positive } a, \text{ and}$$
$$(+\infty) + a = a + (+\infty) = +\infty \ \text{ for finite } a.$$

1.2. For a sequence $(a_n)_n$ of real numbers define *limes superior* as the number

$$\limsup_n a_n = \lim_n \sup_{k\geq n} a_k = \inf_n \sup_{k\geq n} a_k.$$

The second equality is obvious: the sequence $(\sup_{k\geq n} a_k)_n$ is a non-increasing one.

Limes superior is defined for an arbitrary sequence. Furthermore we have

1.2.1. Observation. *If* $\lim_n a_n$ *exists then* $\limsup_n a_n = \lim_n a_n$.

(If $\lim_n a_n = -\infty$ then $(\sup_{k\geq n} a_k)_n$ has no lower bound and if $\lim_n a_n = +\infty$ then $\sup_{k\geq n} a_k = +\infty$ for all n. Let $a = \lim_n a_n$ be finite and let $\varepsilon > 0$. Then $|a_n - a| < \varepsilon$ implies that $|\sup_{k\geq n} a_k - a| \leq \varepsilon$.)

1.3. Proposition. *Suppose* $a_n, b_n \geq 0$*; set* $a = \limsup_n a_n$*. Let there exist a finite and positive* $b = \lim_n b_n$*. Then*

$$\limsup_n a_n b_n = ab.$$

Proof. I. For an $\varepsilon > 0$ choose an n_0 such that

$$n \geq n_0 \quad \Rightarrow \quad b_n < b + \varepsilon \ \text{ and } \ \sup_{k\geq n} a_k \leq a + \varepsilon.$$

Then we have for $n \geq n_0$

$$\sup_{k\geq n} a_k b_k \leq (\sup_{k\geq n} a_k)(b+\varepsilon) \leq (a+\varepsilon)(b+\varepsilon) = ab + \varepsilon(a+b+\varepsilon)$$

and as $\varepsilon > 0$ was arbitrary, we see that $\limsup_n a_n b_n \leq ab$ (this also includes the case of $a = +\infty$ where, of course, the estimate is trivial).

II. For $\varepsilon > 0$ sufficiently small to have $b - \varepsilon > 0$ choose an n_0 such that

$$n \geq n_0 \quad \Rightarrow \quad b_n > b - \varepsilon.$$

Since $\sup_{k\geq n} a_k \geq \inf_m \sup_{k\geq m} a_k = a$ for every n, there exist $k(n) \geq n$ such that

$$a_{k(n)} \geq a - \varepsilon \;\text{ if } a \text{ is finite, and}$$
$$a_{k(n)} \geq n \;\text{ if } a = +\infty.$$

Then for $n \geq n_0$,

$$(a-\varepsilon)(b-\varepsilon) \leq a_{k(n)} b_{k(n)} \quad \text{resp.} \;\; n(b-\varepsilon) \leq a_{k(n)} b_{k(n)} \text{ if } a = +\infty$$

so that

$$ab - \varepsilon(a+b-\varepsilon) \leq \sup_m a_m b_m \quad \text{resp.} \;\; n(b-\varepsilon) \leq \sup_m a_m b_m \text{ if } a = +\infty$$

and since $\varepsilon > 0$ was arbitrary and since $n(b-\varepsilon)$ is arbitrarily large, we also have $ab \leq \limsup_n a_n b_n$. □

1.4. Note. There is a counterpart of the limes superior called *limes inferior* defined for an arbitrary sequence $(a_n)_n$ of real numbers by setting

$$\liminf_n a_n = \lim_n \inf_{k\geq n} a_k = \sup_n \inf_{k\geq n} a_k.$$

Its properties are quite analogous.

2. Power series and the radius of convergence.

Until Chapter XXI we will not systematically treat complex functions of complex variable, but in this section it will be of advantage to consider the coefficients a_n, c and the variable x complex. This is not only because the proof of the theorem on the radius of convergence is literally the same;

what is at the moment perhaps more important, it will explain the seemingly paradoxical behaviour of some *real* power series (see 2.4 below).

2.1. Let a_n and c be complex numbers. A *power series* with coefficients a_n and *center* c is the series

$$\sum_{n=0}^{\infty} a_n(x-c)^n.$$

In this section it will be understood as a function of a complex variable x; the domain will be specified shortly.

2.2. The *radius of convergence* of a power series $\sum_{n=0}^{\infty} a_n(x-c)^n$ is the number

$$\rho = \rho((a_n)_n) = \frac{1}{\limsup_n \sqrt[n]{|a_n|}}\ .$$

2.3.1. Theorem. *Let $\rho = \rho((a_n)_n)$ be the radius of convergence of $\sum_{n=0}^{\infty} a_n(x-c)^n$ and let $r < \rho$. Then the series $\sum_{n=0}^{\infty} a_n(x-c)^n$ converges uniformly and absolutely in the set $\{x \mid |x-c| \le r\}$.*

On the other hand, the series does not converge if $|x-c| > \rho$.

Proof. I. For a fixed $r < \rho$ choose a q such that

$$r \cdot \inf_n \sup_{k \ge n} \sqrt[k]{|a_k|} < q < 1.$$

Then there is an n such that for all $k \ge n$,

$$r \cdot \sup_{k\ge n} \sqrt[k]{|a_k|} < q \quad \text{and hence} \quad r \cdot \sqrt[k]{|a_k|} < q.$$

For a sufficiently large $K \ge 1$ we have, moreover, $r^k \cdot |a_k| < Kq^k$ for all $k \le n$ so that

$$\text{if } |x-c| \le r \text{ then } \ |a_k(x-c)^k| \le Kq^k \text{ for all } k$$

and we see by XVIII.3.5 that $\sum_{n=0}^{\infty} a_n(x-c)^n$ converges uniformly and absolutely in $\{x \mid |x-c| \le r\}$.

II. If $|x-c| > \rho$ then $|x-c| \cdot \inf_n \sup_{k\ge n} \sqrt[k]{|a_k|} > 1$ and hence we have $|x-c| \cdot \sup_{k\ge n} \sqrt[k]{|a_k|} > 1$ for all n. Consequently, for each n there is a $k(n) \ge n$ such that $|x-c| \cdot \sqrt[k(n)]{|a_{k(n)}|} > 1$ and hence $|a_{k(n)}(x-c)^{k(n)}| > 1$ so that the summands of the series do not even converge to zero. □

From 2.3.1 and XVIII.1.5 we obtain

2.3.2. Corollary. *A power series $\sum_{n=0}^{\infty} a_n(x-c)^n$ locally uniformly converges on the open disc $D = \{x \mid |x-c| < \rho((a_n)_n)\}$ and converges in no x with $|x-c| > \rho$. Consequently, the function $f(x) = \sum_{n=0}^{\infty} a_n(x-c)^n$ is continuous on D.*

2.4. Notes. 1. Theorem 2.3.1 is in introductory texts of real analysis often interpreted as a a statement about a real power series and its convergence on the interval $(c-\rho, c+\rho)$. The proofs in the real context and in the complex one (as we have interpreted it) are literally the same (although of course the triangle inequality for the absolute value of a complex number is a much deeper fact than the triangle inequality in $\mathbb{R}$).

2. The domain D of convergence of a power series is bounded by the open and closed discs

$$\{x \mid |x-c| < \rho\} \subseteq D \subseteq \{x \mid |x-c| \leq \rho\}$$

in the complex plane and cannot expand beyond the closed one. This explains the seemingly paradoxical behaviour of the convergence on the real line. Take for instance the real function

$$f(x) = \frac{1}{1+x^2}.$$

In the interval $(-1, 1)$ it can be written as the power series

$$1 - x^2 + x^4 - x^6 + x^8 - \cdots$$

which abruptly stops converging after $+1$ (and for $x < -1$). There is no obvious reason if we think just in real terms: $f(x)$ gets just smaller after the bounds. But in the complex plane the discs $\{x \mid |x| < r\}$ as domains of $f(x)$ have to stop expanding after reaching $r = 1$: there are obstacles in the points i and $-i$ although there is none on the real line.

3. Theorem 2.3.1 speaks about the convergence in the points of $\{x \mid |x| < \rho\}$ and the divergence for $|x| > \rho$. For the points of the circle $C = \{x \mid |x| = \rho\}$ there is no general rule.

2.5. Proposition. *The radius of convergence of the series $\sum_{n=1}^{\infty} na_n(x-c)^{n-1}$ is the same as the radius of convergence of the series $\sum_{n=0}^{\infty} a_n(x-c)^n$.*

Proof. For $x \neq 0$ the series $\mathcal{S} = \sum_{n=1}^{\infty} na_n(x-c)^{n-1}$ obviously converges iff the series $\mathcal{S}_1 = \sum_{n=}^{\infty} na_n(x-c)^n = x(\sum_{n=1}^{\infty} na_n(x-c)^{n-1})$ does. By 1.3 we have

$$\limsup_n \sqrt[n]{n|a_n|} = \limsup_n \sqrt[n]{n}\sqrt[n]{|a_n|} = \lim_n \sqrt[n]{n}\cdot\limsup_n \sqrt[n]{|a_n|} = \limsup_n \sqrt[n]{|a_n|}$$

since $\lim_n \sqrt[n]{n} = \lim_n \mathrm{e}^{\frac{1}{n}\lg n} = \mathrm{e}^0 = 1$. Consequently, the radius of convergence of $\mathcal{S}_1$, and hence of $\mathcal{S}$, is equal to $\rho((a_n)_n)$. $\square$

2.5.1. By XVIII.3.5.1 we now obtain

Theorem. *The series* $f(x) = \sum_{n=0}^{\infty} a_n(x-c)^n$ *has a derivative*

$$f'(x) = \sum_{n=1}^{\infty} na_n(x-c)^{n-1}$$

and also a primitive function

$$(\int f)(x) = C + \sum_{n=0}^{\infty} \frac{a_n}{n+1}(x-c)^{n+1}$$

in the whole interval $J = (c-\rho, c+\rho)$ *where* $\rho = \rho((a_n)_n)$.

In other words, one can differentiate and integrate power series by individual summands.

3. Taylor series.

3.1. Recall VIII.7.3. Let a function f have derivatives $f^{(n)}$ of all orders in an interval $J = (c-\Delta, c+\Delta)$. Then we have for each n and $x \in J$,

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + R_n(f,x)$$

with $R_n(f,x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}$ where ξ is a number between c and x.

3.1.1. Proposition and definition. *Let a function* f *have derivatives* $f^{(n)}$ *of all orders in an interval* $J = (c-\Delta, c+\Delta)$. *Let us have for the remainder* $R_n(f,x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k$

$$\lim_n R_n(f,x) = 0 \ \ \textit{for all } x \in J.$$

Then the function $f(x)$ can be expressed in J as the power series

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

This power series is called the Taylor series *of f.*

Proof. We have

$$\lim_n\sum_{k=0}^{n}\frac{f^{(k)}(c)}{k!}(x-c)^k=\lim_n(f(x)-R_n(f,x)=f(x)-\lim_n R_n(f,x)=f(x).$$

□

3.2. Examples. 1. For an arbitrary large K we have

$$\lim_n\frac{K^n}{n!}=0$$

(indeed, if we put $k_n=\frac{K^n}{n!}$ then for $n>2K$, $k_{n+1}<\frac{k_n}{2}$ and hence $k_{n+m}<2^{-m}k_n$). Consequently for any x the remainder in the Taylor formula VIII.7.3 for e^x, $\sin x$ and $\cos x$ converges to zero and we have the Taylor series

$$\begin{aligned}
e^x&=1+\frac{x}{1!}+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}+\cdots\ ,\\
\sin x&=\frac{x}{1!}-\frac{x^3}{3!}+\frac{x^5}{5!}-\cdots\pm\frac{x^{2n+1}}{(2n+1)!}\mp\cdots\ ,\ \text{and}\\
\cos x&=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots\pm\frac{x^{2n+2}}{(2n+2)!}\mp\cdots
\end{aligned}$$

all of them with the radius of convergence equal to $+\infty$.

2. Just the existence of derivatives of all orders does not suffice: the remainder does not automatically converge to zero. Consider the example from VIII.7.4,

$$f(x)=\begin{cases}e^{-\frac{1}{x^2}} & \text{for } x\neq 0,\\ 0 & \text{for } x=0\end{cases}$$

where$f^{(k)}(0)=0$ for all k.

3.3. Let $f(x) = \sum_{n=0}^{\infty} a_n(x-c)^n$ be a power series with the radius of convergence ρ. Then we have by 2.5.1

$$\begin{aligned} f^{(k)}(x) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)a_n(x-c)^{n-k} = \\ &= k!a_k + \sum_{n=k+1}^{\infty} n(n-1)\cdots(n-k+1)a_n(x-c)^{n-k}. \end{aligned} \qquad (*)$$

3.3.1. Proposition. 1. *The coefficients of a power series $f(x) = \sum_{n=0}^{\infty} a_n(x-c)^n$ are uniquely determined by the function f.*

2. *A power series is its own Taylor series.*

Proof. 1. By $(*)$ we have $a_k = \frac{f^{(k)}(x)}{k!}$.

2. If the series $f(x) = \sum_{n=0}^{\infty} a_n(x-c)^n$ converges we have

$$f(x) = \sum_{n=0}^{k} a_n(x-c)^n + \sum_{n=k+1}^{\infty} a_n(x-c)^n$$

and the remainder $R_k(f,x) = \sum_{n=k+1}^{\infty} a_n(x-c)^n$ converges to zero because of the convergence of the series $\sum_{n=0}^{\infty} a_n(x-c)^n$. Moreover, as we have already observed, we have $a_k = \frac{f^{(k)}(x)}{k!}$. □

3.4. It is not always easy to obtain general formula for the coefficients $\frac{f^{(n)}(c)}{n!}$ of the Taylor series of a function f by taking derivatives. Sometimes, however, we can determine the Taylor series very easily using Proposition 3.3.1 and Theorem 2.5.1.

3.4.1. Example: logarithm. We have $(\lg(1-x))' = \frac{1}{x-1}$. Since

$$\frac{1}{x-1} = -1 - x - x^2 - x^3 - \cdots$$

we have by 2.5.1 (and 3.3.1)

$$\lg(1-x) = C - x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{4}x^4 - \cdots$$

and since $\lg 1 = \lg(1-0) = 0$ we have $C = 0$ and obtain the well known formula $\lg(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$.

3.4.2. Example: arcustangens. We have $\arctan(x)' = \frac{1}{1+x^2}$. Since

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - \cdots$$

we obtain by taking the primitive function

$$\arctan(x) = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \frac{1}{9}x^9 - \cdots \qquad (*)$$

The additive constant is 0, because $\arctan(0) = 0$.

3.4.3. A not very effective but elegant formula for π. The formula $(*)$ suggest that

$$\frac{\pi}{4} = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots .$$

This equation really holds true, but it is not quite immediate. Why: the radius of convergence of the power series $f(x) = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \frac{1}{9}x^9 - \cdots$ is 1 so that the argument 1 is on the border of the disc of convergence $\{x \,|\, |x| < 1\}$ about which the general propositions do not say anything (recall 2.4). The function arctan is continuous and for $|x| < 1$ we have $\arctan(x) = f(x)$. Hence we have to prove that

$$\lim_{x \to 1-} f(x) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots .$$

Consider $\varepsilon > 0$. The series $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$ converges (albeit not absolutely) and hence there is an n such that $|P_n| < \varepsilon$ for $P_n = \frac{1}{2n+1} - \frac{1}{2n+3} + \frac{1}{2n+5} - \cdots$. Now choose a $\delta > 0$ such that for $1 - \delta < x < 1$ and for $P_n(x) = \frac{1}{2n+1}x^{2n+1} - \frac{1}{2n+3}x^{2n+3} + \frac{1}{2n+5}x^{2n+5} - \cdots$ we have

$$\begin{aligned} &|P_n(x)| < \varepsilon \quad \text{and} \\ &|(x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \pm \frac{1}{2n-1}x^{2n-1}) - (1 - \frac{1}{3} + \frac{1}{5} - \cdots \pm \frac{1}{2n-1})| < \varepsilon. \end{aligned}$$

Now we can estimate for $1 - \delta < x < 1$ the difference between $f(x)$ and the alternating sequence $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$:

$$\begin{aligned}
&|f(x) - (1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots)| = \\
&\qquad |(x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \pm \frac{1}{2n-1}x^{2n-1} \mp P_n(x)) - \\
&\qquad\qquad\qquad - (1 - \frac{1}{3} + \frac{1}{5} - \cdots \pm \frac{1}{2n-1} \mp P_n)| \leq \\
&\qquad \leq |(x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \pm \frac{1}{2n-1}x^{2n-1}) - \\
&\qquad\qquad\qquad - (1 - \frac{1}{3} + \frac{1}{5} - \cdots \pm \frac{1}{2n-1})| + |P_n(x)| + |P_n| < 3\varepsilon.
\end{aligned}$$

Note that there is indeed a one-sided limit only: $f(x)$ does not make sense for $x > 1$.

XX. Fourier series

1. Periodic and piecewise smooth functions.

1.1. Piecewise continuous and smooth functions. A real function $f : \langle a, b\rangle :\to \mathbb{R}$ is *piecewise continuous* if there are

$$a = a_0 < a_1 < a_2 < \cdots < a_n = b$$

such that

- f is continuous on each open interval (a_j, a_{j+1}) and
- there exist finite one-sided limits $\lim_{x\to a_j+} f(x)$, $j = 0, \dots, n-1$ and $\lim_{x\to a_j-} f(x)$, $j = 1, \dots, n$.

It is *piecewise smooth* if, moreover,

- f has continuous derivatives on each open interval (a_j, a_{j+1}) and
- there exist finite one-sided limits $\lim_{x\to a_j+} f'(x)$, $j = 0, \dots, n-1$ and $\lim_{x\to a_j-} f'(x)$, $j = 1, \dots, n$.

For $y \in \langle a, b\rangle$ set

$$f(y+) = \lim_{x\to y+} f(x), \quad f(y-) = \lim_{x\to y-} f(x) \quad \text{and} \quad f(y\pm) = \frac{f(y+) + f(y-)}{2}.$$

We will speak of the a_i as of the *exceptional points* of f.

1.1.1. Notes and observations. 1. A piecewise continuous f can be extended to a continuous function on each $\langle a_j, a_{j+1}\rangle$. Consequently it has a Riemann integral.

2. If $y \notin \{a_0, a_1, \dots, a_n\}$ then $f(y+) = f(y-) = f(y\pm) = f(y)$. If $y = a_i$ this may or may not hold. The division points a_i in which $f(a_i+) = f(a_i-) = f(a_i\pm) = f(a_i)$ may be thought of as superfluous in the case of plain piecewise continuity, but not so in the case of piecewise smoothness: we cosider also functions without derivatives of some of the points in which they are continuous.

3. One may ask whether the points in which $f(y+) = f(y-) \neq f(y)$ have some special status. Not really: we will be mostly interested in integrals of

piecewise continuous functions, and values in isolated points will not play any role.

4. Recall VII.3.2.1. The last condition for piecewise smoothness is the same as requiring that f has one-sided derivatives in the exceptional points.

1.2. Periodic functions. A real function $f : \mathbb{R} \to \mathbb{R}$ is said to be *periodic* with *period* p if

$$\forall x \in \mathbb{R}, \quad f(x+p) = f(x).$$

1.2.1. Convention. A periodic function will be called piecewise continuous resp. piecewise smooth if the restriction $f|\langle 0,p\rangle$ is piecewise continuous resp. piecewise smooth.

1.3. A function on a compact interval represented as a periodic function (and vice versa). In this chapter it will be of advantage to represent a real function $f : \langle a,b\rangle \to \mathbb{R}$ as the periodic function $\widetilde{f} : \mathbb{R} \to \mathbb{R}$ with period $p = b - a$ defined by

$$\begin{aligned} &\widetilde{f}(x+kp) = f(x) \quad \text{for} \quad x \in (a,b) \;\; \text{and any integer } k, \\ &\widetilde{f}(a+kp) = \frac{1}{2}(f(a)+f(b)). \end{aligned}$$

If this replacement is obvious, we write simply f instead of $\widetilde{f}$; typically when computing integrals, a possible change of values in a and b does not matter.

On the other hand, we do not loose any information when studying a periodic function with period p restricted to some $\langle a, a+p\rangle$.

1.4. Proposition. *Let f be a piecewise continuous periodic function with period p. Then*

$$\int_0^p f(x)dx = \int_a^{p+a} f(x)dx \quad \textit{for any} \quad a \in \mathbb{R}.$$

Proof. Obviously $\int_b^c f = \int_{b+p}^{c+p} f$ and hence the equality holds for $a = kp$ with k an integer. Now let a be general. Choose an integer k such that $a \le kp \le a + p$. Then

$$\begin{aligned} \int_a^{p+a} f = \int_a^{kp} f + \int_{kp}^{p+a} f = \int_{p+a}^{(k+1)p} f + \int_{kp}^{p+a} f = \\ = \int_{kp}^{p+a} f + \int_{p+a}^{(k+1)p} f = \int_{kp}^{(k+1)p} f = \int_0^p f. \end{aligned}$$

□

Substituting $y = x + C$ and using XI.5.5 we obtain

1.4.1. Corollary. *For an arbitrary real C we have*

$$\int_0^p f(x+C)\,dx = \int_0^p f(x)\,dx.$$

2. A sort of scalar product.

To be able to work with $\sin kx$ and $\cos kx$ without adjustment we will confine ourselves in the following, until 4.4.1, to periodic functions with the period 2π.

2.1. If f, g are piecewise smooth on $\langle -\pi, \pi\rangle$ then obviously $f+g$ and any αf with real α are piecewise smooth. Thus the set of all piecewise smooth functions on $\langle -\pi, \pi\rangle$ constitutes a vector space

$$\mathrm{PSF}(\langle -\pi, \pi\rangle).$$

2.2. For $f, g \in \mathrm{PSF}(\langle -\pi, \pi\rangle)$ define

$$[f, g] = \int_{-\pi}^{\pi} f(x)g(x)\mathrm{d}x.$$

This function $[-,-] : \mathrm{PSF}(\langle -\pi, \pi\rangle) \times \mathrm{PSF}(\langle -\pi, \pi\rangle) \to \mathbb{R}$ behaves almost like a scalar product. See the following

2.2.1. Proposition. *We have*

(1) $[f, f] \geq 0$ *and* $[f, f] = 0$ *iff* $f(x) = 0$ *in all the non-exceptional* x,

(2) $[f+g, h] = [f, h] + [g, h]$, *and*

(3) $[\alpha f, g] = \alpha[f, g]$.

Proof is trivial; the only point that perhaps needs an explanation is the second part of (1). If $f(y) = a \neq 0$ in a non-exceptional point then for some $\delta > 0$, $f(x) > \frac{a}{2}$ for $y - \delta < x < y - \delta$ and we have

$$[f, f] = \int_{-\pi}^{pi} f^2(y)\mathrm{d}x \geq \int_{y-\delta}^{y+\delta} f^2(x)\mathrm{d}x \geq \delta\frac{a^2}{2}.$$

□

2.2.2. Note. The only flaw is in $[f, f]$ not quite implying $f \equiv 0$. But this concerns only finitely many arguments and for our purposes it is inessential.

2.3. A few formulas to recall. From the standard formulas

$$\begin{aligned}&\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha \quad \text{and}\\ &\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta\end{aligned}$$

one immediately obtains (equally standard)

$$\begin{aligned}\sin\alpha\cos\beta &= \frac{1}{2}(\sin(\alpha+\beta) - \sin(\alpha-\beta)),\\ \sin\alpha\sin\beta &= \frac{1}{2}(\cos(\alpha-\beta) - \cos(\alpha+\beta)),\\ \cos\alpha\cos\beta &= \frac{1}{2}(\cos(\alpha+\beta) + \cos(\alpha-\beta)).\end{aligned}$$

2.4. Proposition. *For any two $m, n \in \mathbb{N}$ we have $[\sin mx, \cos nx] = 0$. If $m \neq n$ then $[\sin mx, \sin nx] = 0$ and $[\cos mx, \cos nx] = 0$. Further, $[\cos 0x, \cos 0x] = [1, 1] = 2\pi$ and $[\cos nx, \cos nx] = [\sin nx, \sin nx] = \pi$ for all $n > 0$.*

Thus, the system of functions

$$\frac{1}{2\pi},\ \frac{1}{\pi}\cos x,\ \frac{1}{\pi}\cos 2x,\ \frac{1}{\pi}\cos 3x,\ \dots,\ \frac{1}{\pi}\sin x,\ \frac{1}{\pi}\sin 2x,\ \frac{1}{\pi}\sin 3x,\ \dots$$

is orthonormal in $(PSF(\langle -\pi, \pi\rangle), [-,-])$.

Proof. By 2.3 we have $\sin mx \cos nx = \frac{1}{2}(\sin(m+n)x - \sin(m-n)x)$, $\sin mx \sin nx = \frac{1}{2}(\cos(m-n)x - \cos(m+n)x)$ and $\cos mx \cos mx = \frac{1}{2}(\cos(m+n)x + \cos(m-n)x)$. Primitive function of $\sin kx$ resp. $\cos kx$ is $-\frac{1}{k}\cos kx$ resp. $\frac{1}{k}\sin kx$ and we obtain the values easily from XI.4.3.1. □

3. Two useful lemmas.

3.1. Lemma. *Let g be a piecewise continuous function on $\langle a, b\rangle$. Then*

$$\lim_{y\to+\infty}\int_a^b g(x)\sin(yx)\,dx = 0.$$

Proof. If $a_0, a_1, \dots, a_n$ are the exceptional points of g we have $\int_a^b g = \sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}} g$ and hence it suffices to prove the statement for continuous (and hence uniformly continuous) g.

Since the primitive function of $\sin(yx)$ is $-\frac{1}{y}\cos(yx)$ we have for any bounds u, v,

$$\left|\int_u^v \sin(yx)\mathrm{d}x\right| = \left|\left[-\frac{1}{k}\cos(yx)\right]_u^v\right| \leq \frac{2}{y}.$$

Choose an $\varepsilon > 0$. The function g is uniformly continuous and hence there is a $\delta > 0$ such that for $|x - z| < \delta$, $|g(x) - g(z)| < \varepsilon$. Choose a partition $a = t_1 < t_2 < \cdots < t_n = b$ of $\langle a, b\rangle$ with mesh $< \delta$, that is such that $t_{i+1} - t_i < \delta$ fot all i.

Now let

$$y > \frac{4}{\varepsilon}\sum_{i=1}^n |g(t_i)|.$$

Then we have

$$\left|\int_a^b g(x)\sin(yx)\mathrm{d}x\right| =$$
$$\left|\sum_{i=1}^n\left(\int_{t_{i-1}}^{t_i}(g(x)-g(t_i))\sin(yx)\mathrm{d}x + g(t_i)\int_{t_{i-1}}^{t_i}\sin(yx)\mathrm{d}x\right)\right| \leq$$
$$\leq \sum_{i=1}^n\int_{t_{i-1}}^{t_i}\frac{\varepsilon}{2(b-a)}\mathrm{d}x + \sum_{i=1}^n |g(t_i)|\cdot\left|\int_{t_{i-1}}^{t_i}\sin(yx)\mathrm{d}x\right| \leq \frac{\varepsilon}{2} + \sum|g(t_i)|\frac{2}{y} \leq \varepsilon.$$

□

3.1.1. Note. Lemma 3.1 is in fact a very intuitive statement. Suppose we compute $\int_a^b C\sin(yx)\mathrm{d}x$ with a constant C. Then if y is large we have approximately as much of the function under and over the x-axis. Moreover, if y is much larger still, this happens already on short subintervals of $\langle a, b\rangle$ where g behaves “almost like constant”.

3.2. Lemma. *Let* $\sin\frac{\alpha}{2} \neq 0$*. Then*

$$\frac{1}{2} + \sum_{k=1}^{n} \cos k\alpha = \frac{\sin(2n+1)\frac{\alpha}{2}}{2\sin\frac{\alpha}{2}}.$$

Proof. By the first formula in 2.3 we have

$$2\sin\frac{\alpha}{2}\cos k\alpha = \sin\left(k\alpha + \frac{\alpha}{2}\right) - \sin\left((k-1)\alpha + \frac{\alpha}{2}\right).$$

Thus,

$$2\sin\frac{\alpha}{2}\left(\frac{1}{2} + \sum_{k=1}^{n}\cos k\alpha\right) = \sin\frac{\alpha}{2} + \sum_{k=1}^{n} 2\sin\frac{\alpha}{2}\cos k\alpha =$$
$$= \sin\frac{\alpha}{2} + \sum_{k=1}^{n}\left(\sin\left(k\alpha + \frac{\alpha}{2}\right) - \sin\left((k-1)\alpha + \frac{\alpha}{2}\right)\right) =$$
$$= \sin(2n+1)\frac{\alpha}{2}.$$

□

4. Fourier series.

4.1. Recall from linear algebra representing a general vector as a linear combination of an orthonormal basis.

Let

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$$

be an orthonormal basis, that is a basis such that $\mathbf{u}_i\mathbf{u}_j = \delta_{ij}$, of a vector space V endowed with a scalar product $\mathbf{uv}$. Then a general vector $\mathbf{a}$ is expressed as

$$\mathbf{a} = \sum_{i=1}^{n} a_i\mathbf{u}_i \quad \text{where} \quad a_i = \mathbf{a}\mathbf{u}_i.$$

We will see that something similar happens with the orthonormal system from 2.4.

4.2. Let f be a piecewise smooth periodic function with period 2π. Set

$$a_k = [f, \frac{1}{\pi}\cos kx] = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos kt\mathrm{d}t \quad \text{for } k \geq 0, \quad \text{and}$$
$$b_k = [f, \frac{1}{\pi}\sin kx] = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin kt\mathrm{d}t \quad \text{for } k \geq 1.$$

We will aim at a proof that f is almost equal to

$$\frac{a_0}{2}+\sum_{k=1}^{\infty}(a_k\cos kx+b_k\sin kx).$$

Thus, the orthonormal system from 2.3 behaves smilarly like an orthonormal basis (as recalled in 4.1). There is, of course, the difference that we need *infinite sums* ("infinite linear combinations") to represent the $f\in\mathrm{PSF}(\langle-\pi,\pi\rangle)$ (which is essential) and that the f will be represented up to finitely many values (which is inessential).

4.3. Set

$$s_n(x)=\frac{a_0}{2}+\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx).$$

4.3.1. Lemma. *For every n,*

$$s_n(x)=\frac{1}{\pi}\int_0^\pi (f(x+t)+f(x-t))\cdot\frac{\sin(n+\frac{1}{2})t}{2\sin\frac{1}{2}t}dt.$$

Proof. Using the definitions of a_n and b_n and the standard formula for $\cos k(x-t)=\cos(kx-kt)$, and then using the equality from 3.2 we obtain

$$s_n(x)=\frac{1}{\pi}\int_{-\pi}^{\pi}\left(\frac{1}{2}+\sum_{k=1}^{n}(\cos kt\cdot\cos kx+\sin kt\cdot\sin kx)\right)f(t)\mathrm{d}t=$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi}\left(\frac{1}{2}+\sum_{k=1}^{n}\cos k(x-t)\right)f(t)\mathrm{d}t=\frac{1}{\pi}\int_{-\pi}^{\pi}\left(f(t)\frac{\sin(n+\frac{1}{2})(x-t)}{2\sin\frac{x-t}{2}}\right)\mathrm{d}t$$

Now substitute $t=x+z$. Then $\mathrm{d}t=\mathrm{d}z$ and $z=t-x$, and since $\sin(-u)=-\sin u$ we proceed (using also 1.4)

$$\cdots=\frac{1}{\pi}\int_{-\pi}^{\pi}\left(f(x+z)\frac{\sin(n+\frac{1}{2})z}{2\sin\frac{1}{2}z}\right)\mathrm{d}z=\frac{1}{\pi}\left(\int_0^\pi\cdots+\int_{-\pi}^{0}\cdots\right).$$

Substituting $y=-z$ in the second summand we obtain

$$\cdots=\frac{1}{\pi}\int_0^\pi\left(f(x+z)\frac{\sin(n+\frac{1}{2})z}{2\sin\frac{1}{2}z}\right)\mathrm{d}z+\frac{1}{\pi}\int_{-\pi}^{\pi}\left(f(x-y)\frac{\sin(n+\frac{1}{2})y}{2\sin\frac{1}{2}y}\right)\mathrm{d}y$$

and if we replace in the two integrals t for the respective variables, we conclude

$$\cdots = \frac{1}{\pi}\int_0^\pi (f(x+t)+f(x-t))\frac{\sin(n+\frac{1}{2})t}{2\sin\frac{1}{2}t}\mathrm{d}t.$$

□

4.3.2. Corollary. *For every n,*

$$\frac{1}{\pi}\int_0^\pi \frac{\sin(n+\frac{1}{2})t}{\sin\frac{1}{2}t}dt = 1.$$

Proof. Consider the constant funcion $f = (x \mapsto 1)$. Then $a_0 = 2$ and $a_k = b_k = 0$ for all $k \geq 1$. □

4.4. Theorem. *Let f be piecewise smooth periodic function with period 2π. Then (as $f(x\pm) = \frac{1}{2}(f(x+)+f(x-))$ $\sum_{k=1}^\infty(a_k\cos kx + b_k\sin kx)$ converges in every $x \in \mathbb{R}$ and we have (recall 1.1.)*

$$f(x\pm) = \frac{a_0}{2} + \sum_{k=1}^\infty (a_k\cos kx + b_k\sin kx).$$

Proof. By 4.3.1 and 4.3.2 we obtain

$$\begin{aligned}
&s_n(x) = \\
&= \frac{1}{\pi}\int_0^\pi (2f(x\pm)+f(x+t)-f(x+)+f(x-t)-f(x-))\frac{\sin(n+\frac{1}{2})t}{\sin\frac{1}{2}t}\mathrm{d}t = \\
&= f(x\pm)\cdot\frac{1}{\pi}\int_0^\pi \frac{\sin(n+\frac{1}{2})t}{\sin\frac{1}{2}t}\mathrm{d}t + \\
&+ \frac{1}{\pi}\int_0^\pi \left(\frac{f(x+t)-f(x+)}{t}+\frac{f(x-t)-f(x-)}{t}\right)\frac{\frac{1}{2}t}{\sin\frac{1}{2}t}\sin\left(n+\frac{1}{2}\right)t\mathrm{d}t.
\end{aligned}$$

Set

$$g(t) = \left(\frac{f(x+t)-f(x+)}{t}+\frac{f(x-t)-f(x-)}{t}\right)\frac{\frac{1}{2}t}{\sin\frac{1}{2}t}.$$

This function g is piecewise continuous on $\langle 0,\pi\rangle$: this is obvious for $t > 0$ and in $t = 0$ we have a finite limit because of the left and right derivatives of f in x and the standard $\lim_{t\to 0}\frac{\frac{1}{2}t}{\sin\frac{1}{2}t} = 1$. Thus, we can apply Lemma 3.1 (and Corollary 4.3.2) to obtain

$$\lim_{\to\infty} s_n(x) = f(x\pm).$$

□

4.4.1. Theorem 4.4 can be easily transformed for piecewise smooth periodic function with a general period p. For such f we obtain that

$$f(x\pm) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos\frac{2\pi}{p}kx + b_k \sin\frac{2\pi}{p}kx)$$

where

$$a_k = \frac{2}{p}\int_0^p f(t)\cos\frac{2\pi}{p}kt\,dt \quad \textit{for } k \geq 0, \quad \textit{and}$$
$$b_k = \frac{2}{p}\int_0^p f(t)\sin\frac{2\pi}{p}kt\,dt \quad \textit{for } k \geq 1.$$

Using the representation from 1.3 this can be applied for piecewise smooth functions on a compact interval $\langle a, b\rangle$, setting $p = b - a$.

4.4.2. The series $\frac{a_0}{2}+\sum_{k=1}^{\infty}(a_k\cos kx+b_k\sin kx)$ resp. $\frac{a_0}{2}+\sum_{k=1}^{\infty}(a_k\cos kx+b_k\sin kx)$ is called the Fourier series of f. Note that the sum is equal to $f(x)$ in all the non-exceptional points.

5. Notes.

5.1. The sums $s_n(x)$ are continuous while the resulting f is not necessarily so. Thus, the convergence of the Fourier series in 4.4 is often not uniform (recall XIX.1.3).

If the sums $\sum|a_n|$ and $\sum|b_n|$ converge, then, of course, the Fourier series converges uniformly and absolutely, and if $\sum n|a_n|$ and $\sum n|b_n|$ converge then we can take derivative by the individual summands.

5.2. Differentiating by individual summands may be false even if the resulting sum has a derivative. Here is an example. Consider $f(x) = x$ on $(-\pi, \pi\rangle$ extended to a priodic function with the period 2π. Then we obtain

$$f(x\pm) = 2(\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - \frac{1}{4}\sin 4x + \cdots).$$

$f(x)$ has a derivative 1 in all the $x \neq (2k+1)\pi$. The formal differentiating by summands would yield

$$g(x) = 2(\cos x - \cos 2x + \cos 3x - \cos 4x + \cdots)$$

and if we write $g_n(x)$ for the partial sum up to the n-th summand we obtain $g_n(0) = 2(1-1+1-\cdots+(-1)^{n+1})$, hence $g_n(0) = 0$ for n even and $g_n(x) = 2$ for n odd.

5.3. Note that for f with $f(-x) = f(x)$ all the b_n are zero, and if $f(-x) = -f(x)$ then all the a_n are zero.

5.4. Fourier series have an interesting interpretation in acoustics. A tone is described by a periodic function f. The pitch is determined by the period p (more precisely, it is given by the *frequency* $\frac{1}{p}$). The function f is seldom close to be sinusoidal. The concrete shape of f detremines the *quality* (timbre) making for the character of the sound of that or other musical instrument. In the Fourier interpretation, we see that with the first summand, a (sinusoidal) tone of the basic frequency defining the pitch, we have simultaneously sounding tones of double, triple, etc. frequency. Thus, e.g. when playing flute one gets from the first to the second octave by “blowing away the first basic tone” which results in a tone with twice the basic frequency.

XXI. Curves and line integrals

1. Curves.

In the applications in the following chapter we will need planar curves only. But for the material of the first two sections a restriction of dimension would not make anything simpler.

1.1. Parametrized curve. *A parametrized curve* in $\mathbb{E}^n$ is a continuous mapping

$$\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a, b\rangle \to \mathbb{E}_n$$

(where the compact interval $\langle a, b\rangle$ will be always assumed non-trivial, i.e, with $a < b$).

1.2. Two equivalences. Parametrized curves $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a, b\rangle \to \mathbb{E}_n$ and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n) : \langle c, d\rangle \to \mathbb{E}_n$ are said to be *weakly equivalent* if there is a homeomorphism $\alpha : \langle a, b\rangle \to \langle c, d\rangle$ such that $\boldsymbol{\psi} \circ \alpha = \boldsymbol{\phi}$. We write

$$\boldsymbol{\phi} \sim \boldsymbol{\psi}.$$

(This relation is obviously reflective: it is symmetric because the inverse of a homeomorphism is a homeomorphism, and transitive because a composition of homeomorphisms is a homeomorphism.)

Curves $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are said to be *equivalent* if there is an *increasing* homeomorphism $\alpha : \langle a, b\rangle \to \langle c, d\rangle$ such that $\boldsymbol{\psi} \circ \alpha = \boldsymbol{\phi}$. We write

$$\boldsymbol{\phi} \approx \boldsymbol{\psi}.$$

1.2.1. We will work in particular with

- the curves represented by one-to-one $\boldsymbol{\phi}$, called *simple arcs*, and
- the curves represented by $\boldsymbol{\phi}$ one-to-one with the exception of $\boldsymbol{\phi}(a) = \boldsymbol{\phi}(b)$, called *simple closed curves.*

1.2.2. Proposition. *The $\sim$-equivalence class of a simple arc or a simple closed curve is a disjoint union of precisely two $\approx$-equivalence classes.*

Proof. Since $\boldsymbol{\phi} \approx \boldsymbol{\psi}$ implies $\boldsymbol{\phi} \sim \boldsymbol{\psi}$, a $\sim$-class is a (disjoint) union of $\approx$-classes. The homeomorphism α in $\boldsymbol{\psi} \circ \alpha = \boldsymbol{\phi}$ is (because of the assumption on

$\boldsymbol{\phi}$) uniquely determined (it is uniquely determined on (a,b) and hence on the whole compact interval by IV.5.1 - there are sequences in (a,b) converging to a resp. b) and hence for instance $\boldsymbol{\phi}$ and $\boldsymbol{\phi}\circ\iota$, where $\iota(t)=-t+b+a$, are $\sim$-equivalent but not $\approx$-equivalent. Now let $\boldsymbol{\phi}\sim\boldsymbol{\psi}$, with α such that $\boldsymbol{\psi}\circ\alpha=\boldsymbol{\phi}$. Then α by IV.3.4 either increases or decreases. In the first case, $\boldsymbol{\psi}\approx\boldsymbol{\phi}$, in the second one, $\boldsymbol{\psi}\circ\alpha\circ\iota=\boldsymbol{\phi}\circ\iota$ and $\alpha\circ\iota$ increases so that $\boldsymbol{\psi}\approx\boldsymbol{\phi}\circ\iota$. □

1.3. The $\sim$-equivalence class $L=[\phi]_\sim$ is called a *curve*. The $\approx$-equivalence classes associated with this curve represent its orientations; we speak of *oriented curves* $L=[\phi]_\approx$.

By 1.2.2, a simple arc, or a simple closed curve has two orientations.

A parametrized curve $\boldsymbol{\phi}$ such that $L=[\phi]_\sim$ resp. $L=[\phi]_\approx$ is called a *parametrization* of L.

Often we freely speak of a parametrized curve $\boldsymbol{\phi}:\langle a,b\rangle\to\mathbb{E}_n$ as of a curve resp. oriented curve $\boldsymbol{\phi}$. We have in mind, of course, the associated $\sim$-resp $\approx$-class.

1.3.1. Notes. 1. One may think of a parametrized curve as of a travel on a path with $\boldsymbol{\phi}(t)$ indicating where we are at the instant t. The $\sim$-equivalence gets rid of this extra information (now we have just the railroad and not an information of a concrete train moving on it). The orientation captures the direction of the path.

The reader may think of a simpler description of a curve as of the image $\boldsymbol{\phi}[\langle a,b\rangle]$, the "geometric shape" of $\boldsymbol{\phi}$. In effect, if $\boldsymbol{\phi}$, $\boldsymbol{\psi}$ parametrize a simple arc or a simple closed curve, one can easily prove that $\boldsymbol{\phi}[\langle a,b\rangle]=\boldsymbol{\psi}[\langle c,d\rangle]$ if and only if $\boldsymbol{\phi}\sim\boldsymbol{\psi}$. But using the equivalences classes has a lot of advantages (already orienting a curve is simpler).

2. In the definitions of the equivalences $\sim$ resp. $\approx$ we have parameric curves $\boldsymbol{\phi}:\langle a,b\rangle\to\mathbb{E}_n$, $\boldsymbol{\psi}:\langle c,d\rangle\to\mathbb{E}_n$ with distinct domains. If we choose a fixed interval we can transform the $\boldsymbol{\psi}$ canonocally to $\boldsymbol{\psi}\circ\lambda:\langle a,b\rangle\to\mathbb{E}_n$ with $\lambda(t)=\frac{1}{b-a}((d-c)t+bc-da)$. Sometimes (see e.g. the definition of $\boldsymbol{\phi}*\boldsymbol{\psi}$ in 1.4 below) we freely shift the domain for convenience. This simplifies formulas and does no harm.

3. Proposition 1.2.2 holds for simple arcs and simple closed curves only. Draw a picture with $\boldsymbol{\phi}(x)=\boldsymbol{\phi}(y)$ for some $x\neq a,b$ to see that there are more than two possible orientations.

4. The word "closed" in the expression "simple closed curve" has nothing to do with the closedness of a subset of a metric space. Of course every

$\boldsymbol{\phi}[\langle a,b\rangle]$ is a compact and hence a closed subset of the $\mathbb{E}_n$ in question.

1.4. Composing oriented curves. Let K, L be oriented curves represnted by parametric ones $\boldsymbol{\phi} : \langle a,b\rangle \to \mathbb{E}_n$, $\boldsymbol{\psi} : \langle b,c\rangle \to \mathbb{E}_n$ (if the latter has not originally started in b transform it as indicated in 3.3.1.2) such that $\boldsymbol{\phi}(b) = \boldsymbol{\psi}(b)$. Set

$$(\boldsymbol{\phi} * \boldsymbol{\psi})(t) = \begin{cases} \boldsymbol{\phi}(t) & \text{for } \ t \in \langle a,b\rangle \ \text{ and} \\ \boldsymbol{\psi}(t) & \text{for } \ t \in \langle b,c\rangle. \end{cases}$$

Obviously $\boldsymbol{\phi} * \boldsymbol{\psi}$ is a continuous mapping $\langle a,c\rangle \to \mathbb{E}_n$ and we see that if $\boldsymbol{\phi} \approx \boldsymbol{\phi}_1 : \langle a_1,b_1\rangle \to \mathbb{E}_n$ and $\boldsymbol{\psi} \approx \boldsymbol{\psi}_1 : \langle b_1,c_1\rangle \to \mathbb{E}_n$ then $\boldsymbol{\phi} * \boldsymbol{\psi} \approx \boldsymbol{\phi}_1 * \boldsymbol{\psi}_1$ (note that it is essential that K, L are *oriented curves*, not just curves). Thus, the oriented curve (determined by) $\boldsymbol{\phi} * \boldsymbol{\psi}$ depends on K, L only; it will be denoted by

$$K + L.$$

(Note that the operation $K + L$ is associative.)

1.5. The opposite orientation. For an oriented curve L represented by $\boldsymbol{\phi} : \langle a,b\rangle \to \mathbb{E}_n$ define the *oriented curve with oposite orientation*

$$-L$$

as the $\approx$-class of $\boldsymbol{\phi} \circ \iota : \langle a,b\rangle \to \mathbb{E}_n$ with $\iota(t) = -t + b + a$ (recall the proof of 1.2.2). Obviously $-L$ is determined by L.

1.6. Piecewise smooth curves. Recall XX.1.1. A parametrized curve (oriented curve, or curve) $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a,b\rangle \to \mathbb{E}_n$ is said to be *piecewise smooth* if each of the ϕ_j is piecewise smooth such that, moreover, the system of the exceptional points $a = a_0 < a_1 < a_2 < \cdots < a_n = b$ can be chosen so that

- for each of the open intervals $J = (a_i, a_{i+1})$, there is a j such that $\phi_j'(t)$ is either positive or negative on the whole of J.

However, we will relax the definition of piecewise smoothness by allowing the one-sided limits $\lim_{t\to a_j+} \phi_j'(t)$ and $\lim_{t\to a_j-} \phi_j'(t)$) (in fact, the one-sided derivatives in the exceptional points – recall VII.3.2) infinite.

We will write

$$\boldsymbol{\phi}' \quad \text{for} \quad (\phi_1', \dots, \phi_n')$$

(thus in finitely many points $t \in \langle a, b\rangle$, the value $\boldsymbol{\phi}'(t)$ may be undefined; but the derivative will appear only under an integral so that it does not matter).

1.6.1. Observation. *Let curves $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a, b\rangle \to \mathbb{E}_n$ and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n) : \langle c, d\rangle \to \mathbb{E}_n$ be piecewise smooth and let α be such that $\boldsymbol{\psi} = \boldsymbol{\phi} \circ \alpha$, providing either the $\sim$- or the $\approx$-equivalence of the two parametrizations. Then α is continuous and piecewise smooth.*

(Indeed, between any two exceptional points, some of the ϕ_j is one-to-one. Then we have $\alpha = \phi_j^{-1} \circ \psi_j$ on the interval in question.)

2. Line integrals.

Convention. From now on, the curves will be always piecewise smooth.

Note. The reader may wonder why we will speak first of the line integral of the second kind and only later of the line integral of the first kind. The terminology of "first" resp. "second kind" is traditional. The reason may be in the more obvious geometric sense of the first kind line integral. But the line integral of the second kind is more fundamental (and in fact the first one can be expressed in its terms, which cannot be done reversedly).

2.1. Line integral of the second kind. Let $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a, b\rangle \to \mathbb{E}_n$ be a parametrization of an oriented curve L and let $\mathbf{f} : (f_1, \dots, f_n) : U \to \mathbb{E}_n$ be a continuous vector function defined on a $U \supseteq \boldsymbol{\phi}[\langle a, b\rangle]$. The *line integral of the second kind* over the (oriented) curve L is the number

$$(\mathrm{II})\!\int_L \mathbf{f} = \int_a^b \mathbf{f}(\boldsymbol{\phi}(t)) \cdot \boldsymbol{\phi}'(t)\,\mathrm{d}t = \sum_{j=1}^n \int_a^b f_j(\boldsymbol{\phi}(t))\phi_j'(t)\mathrm{d}t.$$

(Thus the dot in $\int_a^b \mathbf{f}(\boldsymbol{\phi}(t)) \cdot \boldsymbol{\phi}'(t)\,\mathrm{d}t$ indicates the standard scalar product of the n-tuples of reals.) If there is no danger of confusion, we write simply $\int_L$ instead of $(\mathrm{II})\!\int_L$.

Note. The reader may encounter the line integral of the second kind of, say, vector functions (P, Q) or (P, Q, R), denoted by

$$\int_L P\mathrm{d}x + Q\mathrm{d}y \quad \text{or} \quad \int_L P\mathrm{d}x + Q\mathrm{d}y + R\mathrm{d}z.$$

2.2. Proposition. *The value of the line integral $\int_L \mathbf{f}$ does not depend on the choice of parametrization of L.*

Proof. Suppose $\boldsymbol{\phi} = \boldsymbol{\psi} \circ \alpha$, with an increasing homeomorphism $\alpha : \langle a, b\rangle \to \langle c, d\rangle$. By 1.6.1, α is piecewise smooth. Then by XI.5.5

$$\sum_{j=1}^{n} \int_a^b f_j(\boldsymbol{\phi}(t))\phi_j'(t)\mathrm{d}t = \sum_{j=1}^{n} \int_a^b f_j(\boldsymbol{\psi}(\alpha(t)))\psi_j'(\alpha(t))\alpha'(t)\mathrm{d}t =$$
$$= \sum_{j=1}^{n} \int_c^d f_j(\boldsymbol{\psi}(t))\psi_j'(t)\mathrm{d}t.$$

□

2.3. Proposition. *For the operations from 1.5 and 1.4 we have*

$$\textit{(II)}\int_{-L} \mathbf{f} = -\textit{(II)}\int_L \mathbf{f} \quad \textit{and} \quad \textit{(II)}\int_{L+K} \mathbf{f} = \textit{(II)}\int_L \mathbf{f} + \textit{(II)}\int_K \mathbf{f}.$$

Proof. In the proof of 2.2 above we obtained $\int_c^d$ because α was increasing. For a decreasing α the substitution would yield $\int_d^c = -\int_c^d$, hence (II)$\int_{-L} \mathbf{f} = -$(II)$\int_L \mathbf{f}$. The other equation is obvious. □

2.4. Line integral of the first kind: just for information. Sometimes also called *the line integral acording to length*, it is defined for a non-oriented curve parametrized by $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n) : \langle a, b\rangle \to \mathbb{E}_n$. Let $f : U \to \mathbb{R}$ be a continuous real function defined on a $U \supseteq \boldsymbol{\phi}[\langle a, b\rangle]$. The idea is in modifying Riemann integral by computing the sums along a (piecewise smooth) line instead of along an interval. The sums

$$\sum_{i-1}^{k} f(\boldsymbol{\phi}(t_i))\|\boldsymbol{\phi}(t_i)) - \boldsymbol{\phi}(t_{i-1}))\|$$

considered for partitions $a = t_0 < t_1 < \cdots < t_k = b$ converge with the mesh of the partitions converging to 0 to

$$\int_a^b f(\boldsymbol{\phi}(t))\|\boldsymbol{\phi}'(t))\| \,\mathrm{d}t.$$

This integral is called the *line integral of the first kind* over L and denoted by

$$\text{(I)} \int_L f \quad \text{or} \quad \text{(I)} \int_L f(\mathbf{x})\|d\mathbf{x}\|.$$

It has a clear geometrical sense; in particular,

the length of a curve L can be expressed as

$$\textit{(I)}\int_L 1 = \int_a^b \|\boldsymbol{\phi}'(t)\|dt.$$

.

It is easy to see that the line integral of the first kind can be represented as a line integral of the second kind: we have

$$(\mathrm{I})\int_L f = (\mathrm{II})\!\int_L \mathbf{f}$$

where

$$\mathbf{f}(\boldsymbol{\phi}(t)) = \frac{\boldsymbol{\phi}'(t)}{\|\boldsymbol{\phi}'(t)\|}.$$

2.5. Complex line integral. While we will not need the line integral of the first kind in the following text, the complex line integral will be essential.

2.5.1. Complex functions of a real variable. Without much further mentioning we will identify the complex plane $\mathbb{C}$ with the Euclidean plane $\mathbb{E}_2$ (viewing $x + iy$ as (x, y) and taking into account that the absolue value of the difference $|z_1 - z_2|$ coincides with the Euclidean distance). We only must not forget that the structure of $\mathbb{C}$ is richer and that in particular we have the multiplication in the *field* $\mathbb{C}$.

A complex function of one real variable will be decomposed into two real functions,

$$f(t) = f_1(t) + if_2(t)$$

and we will define (unsurprisingly) its derivative $f'(t)$ as $f_1'(t) + if_2(t)$ and its Riemann integral as

$$\int_a^b f(t)\mathrm{d}t = \int_a^b f_1(t)\mathrm{d}t + i\int_a^b f_2(t)\mathrm{d}t.$$

A curve in $\mathbb{C}$ in a parametrized form is a mapping $\phi : \langle a, b\rangle \to \mathbb{C}$, often written as $\phi(t) = \phi_1(t) + i\phi_2(t)$. It will be treated (with respect to the definitions of the equivalence, smoothness, etc.) as the parametrized curve $\boldsymbol{\phi}(t) = (\phi_1(t), \phi_2(t))$; the values in $\mathbb{C}$ can be subjected to complex multiplication.

2.5.2. For an oriented piecewise smooth curve $\phi : \langle a, b\rangle \to \mathbb{C}$ define the *complex line integral* of a complex function of one complex variable by setting

$$\int_L f(z)\mathrm{d}z = \int_a^b f(\phi(t)) \cdot \phi'(t)\mathrm{d}t.$$

The multiplication indicated by $\cdot$ is now (unlike all the multipluications in previous pages) *the multiplication in the field* $\mathbb{C}$.

The invariance on the choice of parametrization will be seen in the following proposition.

2.5.3. Proposition. *Think of a complex function of one complex variable $f(z) = f_1(z) + if_2(z)$ as of a vector function $\mathsf{f} = (f_1, f_2)$. Then the complex line integral over L can be expressed as a line integral of second kind as follows:*

$$\int_L f(z)dz = \textit{(II)}\int_l (f_1, -f_2) + i\,\textit{(II)}\int_L (f_2, f_1).$$

Consequenly,

- $\int_L f(z)dz$ *does not depend on the choice of parametrization, and*
- *we have* $\int_{-L} f(z)dz = -\int_L f(z)dz$ *and* $\int_{L+K} f(z)dz = \int_L f(z)dz + \int_K f(z)dz$.

Proof. We have

$$\int_a^b f(\phi(t))\phi'(t)\mathrm{d}t = \int_a^b (f_1(\phi(t)) + if_2(\phi(t)))(\phi_1'(t) + i\phi_2'(t))\mathrm{d}t =$$
$$= \int_a^b (f_1(\phi(t))\phi_1'(t) - f_2(\phi(t))\phi_2'(t))\mathrm{d}t + i\int_a^b (f_1(t)\phi_2'(t) + f_2(t)\phi_1'(t))\mathrm{d}t =$$
$$= \int_a^b (f_1(\phi(t)), -f_2(\phi(t)))(\phi_1(t), \phi_2(t)) + i\int_a^b (f_2(\phi(t)), f_1(\phi(t)))(\phi_1(t), \phi_2(t))$$

(in the last line we have the scalar products of the pairs). We conclude

$$\cdots = (\mathrm{II})\int_L (f_1, -f_2) + i(\mathrm{II})\int_L (f_2, f_1).$$

□

3. Green's Theorem.

3.1. First, just for information, we will introduce some facts in a generality beyond our technical means. But in the applications in the following text we will need them only for very special cases, for which we will be able to present sufficiently rigorous proofs.

A simple closed curve L divides the plane into two connected regions (by "connected" one can understand that any two points can be connected by a curve, "divided" means that points from distinct regions cannot be so connected), one of them bounded, the other unbounded. This is the famous *Jordan theorem*, very easy to understand and visualize, but not very easy to prove. The bounded region U will be called the *region of L.* The curve C is its boundary, and the closure $\overline{U}$ is equal to $U \cup C$ and (being closed and bounded) it is compact; we will speak of $\overline{C}$ as of the *closed region*[4] *of C.*

In the following we will have to understand also the meaning of the expression "clockwise" resp. "counterclockwise oriented closed curve". This can be given an exact general sense, but we will need it only for very simple figures like circles, (perimeters of) triangles, and similar, where the meaning of the expression will be obvious. The integral over a closed region can be understood as over an interval J containing the region M, with the function extended by values zero on $J \smallsetminus M$.

3.1.1. Theorem. (Green's Theorem, Green's Formula) *Let L be a simple closed piecewise smooth curve oriented counterclockwise, and let M be its closed region. Let $\mathsf{f} = (f_1, f_2)$ be such that both f_j have continuous partial derivatives on the (open) region of L. Then*

$$(II)\!\!\int_L \mathsf{f} = \int_M \left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}\right) dx_1 dx_2 \ .$$

.

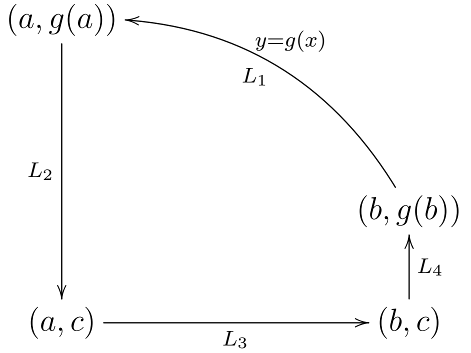
3.2. Lemma. *Let $g : \langle a, b\rangle \to \mathbb{R}$ be a smooth function, let $f(x) \geq c$ for all x. Set*

$$M = \{(x, y) \,|\, a \leq x \leq b,\ c \leq y \leq g(x)\}.$$

Let L be the closed curve which is the perimeter of M. Then the Green formula holds true for L and M.

[4] In the literature one usually speaks of *domains.* We use the term "region" to avoid confusion with domains A of mappings $f : A \to B$.

Proof. Write $L = L_1 + L_2 + L_3 + L_4$ as indicated in the following picture.



Parametrize the curves L_j by

$$\begin{aligned}
-L_1 &: \ \phi_1 : \langle a, b\rangle \to \mathbb{R}_2, \ \phi_1(t) = (t, g(t)),\\
-L_2 &: \ \phi_2 : \langle c, g(a)\rangle \to \mathbb{R}_2, \ \phi_2(t) = (a, t),\\
L_3 &: \ \phi_3 : \langle a, b\rangle \to \mathbb{R}_2, \ \phi_3(t) = (t, c),\\
L_4 &: \ \phi_4 : \langle c, g(b)\rangle \to \mathbb{R}_2, \ \phi_4(t) = (b, t).
\end{aligned}$$

Hence $\phi_1'(t) = (1, g'(t))$, $\phi_2'(t) = \phi_a'(t) = (0, 1)$ and $\phi_3'(t) = (1, 0)$ and we have

$$(\mathrm{II})\!\int_{L_1} = -\int_a^b f_1(t, g(t))\mathrm{d}t - \int_a^b f_2(t, g(t))g'(t)\mathrm{d}t,$$

$$(\mathrm{II})\!\int_{L_2} = -\int_c^{g(a)} f_2(a, t)\mathrm{d}t, \quad (\mathrm{II})\!\int_{L_3} = \int_a^b f_1(t, c)\mathrm{d}t, \quad (\mathrm{II})\!\int_{L_4} = \int_c^{g(b)} f_2(b, t)\mathrm{d}t.$$

Substituting $\tau = g(t)$ in the second integral in the formula for $(\mathrm{II})\!\int_{L_1}$ we obtain

$$(\mathrm{II})\!\int_{L_1} = -\int_a^b f_1(t, g(t))\mathrm{d}t + \int_{g(b)}^{g(a)} f_2(h(\tau), \tau)\mathrm{d}\tau$$

where h is the inverse of g.

Now, to be ready for the statement of the lemma, we will start to write x_1 for the first variable, and x_2 for the second one. For $(\mathrm{II})\!\int_L$ written as $(\mathrm{II})\!\int_{L_1} + (\mathrm{II})\!\int_{L_2} + (\mathrm{II})\!\int_{L_3} + (\mathrm{II})\!\int_{L_4}$ we now obtain (writing the $\int_c^{g(a)}$ in the formula for $(\mathrm{II})\!\int_{L_2}$ as $\int_c^{g(b)} + \int_{g(b)}^{g(a)}$)

$$\begin{aligned}
(\mathrm{II})\!\int_L = \int_c^{g(b)} (f_2(b, x_2) - f_2(a, x_2))\mathrm{d}x_2 + \int_{g(b)}^{g(a)} (f_2(h(x_2), x_2) - f_2(a, x_2))\mathrm{d}x_2 -\\
- \int_a^b (f_1(x_1, g(x_1)) - f_1(x_1, c))\mathrm{d}x_1.
\end{aligned}$$

Extending for the purpose of the integral in two variables the definition of f_j to the interval $J = \langle a, b\rangle \times \langle c, g(a)\rangle$ by values 0 in $J \smallsetminus M$ we obtain

$$f_2(b, x_2) - f_2(a, x_2) = \int_a^b \frac{\partial f_2(x_1, x_2)}{\partial x_1}\mathrm{d}x_1,$$

$$f_2(h(x_2), x_2) - f(a, x_2) = \int_a^{h(x_2)} \frac{\partial f_2(x_1, x_2)}{\partial x_1}\mathrm{d}x_1 = \int_a^b \frac{\partial f_2(x_1, x_2)}{\partial x_1}\mathrm{d}x_1, \text{ and}$$

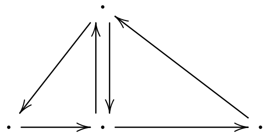
$$f_1(x_1, g(x_1)) - f_1(x_1, c) = \int_c^{g(x_1)} \frac{\partial f_1(x_1, x_2)}{\partial x_2}\mathrm{d}x_2 = \int_c^{g(a)} \frac{\partial f_1(x_1, x_2)}{\partial x_2}\mathrm{d}x_2$$

so that the formula above transforms to

$$(\mathrm{II})\!\!\int_L \boldsymbol{f} = \int_c^{g(a)} \left(\int_a^b \frac{\partial f_2(x_1, x_2)}{\partial x_1}\mathrm{d}x_1\right)\mathrm{d}x_2 - \int_a^b \left(\int_c^{g(a)} \frac{\partial f_1(x_1, x_2)}{\partial x_2}\mathrm{d}x_2\right)\mathrm{d}x_1$$

and the statement follows from Fubini's theorem (XVI.4.1). □

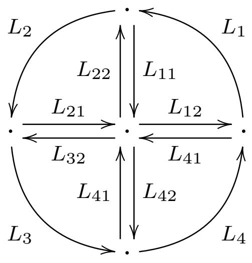
3.3. Now we have the Green's formula in particular also for quadrangles and right-angled triangles with the hypotenuse possibly curved. Using the fact that $(\mathrm{II})\!\int_L = -(\mathrm{II})\!\int_{-L}$ we obtain the formula for any figure that can be cut into such figures. Thus for instance using the decomposition as in the following picture



we infer that

3.3.1. *the Green's formula holds for any triangle.*

Or, using the decomposition of a disc as in



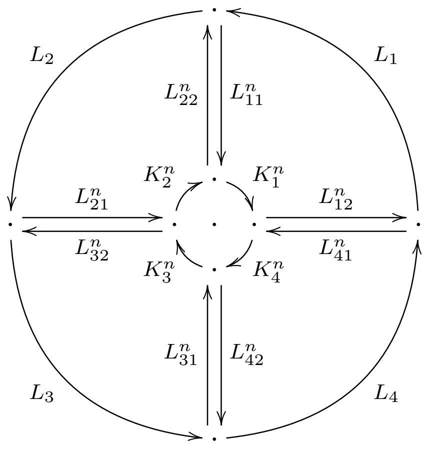
we obtain that

3.3.2. *the Green's formula holds for any disc.*
(Note, however, that in the "curved rectangles" in this decomposition the parametrization from 3.2 would not work: the function g would not have a requested derivative at one of the ends. One can use, for instance, $\phi(t) = (\cos t, \sin t)$. Or, of course, one can cut the disc into more than four pieces.)

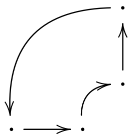
3.3.3. Note. In fact, any region of a piecewise smooth curve can be decomposed into subregions for which the formula follows from Lemma 3.2. Thich is easy tu visualize. But we will need just simple figures for which the decompositions are obvious and a painstaking proof of the general statement is not necessary.

3.4. Proposition. *Let L be a circle with center c and let M be its closed region. Let $\mathbf{f}$ be bounded on M, let partial derivatives of f_j exist and be continuous on $M \smallsetminus \{c\}$, and let $\int_M \left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}\right) dx_1 dx_2$ make sense. Then the Green's formula holds.*

Proof. Denote by K^n the circle with center c and diameter $\frac{1}{n}$ oriented *clockwise*, let $N(n)$ be its region. Let the n be large enough so that K^n (and hence also $N(n)$) is contained in M. In the following picture.



consider the (cunterclockwise oriented) simple closed curves $\widetilde{L}^n_k = L_k + L^n_{k1} + K^n_k + L^n_{k2}$ with regions $M_k(n)$. For these curves the Green's formula obviously holds (suitable carving the shapes



is easy) and we have

$$(\mathrm{II})\!\int_{\widetilde{L}^n_k} \mathbf{f} = \int_{M_k(n)} \left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}\right). \qquad (*)$$

By 2.3,

$$(\mathrm{II})\!\int_{\widetilde{L}^n_1} + (\mathrm{II})\!\int_{\widetilde{L}^n_2} + (\mathrm{II})\!\int_{\widetilde{L}^n_3} + (\mathrm{II})\!\int_{\widetilde{L}^n_4} = (\mathrm{II})\!\int_L + (\mathrm{II})\!\int_{K^n}. \qquad (**)$$

Set $V = V(x_1, x_2)$. Since we assume that the Riemann integral $\int_M V(x_1, x_2)$ exists, V is bounded, that is, we have $|V(x_1, x_2)| < A$ for some A. Since $N(n) \subseteq \langle c - \frac{1}{n}, c + \frac{1}{n}\rangle \times \langle c - \frac{1}{n}, c + \frac{1}{n}\rangle$, we have

$$\left|\int_{N(n)} V\right| < \varepsilon \ \text{ for sufficiently large } n.$$

$\mathbf{f}$ is bounded by assumption and hence we also have (we can parametrize $-K^n$, say, by $\phi(t) + \frac{1}{n}(\cos t, \sin t)$)

$$\left|(\mathrm{II})\!\int_{K^n} \mathbf{f}\right| < \varepsilon \ \text{ for sufficiently large } n.$$

Now we have by $(*)$ and $(**)$

$$(\mathrm{II})\!\int_L + (\mathrm{II})\!\int_{K^n} = \int_{M_1(k)} V + \int_{M_2(k)} V + \int_{M_3(k)} V + \int_{M_4(k)} V = \int_M V - \int_{N(k)} V$$

and hence

$$|(\mathrm{II})\!\int_L \mathbf{f} - \int_M V| \leq (\mathrm{II})\!\int_{K^n} + \int_{N(k)} V$$

and since the right hand side is arbitrarily small the statement follows. □

3.4.1. Note. 1. Proposition 3.4 is only a very special case of a general fact. The same holds for a general piecewise smooth simple closed curve L with region M and an exceptional point $c \in M$.

2. The boundedness of $\mathbf{f}$ is essential as one can see for instance in XXII.4.1 below.

XXII. Basics of complex analysis

1. Complex derivative.

1.1. In the field $\mathbb{C}$ of complex numbers we have not only all the arithmetic operations but also the metric structure allowing to speak about limits. Therefore, given a function f defined in a neighbourhood $U \subseteq \mathbb{C}$ of a point z we can ask whether there exists a limit

$$\lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$$

If it does we will speak of a *derivative* of f at z, and denote the value by

$$f'(z), \quad \frac{\mathrm{d}f(z)}{\mathrm{d}z}, \quad \frac{\mathrm{d}f}{\mathrm{d}z}z, \quad \text{etc.,}$$

similarly like in the real context. Thus for instance, like for the real power x^n we have

$$\begin{aligned}(z^n)' = \lim_{h\to 0}\frac{(z+h)^n - z^n}{h} = \lim_{h\to 0}\frac{\sum_{k=1}^{n}\binom{n}{k}x^{n-k}h^k}{h} = \\ = \lim_{h\to 0}(nz^{n-1} + h\sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^{k-2}) = nz^{n-1}.\end{aligned}$$

Similarly like in VI.1.5 we have

1.1.2. Proposition. *A function f has a derivative A at a $z \in \mathbb{C}$ if and only if there exists for a sufficiently small $\delta > 0$ a complex function $\mu : \{h \,|\, [h| < \delta\} \to \mathbb{C}$ such that*

(1) $\lim_{h\to 0}\mu(h) = 0$, *and*

(2) *for* $0 < |h| < \delta$,

$$f(z+h) - f(z) = Ah + \mu(h)h.$$

($|h|$ is of course the absolute value in $\mathbb{C}$).

(Indeed, similarly like in VI.1.5, if $A = \lim_{h\to 0}\frac{f(z+h)-f(z)}{h}$ exists then $\mu(h) = \frac{f(x+h)-f(x)}{h} - A$ has the required properties, and if a μ satisfying (1) and (2) exists then we have for small $|h|$, $\frac{f(z+h)-f(x)}{h} = A + \mu(h)$, and the limit $f'(x)$ exists and is equal to A.)

1.1.3. Corollary. *Let f have a derivative at z. Then it is continuous at this point.*

1.2. A somewhat surprising example. Proposition 1.1.2 seems to suggest that similarly like in the real case, the existence of a derivative can be interpreted as a "geometric tangent" and expresses a sort of smoothness. But it is a much more special property.

Consider $f(z) = \overline{z}$ (the complex conjugate) and compute the derivative . Writing $h = h_1 + ih_2$ we obtain

$$\frac{\overline{z+h}-\overline{z}}{h} = \frac{\overline{z}+\overline{h}-\overline{z}}{h} = \frac{\overline{h}}{h} = \begin{cases} 1 \text{ for } & h_1 \neq 0 = h_2 \\ -1 \text{ for } & h_1 = 0 \neq h_2. \end{cases}$$

Hence, there is no limit $\lim_{h\to 0} \frac{\overline{z+h}-\overline{z}}{h}$ and our f does not have a derivative at any z whatsoever, while there can be hardly any mapping $\mathbb{C} \to \mathbb{C}$ smoother than this f which is just a mirroring along the real axis.

1.3. *Complex partial derivatives*

$$\frac{\partial f(x,\zeta)}{\partial z} \quad \text{resp.} \quad \frac{\partial f(x,\zeta)}{\partial \zeta}$$

are (similarly as in the real context) derivatives as above with ζ resp. z fixed.

2. Cauchy-Riemann conditions.

Let us write a complex z as $x + iy$ with real x, y and express a complex function $f(z)$ of one complex variable as two real functions of two real variables

$$f(z) = P(x,y) + iQ(x,y).$$

2.1. Theorem. *Let f have a derivative at $z = x + iy$. Then P and Q have partial derivatives at (x, y) and satisfy the equations*

$$\frac{\partial P}{\partial x}(x,y) = \frac{\partial Q}{\partial y}(x,y) \quad \textit{and} \quad \frac{\partial P}{\partial y}(x,y) = -\frac{\partial Q}{\partial x}(x,y).$$

For the derivative f' we then have the formula

$$f' = \frac{\partial P}{\partial x} + i\frac{\partial Q}{\partial x} = \frac{\partial Q}{\partial y} - i\frac{\partial P}{\partial y}.$$

Proof. We have

$$\begin{aligned}\frac{1}{h}(f(z+h)-f(z)) = \frac{1}{h_1+ih_2}&(P(x+h_1,y+h_2)-P(x,y))+\\ &+i\frac{1}{h_1+ih_2}(Q(x+h_1,y+h_2)-Q(x,y)).\end{aligned}$$

If there is a limit $L = \lim_{h\to 0}\frac{1}{h}(f(z+h)-f(z))$ then we have in particular the limits $L = \lim_{h_1\to 0}\frac{1}{h_1}(f(z+h_1)-f(z))$ and $L = \lim_{h_2\to 0}\frac{1}{ih_2}(f(z+ih_2)-f(z)) = -i\lim_{h_2\to 0}\frac{1}{h_2}(f(z+ih_2)-f(z))$. That is,

$$\begin{aligned}L &= \lim_{h_1\to 0}\frac{1}{h_1}(P(x+h_1,y)-P(x,y)) + i\lim_{h_1\to 0}\frac{1}{h_1}(Q(x+h_1,y)-Q(x,y)) =\\ &= \frac{\partial P}{\partial x}(x,y) + i\frac{\partial Q}{\partial x}(x,y)\end{aligned}$$

and in the second case,

$$\begin{aligned}L &= -i\lim_{h_2\to 0}\frac{1}{h_2}(P(x,y+h_2)-P(x,y)) + i\lim_{h_2\to 0}\frac{1}{ih_2}(Q(x,y+h_2)-Q(x,y)) =\\ &= \frac{\partial Q}{\partial y}(x,y) - i\frac{\partial P}{\partial y}(x,y).\end{aligned}$$

□

2.1.1. The (partial differential) equations

$$\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y} \quad \text{and} \quad \frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}$$

are called the *Cauchy-Riemann equations* or the *Cauchy-Riemann conditions.* We have proved that they are necessary for the existence of a derivative. Now we will show that if we, in addition, assume continuity of the partial derivatives, these conditions suffice.

2.2. Theorem. *Let a complex function $f(z) = P(x,y) + iQ(x,y)$ satisfy in an open set $U \subseteq \mathbb{C}$ the Cauchy-Riemann equations and let all the partial derivatives involved be continuous in U. Then f has a derivative in U.*

Proof. By the Mean Value Theorem for real derivatives we have for suitable $0 < \alpha, \beta, \gamma, \delta < 1$,

$$\begin{aligned}
&\frac{1}{h}(f(z+h) - f(z)) = \\
&\quad = \frac{1}{h}(P(x+h_1, y+h_2) - P(x,y) + i(Q(x+h_1, y+h_2) - Q(x,y))) = \\
&\quad = \frac{1}{h}(P(x+h_1, y+h_2) - P(x+h_1, y) + P(x+h_1, y) - P(x,y)) + \\
&\quad\quad + i\frac{1}{h}(Q(x+h_1, y+h_2) - Q(x+h_1, y) + Q(x+h_1, y) - Q(x,y)) = \\
&\quad = \frac{1}{h}\Big(\frac{\partial P(x+h_1, y+\alpha h_2)}{\partial y} h_2 + \frac{\partial P(x+\beta h_1, y)}{\partial x} h_1 + \\
&\quad\quad + i\frac{\partial Q(x+h_1, y+\gamma h_2)}{\partial y} h_2 + i\frac{\partial Q(x+\delta h_1, y)}{\partial x} h_1\Big)
\end{aligned}$$

and using the Cauchy-Riemann equations we proceed

$$\begin{aligned}
\dots = \frac{1}{h}\Big(&- \frac{\partial Q(x+h_1, y+\alpha h_2)}{\partial x} h_2 + \frac{\partial P(x+\beta h_1, y)}{\partial x} h_1 + \\
&+ i\frac{\partial P(x+h_1, y+\gamma h_2)}{\partial x} h_2 + i\frac{\partial Q(x+\delta h_1, y)}{\partial x} h_1\Big) = \\
= \frac{\partial P(x+\beta h_1, y)}{\partial x} &+ F(h_1, h_2, \beta, \gamma)\frac{ih_2}{h} + i\frac{\partial Q(x+\delta h_1, y)}{\partial x} + G(h_1, h_2, \alpha, \delta)\frac{h_2}{h}
\end{aligned}$$

where

$$\begin{aligned}
F(h_1, h_2, \beta, \gamma) &= \frac{\partial P(x+h_1, y+\gamma h_2)}{\partial x} - \frac{\partial P(x+\beta h_1, y)}{\partial x} \quad \text{and} \\
G(h_1, h_2, \alpha, \delta) &= \frac{\partial Q(x+h_1, y+\alpha h_2)}{\partial x} - \frac{\partial Q(x+\delta h_1, y)}{\partial x}.
\end{aligned}$$

Since $|h_2| \leq |h|$ and $F(\cdots)$ and $G(\cdots)$ converge to 0 for $h \to 0$ by continuity, the expression converges to $\frac{\partial P}{\partial x}(x,y) + i\frac{\partial Q}{\partial x}(x,y)$. □

2.3. Complex functions $f : U \to \mathbb{C}$, $U \subseteq \mathbb{C}$, with continuous partial derivatives satisfying the Cauchy-Riemann conditions are said to be *holomorphic* (in U).

3. More about complex line integral. Primitive function.

Recall the complex line integral from XXI.2.5.2

$$\int_L f(z)\mathrm{d}z = \int_a^b f(\phi(t)) \cdot \phi'(t)\mathrm{d}t \tag{$*$}$$

and its representation as a line integral of second kind (XXI.2.5.3)

$$\int_L f(z)\mathrm{d}z = (\mathrm{II})\!\int_L (f_1, -f_2) + i(\mathrm{II})\!\int_L (f_2, f_1).$$

3.1. Theorem. *Let $f(z,\gamma)$ be a continuous complex function of two complex variables defined in $V \times U$, U open, and let for each fixed $z \in V$ the function $f(z,-)$ be holomorphic in U. Let L be a piecewise smooth oriented curve in V. Then for $\gamma \in U$,*

$$\frac{d}{d\gamma}\int_L f(z,\gamma)\,dz = \int_L \frac{\partial f(z,\gamma)}{\partial\gamma}\,dz.$$

Proof. Write $z = x + iy$, $\gamma = \alpha + i\beta$ and

$$f(z,\gamma) = P(x,y,\alpha,\beta) + iQ(x,y,\alpha,\beta).$$

By XXI.2.5.3 we have for $f(\gamma) = \int_L f(z,\gamma)\mathrm{d}z$ by the definition of complex line integral

$$F(\gamma) = \mathcal{P}(\alpha,\beta) + i\mathcal{Q}(\alpha,\beta)$$

where

$$\begin{aligned}\mathcal{P}(\alpha,\beta) &= (\mathrm{II})\!\int_L (P(x,y,\alpha,\beta), -Q(x,y,\alpha,\beta)),\\ \mathcal{Q}(\alpha,\beta) &= (\mathrm{II})\!\int_L (Q(x,y,\alpha,\beta), P(x,y,\alpha,\beta)).\end{aligned}$$

Since f is holomorphic at γ, it satisfies the equations $\frac{\partial P}{\partial\alpha} = \frac{\partial Q}{\partial\beta}$ and $\frac{\partial P}{\partial\beta} = -\frac{\partial Q}{\partial\alpha}$ and we obtain from the definitions of the complex line integral and its expression as in XXI.2.5.3, and from XXVIII.2.4.2 that

$$\begin{aligned}\frac{\partial\mathcal{P}}{\partial\alpha} &= (\mathrm{II})\!\int_L \left(\frac{\partial P}{\partial\alpha}, -\frac{\partial Q}{\partial\alpha}\right) = (\mathrm{II})\!\int_L \left(\frac{\partial Q}{\partial\beta}, \frac{\partial P}{\partial\beta}\right) = \frac{\partial\mathcal{Q}}{\partial\beta},\\ \frac{\partial\mathcal{P}}{\partial\beta} &= (\mathrm{II})\!\int_L \left(\frac{\partial P}{\partial\beta}, -\frac{\partial Q}{\partial\beta}\right) = -(\mathrm{II})\!\int_L \left(\frac{\partial Q}{\partial\alpha}, \frac{\partial P}{\partial\alpha}\right) = -\frac{\partial\mathcal{Q}}{\partial\alpha}\end{aligned} \tag{$*$}$$

and hence the function $F(\gamma)$ is holomorphic in U. Using the formula for the derivative from 2.1 we can conclude that

$$\int_L \frac{\partial f(z,\gamma)}{\partial\gamma}\mathrm{d}z = (\mathrm{II})\!\!\int_L \left(\frac{\partial P}{\partial\alpha}, -\frac{\partial Q}{\partial\alpha}\right) + i(\mathrm{II})\!\!\int \left(\frac{\partial Q}{\partial\alpha}, \frac{\partial P}{\partial\alpha}\right) = \frac{\partial\mathcal{P}}{\partial\alpha} + i\frac{\partial\mathcal{Q}}{\partial\alpha} = \frac{\mathrm{d}F}{\mathrm{d}\gamma}.$$

□

3.2. Theorem. *Let L be an oriented curve parametrized by ϕ and let f_n be continuous complex functions defined (at least) on L. If f_n uniformly converge to f then*

$$\int_L f = \lim_n \int_L f_n.$$

In particular if $\sum_{n=1}^\infty g_n$ is a uniformly convergent series of continuous functions defined on L then

$$\int_L \left(\sum_{n=1}^\infty g_n\right) = \sum_{n=1}^\infty \int_L g_n.$$

Proof. Since ϕ is piecewise smooth, ϕ' is bounded, say by A on L. Consequently we have

$$\begin{aligned}|f_n(\phi(t))\cdot\phi'(t) - f(\phi(t))\cdot\phi'(t)| &= |(f_n(\phi(t)) - f(\phi(t)))\cdot\phi'(t)| = \\ &= |f_n(\phi(t)) - f(\phi(t))|\cdot|\phi'(t)| \le |f_n(\phi(t)) - f(\phi(t))|\cdot A\end{aligned}$$

and hence $f_n \rightrightarrows f$ implies that $(f_n\circ\phi)\cdot\phi' \rightrightarrows (f\circ\phi)\cdot\phi'$ and we can use XVIII.4.1 and the formula $(*)$.

For the second statement it now sufficers to realize that $\int_L(f+g) = \int_L f + \int_L g$. □

The following theorem will be formulated (similarly like XXI.3.1) in a generality we will not have really proved. But we will use it only for curves with easily decomposed regions (recall XXI.3.3 through 3.4.1) which are covered by rigorous proofs.

3.3. Theorem. 1. *Let f have derivatives in an open set $U \subseteq \mathbb{C}$ and let L be an oriented piecewise smooth simple closed curve such that its closed region is contained in U. Then*

$$\int_L f(z)dz = 0.$$

2. *The formula also holds if f is undefined at one of the points of its region provided f is bounded.*

Proof. By XXI.2.5.3 we have for $f(z) = P(x,y) + iQ(x,y)$,

$$\int_L f = (\mathrm{II})\int_L (P, -Q) + i(\mathrm{II})\int_L (Q, P)$$

and by the Green's formula (whether we have in mind the situation from stayement 1, or that from statement 2) we obtain

$$\int_L f = \int_M \left(-\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) + i\int_M \left(\frac{\partial P}{\partial x} - \frac{\partial Q}{\partial y}\right) = 0$$

because by the Cauchy-Riemann equations the functions under the integrals $\int_M$ are zero. □

3.4. Recall that a subset $U \subseteq \mathbb{C}$ is *convex* if for any two points $a, b \in U$ the whole of the line segment $\{z \mid z = a + t(b-a),\ 0 \le t \le 1\}$ is contained in U.

Let f have a derivative in a convex open U. Choose an $a \in U$ and for an arbitrary $u \in U$ define

$$L(a,u)$$

as the oriented curve parametrized by $\phi(t) = a + t(u-a)$. Set

$$F(u) = \int_{L(a,u)} f(z)\mathrm{d}z.$$

3.4.1. Proposition. *The function F is a primitive function of f in U. That is, for each $u \in U$ the (complex) derivative $F'(u)$ exists and is equal to $f(u)$.*

Proof. Let h be such that $u + h \in U$. We have the piecewise smooth closed simple curve

$$L(a,u) + L(u,u+h) - L(a,u+h)$$

and hence by 3.3.1 and XXI.2.4,

$$F(u+h) - F(u) = \int_{L(a,u+h)} f - \int_{L(a,u)} f = \int_{L(u,u+h)} f.$$

Using the parametrization ϕ as above (and writing $f = P + iQ$) we obtain

$$\frac{1}{h}(F(u+h) - F(u)) = \frac{1}{h}\int_0^1 f(u+th)\mathrm{d}t =$$
$$= \frac{1}{h}\int_0^1 P(u+th)\mathrm{d}t + i\frac{1}{h}\int_0^1 Q(u+th)\mathrm{d}t = P(u+\theta_1 h) + iQ(u+\theta_2 h)$$

(for the last equality use the Integral Mean Value Theorem XI.3.3) and this converges to $f(u) = P(u) + iQ(u)$. □

3.4.2. Note. Working with a convex U was just a matter of convenience. More generally, the same can be proved for simply connected open sets U ("open sets without holes"). Instead of the $L(a,u)$ one can take oriented simple arcs L starting with a and ending in u; the integral over such an L depends on a and u only (this is an immediate consequence of 3.3.1 if two such curves L_1, L_2 meet solely in a and u - use the simple closed curve $L_1 - L_2$ - but it can be proved for curves that intersect as well. For connected but not simply connected U the situation is different, though.

4. Cauchy's Formula.

4.1. Lemma *Let K be a circle with center z and an arbitrary radius r, oriented counterclockwise. Then*

$$\int_K \frac{d\zeta}{\zeta - z} = 2\pi i.$$

Proof. Parametrize K by $\phi(t) = z + r(\cos t + i\sin t)$, $0 \le t \le 2\pi$. Then $\phi'(t) = r(-\sin t + i\cos t)$ and hence

$$\int_K \frac{\mathrm{d}\zeta}{\zeta - z} = \int_0^{2\pi} \frac{r(-\sin t + i\cos t)}{r(\cos t + i\sin t)}\mathrm{d}t = \int_0^{2\pi} i\mathrm{d}t = 2\pi i,$$

since $-\sin t + i\cos t = i(\cos t + i\sin t)$. □

4.1.1. Note. Compare this equality with the value 0 in 3.3.2. The function under the integral is holomorphic everywhere with the exception of just one point. But theorem 3.3.2 cannot be applied since f is not bounded in the region of K.

4.2. Theorem. (Cauchy's Formula) *Let a complex function of one variable f have a derivative in a set U containing the closed region of a circle K with center z, oriented counterclockwise. Then*

$$\frac{1}{2\pi i}\int_K \frac{f(\zeta)}{\zeta - z} d\zeta = f(z).$$

Proof. We have

$$\int_K \frac{f(\zeta)}{\zeta - z}\mathrm{d}\zeta = \int_K \frac{f(z)}{\zeta - z}\mathrm{d}\zeta + \int_K \frac{f(\zeta) - f(z)}{\zeta - z}\mathrm{d}\zeta =$$
$$= f(z)\int_K \frac{\mathrm{d}\zeta}{\zeta - z} + \int_K \frac{f(\zeta) - f(z)}{\zeta - z}\mathrm{d}\zeta = 2\pi i f(z) + \int_K \frac{f(\zeta) - f(z)}{\zeta - z}\mathrm{d}\zeta$$

by lemma 4.1. Now the function $g(\zeta) = \frac{f(\zeta)-f(z)}{\zeta-z}$ is holomorphic for $\zeta \neq z$. In the point z it has a limit, namely the derivative $f'(z)$. Thus it can be completed to a continuous function, hence it is bounded and we can apply 3.3.2 to see that the integral is 0. □

4.2.1. Note. Cauchy formula plays in complex differential calculus a central role similar to that played by the Mean Value Theorem in real analysis. We will see some of it in the next chapter.

4.3. Theorem. *If a complex function has a derivative in a neighbourhood of a point z then it has derivatives of all orders in this neighbourhood. More concretely, we have*

$$f^{(n)}(z) = \frac{n!}{2\pi i}\int_K \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta.$$

Proof. This is an immediate consequence of Cauchy's formula and theorem 3.1: take repeatedly partial derivatives behind the integral sign. □

4.3.1. Note. We have already observed that the existence of a derivative in the complex context differs from the differentiability in real analysis. Now we see how much stronger it is. In the next chapter we will see that in fact only power series have complex derivatives.

4.4. Corollary. *A function f is holomorphic in an open set U iff it has a derivative in U.*

In other words f has a derivative in U iff it has continuous partial derivatives satisfying the Cauchy-Riemann equations.

Proof. If f has a derivative f', it also has the second derivative f'' and hence f' has to be continuous. The other implication is trivial. □

4.4.1. Note. In other words, Theorem 2.2 can be reversed.

The question naturally arises whether Theorem 2.1 can be reversed, that is, whether just the Cauchy-Riemann equations suffice (whether they automatically imply continuity). The answer is in the negative.

4.5. Proposition. *A complex function has a primitive function in a convex open set U if and only if it has a derivative in U.*

Proof. If it has a derivative, it has a primitive function by 3.4.1. On the other hand, if F is a primitive function of f, it has by 4.3 the second derivative $F'' = f'$. □

(This is another fact strongly contrasting with real analysis.)

XXIII. A few more facts of complex analysis

1. Taylor formula.

1.1. Theorem. (Complex Taylor Series Theorem) *Let f be holomorphic in a neighbourhood V of a point a. Then in a sufficiently small neighbourhood U of a the function can be written as a power series*

$$f(z) = f(a) + \frac{1}{1!}f'(a)(z-a) + \frac{1}{2!}f''(a)(z-a)^2 + \cdots + \frac{1}{n!}f^n(a)(z-a)^n + \dots\ .$$

Proof. We have

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a} \cdot \frac{1}{1 - \frac{z-a}{\zeta-a}}. \qquad (*)$$

Take a circle K with center a and radius r such that the associated disc (the region of K) is contained in V. Choose a q with $0 < q < 1$ and a neighbourhood U of a sufficiently small such that for $z \in U$, $|z-a| < rq$. Then we have

$$\zeta \in K \quad \Rightarrow \quad \left|\frac{z-a}{\zeta-a}\right| < q < 1. \qquad (**)$$

Now we obtain for $x \in U$ from $(*)$

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a}\left(\sum_{n=0}^{\infty}\left(\frac{z-a}{\zeta-a}\right)^n\right)$$

and hence

$$\frac{f(\zeta)}{\zeta - z} = \sum_{n=0}^{\infty}\frac{f(\zeta)}{\zeta - a}\left(\frac{z-a}{\zeta-a}\right)^n.$$

The continuous function f is bounded on the compact circle K so that by $(**)$ for a suitable A,

$$\left|\frac{f(\zeta)}{\zeta - a}\left(\frac{z-a}{\zeta-a}\right)^n\right| < \frac{A}{r}\cdot q^n$$

and hence by XVIII.4.5 the series $\sum_{n=0}^{\infty}\frac{f(\zeta)}{\zeta-a}\left(\frac{z-a}{\zeta-a}\right)^n$ uniformly converges and we can use XXII.3.2 to obtain

$$\int_K \frac{f(\zeta)}{\zeta - z}\mathrm{d}\zeta = \sum_{n=0}^{\infty}\int_K \frac{f(\zeta)}{\zeta - a}\left(\frac{z-a}{\zeta-a}\right)^n \mathrm{d}\zeta = \sum_{n=0}^{\infty}(z-a)^n\int_K \frac{f(\zeta)}{(\zeta-a)^{n+1}}\mathrm{d}\zeta.$$

Using Cauchy's formula for the first integral and the formula from XXII.4.3 for the last one we conclude that

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

□

1.1.1. Notes. 1. Thus, all complex functions with derivatives can be (locally) written as power series.

2. Compare the proof of 1.1 with its counterpart in real analysis. The complex variant is actually much simpler: we just write $\frac{1}{\zeta - z}$ as a suitable power series and take the integrals of the individual summands (we just have to know we are allowed to do that), and then we apply the Cauchy's formula (and its derivatives). Of course, Cauchy's formula is a very strong tool, but this is not the only reason. In a way, in the real context we are prowing a more general theorem: we have a lot of functions that have just a few derivatives for which the theorem applies.

1.2. The exponential and goniometric functions. Using the techniques of complex analysis we can show that the goniometric functions the existence of which we have so far only assumed really exist. First *define* the exponential function for for complex variable as the power series

$$e^z = \sum_{n=1}^{\infty} \frac{1}{n!} z^n.$$

We already have it in the real context. The (real) logarithm has been proved to exist (see XII.4), e^x is its inverse and can be written as the (real) Taylor series as above.

We will need the addition formula $e^{u+v} = e^u e^v$ for general complex u and v. It is easy:

$$\begin{aligned} e^u e^v &= \left(\sum_{n=0}^{\infty} \frac{1}{n!} u^n\right)\left(\sum_{n=0}^{\infty} \frac{1}{n!} v^n\right) = \sum_{n=0}^{\infty}\left(\sum_{k+r=n} \frac{1}{k!} u^k \frac{1}{r!} v^r\right) = \\ &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} \frac{1}{k!}\frac{1}{(n-k)!} u^k v^{(n-k)}\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\sum_{k=0}^{n} \frac{n!}{k!(n-k)!} u^k v^{(n-k)}\right) = \\ &= \sum_{n=1}^{\infty} \frac{1}{n!}\left(\sum_{k=0}^{n} \binom{n}{k} u^k v^{(n-k)}\right) = \sum_{n=1}^{\infty} \frac{1}{n!}(u+v)^n. \end{aligned}$$

1.2.1. Now define (for general complex z)

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots , \quad \text{and}$$
$$\cos z = \frac{e^{iz} + e^{-iz}}{2i} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots .$$

We obviously have

$$\lim_{z\to 0} \frac{\sin z}{z} = 1$$

and the addition formulas are all we will need. We will prove, say, the formula for sinus:

$$\sin u \cos v + \sin v \cos u = \frac{1}{4i}((e^{iu} - e^{-iu})(e^{iv} + e^{-iv}) + (e^{iv} - e^{-iv})(e^{iu} + e^{-iu})) =$$
$$= \frac{1}{4i}(e^{iu}e^{iv} + e^{iu}e^{-iv} - e^{-iu}e^{iv} - e^{-iu}e^{-iv} + e^{iv}e^{iu} + e^{iv}e^{-iu} - e^{-iv}e^{iu} - e^{-iv}e^{-iu}) =$$
$$= \frac{1}{4i}(2e^{iu}e^{iv} - 2e^{-iu}e^{-iv}) = \frac{1}{2i}(e^{i(u+v)} - e^{-i(u+v)}) = \sin(u+v).$$

2. Uniqueness theorem.

2.1. Recall that two polynomials of degree n agreeing in $n+1$ arguments coincide. Viewing power series as "polynomials of infinite degree" one may for a moment surmise that two series coinciding in infinite many arguments might coincide everywhere. This conjecture is of course immediately refused by such examples as $\sin nx$ and constant 0.

But in effect this conjecture is not all that wrong. The statement holds true if only the set of points of agreement has an accumulation point (recall XVII.3.1).

2.2. First we will prove a local variant of the uniqueness theorem.

Lemma. *Let f and g be holomorphic in an open set U and let c be in U. Let $c_n \neq c$, $c = \lim_n c_n$ and $f(c_n) = g(c_n)$ for all n. Then f coincides with g in an neighbourhood of c.*

Proof. It suffices to prove that if $f(c_n) = 0$ for all n then $f(z) = 0$ in an neighbourhood of c.

Since $c \in U$, the derivative of f in c exists and hence by 1.1 we have in a sufficiently small neighbourhood V of c

$$f(z) = \sum_{k=0}^{\infty} a_k (z-c)^k.$$

If f is not constant zero in V, some of the a_k is not 0. Let a_n be the first of them. Thus,

$$f(z) = (z-c)^n (a_n + a_{n+1}(z-c) + a_{n+2}(z-c)^2 + \cdots)$$

The series $g(z) = a_n + a_{n+1}(z-c) + a_{n+2}(z-c)^2 + \cdots$ is a continuous function and $g(0) = a_n \neq 0$ and hence $g(z) \neq 0$ in a neighbourhood W of c, and $f(z) = (z-c)^n g(z)$ is in W equal to 0 only at c. But for sufficiently large n, c_n is in W, a contradiction. $\square$

2.3. Connectedness: just a few facts. A non-empty metric space X is said to be *disconnected* if there are disjoint non-empty open sets U, V such that $X = U \cup V$. It is *connected* if it is not disconnected.

X is said to be *pathwise connected* if for any two $x, y \in X$ there is a continuous mapping $\phi : \langle a, b \rangle \to X$ such that $\phi(a) = x$ and $\phi(b) = y$.

Of course, we speak of connected resp. pathwise connected *subset* of a metric space if the corresponding *subspace* is connected resp. pathwise connected.

2.3.1. Notes. 1. For good reasons, void space is defined to be disconnected. But all our spaces will be non-void.

2. Since closed sets are precisely the complements of open sets, we see that X is *disconnected* if there are disjoint non-empty closed sets A, B such that $X = A \cup B$.

3. The pathwise connectedness means, of course, connecting of arbitrary pairs of points by curves if we generalize the concept of curve from $\mathbb{E}_n$ to an arbitrary metric space.

4. If we know that a space X is connected we can prove a statement $\mathcal{V}(x)$ about elements $x \in X$ by showing that the set

$$\{x \mid \mathcal{V}(x) \text{ holds}\}$$

is non-empty, open and closed.

2.3.2. Fact. *The compact interval $\langle a,b\rangle$ is connected.*

Proof. Suppose that $\langle a,b\rangle = A\cup B$ with A,B disjoint closed subsets, and let, say, $a\in A$. Set

$$s=\sup\{x\,|\,\langle a,x\rangle\subseteq A\}.$$

Since there are $x\in A$ arbitrary close to s, $s\in\overline{A}=A$. If $s<b$ there are $x\in B$ arbitrary close to s, making $s\in\overline{B}=B$ and contradicting the disjointness. Thus, $s=b$ and B has to be empty. □

2.3.3. Fact. *Each pathwise connected space is connected.*

Proof. Suppose X is pathwise connected but not connected. Then there are non-empty open disjoint U, V such that $X=U\cup V$. Pick $x\in U$ and $y\in V$. There is a continuous $\phi:\langle a,b\rangle\to X$ such that $\phi(a)=x$ and $\phi(b)=y$. Then $U'=\phi^{-1}[U]$, $V'=\phi^{-1}[V]$ are non-empty disjoint open sets such that $U'\cup V'=\langle a,b\rangle$ contradicting 2.3.2. □

2.3.4. Fact. *An open subset of $\mathbb{E}_n$ is connected if and only if it is pathwise connected.*

Proof. Let $U\subseteq\mathbb{E}_n$ be non-empty open. For $x\in U$ define

$$U(x)=\{y\in U\,|\,\exists\phi:\langle a,b\rangle\to U,\ \phi(a)=x,\ \psi(b)=y\}.$$

Sets $U(x)$ and $U(y)$ are either disjoint or equal (if $z\in U(x)\cap U(y)$ choose oriented curves L_1, L_2 connecting x with z and z with y; then L_1+L_2 from XXI.1.4 proves that $y\in U(x)$ and using XXI.1.4 again we see that $U(y)\subseteq U(x)$).

Further, each $U(x)$ is open. Indeed let $y\in U(x)$ and let L be an oriented curve connecting x with y. Since U is open there is an $\varepsilon>0$ such that $\Omega(y,\varepsilon)\subseteq U$. Now for an arbitrary $z\in\Omega(y,\varepsilon)$ we have the oriented line segment K parametrized by $\psi=(t\mapsto y+t(z-y)):\langle 0,1\rangle\to\Omega(y,\varepsilon)$ and hence $L+K$ connecting x with z. Thus, $\Omega(y,\varepsilon)\subseteq U(x)$.

Now if U is not pathwise connected there are x,y with $U(x)\cap U(y)=\emptyset$, the set $V=\bigcup\{U(y)\,|\,y\in U,\ U(x)\cap U(y)=\emptyset\}$ is non-empty open and $U(x)\cup V=U$ and U is not connected. □

2.4. Theorem. *Let f and g be holomorphic in a connected open set U and let there exist c and $c_n\neq c$ in U such that $c=\lim_n c_n$ and $f(c_n)=g(c_n)$ for all n. Then $f=g$.*

Proof. Set

$$V=\{z\,|\,z\in U,\ f(u)=g(u)\text{ for all }u\text{ in a neighbourhood of }z\}.$$

Then V is by definition open and by 2.2 and the assumption on c it is not empty. Now let $z_n \in V$ and $\lim_n z = z$. Then by 2.2, $z \in V$ so that V is also closed, and hence $V = U$ by connectedness (recall 2.3.1.4). $\square$

3. Liouville's Theorem and Fundamental Theorem of Algebra.

3.1. Lemma. *Let f be a complex function defined on a circle K with radius r. If $|f(z)| \le A$ for all z then*

$$\left|\int_L f(z)\,dz\right| \le 8A\pi r.$$

Proof. Let L be parametrized by $\phi : \langle 0, 2\pi\rangle \to \mathbb{C}$ defined by $\phi(t) = c + r\cos t + ir\sin t$ so that $\phi'(t) = -r\sin t + ir\cos t$ and hence $|\phi_1'|, |\phi_1'| \le r$. Let $f = f_1 + if_2$. Then we have

$$\begin{aligned}
\left|\int_L f\right| &= \left|\int_0^{2\pi} f(\phi(t))\phi'(t)\mathrm{d}t\right| = \left|\int_0^{2\pi} f_1\phi_1' - \int_0^{2\pi} f_2\phi_2' + i\int_0^{2\pi} f_1\phi_2' - i\int_0^{2\pi} f_2\phi_1'\right| \le \\
&\le \left|\int_0^{2\pi} f_1\phi_1'\right| + \left|\int_0^{2\pi} f_2\phi_2'\right| + \left|\int_0^{2\pi} f_1\phi_2'\right| + \left|\int_0^{2\pi} f_2\phi_1'\right| \le \\
&\le \int_0^{2\pi} |f_1||\phi_1'| + \int_0^{2\pi} |f_2||\phi_2'| + \int_0^{2\pi} |f_1||\phi_2'| + \int_0^{2\pi} |f_2||\phi_1'| \le \\
&\le 4\int_0^{2\pi} Ar\mathrm{d}t = 4Ar\int_0^{2\pi} \mathrm{d}t = 4Ar2\pi.
\end{aligned}$$

$\square$

Note. This estimate is very rough, but it will do for our purposes.

3.2. Theorem. (Liouville) *If f is bounded and holomorphic in the whole of $\mathbb{C}$ then it is constant.*

Proof. By XXII.4.3 we have for an arbitrary circle K with center z

$$f'(z) = \frac{2!}{2\pi i}\int_K \frac{f(\zeta)}{(\zeta - z)^2}\mathrm{d}\zeta.$$

Let $|f(\zeta)| < A$ for all ζ. If we choose the circle K with diameter r we have $(\zeta - z)^2 = r^2$ for ζ on K, and hence

$$\left|\frac{f(\zeta)}{(\zeta - z)^2}\right| < \frac{A}{r^2}.$$

Hence by lemma 3.1,

$$|f'(z)| < \frac{2!}{2\pi} 8 \frac{A}{r^2} \pi r = \frac{8A}{r}.$$

Since r can be chosen arbitrarily large we see that $f'(z)$ is constant zero, and hence f is a constant. □

3.3. Theorem. (Fundamental Theorem of Algebra) *Each polynomial p of* $\deg(p) > 0$ *with complex coefficients has a complex root.*

Proof. Let a polynomial

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$$

have no root. Then the holomorphic function

$$f(z) = \frac{1}{p(z)}$$

is defined on the whole of $\mathbb{C}$. Set

$$R = 2n \max\{|a_0|, |a_1|, \dots, |a_{n-1}|, 1\}.$$

Then we have for $|z| \geq R$

$$\begin{aligned} |p(z)| &\geq |z|^n - |a_{n-1}z^{n-1} + \cdots + a_1 z + a_0(z)| \geq \\ &\geq |z|^n - |z|^{n-1}\frac{1}{2}R \geq R|z|^{n-1} - |z|^{n-1}\frac{1}{2}R = |z|^{n-1}\frac{1}{2}R \geq \frac{1}{2}R^n. \end{aligned}$$

Thus,

$$|z| \geq R \quad \Rightarrow \quad |f(z)| \leq \frac{2}{R^n}.$$

Finally, the set $\{z \,|\, |z| \leq R\}$ is compact and hence the continuous function f is bounded also for $|z| \leq R$ and hence everywhere. Thus, by Liouville's Theorem, f is constant and hence so is also p. □

4. Notes on conformal maps.

4.1. Recall from analytic geometry the formula for cosinus of the angle α between two (non-zero) vectors $\mathbf{u}, \mathbf{v}$

$$\cos\alpha = \frac{\mathbf{u}\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}.$$

In view of this formula we will understand in this section under the expression "preserving the angle between $\mathbf{u}$ and $\mathbf{v}$" preserving the value $\frac{\mathbf{u}\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$.

4.2. Let U be a connected open subset of $\mathbb{C}$. We will be mostly interested in holomorphic functions f and hence we will use (as before) the notation $f(z) = f(x+iy) = P(x,y) + iQ(x,y)$ for any $f : U \to \mathbb{C}$ with partial derivatives. In this notation we have

4.2.1. Recall the Jacobian from XV.4 and also recall that a mapping $f : U \to \mathbb{C}$ with partial derivatives is said to be regular if

$$\frac{\mathsf{D}(f)}{\mathsf{D}(z)} = \frac{\mathsf{D}(P,Q)}{\mathsf{D}(x,y)} = \det\begin{pmatrix} \frac{\partial P}{\partial x}, \frac{\partial P}{\partial y} \\ \frac{\partial Q}{\partial x}, \frac{\partial Q}{\partial y} \end{pmatrix} = \frac{\partial P}{\partial x}\frac{\partial Q}{\partial y} - \frac{\partial Q}{\partial x}\frac{\partial P}{\partial y} \neq 0. \qquad \text{(reg)}$$

4.2.2. Let $f : U \to \mathbb{C}$ be a holomorphic function. Then by the Cauchy-Riemann equations the condition (reg) transforms to

$$\frac{\partial P}{\partial x}\frac{\partial Q}{\partial y} - \frac{\partial Q}{\partial x}\frac{\partial P}{\partial y} = \frac{\partial P}{\partial x}^2 + \frac{\partial P}{\partial y}^2 = \frac{\partial Q}{\partial x}^2 + \frac{\partial Q}{\partial y}^2$$

and we observe that

A holomorphic f is regular on an open set U iff for all $z \in U$, $f'(z) \neq 0$.

4.3. A mapping $f : U \to \mathbb{C}$ is said to be *conformal* if it is regular and if it preserves angles, by which we mean preserving the angles between tangent vectors of curves when transformed by f.

We will show that conformal regular mappings are closely connected with the holomorphic ones.

4.4. Let ϕ, ψ be curves in U. A regular mapping $f : U \to \mathbb{C}$ transforms them to curves

$$\Phi = f \circ \phi \quad \text{and} \quad \Psi = f \circ \psi$$

in $\mathbb{C}$.

4.4.1. Lemma. *Let f be a holomorphic mapping. Then for the scalar product* $\mathbf{uv}$ *of tangent vectors we have (the dot $\cdot$ designates the multiplication of real numbers)*

$$\Phi'\Psi' = \frac{\mathsf{D}(f)}{\mathsf{D}(z)} \cdot \phi'\psi'.$$

Proof. Using the Cauchy-Riemann equations we obtain

$$\begin{aligned}\Phi_1'\Psi_1' + \Phi_2'\Psi_2' &= \left(\frac{\partial P}{\partial x}\phi_1' + \frac{\partial P}{\partial y}\phi_2'\right)\left(\frac{\partial P}{\partial x}\psi_1' + \frac{\partial P}{\partial y}\psi_2'\right) + \\ &+ \left(-\frac{\partial P}{\partial y}\phi_1' + \frac{\partial P}{\partial x}\phi_2'\right)\left(-\frac{\partial P}{\partial y}\psi_1' + \frac{\partial P}{\partial x}\psi_2'\right) = \\ &= (\phi_1'\psi_1' + \phi_2'\psi_2')\left(\left(\frac{\partial P}{\partial x}\right)^2 + \left(\frac{\partial P}{\partial y}\right)^2\right).\end{aligned}$$

□

4.4.2. Theorem. *A holomorphic mapping $f : U \to \mathbb{C}$ such that $f'(z) \neq 0$ for all $z \in U$ is conformal.*

Proof. From Lemma 4.4.1 we also have for the norm that $\|\Phi'\|^2 = \Psi'\Psi' = \frac{\mathsf{D}(f)}{\mathsf{D}(z)} \cdot \phi'\phi' = \frac{\mathsf{D}(f)}{\mathsf{D}(z)}\|\phi'\|^2$ so that

$$\frac{\Phi'\Psi'}{\|\Phi'\|\|\Psi'\|} = \frac{\frac{\mathsf{D}(f)}{\mathsf{D}(z)}\phi'\psi'}{\sqrt{\frac{\mathsf{D}(f)}{\mathsf{D}(z)}}\|\phi'\|\sqrt{\frac{\mathsf{D}(f)}{\mathsf{D}(z)}}\|\psi'\|} = \frac{\phi'\psi'}{\|\phi'\|\|\psi'\|}.$$

Recall 4.1. □

Note. The condition of regularity, that is, $f'(z) \neq 0$, is essential. For instance the mapping $f(z) = z^2$ redoubles the angles at the point $z = 0$.

4.5. Is, on the other hand, a conformal mapping necessarily a holomorphic one? No, because for instance the mapping

$$\mathsf{conj} = (z \mapsto \overline{z}) : \mathbb{C} \to \mathbb{C}$$

is conformal (even isometric) but not holomorphic (recall XXII.1.2). But if would be a rather cheap answer, if we would leave it at that. In fact, nothing worse than an intervence of conj can happen. We have

Theorem. *Let U be an open subset of $\mathbb{C}$ and let $f : U \to \mathbb{C}$ be a regular mapping. Then the following statements are equivalent.*

(1) *f is conformal.*

(2) *f preserves orthogonality.*

(3) *Either f or* conj $\circ f$ *is holomorphic.*

Proof. (1)⇒(2) is trivial and (3)⇒(1) is in 4.4.2 (the modification by the mapping conj is obvious).

(2)⇒(3): Write (u, v) for the tangent vector $\phi'(t)$ of a parametrization of a curve ϕ. Transformed by f it becomes

$$\left(\frac{\partial P}{\partial x}u + \frac{\partial P}{\partial y}v, \frac{\partial Q}{\partial x}u + \frac{\partial Q}{\partial y}v\right).$$

Now consider for (u, v) two orthogonal vectors (a, b) and $(-b, a)$. Then the scalar product of the transformed vectors

$$\begin{aligned}&\left(\frac{\partial P}{\partial x}a + \frac{\partial P}{\partial y}b, \frac{\partial Q}{\partial x}a + \frac{\partial Q}{\partial y}b\right)\left(-\frac{\partial P}{\partial x}b + \frac{\partial P}{\partial y}a, -\frac{\partial Q}{\partial x}b + \frac{\partial Q}{\partial y}a\right) = \\ &= (a^2 - b^2)\left(\frac{\partial P}{\partial x}\frac{\partial P}{\partial y} + \frac{\partial Q}{\partial x}\frac{\partial Q}{\partial y}\right) + \\ &\qquad + ab\left(\left(\frac{\partial P}{\partial y}\right)^2 + \left(\frac{\partial Q}{\partial y}\right)^2 - \left(\frac{\partial P}{\partial x}\right)^2 - \left(\frac{\partial Q}{\partial x}\right)^2\right)\end{aligned}$$

should be zero. In paricular for the vector $(a, b) = (1, 0)$ this yields

$$\frac{\partial P}{\partial x}\frac{\partial P}{\partial y} + \frac{\partial Q}{\partial x}\frac{\partial Q}{\partial y} = 0 \tag{1}$$

and for $(a, b) = (1, 1)$ we obtain

$$\left(\frac{\partial P}{\partial y}\right)^2 + \left(\frac{\partial Q}{\partial y}\right)^2 - \left(\frac{\partial P}{\partial x}\right)^2 - \left(\frac{\partial Q}{\partial x}\right)^2 = 0. \tag{2}$$

Now since f is regular, some of the partial derivatives, say $\frac{\partial Q}{\partial x}(z)$, is not zero (if we concentrate to a particular argument). Set

$$\lambda = \frac{\partial P}{\partial x}\left(\frac{\partial Q}{\partial x}\right)^{-1}$$

so that we have $\frac{\partial P}{\partial x} = \lambda \frac{\partial Q}{\partial x}$ and the equation (1) yields $\lambda \frac{\partial P}{\partial y} + \frac{\partial Q}{\partial y} = 0$, and substituting these two equalities into (2) we obtain that

$$(1+\lambda^2)\left(\frac{\partial P}{\partial y}\right)^2 = (1+\lambda^2)\left(\frac{\partial Q}{\partial x}\right)^2$$

and since λ is real, $1+\lambda^2 \neq 0$ and we see that

$$\left(\frac{\partial P}{\partial y}\right)^2 = \left(\frac{\partial Q}{\partial x}\right)^2.$$

Now either $\frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}$ and then we obtain from (1) that $\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y}$, and f satisfies the Cauchy-Riemann equations; since the partial derivatives are continuous, f is holomorphic. Or $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ and then (1) yields that $\frac{\partial P}{\partial x} = -\frac{\partial Q}{\partial y}$. Then by the Chain Rule, $\mathsf{conj} \circ f$ satisfies the Cauchy-Riemann equations and hence it is holomorphic. □