Analysis II

Alexander Grigorian
University of Bielefeld

Lecture Notes, October 2006 - February 2007

Contents

1 Integral calculus

In Analysis I we have learned the operation of differentiation. Let us recall that if f is a function defined on an open interval $I \subset \mathbb{R}$ then the *derivative* $f'(x)$ at any point $x \in I$ is defined by

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

The *differential* of f is the expression

$$df = f'(x)\, dx$$

where dx is the increment of the argument x. We have also learned how to evaluate the derivative of a function and to use it for investigation of the function.

Now we start discussing the inverse problem: given a function f, how to find a function F such that $F' = f$ in I. This question as well as more general questions leading to *differential equations*, occur in vast variety of problems both inside and outside Mathematics. We mainly focus on the mathematical aspects of the problem but at some point will give also examples of applications.

1.1 Indefinite integral

Definition. If $F' = f$ on I then the function F is called an *antiderivative* of f or a *primitive function* of f ($Stammfunction$) on the interval I.

As we know, not every function has the derivative. Similarly, not every function has a primitive. Later in this course, we'll prove the following statement.

Theorem. *Any continuous function on an interval $I \subset \mathbb{R}$ has a primitive on this interval.*

What about uniqueness? Can it happen than a function has two different primitives? Yes, this can happen. For example, if $F(x) = C$ - a constant function, then $F' = 0$. Hence, any constant function is a primitive of $f \equiv 0$. However, it is easy to describe the degree of non-uniqueness of a primitive function.

Theorem 1.1 *If F is a primitive of f on an interval $I \subset \mathbb{R}$ then all other primitives of f have the form $F(x) + C$, where C is any constant.*

Proof. If $F' = f$ then also $(F + C)' = F' = f$. Hence, $F + C$ is also a primitive of f. Conversely, if F and G are two primitives of f then $F' = G' = f$ whence $(G - F)' = 0$ on I. By the Constant Test (Theorem 4.10 from Analysis I), the function $G - F$ is constant on I. Denoting this constant by C, we obtain $G(x) = F(x) + C$ for all $x \in I$, which was to be proved. ■

Definition. The family of all primitive function of $f(x)$ is denoted by

$$\int f(x)\, dx.$$

This expression is called also the *indefinite integral of* f. By Theorem 1.1, $\int f(x)\,dx$ is a function up to an additive constant.

The reason for this notation and terminology will become clearer later in the course. Here we only make the following simple observations. By definition, we have

$$\left(\int f(x)\,dx\right)' = f(x),$$

which in terms of differential amounts to

$$d\int f(x)\,dx = f(x)\,dx. \tag{1.1}$$

On the other hand, by Theorem 1.1, we have the identity

$$\int F'(x)\,dx = F(x) + C, \tag{1.2}$$

which can also be written in the form

$$\int dF(x) = F(x) + C. \tag{1.3}$$

Comparing (1.1) and (1.3) we see that the operations d and $\int$ are almost inverse each to other.

The process of finding the primitive is called (indefinite) *integration*. A significant part of the integral calculus consists of the methods of integration. The simplest way of integrating is to reverse the identities obtained by differentiation. For example, since

$$\left(\frac{x^{n+1}}{n+1}\right)' = x^n, \quad n \neq -1,$$

(1.2) yields the following identity on $(-\infty, +\infty)$

$$\boxed{\int x^n dx = \frac{x^{n+1}}{n+1} + C.}$$

In particular, we have

$$\begin{aligned}
\int dx &= x + C, \\
\int x dx &= \frac{x^2}{2} + C, \\
\int \sqrt{x} dx &= \frac{x^{3/2}}{3/2} + .C, \\
\int \frac{dx}{x^2} &= -\frac{1}{x} + C.
\end{aligned}$$

Since $(\ln x)' = \frac{1}{x}$ on $(0, +\infty)$, we obtain

$$\int \frac{dx}{x} = \ln x + C \text{ on } (0, +\infty)$$

and, more generally,

$$\boxed{\int \frac{dx}{x} = \ln|x| + C} \text{ on } (0, +\infty) \text{ and } (-\infty, 0).$$

Indeed, $\ln|x|$ is an even function and, hence, its derivative is odd (see Exercise 1). Since the function $\frac{1}{x}$ is also odd, the identity $(\ln|x|)' = \frac{1}{x}$ extends from $(0, +\infty)$ to $(-\infty, 0)$.

Reversing differentiation of the exponential function, we obtain the following identities:

$$\boxed{\int \exp(x)\, dx = \exp(x) + C,}$$

$$\boxed{\int a^x dx = \frac{a^x}{\ln a} + C} \text{ if } a > 0,\ a \neq 1,$$

Using the derivatives of trigonometric and hyperbolic functions (see Exercises 57 and 59 from Analysis I and Exercise 4 from Analysis II), we obtain the following identities:

$$\boxed{\int \sin x dx = -\cos x + C}$$

$$\boxed{\int \cos x dx = \sin x + C},$$

$$\boxed{\int \frac{dx}{\cos^2 x} = \tan x + C},$$

on any interval where $\cos x$ does not vanish,

$$\boxed{\int \frac{dx}{\sin^2 x} = -\cot x + C},$$

on any interval where $\sin x$ does not vanish,

$$\boxed{\int \sinh x = \cosh x + C}$$

$$\boxed{\int \cosh x dx = \sinh x + C}$$

$$\boxed{\int \frac{1}{\cosh^2 x} = \tanh x + C}$$

$$\boxed{\int \frac{1}{\sinh^2 x} = \coth x + C}.$$

Using the derivatives of the inverse trigonometric and hyperbolic functions (see Exercises 57 and 59 from Analysis I and Exercises 2-4 from Analysis II) we obtain

$$\boxed{\int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C} \text{ on } (-1, 1),$$

$$\boxed{\int \frac{dx}{1+x^2} = \arctan x + C}$$

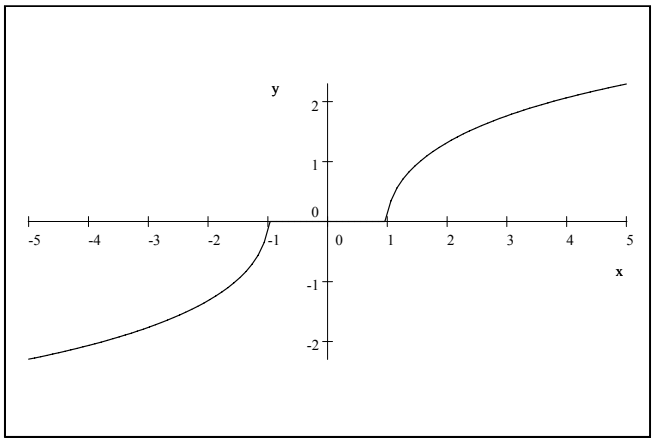
$$\boxed{\int \frac{dx}{\sqrt{x^2+1}} = \sinh^{-1} x + C = \ln\left(x + \sqrt{x^2+1}\right) + C},$$

$$\int \frac{dx}{\sqrt{x^2-1}} = \cosh^{-1} x + C = \ln\left(x + \sqrt{x^2-1}\right) + C \text{ on } (1, +\infty).$$

The latter identity extends to a more general one

$$\boxed{\int \frac{dx}{\sqrt{x^2-1}} = \ln\left|x + \sqrt{x^2-1}\right| + C} \text{ on } (1, +\infty) \text{ and } (-\infty, -1),$$

which follows from the fact that the function $\ln\left|x + \sqrt{x^2-1}\right|$ is odd whereas $\frac{1}{\sqrt{x^2-1}}$ is even. The function $\ln\left|x + \sqrt{x^2-1}\right|$, which is called the *long logarithm*, has the following graph:



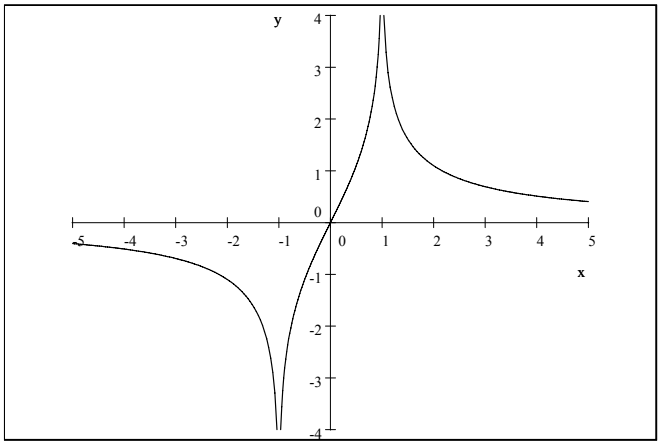
Finally, we have

$$\int \frac{dx}{1-x^2} = \tanh^{-1} x + C = \frac{1}{2} \ln \frac{1+x}{1-x} + C \text{ on } (-1, 1),$$

and this identity extends to any of the intervals $(-\infty, -1)$, $(-1, 1)$, $(1, +\infty)$ as follows:

$$\boxed{\int \frac{dx}{1-x^2} = \frac{1}{2} \ln \left|\frac{1+x}{1-x}\right| + C}.$$

Here is the graph of the *tall logarithm* $\ln\left|\frac{1+x}{1-x}\right|$:



The above boxed formulas form a simplest table of integration. In general, integration may be a difficult operation. Moreover, it is not always possible to express the integral of an elementary function in terms of elementary functions. There are extended tables of integrations containing thousands of indefinite integrals. Nowadays there are also software that are capable of evaluating indefinite integrals.

Our next purpose is to learn the basic methods of integrations. They use the table of integration and some rules of reduction of the given integral to a table integral.

1.2 Linearity of indefinite integral

Theorem 1.2 *If f and g are two functions on an interval I such that the integrals $\int f dx$ and $\int g dx$ exist on I then, for all $a, b \in \mathbb{R}$,*

$$\int (af + bg)\, dx = a \int f dx + b \int g dx. \tag{1.4}$$

Proof. Let F be a primitive of f and G be a primitive of g so that $F' = f$ and $G' = g$. Then, by the linearity of differentiation (Theorem 4.3 from Analysis I), we have

$$(aF + bG)' = aF' + bG' = af + bg$$

whence

$$\int (af + bg)\, dx = aF + bG + \text{const} = a \int f dx + b \int g dx,$$

where in the last identity the constant is absorbed by one of the integrals. ■

Example. 1. Evaluate $\int \left(x + \frac{1}{\sqrt{x}}\right)^2 dx$. We have

$$\left(x + \frac{1}{\sqrt{x}}\right)^2 = x^2 + 2x\frac{1}{\sqrt{x}} + \frac{1}{x} = x^2 + 2\sqrt{x} + \frac{1}{x},$$

whence

$$\begin{aligned} \int \left(x + \frac{1}{\sqrt{x}}\right)^2 dx &= \int x^2 dx + 2\int x^{1/2} dx + \int \frac{1}{x} dx \\ &= \frac{x^3}{3} + \frac{4x^{3/2}}{3} + \ln|x| + C. \end{aligned}$$

2. Evaluate $\int \frac{dx}{\sin^2 x \cos^2 x}$. Observing that

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

and using Theorem 1.2 and the table integrals, we obtain

$$\int \frac{dx}{\sin^2 x \cos^2 x} = \int \frac{dx}{\cos^2 x} + \int \frac{dx}{\sin^2 x} = \tan x - \cot x + C.$$

1.3 Integration by parts

If $u(x)$ and $v(x)$ are two functions then it makes sense to consider the expression $\int udv$, provided the derivative v' exists. Indeed, we have $dv = v'(x)\,dx$ so that

$$\int udv \equiv \int u(x)\,v'(x)\,dx.$$

We say that a function u is continuously differentiable on an interval I if its derivative u' exists on this interval and is a continuous function. Recall that if u is differentiable then u is continuous (Theorem 4.2 from Analysis I).

Theorem 1.3 (Integration-by-parts formula) *If u, v are two continuously differentiable functions on an interval I then*

$$\int udv = uv - \int vdu. \tag{1.5}$$

Proof. The hypothesis of the continuous differentiability of u and v is needed to ensure that the both integrals in (1.5) exist. To prove (1.5), it suffices to verify that the derivatives of the both sides of (1.5) are the same. We have

$$\left(\int udv\right)' = \left(\int uv'dx\right)' = uv'$$

and

$$\left(uv - \int vdu\right)' = (uv)' - vu'.$$

Using the product rule (Theorem 4.3(b) from Analysis I), we obtain

$$(uv)' - vu' = (uv' + u'v) - vu' = uv',$$

whence (1.5) follows. ■

Example. 1. Evaluate $\int \ln x dx$. Taking $u = \ln x$ and $v = x$, we obtain

$$\int \ln x dx = x\ln x - \int x d\ln x = x\ln x - \int x\frac{1}{x}dx = x\ln x - x + C.$$

2. Evaluate $\int x^2 e^x dx$. Note that $e^x dx = de^x$. Hence, taking $u = x^2$ and $v = e^x$, we obtain

$$\int x^2 e^x dx = \int x^2 de^x = x^2e^x - \int e^x dx^2 = x^2e^x - 2\int xe^x dx.$$

To evaluate $\int xe^x dx$, apply Theorem 1.3 again, this time with $u = x$ and $v = e^x$:

$$\int xe^x dx = \int x de^x = xe^x - \int e^x dx = xe^x - e^x + C.$$

Hence,

$$\int x^2 e^x dx = x^2 e^x - 2xe^x + 2e^x + C.$$

3. Evaluate $\int \sqrt{1+x^2}dx$. Taking $u = \sqrt{1+x^2}$ and $v = x$, we obtain

$$\begin{aligned}\int \sqrt{1+x^2}dx &= x\sqrt{1+x^2} - \int \frac{x^2 dx}{\sqrt{1+x^2}} \\ &= x\sqrt{1+x^2} - \int \frac{(1+x^2)\,dx}{\sqrt{1+x^2}} + \int \frac{dx}{\sqrt{1+x^2}} \\ &= x\sqrt{1+x^2} - \int \sqrt{1+x^2}dx + \ln\left(x+\sqrt{x^2+1}\right) + C.\end{aligned}$$

As we see, the same integral appears both in the left hand side and in the right hand side. Moving the latter to the left and dividing by 2, we obtain

$$\int \sqrt{1+x^2}dx = \frac{1}{2}x\sqrt{x^2+1} + \frac{1}{2}\ln\left(x+\sqrt{x^2+1}\right) + C.$$

1.4 Change of variable in the integral

Theorem 1.4 *Let u be a continuously differentiable function on an interval I and $u(I) \subset J$ where J is another interval. If f is a function on J, which has a primitive F on J then*

$$\int f(u(x))\,du = F(u(x)) + C. \tag{1.6}$$

Note that the composite functions $f(u(x))$ and $F(u(x))$ are defined on I. The identity (1.6) can be viewed as follows. The fact that F is a primitive of f can be written as

$$\int f(u)\,du = F(u) + C,$$

where u here is an independent variable (the argument of f and F). The identity (1.6) says that the independent variable u can be replaced here by a function $u = u(x)$. The formula (1.6) is referred to as a *change of variable* (or *substitution*) in the integral.

Proof. Note that

$$\int f(u(x))\,du = \int f(u(x))\,u'(x)\,dx.$$

Hence, all we need to prove is that

$$(F(u(x)))' = f(u(x))\,u'(x).$$

Using the chain rule (Theorem 4.4 from Analysis I) and $F' = f$, we obtain

$$F(u(x))' = F'(u(x))\,u'(x) = f(u(x))\,u'(x),$$

which was to be proved. ■

Example. 1. Evaluate $\int (ax+b)^n\,dx$ where $a \neq 0$. Note that

$$dx = \frac{1}{a} d\,(ax+b)\,.$$

Setting $u = ax+b$, we obtain

$$\int (ax+b)^n\,dx = \frac{1}{a}\int (ax+b)^n\,d\,(ax+b) = \frac{1}{a}\int u^n du.$$

Considering u as an independent variable, we obtain

$$\int u^n du = \begin{cases} \frac{u^{n+1}}{n+1}, & n \neq -1 \\ \ln|u|\,, & n = -1. \end{cases}$$

Hence,

$$\int (ax+b)^n\,dx = \begin{cases} \frac{(ax+b)^{n+1}}{a(n+1)}, & n \neq 1, \\ \frac{1}{a}\ln|ax+b|\,, & n = -1. \end{cases}$$

2. Evaluate $\int \frac{dx}{x^2-1}$. Although this is a table integral, we give here an independent derivation using the linearity of integral and change of variable. We have

$$\frac{1}{x^2-1} = \frac{1}{(x-1)\,(x+1)} = \frac{1}{2}\left(\frac{1}{x-1} - \frac{1}{x+1}\right)$$

whence

$$\begin{aligned} \int \frac{dx}{x^2-1} &= \frac{1}{2}\int \frac{dx}{x-1} - \frac{1}{2}\int \frac{dx}{x+1} \\ &= \frac{1}{2}\int \frac{d\,(x-1)}{x-1} - \frac{1}{2}\int \frac{d\,(x+1)}{x+1} \\ &= \frac{1}{2}\ln|x-1| - \frac{1}{2}\ln|x+1| + C \\ &= \frac{1}{2}\ln\left|\frac{x-1}{x+1}\right| + C. \end{aligned}$$

3. Evaluate

$$\int \frac{xdx}{1+x^2}.$$

Observe that

$$xdx = d\left(\frac{x^2}{2}\right) = \frac{1}{2} d\left(1+x^2\right).$$

Hence, the given integral can be written in the form

$$\frac{1}{2}\int \frac{d\,(1+x^2)}{1+x^2}.$$

Setting $u = 1+x^2$, we obtain

$$\int \frac{xdx}{1+x^2} = \frac{1}{2}\int \frac{du}{u}.$$

In the right hand side, we can consider the integral as if u is an independent variable. By the table integral, we have

$$\int \frac{du}{u} = \ln|u| + C,$$

whence

$$\int \frac{xdx}{1+x^2} = \frac{1}{2}\ln|u(x)| + C = \frac{1}{2}\ln\left(1+x^2\right) + C.$$

4. Evaluate

$$\int \frac{dx}{\sin x}.$$

We have, using $u = \cos x$,

$$\begin{aligned}\int \frac{dx}{\sin x} &= \int \frac{\sin x dx}{\sin^2 x} = -\int \frac{d\cos x}{\sin^2 x} = \int \frac{d\cos x}{\cos^2 x - 1} = \int \frac{du}{u^2-1} \\ &= \frac{1}{2}\ln\left|\frac{u-1}{u+1}\right| + C = \frac{1}{2}\ln\frac{1-\cos x}{1+\cos x} + C.\end{aligned}$$

In fact, the following trigonometric identity takes place:

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

so that

$$\int \frac{dx}{\sin x} = \ln\left|\tan\frac{x}{2}\right| + C.$$

5. Evaluate

$$\int \arcsin x dx.$$

Using the methods described above, we obtain

$$\begin{aligned}\int \arcsin x dx &= x\arcsin x - \int x d\arcsin x \quad \text{(integration by parts)} \\ &= x\arcsin x - \int \frac{x}{\sqrt{1-x^2}}dx \\ &= x\arcsin x + \frac{1}{2}\int \frac{d\left(1-x^2\right)}{\sqrt{1-x^2}} \quad (\text{change } u = 1-x^2) \\ &= x\arcsin x + \frac{1}{2}\int u^{-1/2}du \qquad \text{(table integral)} \\ &= x\arcsin x + u^{1/2} + C \\ &= x\arcsin x + \sqrt{1-x^2} + C.\end{aligned}$$

1.5 Integration of rational functions

A rational function is a function of the form

$$f(x) = \frac{P(x)}{Q(x)}$$

where P and Q are polynomials of x. Integral of any such function can be found using the following statement, which we state without proof.

Theorem. (a) *Any polynomial $Q(x)$ with real coefficients can be uniquely represented in the form*

$$Q(x) = A(x-r_1)^{k_1} ... (x-r_l)^{k_l} \left(x^2+p_1x+q_1\right)^{m_1} ... \left(x^2+p_nx+q_n\right)^{m_n}, \tag{1.7}$$

where the numbers A, r_k, p_k, q_k are real, k_j,m_j are natural, n,l are non-negative integers, and the polynomials $x^2+p_kx+q_k$ have no real roots; also, the numbers r_k are distinct and the couples (p_k, q_k) are distinct.

(b) *Any rational function $\frac{P(x)}{Q(x)}$ can be uniquely represented in the form*

$$\frac{P(x)}{Q(x)} = P_0(x) + \sum_{j=1}^{l}\sum_{k=1}^{k_j} \frac{a_{kj}}{(x-r_j)^k} + \sum_{j=1}^{n}\sum_{m=1}^{m_j} \frac{b_{mj}x+c_{mj}}{(x^2+p_jx+q_j)^m}, \tag{1.8}$$

where $Q(x)$ is as in (1.7), $P_0(x)$ is a polynomial, and a_{kj}, b_{mj}, c_{mj} are real.

The numbers r_k are obviously the roots of $Q(x)$, that is, $Q(r_k)=0$. If complex-valued roots were allowed then, due to the Fundamental Theorem of Algebra, one could always represent $Q(x)$ as a product of the linear terms only:

$$Q(x) = A(x-r_1)^{k_1} ... (x-r_l)^{k_l}. \tag{1.9}$$

Note that the complex roots of a real polynomial come in conjugate couples: if $r = \alpha+i\beta \in \mathbb{C}$ is a root of Q then $\overline{r} = \alpha - i\beta$ is also a root because $Q(\overline{r}) = \overline{Q(r)}$. Moreover, the roots r and $\overline{r}$ have the same multiplicity. Obviously, we have

$$\begin{aligned}(x-r)(x-\overline{r}) &= (x-\alpha-i\beta)(x-\alpha+i\beta)\\ &= (x-\alpha)^2+\beta^2\\ &= x^2+px+q,\end{aligned}$$

where $p=-2\alpha$ and $q=\alpha^2+\beta^2$; note that x^2+px+q has no real roots. Hence, replacing in (1.9) the terms with conjugate roots by the corresponding quadratic polynomials, we obtain (1.7).

Formula (1.8) means that each rational function can be represented as a sum of a polynomial and some number of *elementary rational functions* of the form

$$\frac{a}{(x-r)^k}, \quad \frac{bx+c}{(x^2+px+q)^k},$$

For practical applications, one does not really need the proof of this Theorem, since the splitting (1.8) can be found in each case using some algebraic manipulations rather than a general theory.

Recall that a rational function is a ratio of two polynomials. As was explained in the previous lecture, each rational function can be split into a sum of a polynomial and elementary rational functions of the form

$$\frac{a}{(x-r)^k}, \quad \frac{bx+c}{(x^2+px+q)^k},$$

It is easy to integrate a polynomial because it is a sum of the terms ax^k. Let us explain how to integrate the elementary rational function. A function of the type $\frac{a}{(x-r)^k}$ is integrated as follows:

$$\begin{aligned}\int \frac{a}{(x-r)^k}dx &= a\int \frac{d(x-r)}{(x-r)^k} = (\text{change } u = x-r)\\ &= a\int u^{-k}du = a\begin{cases}\frac{u^{1-k}}{1-k}, & k\neq 1,\\ \ln|u|, & k=1,\end{cases}\\ &= a\begin{cases}\frac{(x-r)^{1-k}}{1-k}, & k\neq 1,\\ \ln|x-r|, & k=1.\end{cases}\end{aligned}$$

Integrating the elementary rational function of the type $\frac{bx+c}{(x^2+px+q)^k}$ is somewhat more involved. Representing x^2+px+q in the form

$$x^2+px+q = (x+p/2)^2 + \left(q-p^2/4\right) = u^2+s^2,$$

where $u = x+p/2$ and $s = \sqrt{q-p^2/4}$, we obtain

$$\begin{aligned}\int \frac{bx+c}{(x^2+px+q)^k}dx &= \int \frac{b'u+c'}{(u^2+s^2)^k}du\\ &= b'\int \frac{udu}{(u^2+s^2)^k} + c'\int \frac{du}{(u^2+s^2)^k}.\end{aligned}$$

Hence, we have to handle the following two integrals:

$$\int \frac{udu}{(u^2+s^2)^k} \text{ and } \int \frac{du}{(u^2+s^2)^k}.$$

The first of them is taken easily as follows:

$$\begin{aligned}\int \frac{udu}{(u^2+s^2)^k} &= \frac{1}{2}\int \frac{d(u^2+s^2)}{(u^2+s^2)^k} = (\text{change } v = u^2+s^2)\\ &= \frac{1}{2}\int \frac{dv}{v^k} = \frac{1}{2}\begin{cases}\frac{v^{1-k}}{1-k}, & k\neq 1,\\ \ln|v|, & k=1,\end{cases}\\ &= \frac{1}{2}\begin{cases}\frac{(u^2+s^2)^{1-k}}{1-k}, & k\neq 1,\\ \ln(u^2+s^2), & k=1.\end{cases}\end{aligned}$$

The second integral will be evaluated inductively in k. Set

$$F_k(u) = \int \frac{du}{(u^2+s^2)^k}$$

and notice that

$$F_1(u) = \int \frac{du}{u^2+s^2} = \int \frac{du}{s^2\left((u/s)^2+1\right)} = \frac{1}{s}\int \frac{d(u/s)}{(u/s)^2+1} = \frac{1}{s}\arctan\frac{u}{s} + C.$$

Integrating by parts in F_k, we obtain

$$\begin{aligned} F_k(u) &= \frac{u}{(u^2+s^2)^k} - \int u d\frac{1}{(u^2+s^2)^k} \\ &= \frac{u}{(u^2+s^2)^k} + 2k\int \frac{u^2 du}{(u^2+s^2)^{k+1}} \\ &= \frac{u}{(u^2+s^2)^k} + 2k\int \frac{u^2+s^2}{(u^2+s^2)^{k+1}} du - 2ks^2\int \frac{du}{(u^2+s^2)^{k+1}} \\ &= \frac{u}{(u^2+s^2)^k} + 2kF_k - 2ks^2F_{k+1}. \end{aligned}$$

It follows that

$$F_{k+1} = \frac{1}{2ks^2}\left(\frac{u}{(u^2+s^2)^k} + (2k-1)F_k\right),$$

which allows to evaluate F_k by induction in k, starting with F_1.

Example. 1. Evaluate

$$\int \frac{dx}{x^3-x}.$$

The denominator is factorized as follows:

$$x^3 - x = x(x-1)(x+1).$$

Hence, function $\frac{1}{x^3-x}$ splits into the sum of elementary rational functions as follows

$$\frac{1}{x^3-x} = \frac{a_1}{x} + \frac{a_2}{x-1} + \frac{a_3}{x+1}, \tag{1.10}$$

where the constants a_1, a_2, a_3 are to be determined. To find a_1, multiply the equation by x:

$$\frac{1}{x^2-1} = a_1 + x\left(\frac{a_2}{x-1} + \frac{a_3}{x+1}\right)$$

and notice that this identity is true for all $x \neq 0, 1, -1$ but by continuity it extends to $x = 0$. Setting $x = 0$, we obtain

$$a_1 = -1.$$

Similarly, multiplying (1.10) by $x-1$ and setting $x = 1$ in the resulting identity

$$\frac{1}{x(x+1)} = a_2 + (x-1)\left(\frac{a_1}{x} + \frac{a_3}{x+1}\right)$$

we obtain

$$a_2 = \frac{1}{2}.$$

Finally, multiplying (1.10) by $x+1$ and setting $x=-1$ in the resulting identity

$$\frac{1}{x\left(x-1\right)}=a_3+\left(x+1\right)\left(\frac{a_1}{x}+\frac{a_2}{x-1}\right),$$

we obtain

$$a_3=\frac{1}{2}.$$

Therefore,

$$\frac{1}{x^3-x}=-\frac{1}{x}+\frac{1}{2}\frac{1}{x-1}+\frac{1}{2}\frac{1}{x+1}$$

whence

$$\begin{aligned}\int\frac{dx}{x^3-x} &= -\ln|x|+\frac{1}{2}\ln|x-1|+\frac{1}{2}\ln|x+1|+C\\ &= \frac{1}{2}\ln\left|\frac{x^2-1}{x^2}\right|+C.\end{aligned}$$

2. Evaluate

$$\int\frac{dx}{\left(x^2+1\right)\left(x-1\right)^2}.$$

Let us split the function $f\left(x\right)=\frac{1}{(x^2+1)(x-1)^2}$ into a sum of elementary rational functions in the form

$$\frac{1}{\left(x^2+1\right)\left(x-1\right)^2}=\frac{a_1}{\left(x-1\right)^2}+\frac{a_2}{x-1}+\frac{bx+c}{x^2+1},\tag{1.11}$$

where the coefficients a_i,b,c are to be found. Multiplying the both sides by $\left(x-1\right)^2$, we obtain

$$\frac{1}{x^2+1}=a_1+a_2\left(x-1\right)+\left(x-1\right)^2R\left(x\right),$$

where $R\left(x\right)=\frac{bx+c}{x^2+1}$. Substituting $x=1$, we obtain

$$a_1=\frac{1}{2}.$$

Subtracting in (1.11) the term $\frac{1}{2}\frac{1}{(x-1)^2}$, we obtain

$$\begin{aligned}\left(\frac{1}{\left(x^2+1\right)}-\frac{1}{2}\right)\frac{1}{\left(x-1\right)^2} &= \frac{a_2}{x-1}+R\left(x\right)\\ -\frac{1}{2}\frac{x^2-1}{\left(x^2+1\right)\left(x-1\right)^2} &= \frac{a_2}{x-1}+R\left(x\right)\\ -\frac{1}{2}\frac{x+1}{\left(x^2+1\right)\left(x-1\right)} &= \frac{a_2}{x-1}+R\left(x\right).\end{aligned}\tag{1.12}$$

Multiplying by $x-1$, we obtain the identity

$$-\frac{1}{2}\frac{x+1}{x^2+1}=a_2+\left(x-1\right)R\left(x\right),$$

and setting here $x=1$, we obtain

$$a_2 = -\frac{1}{2}.$$

It follows from (1.12) that

$$-\frac{1}{2}\frac{x+1}{(x^2+1)(x-1)} + \frac{1}{2}\frac{1}{x-1} = R(x)$$

whence

$$R(x) = \frac{1}{2}\frac{x^2+1-(x+1)}{(x^2+1)(x-1)} = \frac{1}{2}\frac{x(x-1)}{(x^2+1)(x-1)} = \frac{1}{2}\frac{x}{x^2+1}$$

so that $R(x)$ indeed has the form $\frac{bx+c}{x^2+1}$. Hence, we have

$$f(x) = \frac{1}{2}\frac{1}{(x-1)^2} - \frac{1}{2}\frac{1}{x-1} + \frac{1}{2}\frac{x}{x^2+1}.$$

Now let us integrate each term separately:

$$\begin{aligned}\int \frac{dx}{(x-1)^2} &= \int \frac{d(x-1)}{(x-1)^2} = (\text{change } u = x-1)\\ &= \int u^{-2}du = \frac{u^{-1}}{-1} + C\\ &= \frac{1}{1-x} + C,\end{aligned}$$

$$\int \frac{1}{x-1}dx = \int \frac{d(x-1)}{x-1} = \ln|x-1| + C,$$

$$\int \frac{xdx}{x^2+1} = \frac{1}{2}\int \frac{d(x^2+1)}{x^2+1} = \frac{1}{2}\ln\left(x^2+1\right) + C.$$

Combining all the lines together, we obtain

$$\int f(x)\,dx = \frac{1}{2}\frac{1}{1-x} - \frac{1}{2}\ln|x-1| + \frac{1}{4}\ln\left(x^2+1\right) + C.$$

3. Evaluate

$$\int \frac{dx}{(x^2+x+1)^2}.$$

The polynomial x^2+x+1 has no real root. Therefore, function $\frac{1}{(x^2+x+1)^2}$ is already an elementary rational function. Using

$$x^2+x+1 = \left(x+\frac{1}{2}\right)^2 + \frac{3}{4}$$

and making change $u = x+\frac{1}{2}$ we obtain

$$\int \frac{dx}{(x^2+x+1)^2} = \int \frac{du}{\left(u^2+\frac{3}{4}\right)^2}.$$

Recall that

$$\int \frac{du}{u^2+\frac{3}{4}} = \frac{2}{\sqrt{3}} \arctan \frac{2u}{\sqrt{3}} + C.$$

On the other hand, integrating the above integral by parts, we obtain

$$\begin{aligned}
\int \frac{du}{u^2+\frac{3}{4}} &= \frac{u}{u^2+\frac{3}{4}} - \int u d\frac{1}{u^2+\frac{3}{4}} \\
&= \frac{u}{u^2+\frac{3}{4}} + \int u \frac{2u}{\left(u^2+\frac{3}{4}\right)^2} du \\
&= \frac{u}{u^2+\frac{3}{4}} + 2\int \frac{u^2+\frac{3}{4}-\frac{3}{4}}{\left(u^2+\frac{3}{4}\right)^2} du \\
&= \frac{u}{u^2+\frac{3}{4}} + 2\int \frac{1}{u^2+\frac{3}{4}} du - \frac{3}{2}\int \frac{1}{\left(u^2+\frac{3}{4}\right)^2} du.
\end{aligned}$$

It follows that

$$\frac{3}{2}\int \frac{1}{\left(u^2+\frac{3}{4}\right)^2} du = \frac{u}{u^2+\frac{3}{4}} + \int \frac{1}{u^2+\frac{3}{4}} du = \frac{u}{u^2+\frac{3}{4}} + \frac{2}{\sqrt{3}} \arctan \frac{2u}{\sqrt{3}} + C,$$

whence

$$\int \frac{1}{\left(u^2+\frac{3}{4}\right)^2} du = \frac{2u}{3u^2+\frac{9}{4}} + \frac{4}{3\sqrt{3}} \arctan \frac{2u}{\sqrt{3}} + C,$$

and

$$\int \frac{dx}{\left(x^2+x+1\right)^2} = \frac{2x+1}{3\left(x^2+x+1\right)} + \frac{4}{3\sqrt{3}} \arctan \frac{2x+1}{\sqrt{3}} + C.$$

In conclusion, let us mention a method for obtaining the coefficients a_{kj} of the expansion of $\frac{P(x)}{Q(x)}$ in (1.8). Fix one of the roots, say r, of $Q(x)$, and let k be its multiplicity. Then rewrite (1.8) in the form

$$\frac{P(x)}{Q(x)} = \frac{a_1}{x-r} + \frac{a_2}{(x-r)^2} + ... + \frac{a_k}{(x-r)^k} + R(x), \tag{1.13}$$

where $R(x)$ is the remainder term containing other roots of $Q(x)$ as well as the quadratic terms. Note that $R(x)$ has a finite value at $x = r$. Multiplying (1.13) by $(x-r)^k$, we obtain

$$\begin{aligned}
\frac{P(x)}{Q(x)}(x-r)^k &= a_k + a_{k-1}(x-r) + ... + a_1(x-r)^{k-1} + (x-r)^k R(x) \\
&= a_k + a_{k-1}(x-r) + ... + a_1(x-r)^{k-1} + o(x-r)^{k-1} \text{ as } x \to r.
\end{aligned}$$

Hence, by Theorem 4.14 of Analysis I, we conclude that the coefficients $a_k, ..., a_1$ are the Taylor coefficients of the function $f(x) = \frac{P(x)}{Q(x)}(x-r)^k$ at $x = r$, whence it follows that

$$a_{k-l} = \frac{f^{(l)}(r)}{l!}, \quad l = 0, 1, ..., k-1.$$

1.6 Separable differential equations

An *ordinary differential equation* (ODE) is an equation containing an unknown function, say $y = y(x)$ and its derivatives. A general ODE has the form

$$F\left(x, y, y', ..., y^{(n)}\right) = 0$$

where F is a given function of $n+2$ arguments, x is an independent variable, $y = y(x)$ is an unknown function to be found. The *order* of an ODE is the maximal order of the derivative of y that is contained in this equation, which in this case is n.

Consider the following ODE of the 1st order:

$$y' = f(x, y),$$

where f is a given function of two arguments and $y = y(x)$ is unknown. As we already know, the simplest differential equation

$$y' = f(x)$$

is solved by integration:

$$y = \int f(x)\, dx.$$

Consider another example of the 1st order ODE:

$$y' = ay,$$

where a is a constant. One solution is easy to guess: $y = e^{ax}$. Clearly, a function $y = Ce^{ax}$ is also a solution, for any constant C. It turns out that this formula gives all solutions.

Claim. *Any solution to the equation $y' = ay$ in any open interval is given by the formula $y(x) = Ce^{ax}$ where C is a constant.*

Proof. Let $y(x)$ be a solution on an interval I. Assume first that that $y(x)$ does not vanish in a certain open interval $I_0 \subset I$. Then we can divide by y and obtain in I_0 the equation

$$\frac{y'}{y} = a.$$

Observing that $\frac{y'}{y} = (\ln|y|)'$, we rewrite the equation in the form

$$(\ln|y|)' = a,$$

which implies

$$\ln|y| = \int a dx = ax + C,$$

$$|y| = e^C e^{ax}.$$

Therefore, $y(x) = e^C e^{ax}$ or $y(x) = -e^C e^{ax}$. In fact, $y(x)$ must have the same sign on the entire interval I_0 because otherwise, by the intermediate value theorem (Theorem 4.3

from Analysis I), function y would have take also the value 0. Renaming e^C or $-e^C$ by C, we obtain that, for all $x \in I_0$,

$$y(x) = Ce^{ax}. \tag{1.14}$$

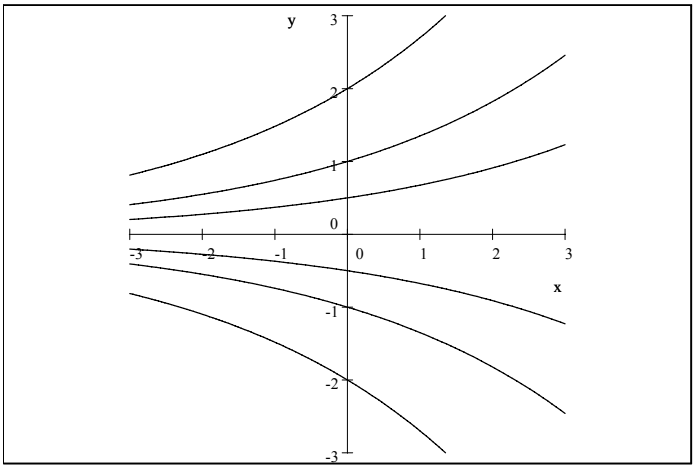
Consider now an arbitrary solution $y(x)$ in I. If $y(x) \equiv 0$ on I then we can write $y(x) = Ce^{ax}$ with $C = 0$. Let us show that the following two cases exhaust all possibilities: either y is identical zero on I or y does not vanish on I. Then the Claim will be proved since in the both cases we have proved (1.14). Assume from the contrary that $y(x_0) \neq 0$ for some $x_0 \in I$ but $y(x) = 0$ for some $x \in I$. Without loss of generality, assume that $x > x_0$ and set

$$x_1 = \inf \{x > x_0 : y(x) = 0\}.$$

By the continuity of y, we have also $y(x_1) = 0$. On the other hand, in the interval $I_0 = (x_0, x_1)$ function y does not vanish, which implies by the first part of the proof that $y = Ce^{ax}$ in this interval, with $C \neq 0$. Again by continuity, the same formula extends to $x = x_1$ whence we conclude that $y(x_1) \neq 0$. This contradiction finishes the proof. ■

In applications, a differential equation frequently comes with additional requirement that $y(x_0) = y_0$ for some given x_0 and y_0, which is called the *initial condition*. For example, in the family of solutions $y(x) = Ce^{ax}$ (where C is any real number), for all x_0,$y_0 \in \mathbb{R}$ there is exactly one solution such that $y(x_0) = y_0$. Indeed, substituting these values, we obtain $y_0 = Ce^{ax_0}$ which holds for $C = y_0 e^{-ax_0}$. Hence, we obtain the unique function $y(x) = y_0 e^{a(x-x_0)}$, which satisfies the equation $y' = ay$ and the initial condition $y(x_0) = y_0$. This solution is called a *particular solution* as opposed to the *general solution* $y = Ce^{ax}$.

Consider the graphs of all solutions $y(x)$ to the equation $y' = ay$. It follows from the above argument that, for any point (x_0, y_0) on the plane, there exists a unique graph of the solution that goes through this point. The graphs of the solutions of an ODE are called the *integral curves* of this equation. Below are shown some integral curves of the equation $y = ay$ with $a = 0.3$:



Similar argument can be used to solve more general equations having the form

$$y' = f(x)\, g(y).$$

Any equation of this form is called a *separable* ODE, because the variables x and y are separated on the right hand side. A separable ODE can be solved as follows. If $g(y)$ vanishes at some point y_0 then $y = y_0$ is a constant solution. In the domain where $g(y) \neq 0$, we can write

$$\frac{y'}{g(y)} = f(x)$$

or, multiplying by dx,

$$\frac{dy}{g(y)} = f(x)\,dx.$$

Integrating the both sides and using the change of variable in the left hand side (namely, considering y in integration as an independent variable), we obtain

$$\int \frac{dy}{g(y)} = \int f(x)\,dx.$$

After evaluating these integrals, we obtain an explicit relation between y and x. Resolving it with respect to y, we obtain $y(x)$. This method is called the method of *separation of variables.*

Example. 1. Solve $y' = -\frac{x}{y}$. Separating the variables, we obtain the equation

$$yy' = -x$$

whence

$$\int ydy = -\int xdx.$$

After integration, we obtain

$$y^2 = -x^2 + C,$$

which can also be written in the form $x^2 + y^2 = C$. The integral curves of this equation are semi-circles $y = \pm\sqrt{C - x^2}$.

2. Consider the following physical problem. A heated body is cooling down in a media of constant temperature T (say, in air or in water). Let us find out the temperature $u(t)$ of the body at time t using the Fourier law of heat conductance: the rate of decrease of $u(t)$ is proportional to the difference $u(t) - T$, that is,

$$u'(t) = -k\,(u(t) - T),$$

where $k > 0$ is the coefficient of thermoconductance of the body. The above equation is a separable ODE. Its obvious solution is $u(t) \equiv T$, but we are interested only in solutions with $u(t) > T$. Assuming that, we can divide by $u - T$ and obtain

$$\frac{u'}{u - T} = -k$$

whence

$$\int \frac{du}{u - T} = -k \int dt,$$
$$\ln(u - T) = -kt + C,$$

$$u = T + Ce^{-kt}.$$

It is clear from this formula that $u(t) \to T$ as $t \to +\infty$.

The constant C can be found from the initial condition $u(0) = u_0$, which gives $u_0 = T + C$, whence $C = u_0 - T$ and, hence,

$$u(t) = T + (u_0 - T)\, e^{-kt}.$$

In fact, the coefficient k can also be found if we have one more measurement of the temperature $u(t)$ at some time $t > 0$.

3. Consider the following mechanical problem. A body falls down along a straight line in a viscous media such as a gas or a liquid. Denoting by $z(t)$ its coordinate at time t on the vertical axis directed downwards, let us find the function $z(t)$ under reasonable physical assumptions. The are two forces acting on the body: the gravity mg directed downwards (where m is the mass of the body and g is the acceleration of the gravity) and the friction force kv^2 directed upwards, where $v = v(t)$ is the velocity of the body and k is the viscosity coefficient. Denoting by $a(t)$ the acceleration of the body at time t, we obtain by the second Newton's law

$$ma = mg - kv^2.$$

For simplification, take $g = 1$ and $m = k$ (which can always be achieved by appropriately changing the units of length and time) and rewrite the equation in the form

$$v' = 1 - v^2, \tag{1.15}$$

where we also have used that $a = v'$. This is a separable ODE, which can be solved by separation of variables. Note that by the physical meaning function $v(t)$ must be non-negative.

First, notice that (1.15) has a constant solution $v \equiv 1$. If $v(t)$ does not take value 1 then we separate variables as follows:

$$\frac{v'}{1 - v^2} = 1,$$

$$\int \frac{dv}{1 - v^2} = \int dt,$$

$$\frac{1}{2} \ln \left| \frac{1 + v}{1 - v} \right| = t + C,$$

$$\frac{1 + v}{1 - v} = \pm e^{2t + 2C}.$$

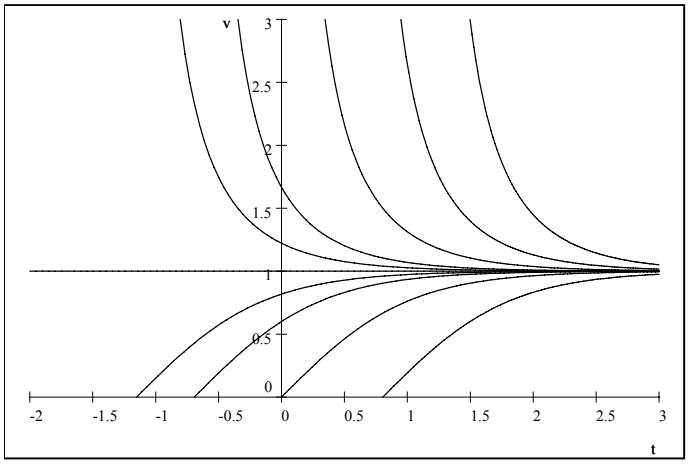
Renaming e^{2C} or e^{-2C} by C, we obtain

$$\frac{1 + v}{1 - v} = Ce^{2t},$$

$$v = \frac{Ce^{2t} - 1}{Ce^{2t} + 1}. \tag{1.16}$$

Note that C is (1.16) is any real number except for 0. For example, we can see from (1.16) that $v(t) \to 1$ as $t \to +\infty$.

The integral curves $v(t)$ of the equation in question look as follows:



As we see, those integral curves that are below $v = 1$, have the domain of the form $(t_0, +\infty)$ where $v(t_0) = 0$ (this occurs when $C > 0$). This corresponds to the physical situation that the body starts falling at time t_0 with the initial velocity 0 and accelerates to the terminal velocity 1 (when $t \to +\infty$). The curves that are above $v = 1$ (which occur for $C < 0$) correspond to the case when the initial velocity of the body is higher than 1 and it slows down (decelerates) so that its terminal velocity is again 1.

Next, notice that $z'(t) = v(t)$ whence

$$z(t) = \int v(t)\, dt.$$

If $v(t) \equiv 1$ then $z(t) = t + C_1$. If $v(t)$ is given by (1.16) then

$$\begin{aligned} z(t) &= \int \frac{Ce^{2t} - 1}{Ce^{2t} + 1} dt = \int \frac{-Ce^{2t} - 1 + 2Ce^{2t}}{Ce^{2t} + 1} dt \\ &= -\int dt + \int \frac{2Ce^{2t}}{Ce^{2t} + 1} dt \\ &= -t + \int \frac{d\left(Ce^{2t} + 1\right)}{Ce^{2t} + 1} \\ &= -t + \ln\left|Ce^{2t} + 1\right| + C_1. \end{aligned} \tag{1.17}$$

The actual values of the constants C, C_1 can be found using the initial data $v(t_0)$ and $z(t_0)$ at the initial time t_0. For example, let us impose the initial conditions

$$z(0) = v(0) = 0. \tag{1.18}$$

Then setting $t = 0$ in (1.16), we obtain

$$0 = \frac{C - 1}{C + 1},$$

whence $C = 1$. Setting $t = 0$ in (1.17), we obtain $0 = \ln 2 + C_1$, whence $C_1 = -\ln 2$. Hence, the solution satisfying the initial conditions (1.18) is

$$z(t) = -t + \ln \frac{e^{2t} + 1}{2} = \ln \frac{e^t + e^{-t}}{2} = \ln \cosh t.$$

2 Riemann integral

2.1 Definition of the Riemann integral

Let $f(x)$ be a function defined on a closed interval $[a,b]$ where $a<b$ are real numbers. Our purpose is to define the notion of the definite integral

$$\int_a^b f(x)\,dx$$

which is a number that can be interpreted as the area under the graph of function f. The procedure for that is due to Riemann and, hence, this notion is called also the Riemann integral.

A *partition* (*Unterteilung*) of the interval $[a,b]$ is any finite strictly increasing sequence $\{x_k\}_{k=0}^n$ such that $x_0=a$ and $x_n=b$, that is

$$a=x_0<x_1<x_2...<x_{n-1}<x_n=b.$$

Normally we denote a partition by p, that is, p is just a sequence $\{x_k\}_{k=0}^n$ as above. For any partition $p=\{x_k\}_{k=0}^n$ define the mesh of partition (*Feinheit*) by

$$m(p)=\max\{\Delta x_k\}_{k=1}^n$$

where

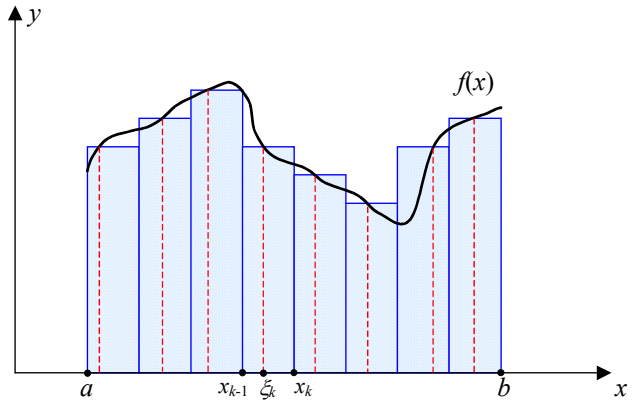
$$\Delta x_k=x_k-x_{k-1}.$$

A *tagged partition* is a partition $\{x_k\}_{k=0}^n$ together with another sequence $\{\xi_k\}_{k=1}^n$ such that $\xi_k\in[x_{k-1},x_k]$. The points ξ_k are called tags (*Stützstellen*). In other words, on each of the intervals $[x_{k-1},x_k]$ we mark one point ξ_k. We denote a tagged partition by (p,ξ) where p is a partition and ξ is the sequence $\{\xi_k\}_{k=1}^n$ of tags.

With any tagged partition (p,ξ) we associate the *Riemann sum* of f defined by

$$S(f,p,\xi)=\sum_{k=1}^n f(\xi_k)\,\Delta x_k.$$

Geometrically, $S(f,p,\xi)$ is equal to the sum of the areas of rectangles with the base $[x_{k-1},x_k]$ and the height $f(\xi_k)$, which approximates the area under the graph of $f(x)$ as on the picture below.



Now we consider the limit of the integral sums when the mesh of the partition tends to 0. Namely, we write that

$$\lim_{m(p)\to 0} S(f,p,\xi) = A$$

where $A \in \mathbb{R}$, if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all tagged partitions (p,ξ) with $m(p) < \delta$ we have

$$|S(f,p,\xi) - A| < \varepsilon.$$

Definition. A function f on $[a,b]$ is called *Riemann integrable* if the limit

$$\lim_{m(p)\to 0} S(f,p,\xi)$$

exists. The value of the limit is called the Riemann (definite) integral of f and is denoted by

$$\int_a^b f(x)\,dx.$$

In other words,

$$\int_a^b f(x)\,dx = \lim_{m(p)\to 0} \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k. \tag{2.1}$$

The notation $\int_a^b f(x)\,dx$ was invented by Leibniz and was chosen to be reminiscent of $\sum_{k=1}^{n} f(\xi_k)\,\Delta x_k$.

The following questions arise in relation to this definition:

1. What functions are integrable and if a function is integrable then how to find the Riemann integral?

2. What the Riemann integral is useful for?

3. What is the relation to the indefinite integral $\int f(x)\,dx$?

These questions will be answered in due course. So far we give two examples.

Example. 1. Let $f(x) \equiv c$ be a constant function. Then f is Riemann integrable because for any tagged partition (p,ξ)

$$S(f,p,\xi) = \sum_{k=1}^{n} f(\xi_k)\,\Delta x_k = c\sum_{k=1}^{n} \Delta x_k = c(b-a)$$

and, hence, the limit in (2.1) exists and is equal to $c(b-a)$. Hence, we can write

$$\int_a^b c\,dx = c(b-a).$$

2. Let f be the function

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & x \notin \mathbb{Q} \end{cases}$$

Then f is not Riemann integrable on any interval. Indeed, whatever is the partition $p = \{x_k\}_{k=0}^n$, we can choose tags ξ_k to be rational so that $f(\xi_k) = 0$ and, hence

$$S(f, p, \xi) = 0.$$

On the other hand, we can choose the tags to be irrational so that $f(\xi_k) = 1$ and, hence,

$$S(f, p, \xi) = b - a.$$

Hence, the limit $\lim_{m(p)\to 0} S(f, p, \xi)$ does not exists.

2.2 Criteria of integrability

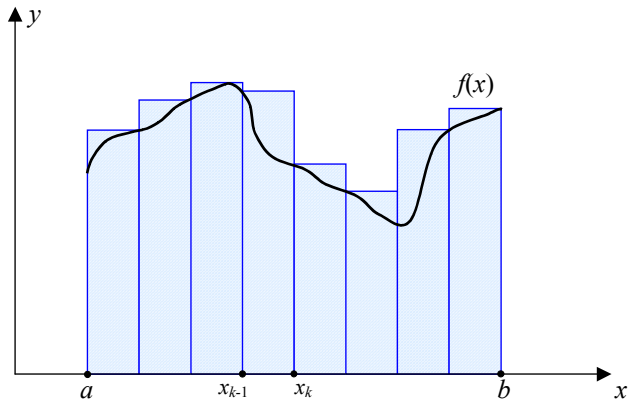
Define the notion of *Darboux* sums of f as follows: for any partition $\{x_k\}_{k=0}^n$, the *upper Darboux sum* is defined by

$$S^*(f, p) = \sum_{k=1}^{n} \sup_{[x_{k-1}, x_k]} f(x) \Delta x_k$$

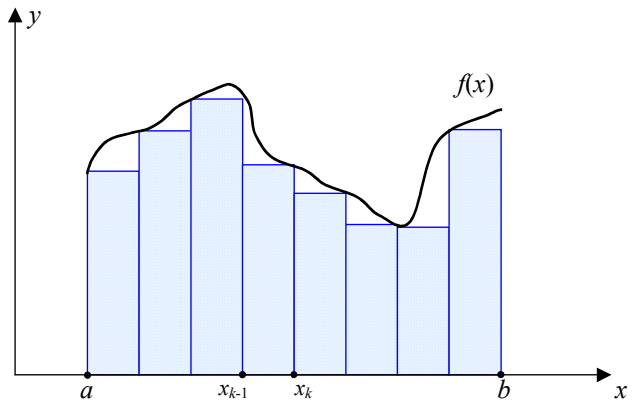
and the *lower Darboux sum* by

$$S_*(f, p) = \sum_{k=1}^{n} \inf_{[x_{k-1}, x_k]} f(x) \Delta x_k.$$

Note that the Darboux sums do not depend on the choice of tags. Geometrically, $S^*(f, p)$ is the sum of the areas of the rectangles that cover the area under the graph of f:



and $S_*(f, p)$ is the sum of the areas of the rectangles that are contained under the graph of f:



Since for any $\xi_k \in [x_{k-1}, x_k]$,

$$\inf_{[x_{k-1},x_k]} f(x) \leq f(\xi_k) \leq \sup_{[x_{k-1},x_k]} f(x),$$

we obtain from the comparison of the Riemann and Darboux sums that, for any choice of tags ξ,

$$S_*(f,p) \leq S(f,p,\xi) \leq S^*(f,p). \tag{2.2}$$

Definition. A function f is called *Darboux integrable* if,

$$\lim_{m(p)\to 0} (S^*(f,p) - S_*(f,p)) = 0.$$

In other words, for any $\varepsilon > 0$ there exists $\delta > 0$ such that, for any partition p with $m(p) < \delta$,

$$S^*(f,p) - S_*(f,p) < \varepsilon.$$

Theorem 2.1 (Darboux criterion of integrability) *Function f is Riemann integrable if and only it is Darboux integrable.*

Proof. Assume that f is Riemann integrable and let

$$A = \int_a^b f(x)\,dx.$$

Then, for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any tagged partition (p,ξ) with $m(p) < \delta$

$$|S(f,p,\xi) - A| < \varepsilon.$$

Note that this is true for any choice of the tags ξ. The tags $\xi_k \in [x_{k-1}, x_k]$ can be chosen so that $f(\xi_k)$ is arbitrarily close to $\sup_{[x_{k-1},x_k]} f(x)$. Then $S(f,p,\xi)$ can be made arbitrarily close to $S^*(f,p)$, which implies that also

$$|S^*(f,p) - A| < \varepsilon.$$

Similarly, we obtain

$$|S_*(f,p) - A| < \varepsilon,$$

whence

$$|S^*(f,p) - S_*(f,p)| < 2\varepsilon.$$

Renaming 2ε to ε, we obtain that f is Darboux integrable.

To prove the opposite implication, we need some properties of Darboux sums. We say that a partition p' is an extension (or refinement) of partition p if, as a set of points, p is contained in p', that is, $p \subset p'$. One can also say that p' is obtained from p by adding more points to the partition p.

Claim 1. *If p' is an extension of p then*

$$S^*(f,p') \leq S^*(f,p)$$

and

$$S_*(f,p') \geq S_*(f,p').$$

In other words, when refining the partition, the upper Darboux sum decreases and the lower Darboux sum increases.

Let $p = \{x_k\}_{k=0}^n$ and $p' = \{x'_k\}_{k=0}^{n'}$. Since $p \subset p'$, any interval $[x_{k-1}, x_k]$ of partition p coincides with some interval $[x'_l, x'_m]$ so that

$$x_{k-1} = x'_l < x'_{l+1} < ... < x'_m = x_k.$$

Therefore,

$$\begin{aligned} \sup_{[x_{k-1},x_k]} f(x)(x_k - x_{k-1}) &= \sum_{i=l+1}^{m} \sup_{[x_{k-1},x_k]} f(x)\left(x'_i - x'_{i-1}\right) \\ &\geq \sum_{i=l+1}^{m} \sup_{\left[x'_{i-1},x'_i\right]} f(x)\left(x'_i - x'_{i-1}\right). \end{aligned}$$

Adding up such inequalities for all k, we obtain $S^*(f,p) \geq S^*(f,p')$. For the lower sums the proof is similar.

Claim 2. *For any two partitions p' and p'',*

$$S_*(f,p') \leq S^*(f,p''). \tag{2.3}$$

In other word, any lower Darboux sum is at most any upper Darboux sum.

Let p be the partition that is obtained by merging p' and p'', that is, as a set of points, $p = p' \cup p''$. Then p is an extension of both p' and p'', and we obtain by (2.2) and Claim 1,

$$S_*(f,p') \leq S_*(f,p) \leq S^*(f,p) \leq S^*(f,p''),$$

whence the claim follows.

It follows from Claim 2 that

$$\sup_p S_*(f,p) \leq \inf_p S^*(f,p). \tag{2.4}$$

Claim 3. *If $f(x)$ is Darboux integrable then*

$$\sup_p S_*(f,p) = \inf_p S^*(f,p).$$

Indeed, set

$$A = \sup_p S_*(f,p) \quad \text{and} \quad B = \inf_p S^*(f,p) \tag{2.5}$$

so that by (2.4) $A \leq B$. On the other hand, by definition of the Darboux integrability, for any $\varepsilon > 0$ there exists $\delta > 0$ such that, for any partition p with $m(p) < \delta$,

$$S^*(f,p) - S_*(f,p) < \varepsilon.$$

In particular, this implies that $B - A < \varepsilon$. Since this is true for any $\varepsilon > 0$, we obtain $A = B$.

Now we can prove that if f is Darboux integrable then f is Riemann integrable. In fact, let us prove that

$$\lim_{m(p)\to 0} S(f,p,\xi) = A,$$

where A is defined by (2.5). Indeed, for any $\varepsilon > 0$ there exists δ such that, for any partition p with $m(p) < \delta$,

$$S^*(f,p) - S_*(f,p) < \varepsilon.$$

By definition of A and B and by $A = B$, we have

$$S_*(f,p) \le A \le S^*(f,p).$$

By (2.2) we have

$$S_*(f,p) \le S(f,p,\xi) \le S^*(f,p).$$

Therefore, both numbers A and $S(f,p,\xi)$ belong to the same interval $[S_*(f,p), S^*(f,p)]$, which implies that

$$|S(f,p,\xi) - A| \le S^*(f,p) - S_*(f,p) < \varepsilon$$

and which finishes the proof. ■

Hence, being Riemann integrable or Darboux integrable is the same. In the future, we'll simply say that a function is integrable if it is Riemann or Darboux integrable.

Corollary. *Function f is integrable if and only if both limits*

$$\lim_{m(p)\to 0} S^*(f,p), \quad \lim_{m(p)\to 0} S_*(f,p) \tag{2.6}$$

exist (as real numbers) and are equal. Moreover, if f is integrable then

$$\lim_{m(p)\to 0} S^*(f,p) = \lim_{m(p)\to 0} S_*(f,p) = \int_a^b f(x)\,dx = \sup_p S_*(f,p) = \inf_p S^*(f,p). \tag{2.7}$$

Proof. If the limits (2.6) exist and are equal then the limit of their difference is 0, which means that f is Darboux integrable. If f is integrable then we have

$$\lim_{m(p)\to 0} (S^*(f,p) - S_*(f,p)) = 0$$

and

$$\lim_{m(p)\to 0} S(f,p,\xi) \text{ exists.}$$

Since

$$S_*(f,p) \le S(f,p,\xi) \le S^*(f,p),$$

$S_*(f,p)$ and $S^*(f,p)$ must have the same limits as $S(f,p,\xi)$. Since the latter is $\int_a^b f(x)\,dx$, we obtain the left identities in (2.7). The right identities were verified in the proof of Theorem 2.1. ■

Corollary. (Necessary condition for integrability) *If a function f is Riemann integrable on $[a,b]$ then f is bounded from above and below on this interval.*

Proof. Indeed, if $\sup_{[a,b]} f = +\infty$ then for any partition p, there is an interval $[x_{k-1}, x_k]$ such that $\sup_{[x_{k-1},x_k]} f = +\infty$, which implies that

$$S^*(f,p) = +\infty.$$

On the other hand, $S_*(f,p) < +\infty$, which makes the inequality

$$S^*(f,p) - S_*(f,p) < \varepsilon$$

impossible. Therefore, $\sup_{[a,b]} f < +\infty$. In the same way, $\inf_{[a,b]} f > -\infty$, which proves the claim. ■

Theorem 2.2 (Sufficient conditions for integrability)
(*a*) *Any continuous function $f(x)$ on $[a,b]$ is integrable on this interval.*
(*b*) *Any monotone function $f(x)$ on $[a,b]$ is integrable on this interval.*

For the proof of part (*a*), we need a notion of *uniform continuity* (*Gleichmäβig Stetigkeit*).

Definition. A function $f(x)$ on an interval I is called *uniformly continuous* if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x, y \in I \text{ and } |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Recall that f is continuous on I if f is continuous at any point $x \in I$, that is, for any $x \in I$ for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$y \in I \text{ and } |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

In the definition of the continuity at x, the parameter δ may depend on x, while in the uniform continuity δ must be the same for all x, which explains the term "uniform".

Example. The function $f(x) = \frac{1}{x}$ is continuous in $(0,1)$ but is not uniformly continuous. Indeed, whatever is δ, choose $0 < x < \delta$ and $y = x/2$ so that $|x-y| < \delta$ whereas the difference

$$|f(x) - f(y)| = \left|\frac{1}{x} - \frac{1}{y}\right| = \frac{1}{x}$$

can be made larger than ε just by taking x small enough.

Lemma 2.3 *If $f(x)$ is a continuous function on a bounded closed interval I then f is uniformly continuous on I.*

Proof. Fix some $\varepsilon > 0$. For any point $x \in I$ there exists $\delta(x) > 0$ such that

$$|y-x| < \delta(x) \implies |f(y) - f(x)| < \varepsilon/2.$$

Denote by J_x the open interval $\left(x - \frac{1}{2}\delta(x), x + \frac{1}{2}\delta(x)\right)$. Obviously, the family of all intervals $\{J_x\}_{x\in I}$ covers I. By the compactness principle (Theorem 1.10 from Analysis I), any family of open intervals covering of a closed bounded interval I contains a finite

subfamily that also covers I. Hence, select finitely many intervals $J_{x_1}, ..., J_{x_n}$ that cover I. Set

$$\delta = \frac{1}{2} \min_{1 \leq k \leq n} \{\delta(x_k)\}$$

and prove that

$$y', y'' \in I \text{ and } |y' - y''| < \delta \implies |f(y') - f(y'')| < \varepsilon, \tag{2.8}$$

which will prove the uniform continuity of f. Indeed, the point y' belongs to some of the intervals J_{x_k} so that

$$|y' - x_k| < \frac{1}{2}\delta(x_k).$$

Since

$$|y'' - y'| < \delta \leq \frac{1}{2}\delta(x_k),$$

we obtain also

$$|y'' - x_k| < \delta(x_k).$$

Hence, by the choice of $\delta(x_k)$, we have

$$|f(y') - f(x)| < \varepsilon/2$$

and

$$|f(y'') - f(x)| < \varepsilon/2$$

whence (2.8) follows. ■

Proof of Theorem 2.2(a)**.** By Lemma 2.3, function f is uniformly continuous on $[a, b]$ that is, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Consider now any partition $p = \{x_k\}_{k=0}^n$ of $[a, b]$ with $m(p) < \delta$. Then, for any two points $x, y \in [x_{k-1}, x_k]$ we have $|x - y| < \delta$ and, hence

$$|f(x) - f(y)| < \varepsilon.$$

In particular, this implies that

$$\sup_{[x_{k-1}, x_k]} f - \inf_{[x_{k-1}, x_k]} f \leq \varepsilon$$

whence

$$\begin{aligned} S^*(f, p) - S_*(f, p) &= \sum_{k=1}^n \left(\sup_{[x_{k-1}, x_k]} f - \inf_{[x_{k-1}, x_k]} f \right) (x_k - x_{k-1}) \\ &\leq \varepsilon \sum_{k=1}^n (x_k - x_{k-1}) \\ &= \varepsilon(b - a). \end{aligned}$$

Renaming $\varepsilon(b - a)$ by ε, we obtain that the function f is Darboux integrable, which implies by Theorem 2.1 that f is Riemann integrable. ■

Proof of Theorem 2.2(b)**.** Assume for simplicity that f is monotone increasing. Obviously, we have for any partition $p = \{x_k\}_{k=0}^n$

$$\sup_{[x_{k-1},x_k]} f = f(x_k)$$

and

$$\inf_{[x_{k-1},x_k]} f = f(x_{k-1})$$

so that

$$\begin{aligned} S^*(f,p) - S_*(f,p) &= \sum_{k=1}^n \left(\sup_{[x_{k-1},x_k]} f - \inf_{[x_{k-1},x_k]} f \right) (x_k - x_{k-1}) \\ &= \sum_{k=1}^n (f(x_k) - f(x_{k-1}))(x_k - x_{k-1}) \\ &\leq m(p) \sum_{k=1}^n (f(x_k) - f(x_{k-1})) \\ &= m(p)(f(b) - f(a)). \end{aligned}$$

Hence, if $m(p) < \delta$ then

$$S^*(f,p) - S_*(f,p) \leq \delta(f(b) - f(a)).$$

If δ is small enough then the right hand side here is $< \varepsilon$, which means that f is Darboux integrable. ■

2.3 Further properties of the Riemann integral

The next theorem establishes the relation between indefinite and definite integrals.

Theorem 2.4 (The fundamental theorem of calculus – *Fundamentalsatz der Differential- und Integralrechnung*). *Let $f(x)$ be a continuous function on a bounded closed interval $[a,b]$ and let $F(x)$ be its primitive on this interval. Then*

$$\int_a^b f(x)\,dx = F(b) - F(a). \tag{2.9}$$

This formula is also called the *Newton-Leibniz formula,* and it is a cornerstone of Analysis. It can also be written in the form

$$\int_a^b F'(x)\,dx = F(b) - F(a).$$

To describe yet another form of (2.9), let us introduce the following notation:

$$[F]_a^b = F(b) - F(a).$$

Then, using the fact that $F = \int f(x)\,dx$ we can write

$$\int_a^b f(x)\,dx = \left[\int f(x)\,dx \right]_a^b.$$

This formula explains why the notation for the definite and indefinite integrals are so similar, despite the fact that the notions are entirely different.

Proof. By Theorem 2.2, the function f is integrable. Recall that, by definition,

$$\int_a^b f(x)\,dx = \lim_{m(p)\to 0} S(f,p,\xi).$$

Fix a partition $p = \{x_k\}_{k=0}^n$ of the interval $[a,b]$ and choose tags $\xi = \{\xi_k\}_{k=1}^n$ as follows. By the mean value theorem (Theorem 4.9 from Analysis I), there exists $\xi_k \in [x_{k-1}, x_k]$ so that

$$F(x_k) - F(x_{k-1}) = F'(\xi_k)(x_k - x_{k-1}).$$

Taking ξ_k as tags, we can evaluate the Riemann sum for the tagged partition (p,ξ) as follows:

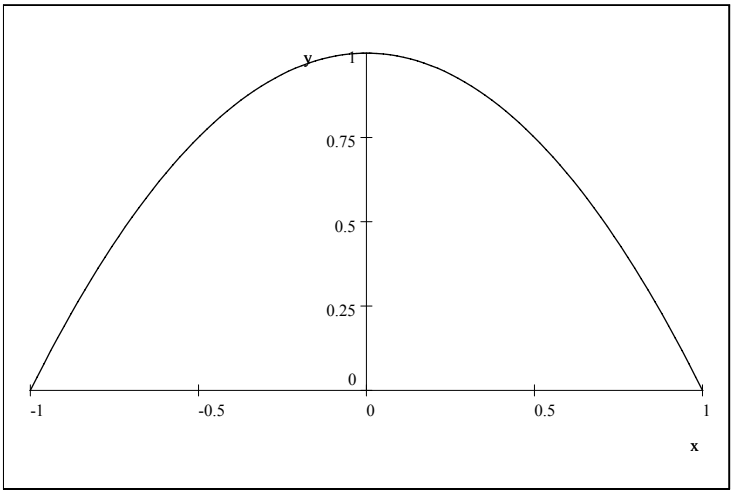
$$\begin{aligned} S(f,p,\xi) &= \sum_{k=1}^n f(\xi_k)(x_k - x_{k-1}) \\ &= \sum_{k=1}^n F'(\xi_k)(x_k - x_{k-1}) \\ &= \sum_{k=1}^n (F(x_k) - F(x_{k-1})) \\ &= F(b) - F(a). \end{aligned}$$

Therefore, the only possible value for $\lim_{m(p)\to 0} S(f,p,\xi)$ is $F(b) - F(a)$ whence the claim follows. ■

Example. 1. Let $f(x) = 1 - x^2$. Then

$$\int_{-1}^{1} \left(1 - x^2\right) dx = \left[\int \left(1 - x^2\right) dx\right]_{-1}^{1} = \left[x - \frac{x^3}{3}\right]_{-1}^{1} = \frac{4}{3}$$

Geometrically the above computation means that the area bounded by the parabola $y = 1 - x^2$ and the x-axis is $4/3$.



In particular, it is $2/3$ of the area of the bounding box $[-1, 1] \times [0, 1]$. The rule that the area under the parabola is $2/3$ of its bounding box was first discovered by Archimedes, who, in the modern terms, was able to evaluate the definite integral directly by definition as the limit of the Riemann sums, without knowing the Newton-Leibniz formula.

2. Let $f(x) = \frac{1}{x}$. Then, for $a > 1$,

$$\int_1^a \frac{dx}{x} = \left[\int \frac{dx}{x}\right]_1^a = [\ln x]_1^a = \ln a.$$

This formula can be used as an independent definition of the natural logarithm, which then leads to definition of $\exp(x)$.

3. Let $f(x) = \frac{1}{1+x^2}$. Then, for $a > 0$,

$$\int_0^a \frac{dx}{1+x^2} = \left[\int \frac{dx}{1+x^2}\right]_0^a = [\arctan x]_0^a = \arctan a.$$

This formula can be used for an independent definition of $\arctan x$, which then leads to definition of $\tan x$.

Note that we did not write the usual constant C in integration because it always cancels out when taking the difference $F(b) - F(a)$.

Theorem 2.5 (Linearity of integral) *If functions f and g are integrable on $[a, b]$ then also $\lambda f + \mu g$ is integrable where λ, μ are constants, and*

$$\int_a^b (\lambda f + \mu g)\, dx = \lambda \int_a^b f dx + \mu \int_a^b g dx. \tag{2.10}$$

Proof. Let (p,ξ) be a tagged partition of $[a,b]$. Then

$$\begin{aligned} S(\lambda f+\mu g,p,\xi) &= \sum_{k=0}^{n}(\lambda f+\mu g)(\xi_k)\Delta x_k \\ &= \lambda\sum_{k=0}^{n} f(\xi_k)\Delta x_k+\mu\sum_{k=0}^{n} g(\xi_k)\Delta x_k, \end{aligned}$$

and when $m(p)\to 0$, the above expression tends to

$$\lambda\int_a^b f(x)\,dx+\mu\int_a^b g(x)\,dx.$$

By definition, this means that $\lambda f+\mu g$ is Riemann integrable and (2.10) holds. ■

Theorem 2.6 (Monotonicity of integral)

(*a*) *If f is integrable on $[a,b]$ and $f\geq 0$ then*

$$\int_a^b f dx\geq 0.$$

(*b*) *If f and g are integrable on $[a,b]$ and $f\geq g$ then*

$$\int_a^b f dx\geq\int_a^b g dx.$$

Proof. (*a*) Indeed, all Riemann sums of f are non-negative, which implies that their limit is also non-negative, which finishes the proof.

(*b*) Since by Theorem 2.5

$$\int_a^b f dx-\int_a^b g dx=\int_a^b (f-g)\,dx$$

and $f-g\geq 0$, the claim follows from part (*a*). ■

Corollary. *If f is integrable on $[a,b]$ then*

$$(b-a)\inf_{[a,b]} f\leq\int_a^b f dx\leq(b-a)\sup_{[a,b]} f.$$

Proof. Let $c=\sup_{[a,b]} f$. Then $f\leq c$ on $[a,b]$ whence by Theorems 2.6 and 2.5

$$\int_a^b f dx\leq\int_a^b c dx=c\int_a^b dx=c\left[\int dx\right]_a^b=c(b-a).$$

The lower bound is proved similarly. ■

Theorem 2.7 *If f is integrable on $[a,b]$ then $|f|$ is also integrable and*

$$\left|\int_a^b f dx\right|\leq\int_a^b |f|\,dx. \tag{2.11}$$

Proof. We claim that, for any interval $I \subset [a,b]$,

$$\sup_I |f| - \inf_I |f| \leq \sup_I f - \inf_I f. \tag{2.12}$$

Denoting the right hand side of (2.12) by M, we have, for all $x, y \in I$,

$$|f(x)| - |f(y)| \leq |f(x) - f(y)| \leq M$$

Taking sup in $x \in I$, we obtain

$$\sup_{x \in I} |f(x)| - |f(y)| \leq M,$$

and taking sup in $y \in I$, we obtain

$$\sup_I |f| - \inf_I |f| \leq M,$$

which proves (2.12).

It follows from (2.12) that, for any partition $p = \{x_k\}_{k=0}^n$ of $[a,b]$,

$$\begin{aligned} S^*(|f|, p) - S_*(|f|, p) &= \sum_{k=0}^n \left(\sup_{[x_{k-1}, x_k]} |f| - \inf_{[x_{k-1}, x_k]} |f| \right) \Delta x_k \\ &\leq \sum_{k=0}^n \left(\sup_{[x_{k-1}, x_k]} f - \inf_{[x_{k-1}, x_k]} f \right) \Delta x_k \\ &= S^*(f, p) - S_*(f, p). \end{aligned}$$

Since

$$S^*(f, p) - S_*(f, p) \to 0 \text{ as } m(p) \to 0,$$

we obtain the same property for $|f|$, which means that $|f|$ is Darboux integrable.

To prove (2.11), observe that $f \leq |f|$ and $-f \leq |f|$, which implies by Theorem 2.6

$$\int_a^b f dx \leq \int_a^b |f| \, dx$$

and

$$-\int_a^b f dx \leq \int_a^b |f| \, dx$$

whence (2.11) follows. Alternatively, for any tagged partition (p, ξ) of $[a,b]$,

$$|S(f, p, \xi)| = \left| \sum_{k=0}^n f(\xi_k) \Delta x_k \right| \leq \sum_{k=0}^n |f(\xi_k) \Delta x_k| = S(|f|, p, \xi).$$

Passing to the limit as $m(p) \to 0$, we obtain (2.11). ■

Theorem 2.8 (Additivity of integral) *Let f be an integrable function on $[a,b]$. Then, for any $c \in (a,b)$, f is integrable on $[a,c]$ and $[c,b]$, and*

$$\int_a^b f dx = \int_a^c f dx + \int_c^b f dx. \tag{2.13}$$

Proof. Let p' and p'' be partitions of $[a,c]$ and $[c,b]$, respectively. Obviously, the set $p = p' \cup p''$ is a partition of $[a,b]$,

$$m(p) = \max(m(p'), m(p''))$$

and

$$S^*(f,p) = S^*(f,p') + S(f,p''), \tag{2.14}$$

and the similar identity holds for S_*. Since f is integrable on $[a,b]$, we have

$$\lim_{m(p)\to 0} (S^*(f,p) - S_*(f,p)) = 0. \tag{2.15}$$

Since

$$S^*(f,p) - S_*(f,p) = (S^*(f,p') - S_*(f,p')) + (S^*(f,p'') - S_*(f,p''))$$

this implies that

$$\lim_{m(p')\to 0} (S^*(f,p') - S_*(f,p')) = 0 = \lim_{m(p'')\to 0} (S^*(f,p'') - S_*(f,p'')).$$

Hence, f is integrable on $[a,c]$ and $[c,b]$. Passing to the limit in identity (2.14) and using (2.7) (see Corollary to Theorem 2.1), we obtain (2.13). ■

Corollary. *If f is integrable on $[A,B]$ then f is integrable on any interval $[a,b] \subset [A,B]$.*

Proof. Indeed, by Theorem 2.8 f is integrable in $[A,b]$. Since $a \in [A,b]$, applying this theorem again, we obtain that f is integrable on $[a,b]$. ■

So far the notion of integral

$$\int_a^b f dx$$

was defined only for the case when $a < b$. Numbers a and b are called, respectively, *lower* and *upper limits* of the integral. Now we define the integral for an arbitrary combination of the lower and upper limit as follows: if $a = b$ then set

$$\int_a^a f(x)\,dx = 0, \tag{2.16}$$

if $a > b$ then set

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx, \tag{2.17}$$

assuming that f is integrable on $[b,a]$. The operation of swapping a and b is referred to as change of the *order of integration*. Hence, changing the order of integration changes the sign of the integral.

Observe that the Newton-Leibniz formula and the linearity of integration remain true for arbitrary upper and lower limits, whereas the monotonicity property requires a specific order (see Exercise 17).

Corollary. *If f is integrable on $[A,B]$ then, for any three points $a,b,c \in [A,B]$, we have*

$$\int_a^b f dx = \int_a^c f dx + \int_c^b f dx \tag{2.18}$$

Proof. By the previous Corollary, all three integrals (2.18) are defined.

If $c = a$ or $c = b$ then (2.18) holds trivially by (2.16). If $a = b$ then (2.18) is equivalent to

$$0 = \int_a^c f dx + \int_c^a f dx,$$

which is true by (2.17). If a, b, c are distinct then there are 6 cases of their mutual location, which can be split into two groups:

1. $c < a < b$, $a < c < b$, $a < b < c$ (in this group $a < b$),
2. $c < b < a$, $b < c < a$, $b < a < c$ (in this group $b < a$).

The case $a < c < b$ was proved in Theorem 2.8.

If $c < a < b$ then rewrite (2.18) in the form

$$\int_a^b f dx = -\int_c^a f dx + \int_c^b f dx. \tag{2.19}$$

Since $a \in (c, b)$, we have by Theorem 2.8

$$\int_c^b f dx = \int_c^a f dx + \int_a^b f dx,$$

whence (2.19) follows.

If $a < b < c$ then rewrite (2.18) in the form

$$\int_a^b f dx = \int_a^c f dx - \int_b^c f dx. \tag{2.20}$$

Since $b \in (a, c)$ we have by Theorem 2.8

$$\int_a^c f dx = \int_a^b f dx + \int_b^c f dx,$$

whence (2.20) follows.

The cases of the second group are considered similarly or obtained from the cases of the first group by changing the order of integration. ■

Before we proceed, let us briefly list (without precise conditions) the main properties of the Riemann integral that we have proved so far.

- The Newton-Leibniz formula:

$$\int_a^b F'(x)\,dx = F(b) - F(a).$$

- Linearity:

$$\int_a^b (\lambda f + \mu g)\,dx = \lambda \int_a^b f dx + \mu \int_a^b g dx.$$

- Monotonicity: if $f \geq g$ and $a < b$ then

$$\int_a^b f dx \geq \int_a^b g dx.$$

  As a consequence, we have proved that

$$(b-a) \inf_{[a,b]} f \ \leq \int_a^b f dx \leq (b-a) \sup_{[a,b]} f \ .$$

  The latter implies the following estimate of the absolute value of the integral

$$\left| \int_a^b f dx \right| \leq |b-a| \sup_{[a,b]} |f|, \tag{2.21}$$

  which is true also when $a \geq b$. Inequality (2.21) is frequently called *LM-inequality* (or ML-inequality). Here L stands for "Length" that is, $|b-a|$, and M stands for "Maximum", which refers to $\sup |f|$.

- Additivity:

$$\int_a^b f dx = \int_a^c f dx + \int_c^b f dx.$$

Using these properties, we are now in position to prove that any continuous function admits a primitive as was promised in section "Indefinite integral".

Theorem 2.9 (Existence of a primitive for a continuous function) *If f is a continuous function on an interval $I \subset \mathbb{R}$ then, for any $c \in I$, the function*

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

is a primitive of f. In particular, f has a primitive on I.

Let us emphasize that I is here an *arbitrary* interval.

Proof. Since f is continuous, the integral $\int_c^x f(t)\,dt$ is defined by Theorem 2.2. We need to prove that $F'(x) = f(x)$ for any $x \in I$, that is,

$$\frac{F(y) - F(x)}{y - x} - f(x) \to 0 \text{ as } y \to x \tag{2.22}$$

(assuming that $y \in I$). We have, using additivity of integral,

$$\begin{aligned} F(y) - F(x) &= \int_c^y f(t)\,dt - \int_c^x f(t)\,dt \\ &= \int_c^y f(t)\,dt + \int_x^c f(t)\,dt \\ &= \int_x^y f(t)\,dt. \end{aligned}$$

Using the fact that $\int_x^y dt = y - x$, we obtain

$$\begin{aligned} \frac{F(y) - F(x)}{y - x} - f(x) &= \frac{1}{y-x}\int_x^y f(t)\,dt - f(x)\frac{1}{y-x}\int_x^y dt \\ &= \frac{1}{y-x}\int_x^y f(t)\,dt - \frac{1}{y-x}\int_x^y f(x)\,dt \\ &= \frac{1}{y-x}\int_x^y (f(t) - f(x))\,dt, \end{aligned}$$

where in the last computation we have used the linearity of the integral. Observe that the variable of integration is t and, hence, $f(x)$ can be regarded as a constant, which can be moved inside the integral.

Next, using LM-inequality, we obtain

$$\begin{aligned} \left|\frac{F(y) - F(x)}{y - x} - f(x)\right| &= \frac{1}{y-x}\left|\int_x^y (f(t) - f(x))\,dt\right| \\ &\leq \frac{1}{y-x}(y-x)\sup_{t\in[x,y]}|f(t) - f(x)| \\ &= \sup_{t\in[x,y]}|f(t) - f(x)|. \end{aligned}$$

Finally, by the continuity of f at x,

$$\sup_{t\in[x,y]}|f(t) - f(x)| \to 0 \text{ as } y \to x,$$

whence (2.22) follows. ■

With Theorem 2.9 at hand, we can restate Theorem 2.4 (the fundamental theorem of calculus) as follows: for any continuous function f on an interval $[a, b]$, a primitive function F exists on $[a, b]$ and

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

If one ignores the question of existence of a primitive then one can end up with wrong results. For example, consider the following computation:

$$\int_{-1}^{1} \frac{dx}{x} = \left[\int \frac{dx}{x}\right]_{-1}^{1} = [\ln|x|]_{-1}^{1} = 0.$$

What is wrong here? Firstly, the function $\frac{1}{x}$ is unbounded on $[-1,1]$ (in fact, it is not defined at 0 but this can be helped by extending it to 0 somehow) and, hence, it is not integrable on $[-1,1]$. Secondly, the primitive $\ln|x|$ is defined away from 0 and, hence, not in the full interval $[-1,1]$ as required by the Newton-Leibniz formula. Hence, the latter is not applicable here.

2.4 Techniques of definite integration

This techniques include the Newton-Leibniz formula as well as integration by parts and change of variable for the definite integral.

Theorem 2.10 (Integration by parts for the definite integral) *If u, v are two continuously differentiable functions on a bounded closed interval $[a,b]$ then*

$$\int_a^b udv = [uv]_a^b - \int_a^b vdu. \tag{2.23}$$

Proof. By Theorem 1.3, we have

$$\int udv = uv - \int vdu \tag{2.24}$$

Note that both integrals here exist by Theorem 2.9. For example, the integral

$$\int udv = \int uv'dx$$

exists because uv' is continuous (the existence of these integrals was used in Theorem 1.3 without proof).

It follows from (2.24) that

$$\left[\int udv\right]_a^b = [uv]_a^b - \left[\int vdu\right]_a^b.$$

By the Newton-Leibniz formula, we have

$$\int_a^b udv = \left[\int udv\right]_a^b,$$

and a similar identity for the second integral in question. Combining these identities, we obtain (2.23). ■

Example. Let us evaluate the integral $\int_0^\pi e^x \cos x dx$. Noticing that $e^x dx = de^x$ and applying (2.23) with $u = \cos x$ and $v = e^x$, we obtain

$$\begin{aligned}\int_0^\pi e^x \cos x dx &= [e^x \cos x]_0^\pi - \int_0^\pi e^x d\cos x \\ &= -e^\pi - 1 + \int_0^\pi e^x \sin x dx \\ &= -(e^\pi + 1) + \int_0^\pi \sin x de^x \\ &= -(e^\pi + 1) + [e^x \sin x]_0^\pi - \int_0^\pi e^x \cos x dx.\end{aligned}$$

Moving the integral to the left hand side and dividing by 2, we obtain

$$\int_0^\pi e^x \cos x dx = -\frac{e^\pi + 1}{2}.$$

Alternatively, one can first evaluate the indefinite integral $\int e^x \cos x$ and then use the Newton-Leibniz formula.

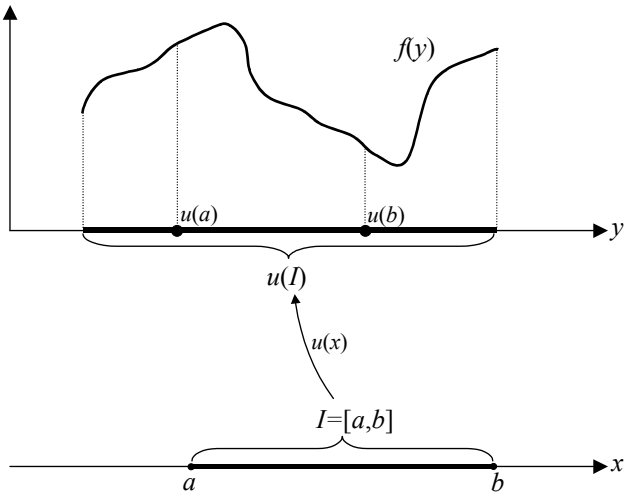
Theorem 2.11 (Change of variable for the definite integral) *Let u be a continuously differentiable function on an interval $I = [a, b]$ and f be a continuous function on the interval $u(I)$. Then*

$$\int_a^b f(u(x))\, du = \int_{u(a)}^{u(b)} f(y)\, dy. \tag{2.25}$$

Recall that the left hand side of (2.25) is a shorthand for

$$\int_a^b f(u(x))\, u'(x)\, dx.$$

Formula (2.25) can be memorized as follows. When making a change $y = u(x)$ in the integral, one has to replace also the limits a and b of integration in variable x by the limits $u(a)$ and $u(b)$ of integration in variable y. Note that $u(a)$ and $u(b)$ do not have to be the endpoints of the interval $u(I)$.



Frequently, one uses letter u instead of y writing

$$\int_a^b f(u(x))\,du = \int_{u(a)}^{u(b)} f(u)\,du.$$

Such notation should be used cautiously since in the integral on the right hand u is an independent variable while the notation $u(a)$ (and $u(b)$) hints that u is still considered as a function of x.

Proof. Note that the image $u(I)$ is a closed bounded interval (Theorem 3.8 from Analysis I). Since f is continuous on $u(I)$, by Theorem 2.9 it has a primitive on this interval, say F. By Theorem 1.4, we have

$$\int f(u(x))\,u'(x)\,dx = F(u(x)) + C,$$

in particular, $F(u(x))$ is a primitive of $f(u(x))\,u'(x)$ on $[a,b]$. In fact, this identity is just a simple application of the chain rule because

$$(F\circ u)'(x) = F'(u(x))\,u'(x) = f(u(x))\,u'(x).$$

Hence, by the Newton-Leibniz formula,

$$\int_a^b f(u(x))\,u'(x)\,dx = F(u(b)) - F(u(a)). \tag{2.26}$$

On the other hand, applying again the Newton-Leibniz formula, we obtain, for any $A, B \in u(I)$,

$$\int_A^B f(y)\,dy = F(B) - F(A),$$

which implies for $A = u(a)$ and $B = u(b)$, that

$$\int_{u(a)}^{u(b)} f(y)\,dy = F(u(b)) - F(u(a)). \tag{2.27}$$

Comparing (2.26) and (2.27), we obtain (2.25). ■

Example. 1. Find $\int_1^2 \frac{dx}{e^x-1}$. To use the change of variable, one needs to represent $\frac{dx}{e^x-1}$ in the form $f(u)\,du$ where $u = u(x)$. If the choice of $u(x)$ is not obvious then one can try to use as $u(x)$ some expression that occurs under the integral. In this case, it is reasonable to take $u = e^x - 1$. Then $du = e^x dx$ whence

$$dx = e^{-x}du = \frac{du}{u+1}.$$

Therefore,

$$\frac{dx}{e^x-1} = \frac{du}{u(u+1)}$$

so that we can take $f(u) = \frac{1}{u(u+1)}$. By (2.25), we have

$$\begin{aligned}\int_1^2 \frac{dx}{e^x-1} &= \int_{u(1)}^{u(2)} \frac{du}{u(u+1)} = \int_{u(1)}^{u(2)} \left(\frac{1}{u} - \frac{1}{u+1}\right) du \\ &= \left[\ln\left|\frac{u}{u+1}\right|\right]_{u(1)}^{u(2)} = \ln\frac{(e^2-1)\,e}{e^2(e-1)} = \ln\frac{e+1}{e} = \ln\left(1+e^{-1}\right).\end{aligned}$$

2.5 Improper integrals

2.5.1 Definition and basic properties of improper integral

So far the notion of the Riemann integral was defined for functions defined on a closed bounded interval $[a,b]$. Assume now that f is defined on a semi-open interval $[a,b)$. Then the integral of f can be still defined as follows.

Definition. If f is defined on an interval $[a,b)$ where $a<b\leq+\infty$ and if f is Riemann integrable on any interval $[a,c]$ with $a<c<b$ then set

$$\int_a^b f(x)\,dx=\lim_{c\to b,c<b}\int_a^c f(x)\,dx,$$

provided the limit exists, finite or infinite. If the limit is finite then one says that the integral $\int_a^b f(x)\,dx$ converges at b. If the limit exists and is $+\infty$ (or $-\infty$) then one says that the integral diverges at b to $+\infty$ (resp. $-\infty$). If the limit does not exist then one says that the integral diverges at b.

The integral $\int_a^b f(x)\,dx$ defined in this way, is called an *improper* (Riemann) integral (the *proper* Riemann integral is the Riemann integral defined as the limit of the Riemann sums). Similarly, if f is defined on $(a,b]$ then one defines the improper integral by

$$\int_a^b f(x)\,dx=\lim_{c\to a,c>a}\int_c^b f(x)\,dx.$$

Example. Consider function $f(x)=x^p$ on $[1,+\infty)$. Consider first the case $p\neq-1$. We have, for any $c>1$,

$$\int_1^c x^p dx=\left[\int x^p dx\right]_1^c=\left[\frac{x^{p+1}}{p+1}\right]_1^c=\frac{c^{p+1}}{p+1}-\frac{1}{p+1}.$$

If $p>-1$ then the above expression goes to $+\infty$ as $c\to+\infty$. Hence,

$$\int_1^{+\infty} x^p dx=+\infty \text{ if } p>-1.$$

If $p<-1$ then $c^{p+1}\to0$ as $c\to+\infty$, and we obtain

$$\int_1^{+\infty} x^p dx=-\frac{1}{p+1} \text{ if } p<-1,$$

so that in this case the integral converges.

In the case $p=-1$, we have

$$\int_1^c \frac{dx}{x}=[\ln x]_1^c=\ln c\to+\infty \text{ as } c\to+\infty$$

so that

$$\int_1^{+\infty} x^{-1}dx=+\infty.$$

Hence, the integral $\int_1^{+\infty} x^p dx$ converges if and only if $p<-1$.

Definition. We say that a function f is *locally integrable* on an interval I if f is defined on I and is Riemann integrable on any bounded closed subinterval of I.

For example, if f is continuous or monotone then f is locally integrable (Theorem 2.2).

Recall that in order to define an improper integral $\int_a^b f(x)\,dx$ of a function f defined on a semi-open interval $[a,b)$, the function f must be locally integrable on $[a,b)$. Then we set

$$\int_a^b f(x)\,dx = \lim_{c\to b}\int_a^c f(x)\,dx,$$

provided the limit exists. The Riemann integral $\int_a^c f(x)\,dx$ exists due to the local integrability of f and because $[a,c]\subset[a,b)$. In this case, we say that the limit of integration b is improper. Similarly, if f is locally integrable on $(a,b]$ then one can consider an improper integral $\int_a^b f(x)\,dx$ with improper limit a.

Most properties of a proper Riemann integral can be extended to improper integral by passing to the limit. For example, let us show the extension of the Newton-Leibniz formula.

Theorem 2.4′ (The Newton-Leibniz formula for improper integrals) *Let $f(x)$ be a continuous function on $[a,b)$ where $-\infty < a < b \le +\infty$ and let F be its primitive on this interval. Then the improper integral $\int_a^b f(x)\,dx$ exists if and only if $\lim_{x\to b} F(x)$ exists, and the following identity holds:*

$$\int_a^b f(x)\,dx = \lim_{x\to b} F(x) - F(a),$$

Proof. By the definition of improper integral and the Newton-Leibniz formula for the proper integral,

$$\int_a^b f(x)\,dx = \lim_{c\to b}\int_a^c f(x)\,dx = \lim_{c\to b}\left(F(c)-F(b)\right) = \lim_{c\to b} F(c) - F(a).$$

We see from this argument that the integral $\int_a^b f(x)\,dx$ exists if and only if $\lim_{c\to b} F(c)$ exists. ■

Let us extend the notation $[F]_a^b$ as follows: if F is defined in $[a,b)$ then set

$$[F]_a^b = \lim_{x\to b} F(x) - F(a),$$

and similarly in the case when F is defined in $(a,b]$. Then the Newton-Leibniz formula can be written in the same way as before:

$$\int_a^b f(x)\,dx = [F]_a^b = \left[\int f dx\right]_a^b.$$

Similarly one obtains extension of integration by parts and of change of variable to improper integrals. The linearity, monotonicity, additivity are also proved by passing to the limit.

In the same way one treats the case when the limit a is improper, that is, when f is locally integrable on $(a,b]$.

Example. 1. Evaluate $\int_0^1 \frac{dx}{\sqrt{x}}$. By the Newton-Leibniz formula, we obtain

$$\int_0^1 \frac{dx}{\sqrt{x}} = \left[\int x^{-1/2}dx\right]_0^1 = \left[2x^{1/2}\right]_0^1 = 2.$$

2. Use integration by parts to evaluate $\int_1^{+\infty} x^{-2}\ln x dx$. We have

$$\begin{aligned}\int_1^{+\infty} x^{-2}\ln x dx &= -\int_1^{+\infty}\ln x d\frac{1}{x} = -\left[\frac{1}{x}\ln x\right]_1^{+\infty} + \int_1^{+\infty}\frac{1}{x}d\ln x \\ &= \int_1^{+\infty}\frac{1}{x^2}dx = \left[\int x^{-2}dx\right]_1^{+\infty} = -\left[x^{-1}\right]_1^{+\infty} = 1.\end{aligned}$$

2. Use change of variable to evaluate $\int_e^{+\infty}\frac{1}{x\ln^3 x}dx$. We have

$$\begin{aligned}\int_e^{+\infty}\frac{1}{x\ln^3 x}dx &= \int_e^{+\infty}\frac{d\ln x}{\ln^3 x} \quad (\text{change } y = u(x) = \ln x) \\ &= \int_{u(e)}^{u(+\infty)}\frac{dy}{y^3} = \left[\int y^{-3}dy\right]_1^{+\infty} = -\left[\frac{y^{-2}}{2}\right]_1^{+\infty} = \frac{1}{2}.\end{aligned}$$

Let us consider now improper integral with both improper limits.

Definition. If f is locally integrable on an open interval (a,b) where $-\infty \le a < b \le +\infty$ then define the improper integral $\int_a^b f(x)\,dx$ with improper limits a and b by

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx, \tag{2.28}$$

where $c \in (a,b)$, provided the both integrals in the right hand side exist as improper integrals with one improper limit, and their sum is defined.

Claim. *The value of $\int_a^b f(x)\,dx$ in (2.28) does not depend on the choice of c.*

Proof. Indeed, if c' is another point in (a,b), then by the additivity integral with one improper limit,

$$\begin{aligned}\int_a^{c'} f dx + \int_{c'}^b f dx &= \left(\int_a^c f dx + \int_c^{c'} f dx\right) + \left(\int_{c'}^c f dx + \int_c^b f dx\right) \\ &= \int_a^c f dx + \int_c^b f dx + \left(\int_c^{c'} f dx + \int_{c'}^c f dx\right) \\ &= \int_a^c f dx + \int_c^b f dx.\end{aligned}$$

■

All the properties of improper integrals, mentioned above remain true with appropriate changes also in the case of two improper limits. For example, to state the Newton-Leibniz

formula, let us extend the notation $[F]_a^b$ to the case when F is defined on an open interval (a,b) as follows:

$$[F]_a^b = \lim_{x\to b} F(x) - \lim_{x\to a} F(x),$$

provided the both limits exist, finite or infinite, and the difference of the limits is defined.

Theorem *2.4″* (The Newton-Leibniz formula for improper integrals with two improper limits) *Let $f(x)$ be a continuous function on (a,b) where $-\infty \le a < b \le +\infty$ and let F be its primitive on this interval. Then the improper integral $\int_a^b f(x)\,dx$ exists if and only if the expression $[F]_a^b$ is defined, and the following identity holds:*

$$\int_a^b f(x)\,dx = [F]_a^b.$$

Proof. Indeed, using the definition (2.28) and Theorem 2.4′, we have

$$\begin{aligned}\int_a^b f(x)\,dx &= -\int_c^a f(x)\,dx + \int_c^b f(x)\,dx \\ &= -\Big(\lim_{x\to a} F(x) - F(c)\Big) + \Big(\lim_{x\to b} F(x) - F(c)\Big) \\ &= [F]_a^b.\end{aligned}$$

■

Example. 1. Consider $\int_{-\infty}^{+\infty} xdx$. By the Newton-Leibniz formula we obtain

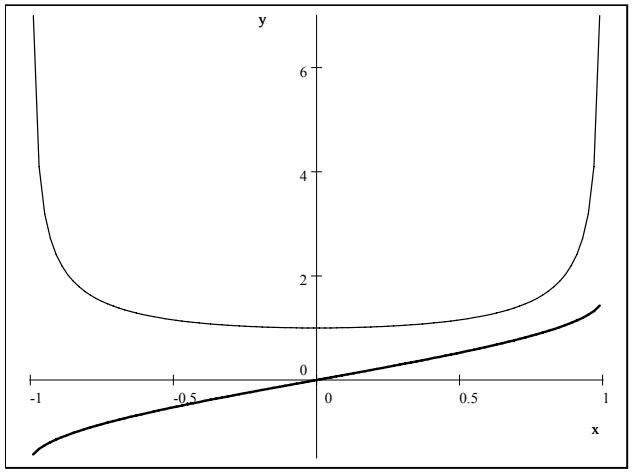
$$\int_{-\infty}^{+\infty} xdx = \left[\frac{x^2}{2}\right]_{-\infty}^{+\infty} = +\infty - (+\infty),$$

which is undefined.

2. Consider $\int_{-1}^{+1} \frac{dx}{\sqrt{1-x^2}}$. The function $\frac{1}{\sqrt{1-x^2}}$ is defined and continuous in $(-1,1)$ (but not at ± 1) so that this integral has two improper limits. Using the Newton-Leibniz formula, we obtain

$$\int_{-1}^{+1} \frac{dx}{\sqrt{1-x^2}} = \left[\int \frac{dx}{\sqrt{1-x^2}}\right]_{-1}^{1} = [\arcsin x]_{-1}^{1} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

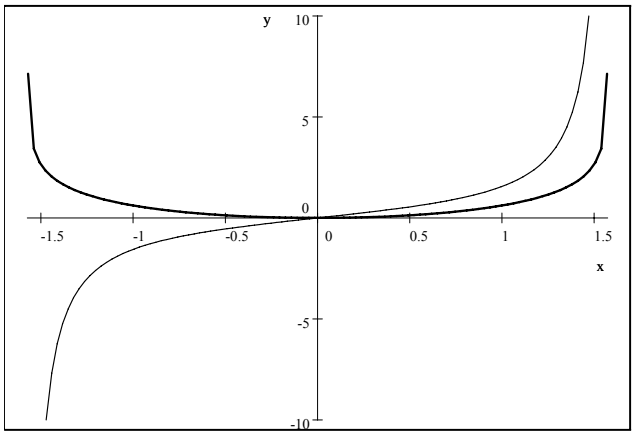
Here are the graphs of functions $\frac{1}{\sqrt{1-x^2}}$ and $\arcsin x$ (thick).



3. Consider $\int_{-\pi/2}^{\pi/2} \tan x$. We have

$$\int_{-\pi/2}^{\pi/2} \tan x dx = \int_{-\pi/2}^{\pi/2} \frac{\sin x}{\cos x} dx = -\int_{-\pi/2}^{\pi/2} \frac{d\cos x}{\cos x} = -\left[\ln|\cos x|\right]_{-\pi/2}^{\pi/2} = -\left(-\infty-(-\infty)\right),$$

since $\cos \pi/2 = 0$ and $\ln y \to -\infty$ as $y \to 0$. Since the difference $-\infty-(-\infty)$ is not defined, the value of the integral $\int_{-\pi/2}^{\pi/2} \tan x$ is not defined either. Here are the graphs of functions $\tan x$ and $-\ln|\cos x|$ (thick):



2.5.2 Convergence of improper integrals

We start with the following simple observation.

Lemma 2.12 *If f is a non-negative locally integrable function on an open interval (a,b) then the improper integral $\int_a^b f(x)\,dx$ exists, finite or infinite.*

Proof. Fix some $c \in (a,b)$ and consider function

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

We claim that the function $F(x)$ is monotone increasing. Indeed, if $y > x$ then

$$F(y) - F(x) = \int_c^y f dt - \int_c^x f dt = \int_x^y f dt \geq 0.$$

Hence, both limits $A = \lim_{x\to a} F(x)$ and $B = \lim_{x\to b} F(x)$ exist, finite or infinite (Theorem 2.6 from Analysis I). By definition, we have

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^c f(x)\,dx + \int_c^b f(x)\,dx = -\int_c^a f(x)\,dx + \int_c^b f(x)\,dx \\ &= -\lim_{x\to a} F(x) + \lim_{x\to b} F(x) = B - A.\end{aligned}$$

We have only to make sure that the difference $B-A$ is defined, that is, A and B cannot be simultaneously $+\infty$ or $-\infty$. Note that $F(x) \leq F(x) = 0$ if $x \leq c$ and $F(x) \geq 0$ if $x \geq c$. Therefore, $A \leq 0$ and $B \geq 0$, which finishes the proof. ■

Hence, if $f \geq 0$ then the integral $\int_a^b f(x)\,dx$ converges if and only if

$$\int_a^b f(x)\,dx < +\infty.$$

Recall for comparison that if $\sum_{k=1}^{\infty} a_k$ is a series of non-negative numbers then it always has a value, finite or infinite, and it converges if and only if

$$\sum_{k=1}^{\infty} a_k < +\infty.$$

The following theorem provides a relation between convergence of series and integrals.

Theorem 2.13 (The integral test for convergence of series) *Let $f(x)$ be a non-negative monotone decreasing function on $[1,+\infty)$. Then*

$$\int_1^{+\infty} f(x)\,dx < \infty \iff \sum_{k=1}^{\infty} f(k) < \infty.$$

Proof. Since $f \geq 0$ and f is locally integrable, both $\int_1^{+\infty} f(x)\,dx$ and $\sum_{k=1}^{\infty} f(k)$ are defined as extended reals. We need to prove that one of them is finite if and only of the other is finite. Since $f(x)$ is decreasing, on any interval $[n-1,n]$ we have

$$f(k) = \inf_{[k-1,k]} f \leq \int_{k-1}^{k} f(x)\,dx \leq \sup_{[k-1,k]} f = f(k-1),$$

whence

$$\int_1^n f(x)\,dx = \sum_{k=2}^{n} \int_{k-1}^{k} f(x)\,dx \leq \sum_{k=2}^{n} f(k-1) = f(1) + ... + f(n-1)$$

and

$$\int_1^n f(x)\,dx = \sum_{k=2}^{n} \int_{k-1}^{k} f(x)\,dx \geq \sum_{k=2}^{n} f(k) = f(2) + ... + f(n).$$

Passing to the limit as $n \to \infty$, we obtain

$$\sum_{k=1}^{\infty} f(k) - f(1) \leq \int_1^{\infty} f(x)\,dx \leq \sum_{k=1}^{\infty} f(k). \tag{2.29}$$

Hence, $\int_1^{\infty} f(x)\,dx$ is finite if and only if $\sum_{k=1}^{\infty} f(k)$ is finite. ■

Example. Let us prove that the series $\sum_{n=1}^{\infty}\frac{1}{n^p}$ converges if and only if $p>1$ (cf. Exercise 41 from Analysis I). If $p\leq 0$ then $\frac{1}{n^p}$ does not do to 0 as $n\to\infty$ so that the series diverges. Let $p>0$. Then the function $f(x)=\frac{1}{x^p}$ is continuous and monotone decreasing on $[1,+\infty)$. Therefore, by Theorem 2.13, the series $\sum_{n=1}^{\infty}\frac{1}{n^p}$ converges if and only if

$$\int_1^{+\infty}\frac{dx}{x^p}<\infty.$$

If $p\neq 1$ then

$$\int_1^{+\infty}\frac{dx}{x^p}=\left[\int x^{-p}dx\right]_1^{+\infty}=\left[\frac{x^{1-p}}{1-p}\right]_1^{+\infty},$$

which is finite and is equal to $\frac{1}{p-1}$ if $1-p<0$, that is, $p>1$. If $p=1$ then

$$\int_1^{+\infty}\frac{dx}{x}=[\ln x]_1^{+\infty}=+\infty.$$

Definition. We say that an improper integral $\int_a^b f(x)\,dx$ *converges absolutely* if

$$\int_a^b |f(x)|\,dx<+\infty.$$

Of course, here $f(x)$ is assumed to be a locally integrable function. Then, by Theorem 2.7, $|f(x)|$ is also integrable. Since $|f|\geq 0$, the integral $\int_a^b |f(x)|\,dx$ exists by Lemma 2.12.

Theorem 2.14 *If an improper integral $\int_a^b f(x)\,dx$ converges absolutely then it converges and*

$$\left|\int_a^b f(x)\,dx\right|\leq\int_a^b |f(x)|\,dx.$$

Proof. Assume for simplicity that the given integral has one improper limit b. Let $F(x)=\int_b^x f(t)\,dt$ and $G(x)=\int_b^x |f(t)|\,dt$. We need to prove that if function $G(x)$ has limit as $x\to b$ then so does $F(x)$. Suffices to prove that $\lim_{n\to\infty}F(x_n)$ exists for any sequence $x_n\to b$. For that, let us show that the sequence $\{F(x_n)\}$ is Cauchy. Indeed, assuming for simplicity that $x_n>x_m$ and using Theorem 2.7, we obtain

$$\begin{aligned}|F(x_n)-F(x_m)| &= \left|\int_a^{x_n}f(t)\,dt-\int_a^{x_m}f(t)\,dt\right|=\left|\int_{x_m}^{x_n}f(t)\,dt\right|\\ &\leq \int_{x_m}^{x_n}|f(t)|\,dt=G(x_n)-G(x_m).\end{aligned}$$

Since by hypothesis $G(x_n)-G(x_m)\to 0$, it follows that also $|F(x_n)-F(x_m)|\to 0$, which finishes the proof. ■

For the next Statement, we need the following notation:

Definition. For functions $f(x)$ and $g(x) > 0$ defined on (a, b), we write that

$$f(x) \sim g(x) \text{ as } x \to b \text{ if } \lim_{x \to b} \frac{f(x)}{g(x)} = 1$$

($f(x)$ is equivalent to $g(x)$ as $x \to b$).

If f is also positive then $f \sim g$ implies $g \sim f$. Also, if $f \sim g$ and $g \sim h$ then $f \sim h$ because

$$\frac{f}{h} = \frac{f}{g}\frac{g}{h}.$$

Also, if $f_1 \sim g_1$ and $f_2 \sim g_2$ then $f_1 f_2 \sim g_1 g_2$ and $\frac{f_1}{f_2} \sim \frac{g_1}{g_2}$.

For example, $\sin x \sim x$ as $x \to 0$ because $\frac{\sin x}{x} \to 1$ as $x \to 0$. Or

$$x^2 + x \sim x^2 \quad \text{as } x \to +\infty$$

because $\frac{x^2+x}{x^2} = 1 + \frac{1}{x} \to 1$ as $x \to +\infty$. On the other hand,

$$x^2 + x \sim x \quad \text{as } x \to 0$$

because $\frac{x^2+x}{x} = x + 1 \to 1$ as $x \to 0$.

Recall for comparison also notation o:

$$f(x) = o(g(x)) \text{ as } x \to b \text{ if } \lim_{x \to b} \frac{f(x)}{g(x)} = 0.$$

Then we have

$$f \sim g \quad \text{is equivalent to } f(x) = g(x) + o(g(x))$$

because $\frac{f}{g} = 1 + \frac{o(g)}{g} \to 1$.

Definition. We write

$$f(x) = O(g(x)) \text{ as } x \to b \text{ if } \limsup_{x \to b} \frac{|f(x)|}{g(x)} < +\infty$$

($f(x)$ is big O of $g(x)$ as $x \to b$). Equivalently,

$$f(x) = O(g(x)) \text{ if } |f(x)| \leq Cg(x) \text{ for large enough } x.$$

Clearly, if $f \sim g$ then $f = O(g)$.

Theorem 2.15 (Comparison test) *Let $f(x)$ and $g(x) > 0$ be two locally integrable functions on $[a, b)$.*

(a) If $f(x) = O(g(x))$ as $x \to b$ then

$$\int_a^b g(x)\,dx < +\infty \implies \int_a^b f(x)\,dx \text{ converges absolutely.}$$

(b) If $f(x) > 0$ and $f(x) \sim g(x)$ as $x \to b$ then the both integrals $\int_a^b f(x)\,dx$ and $\int_a^b g(x)\,dx$ converges or not simultaneously.

Proof. (a) The fact that $f(x) = O(g(x))$ as $x \to b$ means that there is $C > 0$ and $c \in (a, b)$ so that $|f(x)| \leq Cg(x)$ for all $c \leq x < b$. By the additivity of integral, we have

$$\int_a^b |f(x)|\, dx = \int_a^c |f(x)|\, dx + \int_c^b |f(x)|\, dx.$$

By the local integrability of f, the integral $\int_a^c |f(x)|\, dx$ exists as proper. For the second integral, we have

$$\int_c^b |f(x)|\, dx \leq C \int_c^b g(x)\, dx < +\infty.$$

Therefore, $\int_a^b |f(x)|\, dx < \infty$ and, hence, $\int_a^b f(x)\, dx$ converges absolutely.

(b) If $f(x) \sim g(x)$ then $f(x) = O(g(x))$ so that by part (a)

$$\int_a^b g(x)\, dx < +\infty \implies \int_a^b f(x)\, dx < +\infty.$$

For the opposite inequality note that $f \sim g$ implies $g \sim f$ so that we can interchange f and g in this implication and obtain the converse implication. ■

Example. 1. Investigate the convergence of $\int_1^{+\infty} \frac{\sqrt{x}dx}{\sqrt{1+x^4}}$. We have

$$f(x) = \frac{\sqrt{x}}{\sqrt{1+x^4}} = \frac{\sqrt{x}}{x^2\sqrt{x^{-4}+1}} \sim \frac{\sqrt{x}}{x^2} = x^{-3/2} \text{ as } x \to +\infty.$$

Therefore, $f(x) \sim x^{-3/2}$ as $x \to +\infty$ and since

$$\int_1^{+\infty} x^{-3/2} dx < +\infty,$$

we conclude by 2.15 that

$$\int_1^{+\infty} \frac{\sqrt{x}dx}{\sqrt{1+x^4}} < +\infty.$$

2. Investigate the convergence of $\int_1^{+\infty} \frac{\sin x}{x^2} dx$. Indeed, we have

$$\frac{\sin x}{x^2} = O\left(x^{-2}\right) \text{ as } x \to +\infty,$$

and since $\int_1^{\infty} x^{-2} dx < +\infty$, the integral $\int_1^{\infty} \frac{\sin x}{x^2} dx$ converges absolutely.

3. Investigate the convergence of $\int_0^a \frac{dx}{\sqrt{\cos x - \cos a}}$ where $0 < a < \pi/2$. The improper limit is a. We have by the Taylor formula

$$\cos x - \cos a = -\sin a\,(x-a) + o(x-a) \sim (\sin a)\,(a-x) \text{ as } x \to a,$$

whence

$$\frac{1}{\sqrt{\cos x - \cos a}} \sim \frac{1}{\sqrt{\sin a}\sqrt{a-x}}.$$

Since

$$\int_0^a \frac{dx}{\sqrt{a-x}} = \int_0^a \frac{dy}{\sqrt{y}} = \left[2\sqrt{y}\right]_0^a = 2\sqrt{a} < \infty,$$

we conclude that the given integral converges.

2.5.3 Gamma function

Definition. Define function $\Gamma(x)$ for any $x > 0$ by

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt. \tag{2.30}$$

The function $\Gamma(x)$ is called the *gamma* function.

Here we consider some properties of $\Gamma(x)$ related to properties of improper integrals.

Claim 1. *The integral in* (2.30) *converges for any* $x > 0$.

Proof. Note that the both limits 0 and $+\infty$ are improper (to be precise, 0 is improper if $x < 1$). Setting

$$f(t) = t^{x-1} e^{-t},$$

we need to prove that

$$\int_0^1 f(t)\, dt < +\infty \text{ and } \int_1^{+\infty} f(t)\, dt < +\infty.$$

Consider the first integral. Since $f(x) \leq t^{x-1}$ and

$$\int_0^1 t^{x-1} dt = \left[\frac{t^x}{x}\right]_0^1 = \frac{1}{x} < \infty,$$

we conclude that also $\int_0^1 f(t)\, dt < \infty$.

Consider the second integral. If $x \leq 1$ then $t^{x-1} \leq 1$ for all $t \geq 1$ whence $f(t) \leq e^{-t}$. Since

$$\int_1^{+\infty} e^{-t} dt = e^{-1} < \infty,$$

we obtain $\int_1^{+\infty} f(t)\, dt < \infty$. If $x > 1$ then we use the inequality

$$t^{x-1} \leq C e^{\frac{1}{2}t} \text{ for all } t \geq 1,$$

where C is a constant that depends on x. Indeed, the expansion of $e^{\frac{1}{2}t}$ into a power series contains the terms $\frac{1}{n!}\left(\frac{1}{2}t\right)^n$ with any $n \in \mathbb{N}$. Choose $n > x - 1$. Then we have

$$e^{\frac{1}{2}t} \geq \frac{1}{n!}\left(\frac{1}{2}t\right)^n = Ct^n \geq Ct^{x-1},$$

where $C = \frac{1}{n!2^n}$. Therefore, it follows that

$$f(t) = t^{x-1} e^{-t} \leq C e^{-\frac{1}{2}t}.$$

Since $\int_1^{\infty} e^{-\frac{1}{2}t} dt < \infty$, we obtain that also $\int_1^{+\infty} f(t)\, dt < \infty$. ■

Claim 2. *For all* $x > 0$, $\Gamma(x+1) = x\Gamma(x)$.

Proof. We have using integration by parts

$$\begin{aligned} \Gamma(x+1) &= \int_0^{+\infty} t^x e^{-t} dt = -\int_0^{+\infty} t^x de^{-t} = -\left[t^x e^{-t}\right]_0^{+\infty} + \int_0^{+\infty} e^{-t} dt^x \\ &= \int_0^{+\infty} x t^{x-1} e^{-t} dt = x\Gamma(x). \end{aligned}$$

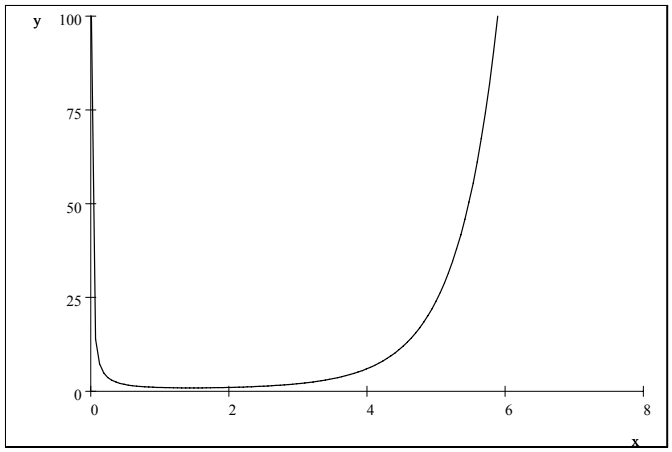
■

Note that

$$\Gamma(1) = \int_0^{+\infty} e^{-t} dt = 1.$$

Using Claim 2, we obtain by induction that $\Gamma(n+1) = n!$ for any non-negative integer n.

Here is the graph[1] of $\Gamma(x)$:



2.5.4 Conditional convergence

Example. Let us show that the integral $\int_0^{+\infty} \frac{\sin x}{x} dx$ converges but not absolutely. Note that the function $\frac{\sin x}{x}$ has limit 1 as $x \to 0$. Hence, it can be extended to 0 as a continuous function so that the limit of integration 0 can be considered as proper. We need only to investigate the convergence at $+\infty$. To prove the convergence, it suffices to prove that the integral $\int_1^{+\infty} \frac{\sin x}{x} dx$ converges. Use integration by parts:

$$\begin{aligned} \int_1^{+\infty} \frac{\sin x}{x} dx &= -\int_1^{+\infty} \frac{d\cos x}{x} = -\left[\frac{\cos x}{x}\right]_1^{+\infty} + \int_1^{+\infty} \cos x \; d\frac{1}{x} \\ &= \cos 1 - \int_1^{+\infty} \frac{\cos x}{x^2} dx. \end{aligned}$$

The last integral converges because

$$\int_1^{+\infty} \left|\frac{\cos x}{x^2}\right| dx \leq \int_1^{+\infty} \frac{1}{x^2} dx < +\infty.$$

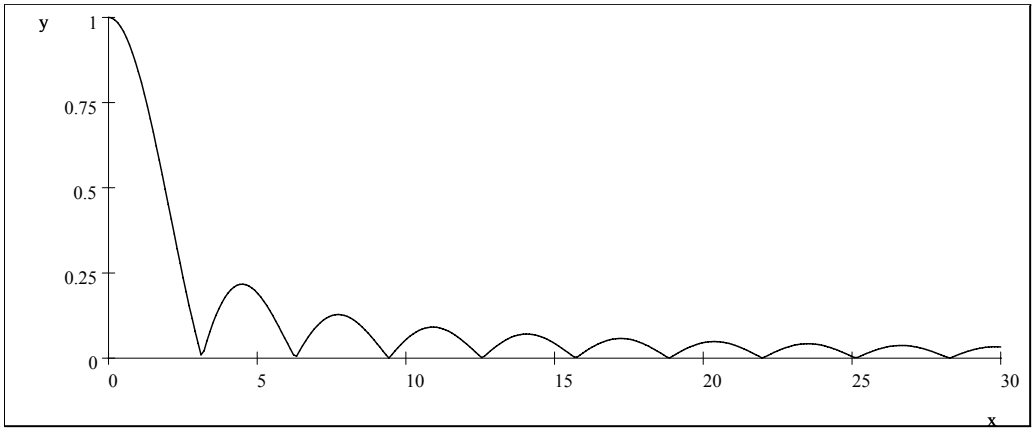
Hence, $\int_1^{+\infty} \frac{\sin x}{x} dx$ converges, too.

[1] In fact, $\Gamma(x)$ can be extended as an analytic function to complex x so that it is defined for all complex x except for non-positive integers.

Now let us prove that

$$\int_1^{+\infty} \frac{|\sin x|}{x} dx = +\infty, \tag{2.31}$$

which will show that the integral $\int_1^{+\infty} \frac{\sin x}{x} dx$ is not absolutely convergent. Here are the graph of function $\frac{|\sin x|}{x}$:



Note that

$$\begin{aligned}
\int_{\pi 2^k}^{\pi 2^{k+1}} \frac{|\sin x|}{x} dx &\geq \int_{\pi 2^k}^{\pi 2^{k+1}} \frac{\sin^2 x}{x} dx \geq \frac{1}{\pi 2^{k+1}} \int_{\pi 2^k}^{\pi 2^{k+1}} \sin^2 x dx \\
&= \frac{1}{\pi 2^{k+1}} \int_{\pi 2^k}^{\pi 2^{k+1}} \frac{1-\cos 2x}{2} dx \\
&= \frac{1}{\pi 2^{k+2}} \int_{\pi 2^k}^{\pi 2^{k+1}} dx - \frac{1}{\pi 2^{k+2}} \int_{\pi 2^k}^{\pi 2^{k+1}} \cos 2x dx \\
&= \frac{\pi 2^{k+1} - \pi 2^k}{\pi 2^{k+2}} - 0 = \frac{1}{4}.
\end{aligned}$$

Since for any $n \in \mathbb{N}$

$$\int_1^{+\infty} \frac{|\sin x|}{x} dx \geq \int_{\pi}^{\pi 2^n} \frac{|\sin x|}{x} dx = \sum_{k=0}^{n-1} \int_{\pi 2^k}^{\pi 2^{k+1}} \frac{|\sin x|}{x} dx \geq \frac{n}{4},$$

we obtain (2.31).

It is possible to prove that $\int_0^{+\infty} \frac{\sin x}{x} dx = \frac{1}{2}\pi$.

Definition. If an improper integral $\int_a^b f(x)\, dx$ converges but does not converge absolutely then we say that it converges *conditionally*.

Hence, we have proved that $\int_0^{+} \frac{\sin x}{x} dx$ converges conditionally.

The next theorem provides useful test for convergence without proving the absolute convergence.

Theorem 2.16 *Let $f(x)$ be a continuous function on $[a,+\infty)$ and $g(x)$ be a continuously differentiable monotone function on $[a,+\infty)$. Then the integral*

$$\int_a^{+\infty} f(x)\, g(x)\, dx$$

converges provided one of the following two conditions is satisfied:

(a) (The Abel test) *Integral $\int_a^{+\infty} f(x)\, dx$ converges and $g(x)$ is bounded..*

(b) (The Dirichlet test) *The function $F(x)=\int_a^x f(t)\, dt$ is bounded and $\lim_{x\to+\infty} g(x)=0$.*

Proof. Consider the function $F(x)$ defined above, which is a primitive of f. Then we have, using integration by parts:

$$\int_a^{+\infty} f(x)\, g(x)\, dx = \int_a^{+\infty} g(x)\, dF(x) = [Fg]_a^{+\infty} - \int_a^{+\infty} F(x)\, g'(x)\, dx. \tag{2.32}$$

We need to show that the both terms in the right hand side are finite.

Let us first prove that the integral $\int_a^{+\infty} F(x)\, g'(x)\, dx$ converges absolutely. Observe that in the both cases function F is **bounded**. Indeed, in the case (b) this is an assumption, while in the case (a), the convergence of $\int_a^{+\infty} f dx$ means that $F(x)$ has a finite limit as $x\to+\infty$, which together with the continuity of F implies that F is bounded. Let $|F(x)|\le C$. Then

$$\int_a^{+\infty} |F(x)\, g'(x)|\, dx \le C\int_a^{+\infty} |g'(x)|\, dx.$$

In the both cases, $g(x)$ has a **finite limit** as $x\to+\infty$. Indeed, in the case (b) this is an assumption, while in the case (a) it follows from the monotonicity and boundedness of g. Since g is monotone, we have that always either $g'(x)\ge 0$ or $g'(x)\le 0$. Assuming that $g'(x)\ge 0$, we obtain

$$\int_a^{+\infty} |g'(x)|\, dx = \int_a^{+\infty} g'(x)\, dx = [g]_a^{+\infty} = \lim_{x\to+\infty} g(x) - g(a) < +\infty,$$

whence the claim follows. Similarly one handles the case $g'(x)\le 0$ using $|g'|=-g'$.

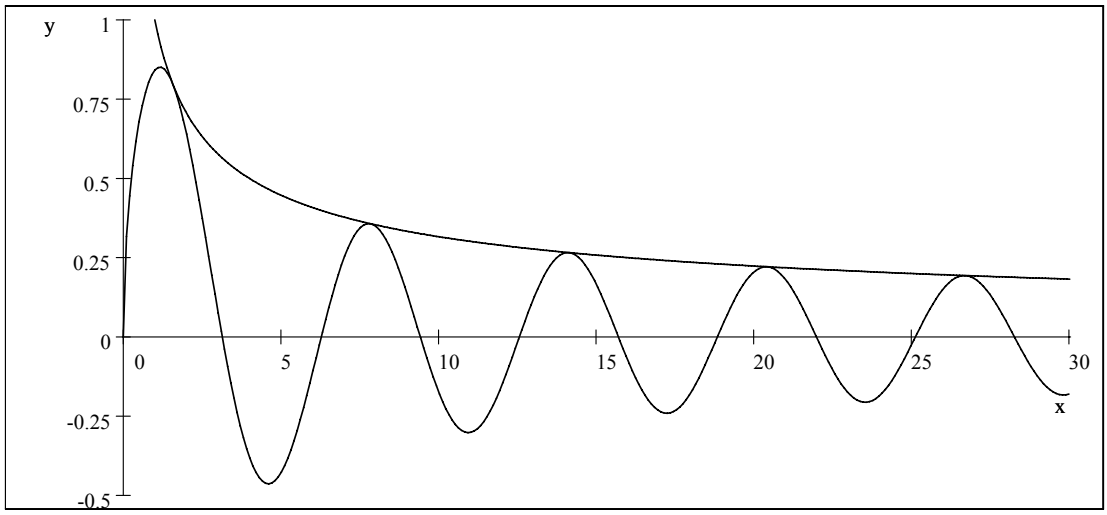
To prove that the expression $[Fg]_a^{+\infty}$ has a finite value, consider two cases separately.

(a) In this case, $F(x)$ has a finite limit at $+\infty$. Since $g(x)$ has also a finite limit, the expression $[Fg]_a^{+\infty}$ is defined and is finite.

(b) If $F(x)$ is bounded and $g(x)\to 0$ as $x\to+\infty$ then $Fg(x)\to 0$ as $x\to+\infty$ so that $[Fg]_a^{+\infty}$ is finite (in fact, it is 0). ■

Example. 1. Consider $\int_1^{+\infty} \frac{\sin x}{x^\alpha} dx$ where $\alpha>0$. Let us show that this integral converges for any $\alpha>0$. Consider function $f(x)=\sin x$ and $g(x)=x^{-\alpha}$. The primitive $F(x)=\int_a^x f(t)\, dt$ is bounded because it is $\cos x + C$, while $g(x)\to 0$ as $x\to+\infty$. Therefore, by the Dirichlet test, the given integral converges.

Below are the graphs of the functions $\frac{\sin x}{\sqrt{x}}$ and $\frac{1}{\sqrt{x}}$:



The area under $\frac{1}{\sqrt{x}}$ is $+\infty$ whereas the signed are under $\frac{\sin x}{\sqrt{x}}$ is finite, which is due to a huge cancellation of the positive and negative parts of the area.

2. Consider $\int_1^{+\infty} \frac{\sin x}{x^\alpha} \arctan x dx$. Set now $f(x) = \frac{\sin x}{x^\alpha}$ and $g(x) = \arctan x$. The function g is bounded and monotone increasing, while $\int_1^{+\infty} f(x)\, dx$ converges by the previous example. Hence, by the Abel test, the given integral converges.

3 Sequences and series of functions

3.1 Uniform convergence

Let $\{f_k\}_{k=1}^{\infty}$ be a sequence of real-valued (or complex valued) functions on a set S. We say that f_k converges to f *pointwise* on S if $f_k(x) \to f(x)$ as $k \to \infty$ for any $x \in S$. We say that f_k converges to f_k *uniformly* (*gleichmässig*) on S if

$$\sup_S |f_k - f| \to 0 \text{ as } k \to \infty.$$

Notation for the uniform convergence: $f_k \rightrightarrows f$.

Claim. *If $f_k \rightrightarrows f$ on S then $f_k(x) \to f(x)$ pointwise on S.*

Proof. Indeed, $\sup_S |f_k - f| \to 0$ implies that $|f_k(x) - f(x)| \to 0$ for any $x \in S$, which exactly means that $f_k(x) \to f(x)$. ■

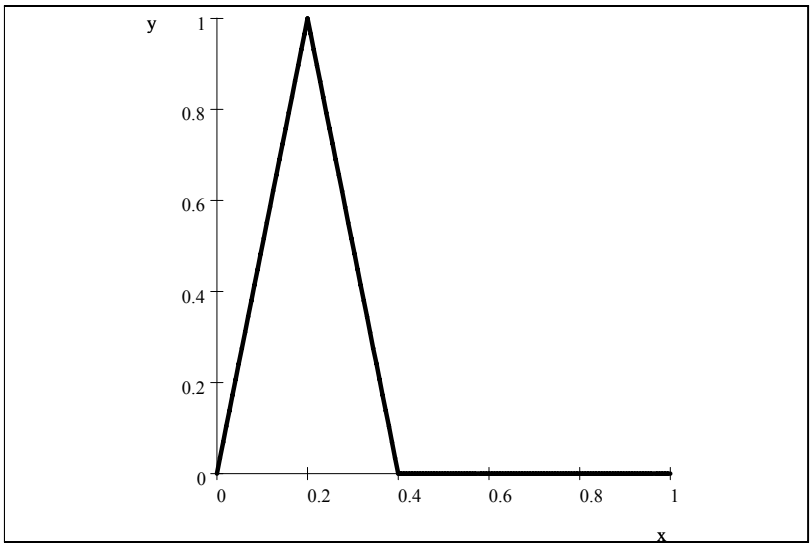
The converse is not true: a sequence may be convergent pointwise but not uniformly.

Example. 1. Let $f_k(x) = \frac{x}{k}$ on $(0, +\infty)$. If $k \to \infty$ then $f_k(x) \to 0$ for any x but not uniformly because $\sup_x |f_k| = \infty$.

2. A more complicated example shows that the same can happen on a closed bounded interval. Indeed, consider on $[0, 1]$ functions

$$f_k(x) = \begin{cases} kx, & 0 \le x \le \frac{1}{k} \\ -k\left(x - \frac{2}{k}\right), & \frac{1}{k} \le x \le \frac{2}{k}, \\ 0, & \frac{2}{k} \le x \le 1. \end{cases}$$

The graph of f_5 is plotted below:



The function f_k is continuous and $\sup f_k = 1$ so that $\{f_k\}$ does not converge uniformly to 0. However, for any $x \in [0,1]$, $f_k(x) \to 0$. Indeed, if $x = 0$ then it is obvious from $f_k(0) = 0$. If $x > 0$ then for large enough k we have $2/k < x$ so that $f_k(x) = 0$.

A major property of the uniform convergence is as follows.

Theorem 3.1 *If $\{f_k\}$ is a sequence of continuous functions on an interval $I \subset \mathbb{R}$ such that $f_k \rightrightarrows f$ on I then f is also continuous on I.*

Proof. Fix a point $x \in I$ and prove that f is continuous at x, that is, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$y \in I,\ |y - x| < \delta \implies |f(y) - f(x)| < \varepsilon.$$

For that, first choose k so big that

$$\sup_I |f_k - f| < \varepsilon/3.$$

Since f_k is continuous at x, there exists δ such that

$$y \in I,\ |y - x| < \delta \implies |f_k(y) - f_k(x)| < \varepsilon/3.$$

Therefore, for such y, we have

$$\begin{aligned} |f(y) - f(x)| &\leq |f(y) - f_k(y)| + |f_k(y) - f_k(x)| + |f_k(x) - f(x)| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \end{aligned}$$

■

Example. It can happen that f_k are continuous, $f_k(x) \to f(x)$ pointwise while f is discontinuous. Indeed, let

$$f_k(x) = \begin{cases} -k\left(x - \frac{1}{k}\right), & 0 \leq x \leq \frac{1}{k}, \\ 0, & \frac{1}{k} \leq x \leq 1, \end{cases}$$

so that $f_k(x)$ is a continuous function on $[0,1]$. If $k \to \infty$ then $f_k(x) \to 0$ for $x > 0$ and $f_k(0) \to 1$ so that the limit function

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & 0 < x \leq 1 \end{cases}$$

is discontinuous at 0.

3.2 Uniform convergence of series

For any function f defined on a set S, define the *norm* of f on S by

$$\|f\|_S := \sup_{x\in S} |f(x)|.$$

Or we just write $\|f\|$ skipping subscript S if it is clear from the context that the sup is taken over S. Then the uniform convergence $f_n \rightrightarrows f$ on S can be stated as follows:

$$\|f_k - f\| \to 0 \text{ as } k \to \infty.$$

For comparison recall that for numerical sequences $x_k \to x$ is equivalent to $|x_n - x| \to 0$.

Consider now a series $\sum_{k=1}^{\infty} f_k(x)$ of functions f_k defined on a set S, and let

$$F_n(x) = \sum_{k=1}^{n} f_k(x)$$

be the partial sums of the series.

Definition. We say that a series $\sum f_k$ converges pointwise/uniformly if the sequence $\{F_n\}$ of partial sums converges as $n \to \infty$ pointwise/uniformly.

Theorem 3.2 (Weierstrass convergence test) *If $\{f_k\}$ is a sequence of functions on a set S such that*

$$\sum_{k=1}^{\infty} \|f_k\| < \infty$$

then the series $\sum_{k=1}^{\infty} f_k(x)$ converges on S absolutely and uniformly.

Hence, this test reduces the question of the uniform convergence of a *functional* series to the question of the convergence of a *numerical* series.

Proof. Obviously, for any $x \in S$,

$$\sum_{k=1}^{\infty} |f_k(x)| \le \sum_{k=1}^{\infty} \|f_k\| < \infty$$

so that the series $\sum f_k(x)$ converges absolutely for any $x \in S$. In particular, the series converges pointwise so that we can set

$$F(x) = \sum_{k=1}^{\infty} f_k(x).$$

We need to prove that the above series converges uniformly, that is, $F_n \rightrightarrows F$ as $n \to \infty$. Let us prove that

$$\|F - F_n\| \to 0 \text{ as } n \to \infty.$$

Indeed, we have, for any $x \in S$

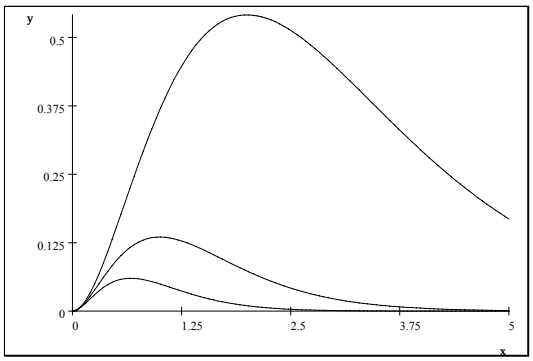
$$|F(x) - F_n(x)| = \left| \sum_{k=n+1}^{\infty} f_k(x) \right| \leq \sum_{k=n+1}^{\infty} |f_k(x)| \leq \sum_{k=n+1}^{\infty} \|f_k\|.$$

Since the right hand side is independent of x, taking sup of x in the left hand side, we obtain

$$\|F - F_k\| \leq \sum_{k=n+1}^{\infty} \|f_k\|.$$

The right hand side tends to 0 as $n \to \infty$ because it is a tail of the convergent series $\sum_{k=1}^{\infty} \|f_k\|$, whence the claim follows. ■

Example. Let us prove that the series $\sum_{k=1}^{\infty} x^2 e^{-kx}$ converges uniformly on $[0, +\infty)$. Function $f_k(x) = x^2 e^{-kx}$ is positive for $x > 0$ but vanishes at $x = 0$ and tends to 0 at $x \to +\infty$. Therefore, it has a maximum at some point $x > 0$. On the next diagram, see the plots of functions $f_k(x)$ for $k = 1, 2, 3$:



At the point of maximum, we have $f_k'(x) = 0$, which is equivalent to $(\ln f_k)' = 0$ that is,

$$(2 \ln x - kx)' = 0$$

whence $x = \frac{2}{k}$. Hence,

$$\|f_k\|_{[1,+\infty)} = \max_{[1,+\infty)} f_k = f_k\left(\frac{2}{k}\right) = \left(\frac{2}{k}\right)^2 e^{-2} = \frac{4e^{-2}}{k^2}.$$

Since the series $\sum \frac{1}{k^2}$ converges, we conclude by the Weierstrass test that the given series converges absolutely and uniformly on $[0, +\infty)$.

Example. Consider a power series $\sum_{k=0}^{\infty} c_k x^k$ (with coefficients $c_k \in \mathbb{R}$) and assume that it converges for some $x = x_0 \neq 0$. We claim that the series converges absolutely in $(-R, R)$ where $R = |x_0|$, and converges uniformly in any interval $[-r, r]$ where $0 < r < R$. Indeed, we have, for any $x \in [-r, r]$,

$$\left|c_k x^k\right| = \left| c_k x_0^k \left(\frac{x}{x_0}\right)^k \right| \leq \left|c_k x_0^k\right| \left(\frac{r}{R}\right)^k.$$

The fact that the series $\sum_{k=0}^{\infty} c_k x^k$ converges implies that $c_k x_0^k \to 0$ as $k \to \infty$. In particular, the sequence $\left\{c_k x_0^k\right\}$ is bounded, say, $\left|c_k x_0^k\right| \leq C$ for all k and some constant C. It follows that

$$\left\|c_k x^k\right\|_{[-r,r]} \leq C \left(\frac{r}{R}\right)^k \text{ for all } x \in [-r, r],$$

Since $\frac{r}{R} < 1$ and, hence, the geometric series $\sum_k \left(\frac{r}{R}\right)^k$ converges, we conclude by the Weierstrass test that $\sum_k c_k x^k$ converges absolutely and uniformly on $[-r, r]$. Consequently, the sum of this series is a continuous function on $(-R, R)$.

3.3 Integration under uniform convergence

Theorem 3.3 *Let $\{f_k\}$ be a sequence of continuous functions on a closed bounded interval $[a, b]$, which converges uniformly on $[a, b]$. Then*

$$\int_a^b \lim_{k\to\infty} f_k(x)\,dx = \lim_{k\to\infty} \int_a^b f_k(x)\,dx. \tag{3.33}$$

Hence, the operations lim and $\int_a^b$ are interchangeable provided the convergence is uniform.

This theorem can be stated as follows: if $f_k \rightrightarrows f$ on $[a, b]$ and f_k are continuous then

$$\int_a^b f_k(x)\,dx \to \int_a^b f(x)\,dx.$$

Proof. Indeed, function $f = \lim_{k\to\infty} f_k$ is continuous by Theorem 3.1 (and, hence, f is Riemann integrable). By the properties of the Riemann integral, we have

$$\left|\int_a^b f_k(x)\,dx - \int_a^b f(x)\,dx\right| = \left|\int_a^b (f_k - f)\,dx\right| \le \sup|f_k - f|\,(b-a).$$

Since $\sup|f_k - f| \to 0$, we obtain

$$\left|\int_a^b f_k(x)\,dx - \int_a^b f(x)\,dx\right| \to 0,$$

which was to be proved. ■

Example. The uniform convergence in Theorem 3.3 is essential. Indeed, consider the following functions on $[0, 1]$:

$$f_k(x) = \begin{cases} k^2 x, & 0 \le x \le \frac{1}{k} \\ -k^2\left(x - \frac{2}{k}\right), & \frac{1}{k} \le x \le \frac{2}{k}, \\ 0, & \frac{2}{k} \le x \le 1. \end{cases}$$

Obviously, $\lim_{k\to 0} f_k = 0$ pointwise. However,

$$\int_0^1 f_k(x)\,dx = \int_0^{1/k} k^2 x dx + \int_{1/k}^{2/k} k^2\left(\frac{2}{k} - x\right)dx = \frac{1}{2} + \frac{1}{2} = 1$$

so that

$$\lim_{k\to\infty} \int_0^1 f_k(x)\,dx = 1 \ne 0 = \int_0^1 \lim_{k\to\infty} f_k dx.$$

Corollary. *If the series* $\sum_{k=1}^{\infty} f_k(x)$ *of continuous functions on* $[a,b]$ *converges uniformly then*

$$\int_a^b \left(\sum_{k=1}^{\infty} f_k(x) \right) dx = \sum_{k=1}^{\infty} \int_a^b f_k(x)\, dx.$$

Proof. Indeed, let

$$F_n = \sum_{k=1}^{n} f_k \text{ and } F = \sum_{k=1}^{\infty} f_k.$$

Then $F_n \rightrightarrows F$ and, by Theorem 3.3,

$$\int_a^b F_n dx \to \int_a^b F dx.$$

Using the linearity of the integral, we obtain

$$\sum_{k=1}^{\infty} \int_a^b f_k dx = \lim_{n\to\infty} \sum_{k=1}^{n} \int_a^b f_k dx = \lim_{n\to\infty} \int_a^b F_n dx = \int_a^b F dx = \int_a^b \left(\sum_{k=1}^{\infty} f_k \right) dx,$$

which was to be proved. ■

Example. Consider a power series $f(x) = \sum_{k=0}^{\infty} c_k x^k$ that converges in $(-R, R)$ for some $R > 0$. As we already know, it converges uniformly in any interval $[-r, r]$ where $0 < r < R$. For any $y \in (-R, R)$, integrating the series from 0 to y, we obtain

$$\int_0^y f(x)\, dx = \sum_{k=0}^{\infty} \int_0^y c_k x^k dx = \sum_{k=0}^{\infty} \frac{c_k}{k+1} y^{k+1}.$$

Hence, a primitive function of f can be obtained by term-by-term integration of the series.

Consider a geometric series

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k.$$

if $|x| < 1$. Integrating it term-by-term from 0 to y, where $|y| < 1$, we obtain

$$\int_0^y \frac{dx}{1-x} = \sum_{k=0}^{+\infty} \frac{y^{k+1}}{k+1}$$

whence

$$-\ln(1-y) = y + \frac{y^2}{2} + \frac{y^3}{3} + \frac{y^4}{4} + ...$$

Or, changing $x = -y$, we obtain the following formula:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots, \tag{3.34}$$

where the series converges for $x \in (-1, 1)$. Note that the partial sums of this series are the Taylor polynomials of the function $\ln(1+x)$.

It is possible to prove that (3.34) extends to the borderline value $x = 1$ so that

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots .$$

Example. Consider $\int_0^1 e^{x^2}dx$. The function e^{x^2} has no primitive in terms of elementary functions. Let us evaluate this integral numerically using expansion

$$e^{x^2} = \sum_{n=0}^{\infty} \frac{x^{2n}}{n!} = 1 + \frac{x^2}{1!} + \frac{x^4}{2!} + ...$$

Since this series converges for any x, it converges uniformly on any bounded closed interval, as was shown in Example in the previous section. By Corollary to Theorem 3.3, we can integrate the series term-by-term, that is,

$$\begin{aligned} \int_0^1 e^{x^2}dx &= \sum_{n=0}^{\infty} \int_0^1 \frac{x^{2n}}{n!}dx = \sum_{n=0}^{\infty} \left[\frac{x^{2n+1}}{(2n+1)\,n!}\right]_0^1 \\ &= \sum_{n=0}^{\infty} \frac{1}{(2n+1)\,(n!)} = 1 + \frac{1}{3\cdot 1!} + \frac{1}{5\cdot 2!} + \frac{1}{7\cdot 3!} + ... \end{aligned}$$

Partial sums of this series can be considered as numerical approximations to $\int_0^1 e^{x^2}dx$. For example, one can compute

$$\sum_{n=0}^{14} \frac{1}{(2n+1)\,n!} \approx 1.462651745\,907\,16$$

$$\sum_{n=0}^{15} \frac{1}{(2n+1)\,n!} \approx 1\,.462651745\,90718$$

$$\sum_{n=0}^{20} \frac{1}{(2n+1)\,(n!)} \approx 1.462\,651745\,90718$$

whence

$$\int_0^1 e^{x^2}dx \approx 1.462\,651745\,90718$$

3.4 Differentiation under uniform convergence

Definition. We say that a functional sequence $\{f_k\}$ converges *locally uniformly* to a function f on in interval I if $f_k \rightrightarrows f$ on any bounded closed subinterval $J \subset I$. We write in this case $f_k \overset{loc}{\rightrightarrows} f$ on I. The same definition applies to a functional series: it converges locally uniformly if the sequence of its partial sums converges locally uniformly.

Remark. The relation with the other types of convergence is obvious:

$$f_k \rightrightarrows f \quad \Longrightarrow \quad f_k \overset{loc}{\rightrightarrows} f \quad \Longrightarrow \quad f_k \to f \text{ pointwise.}$$

Remark. It follows from Theorem 3.1 that if a sequence $\{f_k\}$ of continuous functions converges locally uniformly then the limit is also continuous function, because the limit is continuous on any bounded closed subinterval.

Example. Consider the sequence $f_k(x) = x/k$ on $\mathbb{R}$. Obviously, $f_k \to 0$ pointwise but $f_k \not\rightrightarrows 0$ because $\|f_k\|_{\mathbb{R}} = \infty$. Let us show that $f_k \overset{loc}{\rightrightarrows} 0$. Indeed, in any bounded closed interval J, function $|x|$ is bounded, say, by a constant C_J. Then $\|f_k\|_J \leq C_J/k$ whence $\|f_k\|_J \to 0$ and, hence, $f_k \rightrightarrows 0$ in J. We conclude that $f_k \overset{loc}{\rightrightarrows} f$ on $\mathbb{R}$.

Theorem 3.4 *Let $\{f_k\}$ be a sequence of continuously differentiable functions on an interval $I \subset \mathbb{R}$. Assume that*

1. *$f_k \to f$ pointwise on I,,*
2. *$f_k' \overset{loc}{\rightrightarrows} g$ on I.*

Then $f' = g$.

Equivalently, this theorem can be stated as follows: if the sequence f_k converges pointwise and f_k' converges locally uniformly then

$$\left(\lim_{k\to\infty} f_k\right)' = \lim_{k\to\infty} f_k',$$

that is, the operations of differentiation and limit are interchangeable.

Proof. By the Newton-Leibniz formula, we have for all $x, c \in I$:

$$f_k(x) - f_k(c) = \int_c^x f_k'(t)\,dt. \tag{3.35}$$

Since the convergence of f_k' to g is locally uniform, function g is continuous on I. Since $f_k' \rightrightarrows g$ on $[x, c]$, we have by Theorem 3.3 that

$$\int_c^x f_k'(t)\,dt \to \int_c^x g(t)\,dt \text{ as } k \to \infty.$$

Then, letting in (3.35) $k \to \infty$, we obtain

$$f(x) - f(c) = \int_c^x g(t)\,dt.$$

By Theorem 2.9 we conclude that $f' = g$, which was to be proved. ■

Remark. Note that the statement of Theorem 3.4 is about the derivatives, but the proof uses quite seriously integration.

Corollary. *Let $\{f_k\}$ be a sequence of continuously differentiable functions on an interval $I \subset \mathbb{R}$. Assume that*

1. *$\sum_{k=1}^{\infty} f_k(x)$ converges on I,*
2. *$\sum_{k=1}^{\infty} f_k'(x)$ converges locally uniformly on I.*

Then

$$\left(\sum_{k=1}^{\infty} f_k\right)' = \sum_{k=1}^{\infty} f_k'.$$

The proof is the same as the proof of the Corollary to Theorem 3.3: just apply Theorem 3.4 to partial sums of the series in question.

As an example of application of this Corollary, we prove the following theorem, which is important by itself.

Theorem 3.5 *Let the series $\sum_{k=0}^{\infty} c_k x^k$ converge in $(-R, R)$, where $R \in (0, +\infty]$, and let $F(x)$ be the sum of the series. Then function $F(x)$ is differentiable on $(-R, R)$ and*

$$F'(x) = \sum_{k=1}^{\infty} c_k k x^{k-1}.$$

That is, the power series can be differentiated term-by-term.

For example, consider the series

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

which converges on $(-\infty, +\infty)$. By Theorem 3.5, we have

$$\exp(x)' = \sum_{k=0}^{\infty} \left(\frac{x^k}{k!}\right)' = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \exp(x),$$

so that the derivative of $\exp(x)$ coincides with the function itself. Of course, we already know this from Analysis I.

Proof of Theorem 3.5. By Corollary to Theorem 3.4, we can write

$$\left(\sum_{k=0}^{\infty} c_k x^k\right)' = \sum_{k=0}^{\infty} \left(c_k x^k\right)' = \sum_{k=1}^{\infty} c_k k x^{k-1},$$

provided the series $\sum_k c_k x^k$ converges, which is given, and the series $\sum_k c_k k x^{k-1}$ converges locally uniformly. It suffices to prove that this series converges uniformly on any interval $[-r, r]$ where $0 < r < R$. To prove the latter, observe that, for all $r < a < R$ and for all indices k,

$$\left\|k x^{k-1}\right\|_{[-r,r]} = k r^{k-1} \leq C a^k$$

where C is a constant depending on r, a only. Indeed, write $\frac{a}{r} = 1 + \varepsilon$ where $\varepsilon > 0$. Then, using Bernoulli's inequality, we obtain

$$\left(\frac{a}{r}\right)^k = (1+\varepsilon)^k > k\varepsilon$$

whence

$$k r^{k-1} < r^{-1}\varepsilon^{-1} a^k = C a^k,$$

with $C = r^{-1}\varepsilon^{-1}$. It follows that

$$\left\|kc_k x^{k-1}\right\|_{[-r,r]} \leq C\left|c_k\right| a^k.$$

By Example considered above, the series $\sum c_k a^k$ converges absolutely, that is,

$$\sum_k \left|c_k\right| a^k < \infty$$

whence

$$\sum_k \left\|kc_k x^{k-1}\right\|_{[r,-r]} < \infty.$$

By the Weierstrass test (Theorem 3.2), the series $\sum kc_k x^{k-1}$ converges uniformly on $[-r, r]$, which was to be proved. ■

Example. Knowing that the series $\sum_{k=1}^{\infty} c_k k x^{k-1}$ converges in $(-R, R)$, we can apply Theorem 3.5 to this series, to obtain that F' is differentiable and

$$F''(x) = \sum_{k=2}^{\infty} c_k k(k-1) x^{k-2}.$$

Continuing by induction, we obtain that F is differentiable infinitely many times on $(-R, R)$ and its derivative is obtain by term-by-term differentiation of the series.

Example. Consider again the identity

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k.$$

Differentiating it, we obtain

$$\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1} = \sum_{k=0}^{\infty} (k+1) x^k.$$

Differentiating again,

$$\frac{2}{(1-x)^3} = \sum_{k=1}^{\infty} (k+1) kx^{k-1} = \sum_{k=0}^{\infty} (k+2)(k+1) x^k,$$

etc. All these identities hold for $|x| < 1$ and are particular cases of a binomial series.

3.5 Fourier series

3.5.1 Fourier coefficients

A Fourier series is a series of the form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx), \tag{3.36}$$

where $x \in \mathbb{R}$ and a_k, b_k are the coefficients of the series. Partial sums are

$$S_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx),$$

and they are called a *trigonometric polynomial* for the obvious reason.

In this section, we consider the question whether a given function can be represented as a sum of the Fourier series. We start with the following lemma.

Lemma 3.6 *Let the Fourier series (3.36) converge uniformly on $\mathbb{R}$ to a function $f(x)$. Then, for all k,*

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx dx \quad \text{and} \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx dx. \tag{3.37}$$

Proof. Note that the integrals in (3.37) are defined because function $f(x)$ is continuous as the uniform limit of continuous functions. Fix non-negative integer n and multiply the identity

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) \tag{3.38}$$

by $\cos nx$. The resulting series still converges uniformly (because $|\cos nx| \leq 1$), whence by Theorem 3.3

$$\begin{aligned} \int_0^{2\pi} f(x) \cos nx dx &= \frac{a_0}{2} \int_0^{2\pi} \cos nx dx \\ &+ \sum_{k=1}^{\infty} \left(a_k \int_0^{2\pi} \cos kx \cos nx dx + b_k \int_0^{2\pi} \sin kx \cos nx dx \right). \end{aligned} \tag{3.39}$$

Using the identity

$$\cos a \cos b = \frac{1}{2} (\cos(a-b) + \cos(a+b)),$$

we obtain

$$\int_0^{2\pi} \cos kx \cos nx dx = \frac{1}{2} \int_0^{2\pi} \cos(k-n)x \, dx + \frac{1}{2} \int_0^{2\pi} \cos(k+n)x \, dx. \tag{3.40}$$

Now, use the following formula: for any integer l,

$$\int_0^{2\pi} \cos lx \, dx = \begin{cases} 2\pi, & l = 0 \\ 0, & l \neq 0, \end{cases} \tag{3.41}$$

because in the case $l \neq 0$ the integral is proportional to $[\sin lx]_0^{2\pi} = 0$. It follows from (3.40) that

$$\int_0^{2\pi} \cos kx \cos nx dx = \begin{cases} 2\pi, & k = n = 0 \\ \pi, & k = n \neq 0, \\ 0, & k \neq n. \end{cases}$$

Next, similarly we have, for all k, n,

$$\int_0^{2\pi} \sin kx \cos nx dx = \frac{1}{2} \int_0^{2\pi} (\sin(k+n)x + \sin(k-n)x) \, dx = 0$$

because

$$\int_0^{2\pi} \sin lx \; dx = 0 \text{ for all } l \in \mathbb{Z}. \tag{3.42}$$

It follows from (3.39) that the only non-zero term in the right hand side is the one with $k = n$. If $n = 0$ then we obtain

$$\int_0^{2\pi} f(x)\, dx = \frac{a_0}{2} 2\pi = \pi a_0.$$

If $n > 0$ then

$$\int_0^{2\pi} f(x) \cos nx dx = \pi a_n$$

so that in both cases we have (3.37).

The coefficients b_k are found similarly by multiplying (3.38) by $\sin nx$. ■

Note that all the terms in the Fourier series are 2π-periodic functions on $\mathbb{R}$. Therefore, whenever the sum exists it will be also 2π-periodic. In what follows we'll deal with either functions defined on $[0, 2\pi]$ or 2π-periodic functions on $\mathbb{R}$.

Lemma 3.6 motives the following definition.

Definition. For any Riemann integrable function f on $[0, 2\pi]$, define its *Fourier coefficients* by

$$\boxed{a_k = a_k(f) = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\, dx} \text{ and } \boxed{b_k = b_k(f) = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\, dx} \quad (3.43)$$

for all integers $k \geq 0$. The Fourier series of function f is the series

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx).$$

Since we do not know yet whether this series converges and if so then whether its sum is $f(x)$, we'll write

$$\boxed{f(x) \sim \tfrac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)} \quad (3.44)$$

meaning that the coefficients a_k and b_k are those associated with f. We are going find out under what conditions of f the sign $\sim$ can be replaced by $=$ and in what sense the series converges.

Example. 1. If $f(x) \equiv 1$ then we obtain from (3.43), (3.41), and (3.42), that $a_0 = 1$ while $a_k = b_k = 0$ for all $k \geq 1$. Hence, in this case the Fourier series of $f(x)$ is identically equal to 1 and, hence, coincides with $f(x)$.

2. Consider on $[0, 2\pi]$ a step function

$$f(x) = \begin{cases} 1, & x \leq \pi, \\ 0, & x > \pi. \end{cases}$$

Then

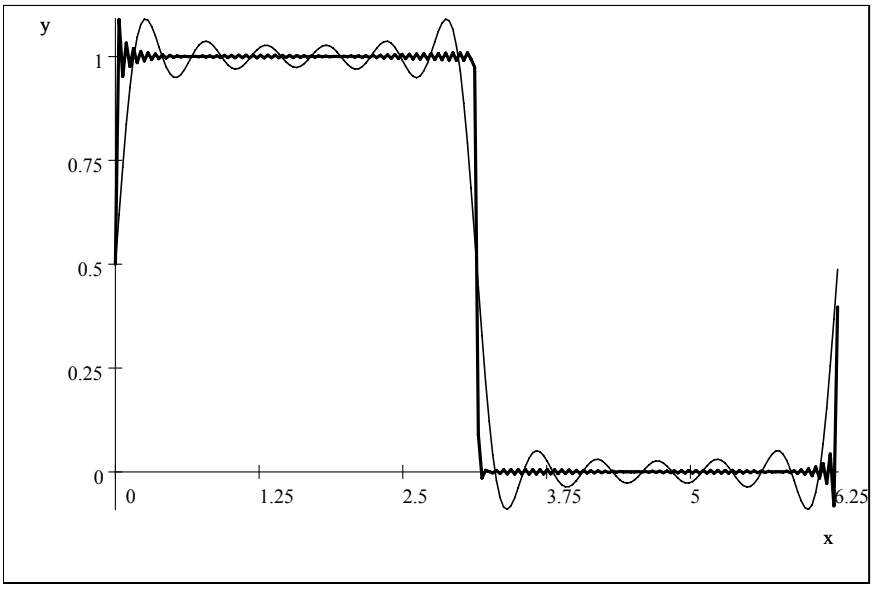
$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_0^{\pi} dx = 1, \\ a_k &= \frac{1}{\pi}\int_0^{\pi} \cos kx\, dx = 0, \ \ k > 0, \\ b_k &= \frac{1}{\pi}\int_0^{\pi} \sin kx\, dx = -\frac{1}{\pi}\left[\frac{\cos kx}{k}\right]_0^{\pi} = \frac{1}{\pi k}\left(1 - (-1)^k\right) = \begin{cases} 0, & k \text{ even}, \\ \frac{2}{\pi k}, & k \text{ odd}. \end{cases} \end{aligned}$$

Hence, the Fourier series has the form

$$f \sim \frac{1}{2} + \sum_{l=0}^{\infty} \frac{2}{\pi(2l+1)} \sin(2l+1)x = \frac{1}{2} + \frac{2}{\pi}\sin x + \frac{2}{3\pi}\sin 3x + \frac{2}{5\pi}\sin 5x + ...$$

It is not obvious at all whether one has equality here. Below are the graphs of the partial sums with $l \leq 5$ and $l \leq 50$ (thick) of this series that do suggest that there is convergence

excepts for the points $x = 0, \pi, 2\pi$:



For what follows it is useful to consider a complex form of the Fourier series. First of all, note that the Riemann integral $\int_a^b f(x)\,dx$ can be defined when f is a complex valued function simply by

$$\int_a^b f(x)\,dx = \int_a^b \operatorname{Re} f(x)\,dx + i\int_a^b \operatorname{Im} f(x)\,dx,$$

provided both $\operatorname{Re} f$ and $\operatorname{Im} f$ are integrable. Most properties of integration are easily transferred to the complex values functions (in particular, linearity, additivity, LM-inequality, the Newton-Leibniz formula, integration-by-parts).

Definition. For any complex valued integrable function f on $[0, 2\pi]$, define its *complex Fourier coefficients* for all $k \in \mathbb{Z}$ by

$$\boxed{c_k = c_k(f) = \tfrac{1}{2\pi}\int_0^{2\pi} f(x)\,e^{-ikx}dx}. \tag{3.45}$$

The complex Fourier series of f is the series

$$f(x) \sim \sum_{k\in\mathbb{Z}} c_k e^{ikx}.$$

Claim. *The complex Fourier series coincides with the Fourier series provided the both converge.*

Proof. Indeed, the complex Fourier series is a double series, that can be written down as follows:

$$\begin{aligned}\sum_{k\in\mathbb{Z}} c_k e^{ikx} &= c_0 + \sum_{k=1}^{\infty} c_k e^{ikx} + \sum_{k=1}^{\infty} c_{-k} e^{-ikx} \\ &= c_0 + \sum_{k=1}^{\infty} \left(c_k e^{ikx} + c_{-k} e^{-ikx}\right).\end{aligned}$$

Using definitions of a_k, b_k, c_k and the Euler formula $e^{ix} = \cos x + i \sin x$, we obtain

$$c_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx = \frac{a_0}{2}$$

and for $k > 0$

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) \cos kx \; dx - i \frac{1}{2\pi} \int_0^{2\pi} f(x) \sin kx dx = \frac{a_k - ib_k}{2},$$

$$c_{-k} = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, e^{ikx} dx = \frac{a_k + ib_k}{2}.$$

Therefore,

$$\begin{aligned} c_k e^{ikx} + c_{-k} e^{-ikx} &= \frac{1}{2}(a_k - ib_k)(\cos kx + i \sin kx) \\ &\quad + \frac{1}{2}(a_k + ib_k)(\cos kx - i \sin kx) \\ &= a_k \cos kx + b_k \sin kx. \end{aligned}$$

It follows that

$$\sum_{k \in \mathbb{Z}} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx),$$

which was to be proved. ■

3.5.2 Bessel's inequality

Theorem 3.7 (Bessel's inequality) *For any integrable function f on $[0, 2\pi]$*

$$\sum_{k \in \mathbb{Z}} |c_k(f)|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 dx. \tag{3.46}$$

Remark. Since function $|f|^2$ is integrable, it follows from (3.46) that the series $\sum_{k \in \mathbb{Z}} |c_k(f)|^2$ converges.

Remark. If f is real values then we have for $k \geq 0$

$$|c_k|^2 = |c_{-k}|^2 = \frac{1}{4}\left(a_k^2 + b_k^2\right),$$

so that by (3.46)

$$\frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2\right) \leq \frac{1}{\pi} \int_0^{2\pi} |f(x)|^2 dx.$$

This is a version of Bessel's inequality for real Fourier series.

Remark. In fact, under the conditions of Theorem 3.7 one has the equality

$$\sum_{k \in \mathbb{Z}} |c_k(f)|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 dx, \tag{3.47}$$

which will be proved later on under some restriction.

Proof. It suffices to prove that, for any positive integer n,

$$\sum_{|k|\leq n} |c_k|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 \, dx.$$

Consider first a function

$$g = \sum_{|k|\leq n} c_k e^{ikx},$$

which is a partial sum of the Fourier series.

Claim 1. *We have the identity*

$$\sum_{|k|\leq n} |c_k|^2 = \frac{1}{2\pi} \int_0^{2\pi} |g|^2 \, dx.$$

The proof is based on the fact that

$$\int_0^{2\pi} e^{ikx} e^{-ilx} dx = \begin{cases} 0, & k \neq l, \\ 2\pi, & k = l. \end{cases}$$

Indeed, the case $k = l$ is trivial, while in the case $k \neq l$ the integral in question is equal to

$$\int_0^{2\pi} \cos(k-l)\, x \; dx + i \int_0^{2\pi} \sin(k-l)\, x \; dx,$$

which vanishes by (3.41) and (3.42).

Next, using

$$\overline{g} = \sum_{|k|\leq n} \overline{c_k e^{ikx}} = \sum_{|k|\leq n} \overline{c_k} e^{-ikx} = \sum_{|l|\leq n} \overline{c_l} e^{-ilx}$$

we obtain

$$|g|^2 = g\overline{g} = \sum_{|k|\leq n} c_k e^{ikx} \sum_{|l|\leq n} \overline{c_l} e^{-ilx} = \sum_{|k|\leq n} \sum_{|l|\leq n} c_k \overline{c_l} e^{i(k-l)x},$$

whence

$$\begin{aligned} \int_0^{2\pi} |g|^2 \, dx &= \sum_{|k|\leq n} \sum_{|l|\leq n} c_k \overline{c_l} \int_0^{2\pi} e^{i(k-l)x} dx \text{ (restricting summation to } l = k) \\ &= \sum_{|k|\leq n} c_k \overline{c_k} 2\pi = 2\pi \sum_{|k|\leq n} |c_k|^2, \end{aligned}$$

which was to be proved.

Claim 2. *We have the identity*

$$\int_0^{2\pi} f\overline{g} dx = \int_0^{2\pi} |g|^2 \, dx.$$

Indeed, using the definition of g and Claim 1, we obtain

$$\begin{aligned}\int_0^{2\pi} f\overline{g}dx &= \int_0^{2\pi} f(x)\,\overline{\sum_{|k|\le n} c_k e^{ikx}}dx = \sum_{|k|\le n}\overline{c_k}\int_0^{2\pi} f(x)\,e^{-ikx}dx \\ &= 2\pi\sum_{|k|\le n}\overline{c_k}c_k = 2\pi\sum_{|k|\le n}|c_k|^2 = \int_0^{2\pi}|g(x)|^2\,dx.\end{aligned}$$

which was to be proved.

Claim 3. *We have inequality*

$$\int_0^{2\pi}|g|^2\,dx \le \int_0^{2\pi}|f|^2\,dx \tag{3.48}$$

Indeed, setting $h = f - g$, we obtain by Claim 2

$$\int_0^{2\pi} h\overline{g}dx = \int_0^{2\pi}(f-g)\,\overline{g}dx = \int_0^{2\pi} f\overline{g}dx - \int_0^{2\pi}|g|^2\,dx = 0. \tag{3.49}$$

Noticing that

$$|f|^2 = |g+h|^2 = (g+h)\left(\overline{g}+\overline{h}\right) = |g|^2 + 2\,\mathrm{Re}\,(h\overline{g}) + |h|^2$$

and integrating this identity using (3.49), we obtain

$$\int_0^{2\pi}|f|^2\,dx = \int_0^{2\pi}|g|^2\,dx + 2\int_0^{2\pi}\mathrm{Re}\,(h\overline{g}) + \int_0^{2\pi}|h|^2 \ge \int_0^{2\pi}|g|^2\,dx,$$

which was to be proved.

Finally, combining Claims 1 and 3, we obtain

$$\sum_{|k|\le n}|c_k|^2 = \frac{1}{2\pi}\int_0^{2\pi}|g|^2\,dx \le \frac{1}{2\pi}\int_0^{2\pi}|f|^2\,dx,$$

which finishes the proof of the theorem. ■

Corollary. (Riemann's lemma) *For any integrable function f on $[0,2\pi]$,*

$$\int_0^{2\pi} f(x)\cos kx\;dx \to 0 \text{ and } \int_0^{2\pi} f(x)\sin kx\;dx \to 0 \text{ as } k\to\infty,$$

where k is a positive integer.

Proof. Since

$$\cos kx = \frac{e^{ikx}+e^{-ikx}}{2} \text{ and } \sin kx = \frac{e^{ikx}-e^{-ikx}}{2i},$$

it suffices to prove that

$$\int_0^{2\pi} f(x)\,e^{-ikx}dx \to 0 \text{ as } |k|\to\infty$$

where $k \in \mathbb{Z}$. The latter is equivalent to $c_k \to 0$ as $|k| \to \infty$, and this is true because by Bessel's inequality the series $\sum_{k\in\mathbb{Z}} |c_k|^2$ converges. ■

In fact, the statement of Riemann's lemma is true when k takes all real values but the proof in this case is more complicated and does not follow from Bessel's inequality. We'll need the following version of Riemann's lemma.

Corollary. *For any integrable function f on $[0, 2\pi]$,*

$$\int_0^{2\pi} f(x) \sin(k + \frac{1}{2})x dx \to 0 \text{ as } k \to \infty,$$

where k is a positive integer.

Proof. Indeed,

$$\sin(k + \frac{1}{2})x = \sin kx \cos\frac{x}{2} + \cos kx \sin\frac{x}{2}$$

whence

$$\int_0^{2\pi} f(x) \sin(k + \frac{1}{2})x dx = \int_0^{2\pi} \left(f(x) \cos\frac{x}{2}\right) \sin kx dx + \int_0^{2\pi} \left(f(x) \sin\frac{x}{2}\right) \cos kx dx.$$

Since the functions $f(x)\cos\frac{x}{2}$ and $f(x)\sin\frac{x}{2}$ are integrable, applying the previous Corollary, we obtain that the both integrals in the right hand side tend to 0, which finishes the proof. ■

3.5.3 Uniform convergence

Theorem 3.8 *Let f be a 2π-periodic function on $\mathbb{R}$ which is continuously differentiable on $\mathbb{R}$. Then the Fourier series of f converges absolutely and uniformly on $\mathbb{R}$.*

So far, we do not claim that the Fourier series converges to $f(x)$ - we just claim that it converges and, moreover, uniformly. Later on we'll prove that the sum of the Fourier series is indeed $f(x)$.

Proof. We start with the following claim that relates the Fourier coefficients of f to those of f'.

Claim. *We have for any $k \in \mathbb{Z}$,*

$$c_k(f') = ikc_k(f). \tag{3.50}$$

Indeed, using integration by parts and 2π-periodicity of f, we obtain

$$\begin{aligned} c_k(f') &= \frac{1}{2\pi}\int_0^{2\pi} f'(x) e^{-ikx} dx \\ &= \frac{1}{2\pi}\int_0^{2\pi} e^{-ikx} df(x) \\ &= \frac{1}{2\pi}\left[e^{-ikx} f(x)\right]_0^{2\pi} + \frac{ik}{2\pi}\int_0^{2\pi} f(x) e^{-ikx} dx \\ &= 0 + ikc_k(f), \end{aligned}$$

which was to be proved.

Since function f' is integrable on $[0, 2\pi]$, we obtain by Bessel's inequality

$$\sum_{k\in\mathbb{Z}} |c_k(f')|^2 < \infty,$$

whence by (3.50)

$$\sum_{k\in\mathbb{Z}} |kc_k(f)|^2 < \infty.$$

Using the inequality $|ab| \le a^2 + b^2$, we obtain

$$\sum_{k\neq 0} |c_k(f)| = \sum_{k\neq 0} |kc_k(f)| \frac{1}{|k|} \le \sum_{k\neq 0} |kc_k(f)|^2 + \sum_{k\neq 0} \frac{1}{k^2} < \infty,$$

which proves that

$$\sum_{k\in\mathbb{Z}} |c_k(f)| < \infty.$$

Finally, since $\left|c_k(f)\, e^{ikx}\right| = |c_k(f)|$, we conclude by the Weierstrass test that the series $\sum_{k\in\mathbb{Z}} c_k(f)\, e^{ikx}$ converges absolutely and uniformly on $\mathbb{R}$. ■

3.5.4 Pointwise convergence

To state the next theorem, we need the following notation. We say that a function f on $\mathbb{R}$ has the *left limit* at point x if the limit

$$\lim_{y\to x, y<x} f(y) \text{ exists and is finite.}$$

In this case, we denote the value of the limit by $f(x-)$ so that

$$f(x-) = \lim_{y\to x, y<x} f(y).$$

Similarly one defines the *right limit* $f(x+)$ replacing above $y < x$ by $y > x$.

Let a function f have the left limit at x. We say that f is *left differentiable* at x if the limit

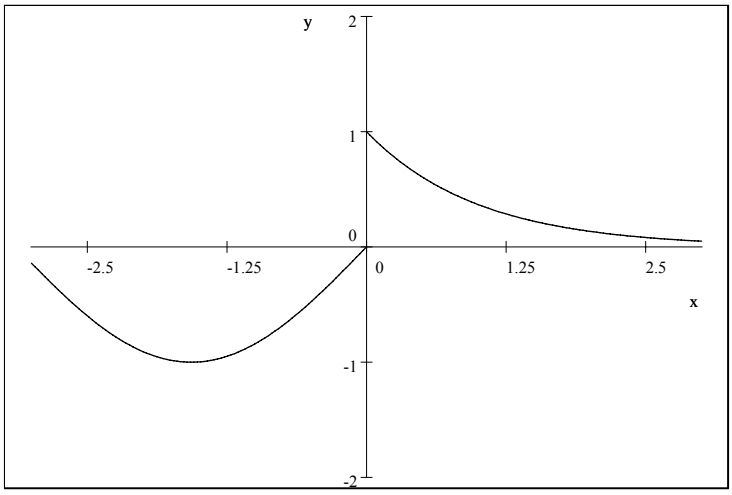
$$\lim_{y\to x, y<x} \frac{f(y) - f(x-)}{y - x} \text{ exists and is finite.}$$

The value of this limit is called the left derivative of f and is denoted by $f'(x-)$. Similarly one defines the right differentiability from the right and the right derivative $f'(x+)$.

For example, for the function

$$f(x) = \begin{cases} e^{-x}, & x \geq 0, \\ \sin x, & x \leq 0 \end{cases}$$

we have $f(0+) = 1$, $f(0-) = 0$, $f'(0+) = -1$, $f'(0-) = 1$.



Of course, if f is differentiable at x then f is left and right differentiable.

Theorem 3.9 *Let f be an 2π-periodic integrable function that is right and left differentiable at some $x \in \mathbb{R}$. Then the Fourier series of f at x converges to $\frac{f(x-)+f(x+)}{2}$. In particular, if $f(x)$ is differentiable at x then the Fourier series of f at x converges to $f(x)$.*

Example. Let $f(x)$ be a 2π-periodic function on $\mathbb{R}$, which is defined on $[0, 2\pi)$ by

$$f(x) = \begin{cases} 1, & 0 \le x \le \pi, \\ 0, & \pi < x < 2\pi. \end{cases}$$

As we have seen in the previous lecture, the Fourier series for this function is

$$f \sim \frac{1}{2} + \sum_{l=0}^{\infty} \frac{2}{\pi(2l+1)} \sin(2l+1)x = \frac{1}{2} + \frac{2}{\pi}\sin x + \frac{2}{3\pi}\sin 3x + \frac{2}{5\pi}\sin 5x + ...$$

If $x = \pi$ then the sum of the Fourier series is $\frac{1}{2}$, which coincides with $\frac{f(\pi-)+f(\pi+)}{2} = \frac{1+0}{2} = \frac{1}{2}$. At all other points in $(0, 2\pi)$ the function f is differentiable; therefore, the Fourier series converges to $f(x)$. For example, take $x = \pi/2$. Replacing $\sim$ by $=$ and using $\sin(2l+1)\frac{\pi}{2} = (-1)^l$, we obtain

$$1 = \frac{1}{2} + \frac{2}{\pi}\sum_{l=0}^{\infty} \frac{(-1)^l}{2l+1},$$

whence

$$\sum_{l=0}^{\infty} \frac{(-1)^l}{2l+1} = \frac{\pi}{4}.$$

Proof of Theorem 3.9. Let

$$S_n(x) = \sum_{|k|\le n} c_k e^{ikx}.$$

We need to prove that

$$S_n(x) \to \frac{f(x-) + f(x+)}{2} \text{ as } n \to \infty.$$

Substituting

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(t)\, e^{-ikt} dt,$$

into $S_n(x)$, we obtain

$$S_n(x) = \frac{1}{2\pi}\sum_{|k|\le n}\int_0^{2\pi} f(t)\, e^{ik(x-t)} dt = \frac{1}{2\pi}\int_0^{2\pi} f(t)\left(\sum_{|k|\le n} e^{ik(x-t)}\right) dt.$$

Let us compute the sum in the brackets setting $u = x - t$. Denoting this sum by $D(u)$ (it is called the Dirichlet kernel) we obtain for $u \neq 0$, using the formula for the sum of a geometric series,

$$\begin{aligned} D(u) &= \sum_{k=-n}^{n} e^{iku} = \sum_{k=-n}^{n} \left(e^{iu}\right)^k = e^{-inu}\sum_{k=0}^{2n}\left(e^{iu}\right)^k \\ &= e^{-inu}\frac{\left(e^{iu}\right)^{2n+1} - 1}{e^{iu} - 1} = \frac{e^{iu\left(n+\frac{1}{2}\right)} - e^{-iu\left(n+\frac{1}{2}\right)}}{e^{iu/2} - e^{-iu/2}} = \frac{\sin\left(n+\frac{1}{2}\right)u}{\sin\frac{u}{2}}. \end{aligned}$$

If $u=0$ then $D(u)=2n+1$, which is equal to $\lim_{u\to 0} D(u)$. Hence, $D(u)$ is continuous function of u, and we obtain

$$\begin{aligned} S_n(x) &= \frac{1}{2\pi}\int_0^{2\pi} f(t)\, D(x-t)\, dt \text{ (change } u=x-t) \\ &= -\frac{1}{2\pi}\int_x^{x-2\pi} f(x-u)\, D(u)\, du = \frac{1}{2\pi}\int_{x-2\pi}^{x} f(x-u)\, D(u)\, du. \end{aligned}$$

Since the functions f and D are 2π-periodic, the integral in the right hand side is the same over any interval of length 2π. Therefore,

$$\begin{aligned} S_n(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)\, D(u)\, du \text{ (using } D(u)=D(-u)) \\ &= \frac{1}{2\pi}\int_0^{\pi} (f(x-u)+f(x+u))\, D(u)\, du. \end{aligned}$$

Applying this formula for function $f\equiv 1$ and using that $S_n(x)=1$ we obtain

$$\frac{1}{\pi}\int_0^{\pi} D(u)\, du = 1.$$

Therefore,

$$\begin{aligned} S_n(x) - \frac{f(x-)+f(x+)}{2} &= \frac{1}{2\pi}\int_0^{\pi} [f(x-u)+f(x+u)]\, D(u)\, du \\ &\quad -\frac{1}{\pi}\int_0^{\pi} \frac{f(x-)+f(x+)}{2} D(u)\, du \\ &= \frac{1}{2\pi}\int_0^{\pi} [(f(x-u)-f(x-)) \\ &\quad +(f(x+u)-f(x+))] D(u)\, du. \end{aligned}$$

Define function $F_+(u)$ for $u>0$ by

$$F_+(u) = \frac{f(x+u)-f(x+)}{u}$$

so that F_+ is locally integrable in $(0,+\infty)$. By the hypotheses, $\lim_{u\to 0} F_+(u)$ exists and is finite so that F_+ can be extended by continuity at $u=0$. This implies that F_+ is integrable on any interval $[0,a]$, in particular, on $[0,\pi]$ (see Exercises). Similarly, define for $u>0$ the function

$$F_-(u) = \frac{f(x-u)-f(x-)}{u}$$

and observe that F_- is integrable on $[0,\pi]$.

Next, we write

$$\begin{aligned} [(f(x-u)-f(x-))+(f(x+u)-f(x+))]\, D(u) &= \frac{F_-(u)\, u + F_+(u)\, u}{\sin\frac{u}{2}} \sin(n+\frac{1}{2})u \\ &= G(u)\sin(n+\frac{1}{2})u, \end{aligned}$$

where

$$G(u) = \frac{F_-(u) + F_+(u)}{\sin\frac{u}{2}}u.$$

Function $G(u)$ is integrable on $[0,\pi]$ because so are functions $F_-(u)$, $F_+(u)$ while the function $\frac{u}{\sin u/2}$ can be considered as continuous on $[0,\pi]$: its is obviously continuous on $(0,\pi]$ and extends continuously to $u=0$ since it tends to 2 as $u \to 0$. Since

$$S_n(x) - \frac{f_+(x) + f_-(x)}{2} = \frac{1}{2\pi}\int_0^\pi G(u)\sin(n+\frac{1}{2})udu,$$

we conclude by Riemann's lemma that the both sides of this identity go to 0 as $n \to \infty$, which finishes the proof. ■

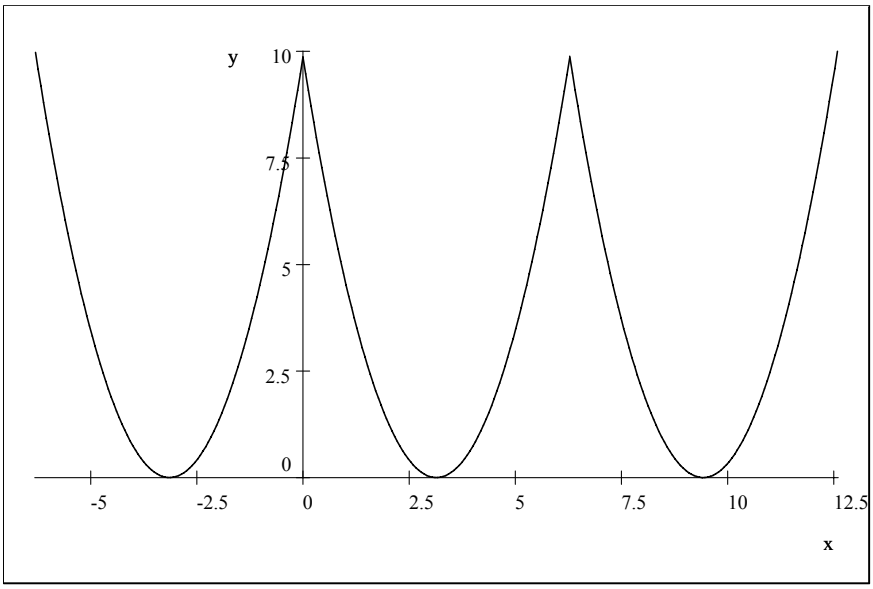
3.5.5 Uniform convergence revisited

It follows from Theorem 3.9 that if $f(x)$ is 2π-periodic and differentiable at all $x \in \mathbb{R}$ then the Fourier series of f converges to f pointwise on $\mathbb{R}$. On the other hand, if f is in addition continuously differentiable then by Theorem 3.8 the Fourier series of f converges uniformly. Combining these two results, we obtain that in this case the Fourier series converges to f *uniformly*.

This result will be extended in the next theorem to the case when f may be not differentiable at some points.

Theorem 3.10 *Let f be a 2π-periodic continuous function on $\mathbb{R}$. Assume that f is differentiable in $[0,2\pi]\setminus A$ where A is a finite set, and that the derivative f' is continuous in $[0,2\pi]\setminus A$ and has finite left and right limits at the points of A. Then the Fourier series of f converges to f uniformly on $\mathbb{R}$.*

Example. Consider function $f(x) = (x-\pi)^2$ on $[0,2\pi]$ that is extended 2π-periodically to $\mathbb{R}$ – see the graph below.



This function satisfies the conditions of Theorem 3.10: Function f is continuous, differentiable in $[0,2\pi]\setminus A$ where $A=\{0,2\pi\}$, the derivative f' is continuous away from 0 and 2π and has right and left limits at these points. Let us compute the real Fourier coefficients of function f:

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_0^{2\pi}(x-\pi)^2\,dx = (\text{change } y = x-\pi) \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} y^2 dy = \frac{1}{\pi}\frac{2\pi^3}{3} = \frac{2\pi^2}{3}, \end{aligned}$$

and for $k \geq 1$

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_0^{2\pi}(x-\pi)^2\cos kx dx = \frac{1}{\pi}\int_{-\pi}^{\pi} y^2\cos k(y+\pi)\,dx \\ &= (-1)^k\frac{1}{\pi}\int_{-\pi}^{\pi} y^2\cos ky dx. \end{aligned}$$

The latter integral is evaluated by twice integrating by parts:

$$\begin{aligned} \int_{-\pi}^{\pi} y^2\cos ky dx &= \frac{1}{k}\int_{-\pi}^{\pi} y^2 d\sin ky = \frac{1}{k}\left[y^2\sin ky\right]_{-\pi}^{\pi} - \frac{2}{k}\int_{-\pi}^{\pi} y\sin ky dy \\ &= 0 + \frac{2}{k^2}\int_{-\pi}^{\pi} y d\cos ky = \frac{2}{k^2}\left[y\cos ky\right]_{-\pi}^{\pi} - \frac{2}{k^2}\int_{-\pi}^{\pi}\cos ky dy \\ &= \frac{4\pi}{k^2}(-1)^k + 0, \end{aligned}$$

whence

$$a_k = \frac{4}{k^2}.$$

Finally,

$$b_k = \frac{1}{\pi}\int_0^{2\pi}(x-\pi)^2\sin kx dx = (-1)^k\frac{1}{\pi}\int_{-\pi}^{\pi} y^2\sin ky dx = 0$$

because the function $y^2\sin ky$ is odd (obviously, the integral of an odd function over a symmetric interval $[-a,a]$ is zero). Hence, we obtain the Fourier series

$$(x-\pi)^2 \sim \frac{\pi^2}{3} + 4\sum_{k=1}^{\infty}\frac{\cos kx}{k^2}. \tag{3.51}$$

By Theorem 3.10, we have here equality for all x and, moreover, the convergence is uniform (the pointwise convergence holds also by Theorem 3.9).

For example, setting $x = 0$ we obtain a remarkable identity

$$\frac{\pi^2}{6} = \sum_{k=1}^{\infty}\frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \dots .$$

Setting $x = \pi$, we obtain

$$\frac{\pi^2}{12} = \sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k^2} = 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \dots .$$

Proof of Theorem 3.10. Let S_n be as before the n-th partial sum of the complex Fourier series of f. It suffices to prove the following two statements:

1. $S_n(x) \to f(x)$ pointwise in $\mathbb{R}$
2. The sequence $\{S_n(x)\}$ convergences uniformly in $\mathbb{R}$.

Then S_n converges uniformly to its pointwise limit f.

By Theorem 3.9, to prove the first statement, it suffices to prove that f is right and left differentiable at any point $x \in [0, 2\pi]$. If $x \notin A$ then f is differentiable at x. Let

$x \in A$. Then, for y close enough to x, the interval (x, y) contains no points from A so that f is differentiable in (x, y). Since f is continuous, in particular, on $[x, y]$, we can apply the Lagrange mean value theorem, which say that there is a point $\xi \in (x, y)$ such that

$$f'(\xi) = \frac{f(y) - f(x)}{y - x}.$$

If $y \to x$ from the left or from the right then also $\xi \to x$ from the same side and, hence, $f'(\xi)$ has the limit. Therefore, also quotient $\frac{f(y)-f(x)}{y-x}$ has the limit, which means that f is right and left differentiable.

The proof of the second statement, that is, of the fact that the Fourier series of f converges uniformly, will follow the same approach as the proof of Theorem 3.8. For that we need to consider the Fourier coefficients of the derivative f' while f' is not yet defined in A. So, we are going to extend f' to A, and to do so, we need the following simple property of the integrable functions.

Claim 1 *If a function g is Riemann integrable on some interval $[a, b]$ and a function h is equal to g everywhere except for a single point, then h is also Riemann integrable in $[a, b]$ and*

$$\int_a^b g dx = \int_a^b h dx.$$

Consequently, the same is true if h is equal to g everywhere except for a finite set.

Indeed, renaming $h - g$ by h it suffices to consider the case when $g \equiv 0$. Then $h(x) = 0$ for all $x \in [a, b]$ except for some $x = c$, and the integrability of h follows from Exercise 21. Alternatively, we can argue as follows. For any partition $p = \{x_k\}$ of $[a, b]$ and any tags $\xi = \{\xi_k\}$ associated with p, the Riemann sum $S(f, p, \xi)$ is either 0 when $c \notin \xi$ or it is

$$S(f, p, \xi) = h(c)(x_{k-1} - x_k)$$

for some interval $[x_{k-1}, x_k]$ of the partition. Since $h(c)(x_{k-1} - x_k) \to 0$ as $m(p) \to 0$, we obtain that h is integrable and $\int_a^b h(x) dx = 0$.

Let the set A ordered in the increasing order be $\{a_l\}_{l=0}^m$, and let $a_0 = 0$ and $a_m = 2\pi$ (clearly, we can add to A any point). Extend the derivative f', which is initially defined outside A, to any point a_l *arbitrarily*, so that the function $f'(x)$ is now defined on $[0, 2\pi]$. Let us show that this function is integrable on $[0, 2\pi]$.

Indeed, within any interval (a_{l-1}, a_l) function $f'(x)$ is continuous and has limits at the endpoints of the interval. Obviously, f' can be extended to $[a_{l-1}, a_l]$ as a continuous function, and the continuous extension of f' is Riemann integrable in $[a_{l-1}, a_l]$. The global extension of f' defined above, may be different from the continuous extension only at the endpoints a_{l-1}, a_l. Therefore, it is also integrable on $[a_{l-1}, a_l]$ by Claim 1. By Exercise 18, function f' is integrable in the whole interval $[0, 2\pi]$. In particular, the Fourier coefficients of f' are defined.

Claim 2. *For all $k \in \mathbb{Z}$*

$$c_k(f') = ikc_k(f). \tag{3.52}$$

Recall that this identity was used in the proof of Theorem 3.8 when f' was continuous on $[0, 2\pi]$. To prove it in the present setting, we split the domain of integration as follows:

$$2\pi c_k(f') = \int_0^{2\pi} f'(x) e^{-ikx} dx = \sum_{l=1}^m \int_{a_{l-1}}^{a_l} f'(x) e^{-ikx} dx.$$

Since in any interval $[a_{l-1}, a_l]$ function f' can be replaced by its continuous modification, we obtain by the the integration by parts formula:

$$\int_{a_{l-1}}^{a_l} f'(x) e^{-ikx} dx = \int_{a_{l-1}}^{a_l} e^{-ikx} df(x) = \left[e^{-ikx} f(x)\right]_{a_{l-1}}^{a_l} + ik \int_{a_{l-1}}^{a_l} f(x) e^{-ikx} dx.$$

Therefore, summing up these identities and using the 2π-periodicity of $e^{-ikx} f(x)$, we obtain

$$\begin{aligned} 2\pi c_k(f') &= \sum_{l=1}^{m} \left[e^{-ikx} f(x)\right]_{a_{l-1}}^{a_l} + ik \sum_{l=1}^{m} \int_{a_{l-1}}^{a_l} f(x) e^{-ikx} dx \\ &= \left[e^{-ikx} f(x)\right]_0^{2\pi} + ik \int_0^{2\pi} f(x) e^{ikx} dx = 0 + ik 2\pi c_k(f), \end{aligned}$$

whence (3.52) follows.

Now we can finish the proof as in Theorem 3.8. The identity (3.52) together with the Bessel inequality yields

$$\begin{aligned} \sum_{k\neq 0} |c_k(f)| &\leq \sum_{k\neq 0} \left(k^2 |c_k(f)|^2 + \frac{1}{k^2}\right) = \sum_{k\neq 0} |c_k(f')|^2 + \sum_{k\neq 0} \frac{1}{k^2} \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} |f'(x)|^2 dx + \sum_{k\neq 0} \frac{1}{k^2} < \infty. \end{aligned}$$

Then by the Weierstrass test the series $\sum_{k\in\mathbb{Z}} c_k e^{ikx}$ converges absolutely and uniformly. ■

3.5.6 Parseval's identity

Theorem 3.11 (Parseval's identity) *Let f be an integrable function on $[0, 2\pi]$. If the Fourier series of f converges to f uniformly on $[0, 2\pi]$ then*

$$\sum_{k\in\mathbb{Z}} |c_k(f)|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 dx.$$

Remark. In fact, the Parseval identity holds for any integrable function f on $[0, 2\pi]$, without the assumption that the Fourier series converges to f uniformly, but the proof is more complicated and will not be presented here. Theorem 3.11 applies to any function f satisfying the hypotheses of Theorem 3.10.

Proof. Denote again by S_n the partial sum of the Fourier series, that is,

$$S_n(x) = \sum_{|k|\leq n} c_k e^{ikx},$$

and recall that by Claim 1 from the proof of Theorem 3.7,

$$\sum_{|k|\leq n} |c_k|^2 = \frac{1}{2\pi} \int_0^{2\pi} |S_n(x)|^2 dx.$$

Since $S_n \rightrightarrows f$, it follows that also $|S_n|^2 \rightrightarrows |f|^2$ because

$$\left\| |S_n|^2 - |f|^2 \right\| \leq \| |S_n| - |f| \| \, \| |S_n| + |f| \| \leq \|S_n - f\| \left(\|S_n\| + \|f\| \right) \to 0 \text{ as } n \to \infty$$

(see Exercise 31 for the properties of the norm used above). By Theorem 3.3, we obtain

$$\int_0^{2\pi} |S_n(x)|^2 \, dx \to \int_0^{2\pi} |f(x)|^2 \, dx \text{ as } n \to \infty.$$

Therefore,

$$\sum_{k \in \mathbb{Z}} |c_k|^2 = \lim_{n \to \infty} \sum_{|k| \leq n} |c_k|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 \, dx,$$

which was to be proved. ■

Remark. If function f is real-value then as we have shown in the previous lectures, the relation between the complex and real Fourier coefficients is as follows:

$$2 \sum_{k \in \mathbb{Z}} |c_k|^2 = \frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2 \right).$$

In this case, the Parseval identity takes the form

$$\frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2 \right) = \frac{1}{\pi} \int_0^{2\pi} f^2(x) \, dx. \tag{3.53}$$

Example. For the function $f(x) = (x - \pi)^2$ on $[0, 2\pi]$ extended 2π-periodically, we have the expansion into Fourier series (3.51), that is,

$$(x - \pi)^2 = \frac{\pi^2}{3} + 4 \sum_{k=1}^{\infty} \frac{\cos kx}{k^2}.$$

Substituting the values of a_k and b_k into (3.47) and noticing that

$$\int_0^{2\pi} f(x)^2 \, dx = \int_{-\pi}^{\pi} y^4 dy = \frac{2}{5} \pi^5,$$

we obtain

$$\frac{2}{9} \pi^4 + \sum_{k=1}^{\infty} \frac{16}{k^4} = \frac{2}{5} \pi^4,$$

whence

$$\sum_{k=1}^{\infty} \frac{1}{k^4} = \frac{\pi^4}{90}.$$

In conclusion let us mention yet another type of convergence, which is called the convergence in *quadratic mean* and which means the following: $f_k \to f$ in quadratic mean on $[a, b]$ if

$$\int_a^b |f_k - f|^2 \, dx \to 0 \quad \text{as } k \to \infty.$$

Of course, the uniform convergence implies the convergence in quadratic mean, but not vice versa. The following theorem is presented here without proof.

Theorem. *For any integrable function f on $[0, 2\pi]$, its Fourier series converges to f in quadratic mean on $[0, 2\pi]$.*

4 Metric spaces

Our aim here is to develop necessary tools for analysis of functions of several variables. The crucial role in analysis of functions of a single variable was played by the notion of the distance between two reals x, y, that is $|x-y|$. An analogous notion will be developed here in an abstract context.

4.1 Notion of a distance function

Definition. Let X be an arbitrary set. A *distance function* (or a *metric*) on X is a function $d(x,y)$ of two variables $x, y \in X$ such that the following properties are satisfied:

1. Positivity: $d(x,y)$ is a non-negative real, and $d(x,y)=0$ if and only if $x=y$ (hence, $d(x,y)>0$ if $x \neq y$).
2. Symmetry: $d(x,y)=d(y,x)$ for all $x, y \in X$.
3. The triangle inequality: $d(x,y) \leq d(x,z)+d(y,z)$ for all $x, y, z \in X$.

If $d(x,y)$ is a metric on X then the couple (X,d) is called a *metric space*.

Example. 1. Let $X=\mathbb{R}$. Then $d(x,y)=|x-y|$ is a distance function on $\mathbb{R}$.

2. Let X be an arbitrary set and define $d(x,y)=1$ if $x \neq y$ and $d(x,y)=0$ if $x=y$. Then $d(x,y)$ is a distance function. This particular $d(x,y)$ is called the *discrete metric* on X.

Our main example of a metric space will be the set $\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times ... \times \mathbb{R}}_{n \text{ times}}$ that consists of all n-tuples $(x_1, ..., x_n)$ – sequences of n reals. The elements of $\mathbb{R}^n$ are also referred to as *vectors*.

If $x=(x_1, x_2, ..., x_n) \in \mathbb{R}^n$ then we refer to x_k as the components (or the coordinates) of the vector x. For any two vectors $x, y \in \mathbb{R}^n$ their sum is defined as follows:

$$x+y=(x_1+y_1, x_2+y_2, ..., x_n+y_n).$$

For any real λ and any vector x, the product λx is defined by

$$\lambda x=(\lambda x_1, \lambda x_2, ..., \lambda x_n).$$

The set $\mathbb{R}^n$ with these two operation becomes a vector space. Recall the definition of a vector space with the scalar field $\mathbb{R}$. Real numbers will also be called *scalars*.

Definition. Let V be a set where two operations are defined: addition $x+y \in V$ for all $x, y \in V$ and multiplication by scalar $\lambda x \in V$ for all scalars $\lambda \in \mathbb{R}$ and vectors $x \in V$. The set V with these operations is called a *vector space* (or a *linear space*) if the following properties are satisfied:

1. Neutral element: there exists $0 \in V$ such that $x+0=0+x=x$ for all $x \in V$.

2. Negative element: for any $x \in V$ there is a vector denoted by $-x \in V$ such that $x + (-x) = (-x) + x = 0$.

3. Associative law for addition: $(x + y) + z = x + (y + z)$

4. Commutative law for addition: $x + y = y + x$

5. Neutral element for scalar multiplication: $1x = x$.

6. Associative law for scalar multiplication: $(\lambda\mu) x = \lambda (\mu x)$.

7. Distributive law for addition of scalars: $(\lambda + \mu) x = \lambda x + \mu x$.

8. Distributive law for addition of vectors: $\lambda (x + y) = \lambda x + \lambda y$.

Example. 1. It is straightforward to check that $\mathbb{R}^n$ with the above operations is a vector space. Note that the neutral element is the zero vector $0 = (0, 0, ..., 0)$ and the negative to x is $-x = (-x_1, ..., -x_n)$.

2. Let S be any set and $F(S)$ the set of all real valued function on S. For any two functions $f, g \in F(S)$, addition $f + g$ is defined by

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

and the multiplication by scalar by $(\lambda f)(x) = \lambda f(x)$ where $\lambda \in \mathbb{R}$. Zero element is $f \equiv 0$. Clearly, $F(S)$ is a vector space.

If $S = \{1, 2, ..., n\}$, that is, S consists of n elements, then any function f on S can be identified with the sequence $\{f(1), ..., f(n)\}$ of its values on S. Considering this sequence as an element of $\mathbb{R}^n$, we see that $\mathbb{R}^n$ is identical with the vector space $F(S)$.

A natural way to introduce a distance function on a vector space is to use the notion of a *norm*.

Definition. A function N on a vector space V is called a *norm* if it satisfies the following properties:

1. Positivity: $N(x) \geq 0$ and $N(x) = 0$ if and only if $x = 0$ (hence, $N(x) > 0$ if $x \neq 0$).

2. The scaling property: for any $\lambda \in \mathbb{R}$, $N(\lambda x) = |\lambda| N(x)$

3. The triangle inequality: $N(x + y) \leq N(x) + N(y)$.

A vector space V endowed with a norm N, is called a *normed space*.

Example. 1. Function $N(x) = |x|$ is a norm in $\mathbb{R}$. To resemble $|x|$, the norm in arbitrary vector space is normally denoted by $\|x\|$.

2. Let $V = B(S)$ be the set of all bounded functions on S, which is obviously a vector space. The following is a norm in $B(S)$:

$$\|f\| = \sup_{x \in S} |f(x)|$$

(see Exercise 31). To distinguish this norm from other norms, let us call it the sup-norm and denote by $\|f\|_{\sup}$.

3. Let $V = \mathbb{R}^n$. Considering $\mathbb{R}^n$ as the space $B(S)$ with $S = \{1, 2, ..., n\}$ and using the sup-norm, we obtain the following norm in $\mathbb{R}^n$:

$$\|x\|_{\sup} = \sup_{1 \le k \le n} |x_k| = \max_{1 \le k \le n} |x_k| .$$

Another norm in $\mathbb{R}^n$, which is called the 1-norm, is given by

$$\|x\|_1 = |x_1| + |x_2| + ... + |x_n| .$$

Claim. *If (V, N) is a normed space then $d(x, y) = N(x - y)$ is a distance function on V.*

Proof. Let us check the three properties of a metric:

Positivity: $d(x, y) = N(x - y) \ge 0$ and $N(x - y) = 0$ if and only if $x = y$, follows from the positivity of the norm.

Symmetry:

$$d(x, y) = N(y - x) = N((-1)(x - y)) = N(x - y) = d(y, x),$$

where we have used the scaling property of the norm.

The triangle inequality:

$$d(x, y) = N(x - y) = N((x - z) + (z - y)) \le N(x - z) + N(z - y) = d(x, z) + d(z, y),$$

where we have used the triangle inequality for the norm. ■

Hence, any normed space is a metric space.

Example. Consider examples of a metric defined via a norm.

1. For $V = B(S)$ with the sup-norm

$$d_{\sup}(f, g) = \|f - g\|_{\sup} = \sup_{x \in S} |f(x) - g(x)| .$$

2. For $V = \mathbb{R}^n$ with the sup-norm

$$d_{\sup}(x, y) = \|x - y\|_{\sup} = \max_{1 \le k \le n} \{|x_k - y_k|\}$$

and for $V = \mathbb{R}^n$ with 1-norm:

$$d_1(x, y) = \|x - y\|_1 = \sum_{k=1}^{n} |x_k - y_k| .$$

Theorem 4.1 *For any $1 \le p < \infty$, define the p-norm in $\mathbb{R}^n$ by*

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} .$$

Then $\|\cdot\|_p$ is a norm in $\mathbb{R}^n$.

Proof. If $p = 1$ then we already know that $\|\|_1$ is a norm. Let us assume in the sequel that $p > 1$.

Positivity. Obviously, $\|x\|_p \geq 0$ and the equality takes places if and only if all $x_k = 0$ that is, when $x = 0$.

The scaling property:

$$\|\lambda x\|_p = \left(\sum_{k=1}^{n} |\lambda|^p |x_p|^p\right)^{1/p} = |\lambda| \, \|x\|.$$

The triangle inequality is much more involved and will be proved after some preparation.

Claim 1. (The Hölder inequality) *For any reals $p, q > 1$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{4.1}$$

the following inequality holds for all $x, y \in \mathbb{R}^n$:

$$\|x\|_p \|y\|_q \geq |x_1 y_1| + |x_2 y_2| + ... + |x_n y_n| \, . \tag{4.2}$$

If $x = 0$ or $y = 0$ then (4.2) holds trivially. Assume further that both $x, y \neq 0$. Notice that the inequality (4.2) does not change if we multiply x by a scalar λ, since the both sides multiply by $|\lambda|$. Therefore, multiplying x by $\lambda = \frac{1}{\|x\|_p}$ and renaming λx by x, we can assume that $\|x\|_p = 1$. Similarly, we assume that $\|y\|_q = 1$.

Next, we use the Young inequality which was proved in Analysis I: for all non-negative a, b,

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab.$$

Applying it with $a = |x_k|$ and $b = |y_k|$ and summing up in $k = 1, 2, ..., n$, we obtain

$$\sum_{k=1}^{n} \left(\frac{|x_k|^p}{p} + \frac{|y_k|^q}{q}\right) \geq \sum_{k=1}^{n} |x_k| \, |y_k| \, .$$

Using the hypotheses that $\|x\|_p = \|y\|_q = 1$ and (4.1) we obtain that the left hand side is equal to

$$\frac{1}{p} \sum_{k=1}^{n} |x_k|^p + \frac{1}{q} \sum_{k=1}^{n} |y_k|^q = \frac{1}{p} \|x\|_p^p + \frac{1}{q} \|y\|_q^q = \frac{1}{p} + \frac{1}{q} = 1 = \|x\|_p \|y\|_q,$$

whence (4.2) follows.

For any two vectors $x, y \in \mathbb{R}^n$, set

$$x \cdot y = \sum_{k=1}^{n} x_k y_k.$$

The expression $x \cdot y$ is called the *dot product* (or the *inner product*) of x and y. Note that $x \cdot y$ is a real number. The obvious properties of the dot product are:

1. symmetry: $x \cdot y = y \cdot x$

2. linearity in each argument:
$$(\lambda x) \cdot y = \lambda (x \cdot y)$$
and
$$(x + y) \cdot z = x \cdot z + y \cdot z.$$
Also, it follows from (4.2) that
$$\|x\|_p \|y\|_q \geq x \cdot y \tag{4.3}$$
whenever p and q satisfy (4.1). Positive numbers p, q satisfying (4.1) are called *Hölder conjugate*.

Claim 2. *If p and q are Hölder conjugate then, for any $x \in \mathbb{R}^n$,*
$$\|x\|_p = \sup_{y \in \mathbb{R}^n \setminus \{0\}} \frac{x \cdot y}{\|y\|_q}. \tag{4.4}$$

It follows from (4.3) that
$$\|x\|_p \geq \frac{x \cdot y}{\|y\|_q}$$
whence
$$\|x\|_p \geq \sup_{y \neq 0} \frac{x \cdot y}{\|y\|_q}.$$
Let us prove the opposite inequality. If $x = 0$ then it is obvious. Otherwise, choose y as follows:
$$y_k = x_k |x_k|^{p-2}$$
so that
$$x \cdot y = \sum_{k=1}^{n} |x_k|^p = \|x\|_p^p$$
and, using $q = \frac{p}{p-1}$,
$$\|y\|_q = \left(\sum_{k=1}^{n} |y_k|^q \right)^{1/q} = \left(\sum_{k=1}^{n} |x_k|^{(p-1)\frac{p}{p-1}} \right)^{\frac{p-1}{p}} = \|x\|_p^{p-1}.$$
Therefore, for this y,
$$\frac{x \cdot y}{\|y\|_q} = \frac{\|x\|_p^p}{\|x\|_p^{p-1}} = \|x\|_p.$$
It follows that
$$\sup_{y \neq 0} \frac{x \cdot y}{\|y\|_q} \geq \|x\|_p,$$
which finishes the proof of this claim. This argument implies that sup in (4.4) can be replaced by max.

Now we are ready to prove the triangle inequality. By Claim 2, we have

$$\begin{aligned}\|x+y\|_p &= \sup_{z\neq 0}\frac{(x+y)\cdot z}{\|z\|^q} = \sup_{z\neq 0}\left(\frac{x\cdot z}{\|z\|^q}+\frac{y\cdot z}{\|z\|^q}\right)\\ &\leq \sup_{z\neq 0}\frac{x\cdot z}{\|z\|^q}+\sup_{z\neq 0}\frac{y\cdot z}{\|z\|^q}\\ &= \|x\|_p+\|y\|_p,\end{aligned}$$

which was to be proved. ■

A special role is played by the 2-norm:

$$\|x\|_2 = \left(\sum_{k=1}^{n}|x_k|^2\right)^{1/2},$$

which is obviously related to the dot product as follows:

$$x\cdot x = \|x\|_2^2.$$

The Hölder inequality in this case becomes

$$x\cdot y \leq \|x\|_2\|y\|_2.$$

This particular case of the Hölder inequality is called the *Cauchy-Schwarz* inequality. Let us give an independent proof of the Cauchy-Schwarz inequality. Note that $x\cdot x\geq 0$ for any $x\in\mathbb{R}^n$. In particular, for all $x,y\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}$,

$$(x+\lambda x)\cdot(x+\lambda y)\geq 0.$$

Expanding the left hand side using the linearity of the dot product, we obtain

$$x\cdot x+2\lambda\,(x\cdot y)+\lambda^2\,(y\cdot y)\geq 0.$$

Since this quadratic polynomial of λ is non-negative for all $\lambda\in\mathbb{R}$, its discriminant must be non-positive, that is,

$$(x\cdot y)^2\leq (x\cdot x)\,(y\cdot y) = \|x\|_2^2\|y\|_2^2,$$

whence the claim follows.

Let us show how the sup-norm is related to the p-norm.

Claim. *We have*

$$\|x\|_p \to \|x\|_{\sup} \text{ as } p \to \infty.$$

Proof. Let $m = \|x\|_{\sup}$ so that $|x_k| \leq m$ for all $k = 1, 2, ..., l$ but $|x_k| = m$ for some m. Then

$$\|x\|_p^p = \sum_{k=1}^n |x_k|^p \geq m^p$$

and

$$\|x\|_p^p \leq nm^p$$

whence

$$m \leq \|x\|_p \leq n^{1/p} m.$$

Clearly, $n^{1/p} \to 1$ as $p \to \infty$ whence $\|x\|_p \to m$, which was to be proved. ■

For this reason, the sup-norm is also called the ∞-norm and is denoted by $\|x\|_\infty$ so that

$$\|x\|_\infty = \|x\|_{\sup} = \lim_{p \to +\infty} \|x\|_p.$$

Note that $p = \infty$ and $q = 1$ are Hölder conjugate because $\frac{1}{\infty} + \frac{1}{1} = 1$, and the Hölder inequality extends to this case as follows:

$$\|x\|_\infty \|y\|_1 \geq x \cdot y,$$

because

$$x \cdot y = \sum_{k=1}^n x_k y_k \leq \max |x_k| \sum_{k=1}^n |y_k| = \|x\|_\infty \|y\|_1.$$

Summarizing the above, we can say that $\mathbb{R}^n$ possesses the following family of distance functions: for any $p \in [1, +\infty]$

$$d_p(x, y) = \|x - y\|_p = \begin{cases} \left(\sum_{k=1}^n |x_k - y_k|^p\right)^{1/p}, & p < \infty, \\ \max_{1 \leq k \leq n} |x_k|, & p = \infty. \end{cases}$$

4.2 Metric balls

In any metric space (X, d) one can consider *metric balls* defined as follows.

Definition. For any $x_0 \in X$ and $r > 0$ define the ball $B(x, r)$ of radius r centered at x_0 by

$$B(x_0, r) = \{x \in X : d(x, x_0) < r\}.$$

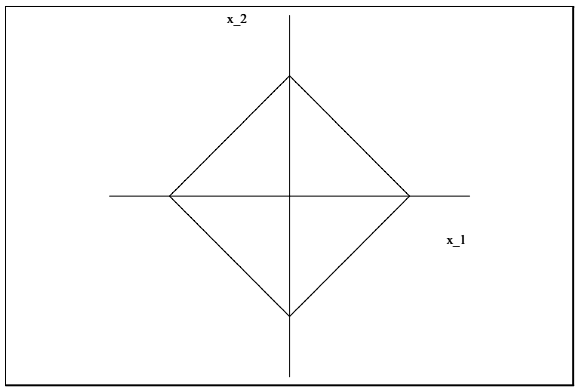
Example. In $\mathbb{R}$ with the distance $d(x, y) = |x - y|$ the ball $B(x_0, r)$ is the symmetric open interval $(x_0 - r, x_0 + r)$.

Example. Consider $\mathbb{R}^2$ with the metric d_p where $1 \leq p \leq \infty$ and describe the metric ball $B(0, r)$ for various p.

If $p = 1$ then

$$B(0,r) = \left\{x \in \mathbb{R}^2 : \|x\|_1 < r\right\} = \left\{(x_1, x_2) \in \mathbb{R}^2 : |x_1| + |x_2| < r\right\}$$

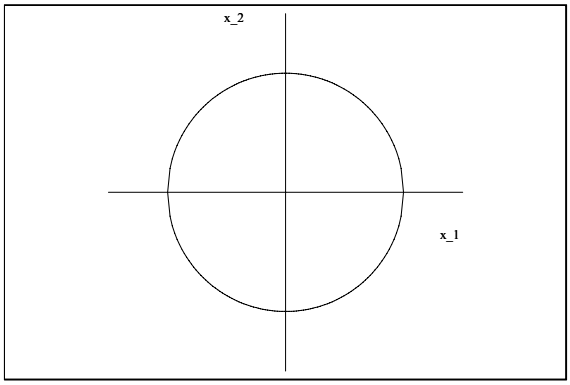
Hence, the metric ball $B(0,r)$ is a rhombus (rotated square) as on the diagram below:



If $p = 2$ then

$$B(0,r) = \left\{x \in \mathbb{R}^2 : \|x\|_2 < r\right\} = \left\{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < r^2\right\}$$

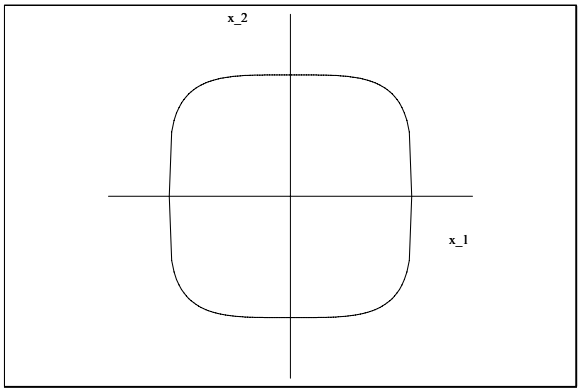
which is the circle:



If $p = 4$ then

$$B(0,r) = \left\{x \in \mathbb{R}^2 : \|x\|_4 < r\right\} = \left\{(x_1, x_2) \in \mathbb{R}^2 : x_1^4 + x_2^4 < r^4\right\},$$

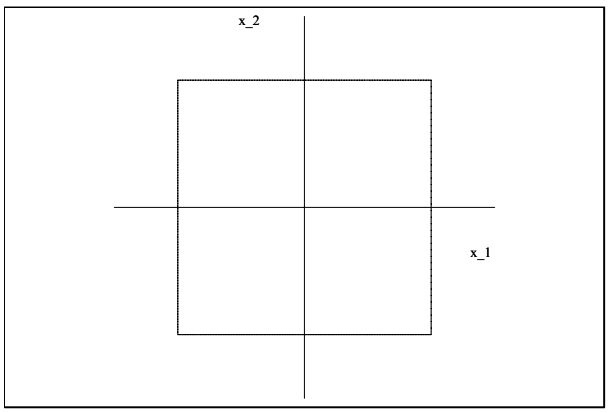
which is shown on the diagram:



If $p = \infty$ then

$$B(0,r) = \left\{x \in \mathbb{R}^2 : \|x\|_\infty < r\right\} = \left\{(x_1, x_2) \in \mathbb{R}^2 : \max\{|x_1|, |x_2|\} < r\right\},$$

which is the square:



Let us prove the following general property of metric balls, which will be frequently used.

Lemma 4.2 *Let $B(x_1, r_1)$ and $B(x_2, r_2)$ be two metric balls in a metric space (X, d).*
(a) If $r_1 + r_2 \leq d(x_1, x_2)$ then the balls are disjoint.
(b) If $r_1 - r_2 \geq d(x_1, x_2)$ then $B(x_1, r_1) \supset B(x_2, r_2)$.

Proof. (a) If x belongs to the both balls then $d(x, x_1) < r_1$ and $d(x, x_2) < r_2$ whence by the triangle inequality

$$d(x_1, x_2) \leq d(x, x_1) + d(x, x_2) < r_1 + r_2$$

which contradicts the assumption.

(b) For any $x \in B(x_2, r_2)$, we have

$$d(x, x_1) \leq d(x, x_2) + d(x_1, x_2) < r_2 + d(x_1, x_2) \leq r_1$$

whence $x \in B(x_1, r_1)$. Hence, $B(x_2, r_2) \subset B(x_1, r_1)$. ■

4.3 Limits and continuity

Definition. Let (X, d) be a metric space. We say that a sequence $\{x_n\}_{n=1}^\infty$ of points in X converges to a point $a \in X$ if $d(x_n, a) \to 0$ as $n \to \infty$. The point a is called the limit of $\{x_n\}$ and we write in this case $\lim_{n\to\infty} x_n = a$ or $x_n \to a$ or $x_n \xrightarrow{d} a$ indicating that the convergence relates to the distance function d.

Equivalently, $x_n \to a$ if, for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies x_n \in B(a, \varepsilon),$$

because the latter is equivalent to $d(x_n, a) < \varepsilon$.

Claim. *Any sequence in a metric space can have at most one limit.*

Proof. If $x_n \to a$ and $x_n \to b$ then $x_n \in B(a,\varepsilon)$ and $x_n \in B(b,\varepsilon)$ for any $\varepsilon > 0$ and for large enough n. But we can choose $\varepsilon = \frac{1}{2}d(a,b)$ so that by Lemma 4.2 the balls $B(a,\varepsilon)$ and $B(b,\varepsilon)$ are disjoint. Hence, x_n cannot belong to both balls. ■

Definition. Let (X, d_X) and (Y, d_Y) be two metric spaces and $f : X \to Y$ be a function (a mapping) from X to Y. We say that $f(x)$ converges to $b \in Y$ as $x \to a \in X$ and write $f(x) \to b$ or

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

if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$d_X(x,a) < \delta,\ x \neq a \implies d_Y(f(x), b) < \varepsilon.$$

Denoting by B_X and B_Y the metric balls in X and Y, respectively, rewrite the above condition in the form

$$x \in B_X(a,\delta) \setminus \{a\} \implies f(x) \in B_Y(b,\varepsilon).$$

Claim. *If the limit of a function exists then it is unique.*

The proof is the same as for sequences.

Definition. Function $f : X \to Y$ is called continuous at $a \in X$ if $\lim_{x \to a} f(x) = f(a)$.

In other words, for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$x \in B_X(a,\delta) \implies f(x) \in B_Y(f(a), \varepsilon). \tag{4.5}$$

Function $f : X \to Y$ is called continuous if it is continuous at all points of X.

Example. Fix a point $x_0 \in X$ and consider a function $f : X \to \mathbb{R}$ defined by $f(x) = d(x, x_0)$. We claim that this function is continuous on X. For that, we need to check that for any $a \in X$ and any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d(x,a) < \delta \implies |d(x,x_0) - d(a,x_0)| < \varepsilon.$$

It follows from the triangle inequality that

$$|d(x,x_0) - d(a,x_0)| \leq d(x,a).$$

Therefore, it suffices to take $\delta = \varepsilon$.

4.4 Open and closed sets

Let (X, d) be a metric space.

Definition. A set $U \subset X$ is called *open* if for any $x \in U$ there exists $r > 0$ such that $B(x,r) \subset U$. A set $F \subset X$ is called *closed* if its complement $X \setminus F$ is open.

For example, the empty set $\emptyset$ and the full set X are open. Hence, their complements, that is, X and $\emptyset$, are closed.

In $\mathbb{R}$ the open intervals are open sets and the closed intervals are closed sets.

Theorem 4.3

(a) The union of any family of open sets is open.

(b) The intersection of a finite family of open sets is open.

(c) The intersection of any family of closed sets is closed.

(d) The union of a finite family of closed sets is closed.

(e) A set $U \subset X$ is open if and only if U is the union of a family of metric balls.

(f) A set $F \subset X$ is closed if and only if any convergence sequence from F has the limit in F.

Proof. (a) Let $\mathcal{F}$ be a family of open sets. To prove that the union $U = \bigcup_{S \in \mathcal{F}} S$ is open let us verify that for any $x \in U$ there is $r > 0$ such that $B(x, r) \subset U$. Indeed, x must belong to some set $S \in \mathcal{F}$. Since S is open, we have $B(x, r) \subset S$ for some $r > 0$. Hence, $B(x, r) \subset U$, which was to be proved.

(b) Let $S_1, S_2, ..., S_n$ be a finite family of open sets and set $U = \bigcap_{k=1}^{n} S_k$. If $x \in U$ then $x \in S_k$ for any k. Then, for any k, there is $r_k > 0$ such that $B(x, r_k) \subset S_k$. Set

$$r = \min(r_1, r_2, ..., r_n) > 0.$$

Then $B(x, r) \subset S_k$ for any k, whence $B(x, r) \subset U$.

Remark. The intersection of infinitely many open sets may be not open. For example, the intersection of all open intervals $\left(-\frac{1}{n}, \frac{1}{n}\right)$ in $\mathbb{R}$, where $n \in \mathbb{N}$, is $\{0\}$ which is not open.

(c) If $\mathcal{F}$ is a family of closed sets then the set

$$\left(\bigcap_{S \in \mathcal{F}} S\right)^c = \bigcup_{S \in \mathcal{F}} S^c$$

is open by part (a). Hence, $\bigcap_{S \in \mathcal{F}} S$ is closed.

(d) If $S_1, ..., S_n$ is a family of closed sets then

$$\left(\bigcup_{k=1}^{n} S_k\right)^c = \bigcap_{k=1}^{n} S_k^c$$

is open by part (b). Hence, $\bigcup_{k=1}^{n} S_k$ is closed.

(e) If U is open then for any $x \in U$ there is $r_x > 0$ such that $B(x, r_x) \subset U$. Clearly, U is the union of all balls $B(x, r_x)$ as x varies in U.

To prove the converse statement, it suffices to prove that any metric ball is open (and then use (a)). Let us show that the metric ball $B(a, r)$ is open for any $a \in X$ and $r > 0$. Indeed, for any $x \in B(a, r)$ set

$$\varepsilon = r - d(a, x) > 0.$$

Then $r - \varepsilon = d(a, x)$, which implies by Lemma 4.2 that $B(x, \varepsilon) \subset B(a, r)$. Hence, $B(a, r)$ is open.

(f) Let F be closed and let $\{x_n\}$ be a sequence in F that converges to $a \in X$. Let us show that $a \in F$. Assuming the contrary, that is, a belongs to the open set F^c, we obtain that there is $r > 0$ such that $B(a,r) \subset F^c$. Since $x_n \in F$, it follows that $x_n \notin B(a,r)$. The latter means that $x_n \not\to a$, which contradicts the assumption.

Assume now that F contains the limits of all its convergence sequences and prove that F is closed. We need to show that F^c is open, that is, for any $a \in F^c$ there is $r > 0$ such that $B(a,r) \subset F^c$. Assume from the contrary that F^c is not open, that is, for some $a \in F^c$, no ball $B(a,r)$ is contained in F^c. This means that, for any $n \in \mathbb{N}$ there is $x_k \in B\left(a, \frac{1}{k}\right)$ such that $x_k \notin F^c$, that is, $x_k \in F$. Hence, we obtain a sequence $\{x_k\}$ of points in F such that $d(x_k, a) < \frac{1}{k}$ whence $x_k \to a$ as $k \to \infty$. By hypothesis, we must have $a \in F$, which contradicts the assumption $a \in F^c$. ■

Remark. One can start with definition of the family of open sets in an axiomatic manner. Namely, consider a set X and a family $\mathcal{O}$ of subsets of X that are called open sets, which satisfies the following axioms:

1. $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$.
2. The union of any family of sets from $\mathcal{O}$ is in $\mathcal{O}$
3. The intersection of any finite family of sets from $\mathcal{O}$ is in $\mathcal{O}$.

Any family $\mathcal{O}$ with such properties is called a *topology* in X, and the couple $(X, \mathcal{O})$ is called a *topological space*. The topology in X can be used to define the notion of convergent sequences, continuous functions, etc.

In this course, we consider only the topology in a metric space, which is defined via the distance function.

Example. As we already know, and metric ball is an open set. This implies that the complement of any ball is a closed set. Consider the notion of a *closed* ball

$$\overline{B}(x_0, r) = \{x \in X : d(x_0, x) \leq r\}$$

and prove that a closed ball is a closed set. It suffices to prove that the complement of the ball

$$C = \{x \in X : d(x_0, x) > r\}$$

is open. Let $x \in C$ and let us show that C contains a ball $B(x, \varepsilon)$ for some $\varepsilon > 0$. Indeed, just take

$$\varepsilon = d(x_0, x) - r > 0$$

so that

$$r + \varepsilon = d(x_0, x),$$

which implies similarly to Lemma 4.2 that the balls $\overline{B}(x_0, r)$ and $B(x, \varepsilon)$ are disjoint. It follows that $B(x, \varepsilon) \subset C$, which proves the openness of C.

As a consequence, we obtain that a closed ball contains limits of all convergent sequences from this ball.

Theorem 4.4 *Let X, Y be metric spaces and $f : X \to Y$ be a mapping.*

(a) The mapping f is continuous if and only if $f^{-1}(U)$ is open in X for any open subset $U \subset Y$.

(*b*) *The mapping f is continuous if and only if $f^{-1}(F)$ is closed in X for any closed subset $F \subset Y$.*

(*c*) *The mapping f is continuous at $a \in X$ if $f(x_n) \to f(a)$ for any sequence $\{x_n\} \subset X$ such that $x_n \to a$.*

Proof. (*a*) Assume that f is continuous and prove that the inverse image $f^{-1}(U)$ is open in X for any open set $U \subset Y$. Take $a \in f^{-1}(U)$. Then $f(a) \in U$ and by the openness of U there exists $\varepsilon > 0$ such that $B_Y(f(a), \varepsilon) \subset U$. By the continuity of f at a, there is $\delta > 0$ such that

$$x \in B_X(a, \delta) \implies f(x) \in B_Y(f(a), \varepsilon). \tag{4.6}$$

This implies that $f(B_X(a, \delta)) \subset U$ whence $B_X(a, \delta) \subset f^{-1}(U)$, which proves the openness of $f^{-1}(U)$.

Assume that $f^{-1}(U)$ is open for any open set $U \subset X$ and prove that f is continuous. To prove the continuity of f at $a \in X$, we need to show that, for any $\varepsilon > 0$ there is $\delta > 0$ such that (4.6) holds. Consider the ball $B_Y(f(a), \varepsilon)$, which is an open subset of Y. Hence, its inverse image is open in X. Since a belongs to the inverse image, there is $\delta > 0$ that also $B_X(a, \delta)$ is contained in the inverse image, which implies that

$$f(B_X(a, \delta)) \subset B_Y(f(a), \varepsilon),$$

which is equivalent to (4.6).

(*b*) Let $f^{-1}(F)$ be closed in X for any closed set $F \subset Y$. Then $f^{-1}(U)$ is open for any open $U \subset Y$, because U^c is closed, $f^{-1}(U^c)$ is closed, and

$$f^{-1}(U)^c = f^{-1}(U^c).$$

The converse statement is proved similarly and, hence, the result follows by (*a*).

(*c*) Assume that f is continuous at a and prove that $f(x_n) \to f(a)$ for any sequence $x_n \to a$. By the continuity, for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$x \in B_X(a, \delta) \Longrightarrow f(x) \in B_Y(f(a), \varepsilon).$$

The hypothesis $x_n \to a$ implies that, for any $\delta > 0$ there is $N \in \mathbb{N}$ such that

$$n \geq \mathbb{N} \implies x_n \in B_X(a, \delta).$$

Combining the two statements, we see that for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that

$$n \geq \mathbb{N} \implies f(x_n) \in B_Y(f(a), \varepsilon),$$

which means that $f(x_n) \to f(a)$.

Conversely, assume that $f(x_n) \to f(a)$ for any sequence $x_n \to a$ and prove that f is continuous at a. Assume the contrary that f is not continuous at a, that is, there exists $\varepsilon > 0$ such that for any $\delta > 0$ there is $x \in B_X(a, \delta)$ such that

$$f(x) \notin B_Y(f(a), \varepsilon).$$

Applying this with $\delta = \frac{1}{k}$ where $k \in \mathbb{N}$, we find $x_k \in B_X\left(a, \frac{1}{k}\right)$ such that

$$f(x_k) \notin B_Y(f(a), \varepsilon)$$

Therefore, $x_k \to a$ while $f(x_k) \not\to f(a)$, which contradicts the hypothesis. ■

Corollary. *If $f : X \to Y$ and $g : Y \to Z$ are two continuous mappings of metric spaces then the composite mapping $g \circ f : X \to Z$ is also continuous.*

Proof. Indeed, for any open set $U \subset Z$, we have

$$(g \circ f)^{-1}(U) = f^{-1}\left(g^{-1}(U)\right).$$

Since $g^{-1}(U)$ is open in Y and, hence, $f^{-1}(g^{-1}(U))$ is open in X, we conclude that $(g \circ f)^{-1}(U)$ is open in X. Therefore, $g \circ f$ is continuous by Theorem 4.4. ■

Corollary. *If $f, g : X \to \mathbb{R}$ are continuous mappings from a metric space X to $\mathbb{R}$ then $f + g$, fg, f/g are also continuous (in the case f/g assume that $g \neq 0$).*

Proof. Let us, for example, prove that $f+g$ is continuous at any point $a \in X$. Indeed, for any sequence $x_n \to a$ we have by Theorem 4.4 $f(x_n) \to f(a)$ and $g(x_n) \to g(a)$. Therefore, by the properties of convergent sequences, $f(x_n) + g(x_n) \to f(a) + g(a)$ which implies again by Theorem 4.4 that $f + g$ is continuous at a.

■

Consider now $\mathbb{R}^n$. As we know there are many choices of distance functions in $\mathbb{R}^n$ even among those induced by a norm.

Definition. Let N' and N'' be two norms in $\mathbb{R}^n$. We say that N' and N'' are *equivalent* if there are positive constants C_1, C_2 such that, for all $x \in \mathbb{R}^n$,

$$C_2 N'(x) \leq N''(x) \leq C_1 N'(x). \tag{4.7}$$

It is easy to verify that the equivalence of norms in an equivalence relation.

Claim. *Let the two norms N' and N'' be equivalent and let d' and d'' be the distance functions of the norms N' and N'' respectively. Then:*

1. *Convergence in d' and in d'' is the same.*
2. *Continuity in d' and in d'' is the same.*
3. *Open sets in d' and in d'' are the same.*

Proof. 1. If $x_k \xrightarrow{d'} a$ then $d'(x_k, a) \to 0$ as $k \to \infty$, that is, $N'(x_k - a) \to 0$. By (4.7), we have also $N''(x_k - a) \to 0$ whence $d''(x_k, a) \to 0$ and $x_k \xrightarrow{d''} a$.

2. The continuity amounts to convergence of sequences by Theorem 4.4.

3. The closed sets are characterized in terms of convergent sequences by Theorem 4.3 and, hence, are the same, which implies the identity of the open sets. ■

Claim. *All p-norms in $\mathbb{R}^n$ with $p \in [1, +\infty]$ are equivalent.*

In the sequel, we always assume that the topology of $\mathbb{R}^n$ is defined using a p-norm.

Proof. It suffices to prove that any p-norm is equivalent to ∞-norm (=sup-norm). Indeed, for any $1 \leq p < \infty$, we have

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \geq \left(\max\{|x_k|\}^p\right)^{1/p} = \|x\|_\infty$$

and

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \leq \left(n \max\{|x_k|\}^p\right)^{1/p} = n^{1/p}\|x\|_\infty$$

whence

$$\|x\|_\infty \leq \|x\|_p \leq n^{1/p}\|x\|_\infty,$$

which was to be proved. ■

4.5 Complete spaces

Let (X, d) be any metric space. A sequence $\{x_n\} \subset X$ is called Cauchy if

$$d(x_n, x_m) \to 0 \text{ as } n, m \to \infty.$$

If $x_n \to a$ then

$$d(x_n, x_m) \leq d(x_n, a) + d(x_m, a) \to 0 \text{ as } n, m \to \infty,$$

which means that any convergent sequence is Cauchy. The converse is true in $\mathbb{R}$ but not in general.

Example. Let $X = (0, +\infty)$ be the half-line with the distance function $d(x, y) = |x - y|$. Then (X, d) is a metric space. Consider a sequence $x_n = \frac{1}{n}$, which is obviously Cauchy. However, this sequence does not converge in X (although it converges to 0 in a larger space $\mathbb{R}$).

Definition. A metric space (X, d) is called *complete* (*Vollständing*) if any Cauchy sequence is convergent. Otherwise, the metric space is called incomplete.

The above example shows that the open half-line is incomplete.

Theorem 4.5 *Let S be an arbitrary set and let $X = B(S)$ be the vector space of all bounded functions on S endowed with the* sup*-norm. Then X is a complete metric space.*

Proof. Recall that the sup-norm is defined by

$$\|f\| = \sup_S |f|.$$

Let $\{f_n\} \subset X$ be a Cauchy sequence, that is,

$$\|f_n - f_m\| \to 0 \text{ as } n, m \to \infty,$$

and let us show that the sequence $\{f_n\}$ converges. For any $x \in S$, we have

$$|f_n(x) - f_m(x)| \leq \sup |f_n - f_m| = \|f_n - f_m\| \to 0$$

which means that the numerical sequence $\{f_n(x)\}$ is Cauchy, for any fixed x. By Theorem 2.4 from Analysis I, the sequence $\{f_n(x)\}$ converges. Denote its limit by $f(x)$ so that

$$f_n(x) \to f(x) \text{ as } n \to \infty, \text{ for any } x \in S.$$

We have obtained a function $f(x)$ on S. Let us show that f is bounded and that $f_n \to f$ in the sup-norm. By hypothesis, we have that, for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$n, m \geq N \implies \sup |f_n - f_m| < \varepsilon.$$

It follows that, for any $x \in S$,

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Passing to the limit as $m \to \infty$, we obtain

$$|f_n(x) - f(x)| \leq \varepsilon$$

whence

$$\sup |f_n - f| \leq \varepsilon.$$

It follows that f is bounded and that $f_n \to f$ in the sup-norm as $n \to \infty$. ■

Corollary. *The space $\mathbb{R}^n$ is complete with respect to any p-norm.*

Proof. Indeed, we have $\mathbb{R}^n = B(S)$ with $S = \{1, 2, ..., n\}$. By Theorem 4.5, $\mathbb{R}^n$ is complete with respect to the sup-norm, that is, ∞-norm. Then $\mathbb{R}^n$ is complete with respect to the p-norm because the two norms are equivalent. ■

Consider a metric space (X, d) and a mapping $f : X \to X$. We say that a point $a \in X$ is a *fixed point* (*Fixpunkt*) if $f(a) = a$. Many problems in Analysis amount to obtaining a fixed point of a certain function or mapping.

Example. Let us show that any continuous mapping $f : [0, 1] \to [0, 1]$ has a fixed point. Indeed, the function $f(x) - x$ is non-negative at $x = 0$ and non-positive at $x = 1$, which implies by the intermediate value theorem that it vanishes at some point x, which is hence a fixed point. On the other hand, the mapping $f(x) = x + 1$ on $\mathbb{R}$ has obviously no fixed point.

The next theorem ensures the existence of a fixed point under certain assumptions of the space and the mapping. We say that a mapping $f : X \to X$ is a *contraction mapping* if there is $0 < q < 1$ such that

$$d(f(x), f(y)) \leq q d(x, y)$$

for all $x, y \in X$.

Theorem 4.6 (The Banach fixed point theorem) *Let (X, d) be a complete metric space. Then any contraction mapping $f : X \to X$ has exactly one fixed point.*

Proof. Choose an arbitrary point $x \in X$ and define a sequence $\{x_n\}_{n=0}^{\infty}$ by induction using

$$x_0 = x \text{ and } x_{n+1} = f(x_n) \text{ for any } n \geq 0.$$

Our purpose will be to show that the sequence $\{x_n\}$ converges and that the limit is a fixed point of f. We start with the observation that

$$d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq q d(x_n, x_{n-1}).$$

It follows by induction that

$$d(x_{n+1}, x_n) \leq q^n d(x_1, x_0).$$

Claim. *If $\{x_n\}$ is a sequence of points in a metric space such that, for some $C > 0$ and $q \in (0, 1)$,*

$$d(x_{n+1}, x_n) \leq C q^n \text{ for all } n,$$

then $\{x_n\}$ is Cauchy.

Indeed, for any $m > n$, we obtain using the triangle inequality

$$\begin{aligned} d(x_m, x_n) &\leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + ... + d(x_{m-1}, x_m) \\ &\leq C\left(q^n + q^{n+1} + ... + q^{m-1}\right) \\ &\leq \frac{Cq^n}{1-q}. \end{aligned}$$

Therefore, $d(x_m, x_n) \to 0$ as $n, m \to \infty$, that is, the sequence $\{x_n\}$ is Cauchy.

Applying this Claim and the hypothesis that the space (X, d) is complete, we conclude that $\{x_n\}$ converges, say to a. Then

$$d(f(x_n), f(a)) \leq qd(x_n, a) \to 0$$

so that $f(x_n) \to f(a)$. On the other hand, $f(x_n) = x_{n+1} \to a$ as $n \to \infty$, whence it follows that $f(a) = a$, that is, a is a fixed point.

If b is another fixed point then

$$d(a, b) = d(f(a), f(b)) \leq qd(a, b),$$

which is only possible if $d(a, b) = 0$ and, hence, $a = b$. ■

Example. Fix $c > 0$ and consider a function $f(x) = \frac{1}{2}\left(x + \frac{c}{x}\right)$, $x > 0$. It is easy to see that a unique fixed point of this function is $x = \sqrt{c}$. It is the contents of Exercise 49 to show that f is, in fact, a contraction mapping in $X = [\sqrt{c}, +\infty)$. The procedure of the proof of Theorem 4.6 applies in this case and gives a sequence $\{x_n\}$ converging to $\sqrt{c}$, which can be used for numerical approximation to $\sqrt{c}$.

Let us give a numerical example. Set $c = 2$ and, hence, $f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right)$. Define $\{x_n\}$ by $x_0 = 1$ and

$$x_{n+1} = f(x_n) = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right).$$

Then $x_1 = f(1) = \frac{3}{2}$, $x_2 = f\left(\frac{3}{2}\right) = \frac{17}{12}$, $x_3 = f\left(\frac{17}{12}\right) = \frac{577}{408}$, $x_4 = f\left(\frac{577}{408}\right) = \frac{665\,857}{470\,832}$ and

$$x_5 = f\left(\frac{665\,857}{470\,832}\right) = \frac{886\,731\,088\,897}{627\,013\,566\,048} \approx 1.41421356237309505,$$

which is already a very good approximation for $\sqrt{2}$ (all the displayed digits are correct).

4.6 Compact spaces

Let (X, d) be a metric space and K be a subset of X. We consider families of open sets that covers K. Namely, a family $\mathcal{F}$ of sets is called an *open cover* ($offene\ Überdeckung$) of K if all members of $\mathcal{F}$ are open sets and the union of these sets contains K, that is

$$\bigcup_{U \in \mathcal{F}} U \supset K.$$

Any subfamily of $\mathcal{F}$ that also covers K is called a *subcover*.

Definition. A subset K of a metric space X is called *compact* if any open cover of K contains a finite subcover.

Any finite set is obviously compact. Also, a closed bounded interval in $\mathbb{R}$ is a compact set, as follows from Theorem 1.10 from Analysis I (later on we will obtain an independent proof of this fact), while $\mathbb{R}$ itself is not compact.

Any non-empty subset $Y \subset X$ of a metric space X can be considered itself as a metric space, with the same distance function d. The metric space (Y, d) is called a *subspace* of (X, d). In general, the properties of a subset $K \subset Y$ may depend on whether K is considered as a part of (Y, d) or (X, d). A certain property of a set K is called *intrinsic* if it does not depend on the choice of the ambient space Y. For example, the property to be open (or closed) is not intrinsic because if we choose $K = Y$ then K is always open in Y whereas K does not have to be open in X.

Claim. *The compactness of a set K is an intrinsic property.*

Proof. We need to prove that if $K \subset Y \subset X$ then K is compact in a metric space (X, d) if and only if it is compact in (Y, d). Assume that K is compact in (Y, d) and let $\mathcal{F}$ be an open cover of K in X. For any set U, which is open in X, the set $U' = U \cap Y$ is open in Y, which simply follows from

$$B_Y(x, r) = B_X(x, r) \cap Y.$$

Therefore, taking intersections of sets $U \in \mathcal{F}$ with Y, we obtain the family $\mathcal{F}'$ that is an open cover of K in Y. By the compactness of K in Y, there is a finite subcover of $\mathcal{F}'$. Taking the corresponding members of $\mathcal{F}$, we obtain a finite subcover of $\mathcal{F}$.

Conversely, let K be a compact in (X, d) and $\mathcal{F}$ be an open cover of K in Y. Any set $U \in \mathcal{F}$ is open and, hence, is a union of balls in Y. Taking the union of the corresponding balls in X, we obtain an open set $\widetilde{U}$ in X such that $U = \widetilde{U} \cap Y$. The union of all sets $\widetilde{U}$ is an open cover $\widetilde{\mathcal{F}}$ of K in X. By the compactness of K in X, there is a finite subcover of $\widetilde{\mathcal{F}}$, which yields the corresponding finite subcover of $\mathcal{F}$. ■

One of the important features of compactness is the following relation to continuous mappings.

Theorem 4.7 *Let (X, d_X) and (Y, d_Y) be two metric spaces and $f : X \to Y$ be a continuous mapping. If K is a compact subset of X then the image $f(K)$ is compact in Y.*

Hence, a continuous image of a compact set is compact.

Proof. Let $\mathcal{F}$ be an open cover of $f(K)$ so that

$$f(K) \subset \bigcup_{U \in \mathcal{F}} U.$$

Taking inverse image of these sets, we obtain that

$$K \subset \bigcup_{u \in \mathcal{F}} f^{-1}(U).$$

The sets $f^{-1}(U)$ are open in X by Theorem 4.4. Hence, the family $\{f^{-1}(U)\}_{U \in \mathcal{F}}$ is an open cover of K in X. By the compactness of K, there is a finite subcover $\{f^{-1}(U_k)\}_{k=1}^{N}$. Then the family $\{U_k\}_{k=1}^{N}$ is a finite subcover of $\mathcal{F}$, which finishes the proof. ■

Definition. A subset K of a metric space X is called *sequentially compact* ($Folgenkompact$) if every sequence in K contains a convergent subsequence with the limit in K.

For example, a bounded closed interval in $\mathbb{R}$ is sequentially compact by the theorem of Bolzano-Weierstrass (Theorem 2.3 from Analysis I), while $\mathbb{R}$ is not sequentially compact.

Definition. A subset K of a metric space X is called *totally bounded* if, for any $\varepsilon > 0$, there is a finite cover of K by balls of radii ε.

The latter means that for any $\varepsilon > 0$ there is a finite sequence $\{x_1, ..., x_l\}$ of points in X such that the union of the balls $B(x_i, \varepsilon)$ contains K. Any sequence $\{x_i\}$ with this property is called an *ε-net* of K in X. One can regard a finite ε-net as an approximation of the set K by a finite set with error $\leq \varepsilon$.

Clearly, any bounded interval in $\mathbb{R}$ is totally bounded, while $\mathbb{R}$ is not totally bounded.

It is easy to see that both sequential compactness and total boundedness are intrinsic properties of K (cf. Exercise 50). We use these notions for the following useful criteria for compactness.

Theorem 4.8 *Let (X, d) be a metric space. Then the following are equivalent:*

1. *X is compact.*
2. *X is sequentially compact.*
3. *X is totally bounded and complete.*

Proof. 1 $\implies$ 2. Let $\{x_n\}$ be a sequence in X and let us show that there is a convergent subsequence. Assume the contrary, that is, no point $x \in X$ is a limit of a subsequence of $\{x_n\}$. This means that for any $x \in X$ there is $\varepsilon_x > 0$ such that the ball $B(x, \varepsilon_x)$ contains only finitely many terms of the sequence $\{x_n\}$ (indeed, if $B(x, \varepsilon)$ contains infinitely many terms for any $\varepsilon > 0$ then setting $\varepsilon = \frac{1}{k}$ and choosing $x_{n_k} \in B\left(x, \frac{1}{k}\right)$ we obtain a subsequence $\{x_{n_k}\}$ that converges to x). The family of balls $\{B(x, \varepsilon_x)\}_{x \in X}$ is an open cover of X, whence it follows that there is a finite family of such balls that covers X. Since each ball contains only finitely many terms of $\{x_n\}$, it follows that the whole sequence $\{x_n\}$ contains only finitely many terms, which contradicts the definition of a sequence.

2 $\implies$ 3. Let us first prove that X is complete. Let $\{x_n\}$ be a Cauchy sequence in X. By the sequential compactness of X, $\{x_n\}$ has a convergent subsequence. Then use the following general fact.

Claim. *If a Cauchy sequence $\{x_n\}$ has a convergent subsequence then $\{x_n\}$ is also convergent.*

Indeed, assume that a subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ converges to $a \in X$ and prove that also $\{x_n\}_{n=1}^{\infty}$ converges to a. By the fact that $\{x_n\}$ is Cauchy, for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that

$$n, m \geq N \implies d(x_n, x_m) < \varepsilon.$$

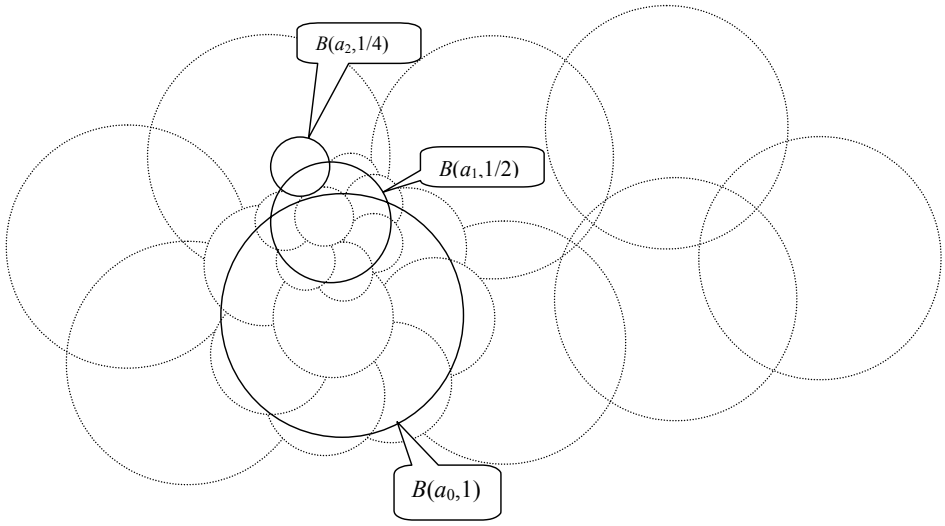
Choose k so big that $n_k \geq N$ and $d(x_{n_k}, a) < \varepsilon$. Then, for any $n \geq N$,

$$d(x_n, a) \leq d(x_n, x_{n_k}) + d(x_{n_k}, a) \leq 2\varepsilon$$

which implies that $x_n \to a$ as $n \to \infty$.

Let us prove that X is totally bounded, that is, for any $\varepsilon > 0$ there is a finite ε-net for X. Assume from the contrary that for some $\varepsilon > 0$ there is no finite ε-net. Let us construct a sequence $\{x_n\}_{n=1}^{\infty} \subset X$ as follows. Choose x_1 arbitrarily. If $x_1, ..., x_{n-1}$ have been already constructed, then choose x_n as follows. The balls $\{B(x_i, \varepsilon)\}_{i=1}^{n-1}$ do not cover all X by hypothesis. Therefore, there is a point in X that does not belong to any of these balls; denote this point by x_n. By construction, we obtain that $d(x_n, x_m) \geq \varepsilon$ for any two distinct indices n, m. This implies that the sequence $\{x_n\}$ is not Cauchy; moreover, any subsequence of it is not Cauchy either. Hence, no subsequence of $\{x_n\}$ converges, which contradicts the sequential compactness of X.

$3 \implies 1$. Let $\mathcal{F}$ be a family of open sets that covers X and let us show that it has a finite subcover. Assume from the contrary that no finite subfamily of $\mathcal{F}$ covers X. Choose a finite 1-net $\{x_1, ..., x_k\}$ and notice that one of the balls $B(x_i, 1)$ admits no finite subcover of $\mathcal{F}$ (indeed, if every ball $B(x_i, 1)$ admits a finite subcover then X does too). Let this ball be $B(a_0, 1)$ (where a_0 is one of the points $x_1, ..., x_k$). Choose a finite $\frac{1}{2}$-net, say $\{y_1, ..., y_l\}$ and notice that at least one of the balls $B\left(y_i, \frac{1}{2}\right)$ that intersect $B(a_0, 1)$ admits no finite subcover of $\mathcal{F}$; let this ball be $B\left(a_1, \frac{1}{2}\right)$.



Continuing this way, we construct a sequence of balls $\{B(a_n, 2^{-n})\}_{n=0}^{\infty}$ such that

(i) any two consecutive balls in this sequence has a non-empty intersection

(ii) none of the balls admits a finite subcover of $\mathcal{F}$.

Indeed, if $B\left(a_{n-1}, 2^{-(n-1)}\right)$ is already constructed then choose a finite ε-net $\{z_1, ..., z_m\}$ with $\varepsilon = 2^{-n}$. Then one of the balls $B(z_i, 2^{-n})$ that intersect $B\left(a_{n-1}, 2^{-(n-1)}\right)$, admits no finite subcover of $\mathcal{F}$; let it be $B(a_n, 2^{-n})$.

The fact that $B\left(a_{n-1}, 2^{-(n-1)}\right)$ and $B(a_n, 2^{-n})$ intersect each other, implies by Lemma 4.2 that

$$d(a_n, a_{n-1}) \leq 2^{-(n-1)} + 2^{-n} < 2^{-(n-2)}.$$

It follows that the sequence $\{a_n\}$ is Cauchy (cf. the Claim in the proof of Theorem 4.6). By the compactness of X, the sequence $\{a_n\}$ converges, say to $a \in X$. The point a belongs to one of the sets from $\mathcal{F}$, say to $U \in \mathcal{F}$. Since U is open, there is $r > 0$ such that $B(a, r) \subset U$. For a large enough n, we have

$$d(a, a_n) + 2^{-n} < r$$

because $d(x, a_n) \to 0$ and $2^{-n} \to 0$ as $n \to \infty$. This implies $B(a_n, 2^{-n}) \subset B(a, r)$ and, hence $B(a_n, 2^{-n}) \subset U$, which contradicts the assumption that $B(a_n, 2^{-n})$ admits no finite subcover of $\mathcal{F}$. ■

Definition. We say that a set K in a metric space is *bounded* if it is contained in a metric ball.

For example, any bounded interval in $\mathbb{R}^n$ is a bounded set. It is easy to prove that in general any totally bounded set is bounded, while the converse may not true (see Exercise 50). Now we can give a simple characterization of compact sets in $\mathbb{R}^n$.

Theorem 4.9 *A subset $K \in \mathbb{R}^n$ is compact if and only K is bounded and closed*

Proof. First note that a compact set K in any metric space X must be closed and bounded. Indeed, if K is not closed then by Theorem 4.3 there is a convergent sequence in K whose limit is outside K. Clearly, this sequence has no convergent subsequence in K so that K is not sequentially compact and, hence, is not compact, by Theorem 4.8. If K is not bounded then K is not totally bounded and, hence, not compact by Theorem 4.8.

Conversely, let now K be a bounded closed subset of $\mathbb{R}^n$ and let us prove that K is compact. Since $\mathbb{R}^n$ is a complete space (by Theorem 4.5) and K is a closed subset of $\mathbb{R}^n$, K is a complete space itself (cf. Exercise 47). In $\mathbb{R}^n$ any bounded set is totally bounded because any ball can be covered by finitely many balls of arbitrarily small radii. This particularly easy to see with the ∞-norm since in this norm a ball has a shape of a box:

$$B(a, r) = \{x \in \mathbb{R}^n : a_k - r < x_k < a_k + r\},$$

and any box can be covered by a finite family by arbitrarily small boxes. By Theorem 4.8, we conclude that K is compact. ■

Corollary. (The maximal/minimal value theorem) *Let f be a continuous function on a closed bounded set $K \subset \mathbb{R}^n$. Then both $\max_K f$ and $\min_K f$ exist.*

Proof. Indeed, since K is a compact set by Theorem 4.9, $f(K)$ is also compact by Theorem 4.7. Hence, $f(K)$ is bounded and closed subset of $\mathbb{R}$. In particular, $\sup_K f$ and $\inf_K f$ are finite. Since $\sup_K f$ and $\inf_K f$ are the limits of some sequences in $f(K)$, they must belong to $f(K)$ by the closedness of this set. Hence, $f(K)$ has both max and min. ■

Corollary. *All norms in $\mathbb{R}^n$ are equivalent* (see Exercise 51).

5 Differential calculus of functions in $\mathbb{R}^n$

We consider functions $f : \Omega \to \mathbb{R}^m$ where Ω is an open subsets of $\mathbb{R}^n$ and define what is means to differentiate such a function.

5.1 Differential and partial derivatives

Let us use the following notation for the components of x and $f(x)$:

$$x = (x_1, ..., x_n)$$

and

$$f(x) = (f_1(x), ..., f_m(x)),$$

where each $f_k(x)$ is a real valued function from Ω to $\mathbb{R}$. Each component $f_k(x)$ can also be written in the form $f_k(x_1, ..., x_n)$ and can be considered as a real valued function of n real variables.

We can use the notion of the derivative of a function of a single variable to obtain derivatives of f_k as follows. Fix some index $j = 1, 2, ..., n$ and consider x_j as a variable while all other x_i with $i \neq j$ are fixed. Then $f_k(x_1, ..., x_n)$ can be considered as a function of x_j alone so that we can take the derivative with respect to x_j:

$$\frac{\partial f_k}{\partial x_j}(x) = \lim_{h \to 0} \frac{f_k(x_1, ...x_j + h, ...x_n) - f(x_1, ..., x_j, ..., x_n)}{h}.$$

This derivative is called a *partial derivative of f of the 1^{st} order*. Note that in the notation $\dfrac{\partial f_k}{\partial x_j}$ one uses a round ∂ instead of a straight d to indicate that f_k has also other arguments apart from x_k. The term "partial" refers to the fact that only one argument varies while the others remain constant.

Sometimes one uses also the notation $\partial_j f_k$ or $D_j f_k$ instead of $\frac{\partial f_k}{\partial x_j}$. The set of all partial derivatives of the 1^{st} order can be arranged as a $m \times n$ matrix

$$\left(\frac{\partial f_k}{\partial x_j}\right) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & ... & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & ... & \frac{\partial f_2}{\partial x_n} \\ ... & ... & ... & ... \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & ... & \frac{\partial f_m}{\partial x_n} \end{pmatrix},$$

where $k = 1, ..., m$ is the index of rows and $j = 1, ..., n$ is the index of columns. This matrix is called the *Jacobian matrix* of f and is denoted by J_f (or $J_f(x)$), so that

$$J_f = \left(\frac{\partial f_k}{\partial x_j}\right).$$

Example. Consider a function $f(x) : \mathbb{R}^2 \to \mathbb{R}^1$ given by

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

It any point $x \neq 0$, this function is obviously differentiable both in x_1 and x_2 and

$$\frac{\partial f}{\partial x_1} = \frac{x_2(x_1^2 + x_2^2) - x_1 x_2 2x_1}{(x_1^2 + x_2^2)^2} = \frac{x_2^3 - x_2 x_1^2}{(x_1^2 + x_2^2)^2}$$

and a similar identity holds for $\frac{\partial f}{\partial x_2}$. Let us show that the partial derivatives exist also at $x = 0$. Indeed, by definition,

$$\frac{\partial f}{\partial x_1}(0) = \lim_{h \to 0} \frac{f(h, 0) - f(0, 0)}{h} = 0$$

because $f(h, 0) = 0$. Similarly, $\frac{\partial f}{\partial x_2}(0) = 0$. Hence, f has partial derivatives of the first order at all points in $\mathbb{R}^2$.

However, the function f is not continuous at 0. Indeed, if $x_1 = x_2$ then $f(x_1, x_2) = \frac{1}{2}$ so that the limit of $f(x)$ at 0 along the line $x_1 = x_2$ is equal to $\frac{1}{2}$, whereas $f(0) = 0$.

So, one needs a better definition of differentiability.

Definition. We say that the function f is differentiable at a point $x \in \Omega$ if there exists a linear mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(x + h) - f(x) = Ah + o(h) \text{ as } h \to 0. \tag{5.1}$$

Let us explain all the entries here. A vector $h \in \mathbb{R}^n$ is an increment of the argument x. Fix some norm $\|\cdot\|$ in $\mathbb{R}^n$ and consider the metric balls with respect to the corresponding distance function. By the openness of Ω, there is a ball $B(x, \varepsilon)$ that is contained in Ω. Therefore, if $\|h\| < \varepsilon$ then $x + h \in \Omega$ and $f(x + h)$ makes sense.

Next, Ah means the action of the linear mapping A at h so that $Ax \in \mathbb{R}^m$. Finally, $o(h)$ stands for any function $\varphi(h)$ with values in $\mathbb{R}^m$ such that

$$\frac{\|\varphi(h)\|}{\|h\|} \to 0 \text{ as } h \to 0.$$

Shortly, the identity (5.1) means that the increment $f(x+h) - f(x)$ of the function f for small h can be split into two terms - the leading linear term Ah and a small term $o(h)$.

The variable h is also called the *differential* of x and is denoted by dx (so that $dx \in \mathbb{R}^n$ is an independent variable). The function $h \mapsto Ah$ is called the *differential* of f at the point x and is also denoted by $df(x)$ so that $df = Adx$. The mapping A is called the (full) *derivative* of f at the point x and is denoted by $\frac{df}{dx}(x)$ or by $f'(x)$. We can rewrite (5.1) in the form

$$f(x+h) - f(x) = f'(x)h + o(h).$$

Recall that the differentiability of a function $f : \mathbb{R} \to \mathbb{R}$ is equivalent to

$$f(x+h) - f(x) = ah + o(h)$$

for some $a \in \mathbb{R}$ and $f'(x) = a$. Obviously, this is fully compatible with the more general definition above since in this case A is a linear mapping from $\mathbb{R}$ to $\mathbb{R}$ that is given by multiplication by a real number a.

Recall that any linear mapping $A : \mathbb{R}^m \to \mathbb{R}^n$ is represented by a $m \times n$ matrix, which will also be denoted by A. Consider h as a column vector in $\mathbb{R}^n$, that is, a $n \times 1$ matrix. Then Ah is the product of $m \times n$ and $n \times 1$ matrices, which results in a $m \times 1$ matrix, that is, is a column vector in $\mathbb{R}^m$, as expected. Hence, $f'(x)$ can be considered as a $m \times n$ matrix.

Example. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping given by $f(x) = Ax$. Then we have

$$f(x+h) - f(x) = Ah$$

so that $f(x)$ is differentiable at any point x and $f'(x) = A$.

Lemma 5.1 *If a function f is differentiable at x then f is continuous at x.*

Proof. Since

$$f(x+h) - f(x) = Ax + o(h)$$

as $h \to 0$ it suffices to prove that the right hand side here goes to 0 as $h \to 0$. The fact that $o(h) \to 0$ is obvious. That $Ah \to 0$ follows from the following claim.

Claim. *For any $m \times n$ matrix A there is a constant C such that*

$$\|Ah\| \leq C\|h\| \text{ for any } h \in \mathbb{R}^n. \tag{5.2}$$

(Exercise 46). ■

Remark. The best constant C such that (5.2) is called the norm of the mapping (or matrix) A and is denoted by $\|A\|$. Equivalently, we have

$$\|A\| = \sup_{h \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ah\|}{\|h\|}$$

(note that the norms in $\mathbb{R}^n$ and $\mathbb{R}^m$, which are used here, are arbitrary but fixed). Then the above Claim means that $\|A\| < \infty$ for any linear mapping.

Lemma 5.2 *Let Ω be an open subset of $\mathbb{R}^n$. If function $f : \Omega \to \mathbb{R}^m$ is differentiable at $x \in \Omega$ then all partial derivatives $\frac{\partial f_k}{\partial x_j}(x)$ exist and*

$$f'(x) = J_f(x).$$

Note that both $f'(x)$ and $J_f(x)$ are $m \times n$ matrices. This lemma implies, in particular, that the derivative $f'(x)$ is uniquely defined.

Proof. Let $f'(x) = A = (a_{kj})$ where k is the index of rows and j is the index of columns. In the identity

$$f(x+h) - f(x) = Ah + o(h) \tag{5.3}$$

set $h = (0, ..., h_j, ...0)$ so that h has the only non-vanishing component h_j. Then

$$Ah = \begin{pmatrix} a_{11} & ... & a_{1j} & ... & a_{1n} \\ ... & ... & ... & ... & ... \\ a_{k1} & ... & a_{kj} & ... & a_{kn} \\ ... & ... & ... & ... & ... \\ a_{m1} & ... & a_{mj} & ... & a_{mn} \end{pmatrix} \begin{pmatrix} 0 \\ ... \\ h_j \\ ... \\ 0 \end{pmatrix} = \begin{pmatrix} a_{1j}h_j \\ ... \\ a_{kj}h_j \\ ... \\ a_{mj}h_j \end{pmatrix}.$$

Taking in the identity (5.3) the k-th component, we obtain

$$f_k(x+h) - f(x) = a_{kj}h_j + o(h_j).$$

Dividing by h_j and passing to the limit as $h_j \to 0$, we obtain

$$\frac{\partial f_k}{\partial x_j}(x) = a_{kj},$$

which was to be proved. ■

Evaluating the partial derivatives is relatively easy since it amounts to familiar differentiation of functions of single variable. Lemma 5.2 suggests that the full derivative $f'(x)$ can be evaluated via the partial derivatives, although one has to know a priori that f is differentiable. Our next purpose is to state a condition of differentiability in terms of partial derivatives.

Theorem 5.3 *Let $\Omega \subset \mathbb{R}^n$ be an open set and assume that a function $f : \Omega \to \mathbb{R}^m$ has all partial derivatives of the first order at all points of Ω. If all partial derivatives $\frac{\partial f_k}{\partial x_j}$ are continuous at a point $x \in \Omega$ then f is differentiable at x.*

Proof. Consider first the case when $m = 1$, that is, f is a real valued function on Ω. For simplicity of notation, assume that the point x where the partial derivatives of f are continuous, is the origin 0. Note that the Jacobian matrix J_f is a $1 \times n$ matrix, that is, a row

$$J_f = (\partial f_1, ..., \partial f_n).$$

Let us prove that $f'(0) = J_f(0)$, that is,

$$f(h) - f(0) = J_f(0)\, h + o(h) \text{ as } h \to 0.$$

We assume that $\|h\|_\infty < \varepsilon$ where ε is chosen so that $B(0, \varepsilon) \subset \Omega$ (where the ball is also taken in the ∞-norm). Consider a sequence of points $\{a_k\}_{k=0}^n$ defined as follows:

$$a_k = (h_1, h_2, ..., h_k, 0, ..., 0).$$

That is, $a_0 = 0$, $a_1 = (h_1, 0, ..., 0)$,..., $a_n = (h_1, ..., h_n) = h$. Clearly, $\|a_k\|_\infty \leq \|h\|_\infty < \varepsilon$ so that all points a_k are contained in Ω. Then we have

$$f(h) - f(0) = f(a_n) - f(a_0) = \sum_{k=1}^{n} (f(a_k) - f(a_{k-1})).$$

To estimate the difference $f(a_k) - f(a_{k-1})$, consider the function

$$g(t) = f(h_1, ..., h_{k-1}, t, 0, ..., 0)$$

so that $g(0) = f(a_{k-1})$ and $g(h_k) = f(a_k)$. Clearly, $g(t)$ is defined for any $t \in [0, h_k]$ and is differentiable in t because

$$g'(t) = \partial_k f(h_1, ..., h_{k-1}, t, 0, ..., 0).$$

Applying the mean value theorem to function g, we obtain that

$$f(a_k) - f(a_{k-1}) = g(h_k) - g(0) = g'(\xi) h_k = \partial_k f(b_k) h_k$$

where $\xi \in (0, h_k)$ and

$$b_k = (h_1, ..., h_{k-1}, \xi, 0, ..., 0).$$

Note that $\|b_k\|_\infty \leq \|h\|_\infty$. Therefore, we have

$$\begin{aligned} f(h) - f(0) &= \sum_{k=1}^{n} \partial_k f(b_k) h_k \\ &= \sum_{k=1}^{n} \partial_k f(0) h_k + \sum_{k=1}^{n} (\partial_k f(b_k) - \partial_k f(0)) h_k \\ &= Jh + o(h), \end{aligned}$$

because $b_k \to 0$ as $h \to 0$ and, hence, $\partial_k f(b_k) - \partial_k f(0) \to 0$ by the continuity of the partial derivatives at 0.

Consider now the general case when m is arbitrary. By the above argument, we obtain the differentiability at x of all components f_k separately so that

$$f_k(x + h) - f_k(x) = A_k h + o(h)$$

where $A_k = f_k'(x)$ is an $1 \times n$ matrix, that is, a row. Combining together all rows A_k into a matrix A, we obtain

$$f(x + h) - f(x) = Ah + o(h),$$

which means that f is differentiable. ■

We say that a function $f : \Omega \to \mathbb{R}^n$ is *continuously differentiable* in Ω if all the partial derivatives of f of the first order exist in Ω are are continuous in Ω.

Corollary. *If f is continuously differentiable in Ω then f is differentiable in Ω (that is, at any point in Ω).*

Proof. Indeed, if f is continuously differentiable in Ω then Theorem 5.3 applies to any point of Ω and gives the differentiability in Ω. ■

5.2 The rules of differentiation

5.2.1 Linearity

Theorem 5.4 (The linearity of differentiation) *If $f, g : \Omega \to \mathbb{R}^n$ are two functions that are differentiable at a point $x \in \Omega$ then also their linear combination $af + bg$ (where $a, b \in \mathbb{R}$) is differentiable at x and*

$$(af + bg)'(x) = af'(x) + bg'(x).$$

Proof. We have

$$f(x+h) - f(x) = f'(x)h + o(h)$$

and

$$g(x+h) - g(x) = g'(x)h + o(h),$$

whence it follows that the function $F = af + bg$ satisfies

$$F(x+h) - F(x) = (af'(x) + bg'(x))h + o(h)$$

whence $F'(x) = af'(x) + bg'(x)$. ■

5.2.2 The chain rule

Theorem 5.5 (The chain rule) *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open sets and let a function $f : U \to V$ be differentiable at $x \in U$ and $g : V \to \mathbb{R}^l$ be differentiable at $y = f(x) \in V$. Then the composite function $g \circ f : U \to \mathbb{R}^l$ is differentiable at x and*

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

Note that

$$\mathbb{R}^n \xrightarrow{f'(x)} \mathbb{R}^m \xrightarrow{g'(y)} \mathbb{R}^l$$

so that the product (composition) $g'(y) f'(x)$ is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^l$, and so is the derivative $(g \circ f)'$.

Proof. We have

$$f(x+a) - f(x) = f'(x)\, a + \varphi(a),$$

where $\varphi(a) = o(a)$ as $a \to 0$ and

$$g(y+b) - g(y) = g'(y)\, b + \psi(b),$$

where $\psi(b) = o(b)$ as $b \to 0$. Set here $y = f(x)$ and $y + b = f(x+a)$, that is,

$$b = f(x+a) - f(x) = f'(x)\, a + \varphi(a), \tag{5.4}$$

we obtain

$$\begin{aligned} g \circ f(x+a) - g \circ f(x) &= g'(y)\, b + \psi(b) \\ &= g'(y)\, f'(x)\, a + g'(y)\, \varphi(a) + \psi(b). \end{aligned} \tag{5.5}$$

We are left to prove that

$$g'(y)\, \varphi(a) + \psi(b) = o(a) \text{ as } a \to 0. \tag{5.6}$$

Using the finiteness of the norm of the matrix $g'(y)$ and we obtain, for some constant C,

$$\|g'(y)\, \varphi(a)\| \le C\|\varphi(a)\| = o(\|a\|)$$

whence

$$g'(y)\, \varphi(a) = o(a) \text{ as } a \to 0.$$

To handle the term $\psi(b)$ in (5.6), note that by (5.4) and the finiteness of the norm of the matrix $f'(x)$, there exists a positive constant C, such that

$$\|b\| \le \|f'(x)\, a\| + \|\varphi(a)\| \le C\|a\|,$$

whence

$$\|\psi(b)\| = o(\|b\|) = o(\|a\|)$$

and $\psi(b) = o(a)$ as $a \to 0$. ■

Corollary. (The chain rule in terms of partial derivatives) *Under the conditions of* Theorem 5.5,

$$\frac{\partial (g \circ f)_k}{\partial x_j}(x) = \sum_{i=1}^{m} \frac{\partial g_k}{\partial y_i}(y)\, \frac{\partial f_i}{\partial x_j}(x),$$

where $y = f(x)$.

Proof. By Lemma 5.2, the partial derivative $\frac{\partial (g \circ f)_k}{\partial x_j}$ is the (k,j)-entry of the matrix $(g \circ f)'$. Since this matrix is the product of the matrices

$$g'(y) = \left(\frac{\partial g_k}{\partial y_i} \right)$$

and

$$f'(x) = \left(\frac{\partial f_i}{\partial x_j} \right),$$

the result follows by the rule of multiplication of the matrices. ■

Corollary. (The derivative of the inverse mapping) *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^n$ be open sets and assume that the function $f : U \to V$ has the inverse $g : V \to U$. If f is differentiable at $x \in U$ and g is differentiable at $y = f(x)$ then*

$$g'(y) = f'(x)^{-1}.$$

Note that both mappings $f'(x)$ and $g'(y)$ are from $\mathbb{R}^n$ to $\mathbb{R}^n$ and, hence, are represented by $n \times n$ matrices. Therefore, to compute $g'(y)$ one needs to evaluate the inverse matrix $f'(x)^{-1}$.

Proof. The composition $g \circ f$ is the identity mapping $I : U \to U$. This mapping is obviously differentiable and $I'(x)$ is the identity mapping $\mathrm{id} : \mathbb{R}^n \to \mathbb{R}^n$. Therefore, by the chain rule,

$$\mathrm{id} = I'(x) = (g \circ f)'(x) = g'(y) f'(x),$$

whence the result follows. ■

Example. Let F be a mapping making the change of the polar coordinates to Cartesian:

$$F(r, \theta) = (x, y) = (r \cos\theta, r \sin\theta),$$

that is, F maps some open set $U \subset \mathbb{R}^2$ (any set where the polar coordinates are defined) to $\mathbb{R}^2$. Its full derivative exists and equal to the Jacobian matrix

$$F' = J_F = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \tag{5.7}$$

because J_F is continuous in (r, θ). If G is the inverse to F, that is, G makes the change of the Cartesian coordinates to polar $G(x, y) = (r, \theta)$ then

$$G'(x) = (F')^{-1} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\frac{1}{r}\sin\theta & \frac{1}{r}\cos\theta \end{pmatrix} = \begin{pmatrix} x/r & y/r \\ -y/r^2 & x/r^2 \end{pmatrix}.$$

Since

$$G'(x) = \begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{pmatrix}$$

we obtain

$$\begin{aligned} \frac{\partial r}{\partial x} &= \frac{x}{r}, \quad \frac{\partial r}{\partial y} = \frac{y}{r} \\ \frac{\partial \theta}{\partial x} &= -\frac{y}{r^2}, \quad \frac{\partial \theta}{\partial y} = \frac{x}{r^2}. \end{aligned} \tag{5.8}$$

Of course, this can also be found directly using that $r = \sqrt{x^2 + y^2}$ and $\tan\theta = y/x$.

5.2.3 Change of variables

Let $f = f(y)$ be a real-valued function of $y \in V$ where V is an open subset of $\mathbb{R}^n$. Let U be another open subset of $\mathbb{R}^n$ and assume that we have a function $y = y(x)$ that maps U to V. We consider $y = y(x)$ as a change of coordinates

$$\begin{cases} y_1 = y_1(x_1, ..., x_n) \\ y_2 = y_2(x_1, ..., x_n) \\ ... \\ y_n = y_n(x_1, ..., x_n) \end{cases}$$

so that $f(y)$ can be considered as a function of x as follows: $f(y(x))$. In fact, this function is just composition of the two mappings $U \xrightarrow{y} V \xrightarrow{f} \mathbb{R}$. By the chain rule, we then have

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial x_j},$$

provide the both functions $f(y)$ and $y(x)$ are differentiable.

Consider again the change of the Cartesian coordinates (x, y) in $\mathbb{R}^2$ to the polar coordinates (r, θ). Using (5.7), we obtain

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \frac{\partial f}{\partial x}\cos\theta + \frac{\partial f}{\partial y}\sin\theta$$

and

$$\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = r\left(-\frac{\partial f}{\partial x}\sin\theta + \frac{\partial f}{\partial y}\cos\theta\right).$$

For the inverse change, we obtain using (5.8),

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial x} = \frac{\partial f}{\partial r}\frac{x}{r} - \frac{\partial f}{\partial \theta}\frac{y}{r^2}$$

and

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial y} = \frac{\partial f}{\partial r}\frac{y}{r} + \frac{\partial f}{\partial \theta}\frac{x}{r^2}.$$

For example, if $f = r^a \sin b\theta$ then

$$\frac{\partial f}{\partial x} = ar^{a-1}\sin b\theta\frac{x}{r} - r^a b\cos b\theta\frac{y}{r^2} = r^{a-1}\left(a\sin b\theta\cos\theta - b\cos b\theta\sin\theta\right).$$

5.3 Mean value theorem

Definition. Let f be a real valued function on an open set $U \subset \mathbb{R}^n$. For any $x \in U$ and any vector $v \in \mathbb{R}^n$, define the *directional derivative* $\frac{\partial f}{\partial v}(x)$ by

$$\frac{\partial f}{\partial v}(x) = \lim_{t\to 0}\frac{f(x+tv) - f(v)}{t} = \left.\frac{df(x+tv)}{dt}\right|_{t=0}.$$

where t is a real valued variable. Alternative notation: $\partial_v f = D_v f = \frac{\partial f}{\partial v}$.

Note that the partial derivatives are particular cases of directional derivatives. Let e_j be the unit vector in the direction of x_j, that is,

$$e_j = (0, ..., 0, 1, 0, .., 0)$$

where the only 1 is at position j. We claim that $\frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial e_j}$. Indeed,

$$\begin{aligned} \frac{\partial f}{\partial e_j} &= \lim_{t \to 0} \frac{f(x + te_j) - f(x)}{t} = \\ &= \lim_{t \to 0} \frac{f(x_1, ..., x_j + t, ..., x_n) - f(x_1, ..., x_n)}{t} = \frac{\partial f}{\partial x_j}. \end{aligned}$$

Lemma 5.6 *If f is differentiable at x then all directional derivatives of f at x exist and, for any $v \in \mathbb{R}^n$,*

$$\frac{\partial f}{\partial v}(x) = f'(x) v,$$

where the right hand side is the product of $1 \times n$ and $n \times 1$ matrices.

Proof. Indeed, by the definition of differentiability,

$$f(x + tv) - f(x) = f'(x)(tv) + o(tv).$$

Dividing by t and passing to the limit, we obtain the claim. ■

As a consequence, $\frac{\partial f}{\partial v}(x)$ is a linear functional of v, which is not obvious from the definition.

Definition. If x, y are two point in $\mathbb{R}^n$ then denote by $[x, y]$ the closed interval between x, y, that is,

$$[x, y] = \{(1 - t) x + ty : 0 \leq t \leq 1\}.$$

Theorem 5.7 (Mean-value theorem) *Let f be a differentiable function in an open set U and let x, y be two distinct points in U such that the interval $[x, y]$ is contained in U. Then there exists a point $\xi \in [x, y]$ such that*

$$f(y) - f(x) = f'(\xi)(y - x).$$

Proof. Set $v = y - x$ and consider a function $g(t) = f(x + tv)$ so that $g(0) = f(x)$ and $g(1) = f(y)$. Then g is differentiable in $[0, 1]$ as composition of functions $t \mapsto x + tv$ and f. By the mean value theorem from Analysis I, there exists $s \in [0, 1]$ such that

$$g(1) - g(0) = g'(s).$$

Set $\xi = x + sv$ so that $\xi \in [x, y]$ and notice that

$$\partial_v f(\xi) = \left.\frac{df(\xi + tv)}{dt}\right|_{t=0} = \left.\frac{df(x + tv)}{dt}\right|_{t=s} = g'(s).$$

Using Lemma 5.6, we obtain

$$f(y) - f(x) = g(1) - g(0) = g'(s) = \partial_v f(\xi) = f'(x) v = f'(x)(y - x),$$

which was to be proved. ■

5.4 Higher order partial derivatives

5.4.1 Changing the order

Let $U \subset \mathbb{R}^n$ be an open set and consider a function $f : U \to \mathbb{R}$. If the partial derivative $\frac{\partial f}{\partial x_i}$ exists then we can try to differentiate it considering the its partial derivative:

$$\frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}.$$

If this derivative exists then it is called a partial derivative of the second order and is denoted by $\frac{\partial^2 f}{\partial x_j \partial x_i}$ or by $\partial_{ji} f$ or by $D_{ji} f$. If $i = j$ then one uses notation $\frac{\partial^2 f}{\partial x_i^2} = \partial_{ii} f = D_{ii} f$. If $i \neq j$ then $\frac{\partial^2 f}{\partial x_j \partial x_i}$ is called a *mixed* derivative.

Similarly, one can consider partial derivatives of an arbitrary order $k \in \mathbb{N}$:

$$\frac{\partial}{\partial x_{i_1}}\left(\frac{\partial}{\partial x_{i_2}}\left(...\frac{\partial f}{\partial x_{i_k}}\right)\right) = \frac{\partial^k f}{\partial x_{i_1}\partial x_{i_2}...\partial x_{i_k}} = \partial_{i_1 i_2 ... i_k} f = D_{i_1 i_2 ... i_k} f.$$

Theorem 5.8 (Hermann Schwarz's Theorem) *If a function $f : U \to \mathbb{R}$ has in U both partial derivatives $\partial_{ij} f$ and $\partial_{ji} f$ and they are both continuous at a point $x \in U$ then $\partial_{ij} f(x) = \partial_{ji} f(x)$.*

Proof. Assume for simplicity that $x = 0$ is the origin of $\mathbb{R}^n$ and also that $i = 1$ and $j = 2$. In the course of the proof, we will not vary the variables $x_3, ..., x_n$ so that we can consider them constantly equal to 0. Therefore, f can be regarded as a function $f(x_1, x_2)$ of two variables x_1 and x_2, defined for $|x_1| < \varepsilon$ and $|x_2| < \varepsilon$ for small enough $\varepsilon > 0$. Consider an auxiliary function

$$\begin{aligned} F(x_1, x_2) &= f(x_1, x_2) - f(x_1, 0) - f(0, x_2) + f(0, 0) \\ &= \varphi(x_1) - \varphi(0) \end{aligned}$$

where

$$\varphi(t) = f(t, x_2) - f(t, 0).$$

By hypothesis, function φ is differentiable on $[0, x_1]$. By the mean value theorem, there exists $s_1 \in [0, x_1]$ such that

$$\varphi(x_1) - \varphi(0) = \varphi'(s_1) x_1 = (\partial_1 f(s_1, x_2) - \partial_1 f(s_1, 0)) x_1.$$

The function

$$\psi(t) = \partial_1 f(s_1, t)$$

is differentiable in $t \in [0, x_2]$ so that we obtain that there exists $s_2 \in [0, x_2]$ such that

$$\psi(x_2) - \psi(0) = \psi'(s_2) x_2$$

that is,

$$\partial_1 f(s_1, x_2) - \partial_1 f(s_1, 0) = \partial_2 \partial_1 f(s_1, s_2) x_2.$$

Combining the above lines, we obtain

$$F(x_1, x_2) = \partial_2 \partial_1 f(s_1, s_2) x_1 x_2.$$

If we represent F in the form

$$\begin{aligned} F(x_1, x_2) &= f(x_1, x_2) - f(0, x_2) - f(x_1, 0) + f(0, 0) \\ &= \widetilde{\varphi}(x_2) - \widetilde{\varphi}(0) \end{aligned}$$

where

$$\widetilde{\varphi}(t) = f(x_1, t) - f(0, t)$$

then a similar argument shows that there are $\widetilde{s}_1 \in [0, x_1]$ and $\widetilde{s}_2 \in [0, x_2]$ such that

$$F(x_1, x_2) = \partial_1 \partial_2 f(\widetilde{s}_1, \widetilde{s}_2) x_1 x_2.$$

Assuming that x_1 and x_2 are $\neq 0$, we obtain from the comparison of the two expressions for F, that

$$\partial_2 \partial_1 f(s_1, s_2) = \partial_1 \partial_2 f(\widetilde{s}_1, \widetilde{s}_2).$$

Note that all variables $s_1, s_2, \widetilde{s}_1, \widetilde{s}_2$ are functions of x_1, x_2 and by construction they all go to 0 as $x_1, x_2 \to 0$. Passing to the limit in the above identity and using the continuity of the mixed derivatives at 0, we obtain

$$\partial_2 \partial_1 f(0, 0) = \partial_1 \partial_2 f(0, 0),$$

which was to be proved. ■

Without the continuity assumption, the derivatives $\partial_{ij} f$ and $\partial_{ji} f$ can be different as it is shown in the next example.

Example. Consider the function in $\mathbb{R}^2$

$$f(x, y) = \begin{cases} xy \frac{x^2 - y^2}{x^2 + y^2}, & (x, y) \neq 0, \\ 0, & x = y = 0, \end{cases}$$

and evaluate $\partial_{12} f(0)$ and $\partial_{21} f(0)$. By definition,

$$\partial_1 \partial_2 f(0, 0) = \lim_{x \to 0} \frac{\partial_2 f(x, 0) - \partial_2 f(0, 0)}{x}.$$

Let us find $\partial_2 f(x, 0)$:

$$\partial_2 f(x, 0) = \lim_{y \to 0} \frac{f(x, y) - f(x, 0)}{y} = \lim_{y \to 0} x \frac{x^2 - y^2}{x^2 + y^2} = x.$$

Since also $\partial_2 f(0, 0) = 0$, we conclude

$$\partial_1 \partial_2 f(0, 0) = \lim_{x \to 0} \frac{x - 0}{x} = 1.$$

Similarly, we have

$$\partial_1 f(0, y) = \lim_{x \to 0} \frac{f(x, y) - f(0, y)}{x} = \lim_{x \to 0} y \frac{x^2 - y^2}{x^2 + y^2} = -y$$

and

$$\partial_{21} f(0,0) = \lim_{y\to 0} \frac{\partial_1 f(0,y) - \partial_1 f(0,0)}{y} = \lim_{y\to 0} \frac{-y-0}{y} = -1.$$

Hence, $\partial_{12} f(0) \neq \partial_{21} f(0)$.

Definition. We say that a function $f : U \to \mathbb{R}$ belongs to the class $C^k(U)$ (where k is a non-negative integer) if all partial derivatives of f of the order $\leq k$ exist in U and are continuous in U.

Note that if $k = 0$ then $C^0(U)$ is the class of all continuous functions in U.

Corollary. *If $f \in C^k(U)$ then the value of any partial derivative of f the order $\leq k$ does not depend on the order of differentiation. More precisely, if $i_1, ..., i_l$ a sequence of $l \leq k$ indices and $j_1, ..., j_l$ is a permutation of $i_1, ..., i_l$ then $\partial_{i_1....i_l} f = \partial_{j_1....j_l} f$.*

Proof. By Theorem 5.8, any two neighboring indices in the sequence $i_1, ...i_l$, say i_m and i_{m+1}, can be interchanged without changing the value of the derivative:

$$\begin{aligned} \partial_{i_1...i_m i_{m+1}...i_l} f &= \partial_{i_1...i_{m-1}} \left(\partial_{i_m} \partial_{i_{m+1}}\right) \partial_{i_{m+2}....i_l} f \\ &= \partial_{i_1...i_{m-1}} \left(\partial_{i_{m+1}} \partial_{i_m}\right) \partial_{i_{m+2}....i_l} f \\ &= \partial_{i_1...i_{m+1} i_m...i_l} f. \end{aligned}$$

Now the claim follows from the fact that the sequence $j_1, ..., j_l$ can be obtained from $i_1, ..., i_l$ using finitely many interchanging of neighboring indices. ■

5.4.2 Taylor's formula

The previous Corollary allows to introduce another notation for higher order derivatives: if $f \in C^k(U)$ then any partial derivative of the order $l \leq k$ can be written as

$$\frac{\partial^l f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}},$$

where $\alpha_1 + ... + \alpha_n = l$. This notation means that we differentiate f α_1 times in x_1, α_2 times in x_2, etc (while the order of differentiation does not matter), that is,

$$\frac{\partial^l f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} = \partial_{\underbrace{1...1}_{\alpha_1} \underbrace{2...2}_{\alpha_2} ... \underbrace{n...n}_{\alpha_n}} f \ .$$

Definition. A *multiindex* α of dimension n is any sequence $\alpha = (\alpha_1, ..., \alpha_n)$ of non-negative integers. For any multiindex α, define its *order*

$$|\alpha| = \alpha_1 + ... + \alpha_n$$

and *factorial*

$$\alpha! = \alpha_1! ... \alpha_n!.$$

For any vector $x \in \mathbb{R}^n$, set

$$x^\alpha = x_1^{\alpha_1} ... x_n^{\alpha_n}.$$

For any function $f \in C^k(U)$ where U is an open set in $\mathbb{R}^n$ and for any multiindex α of dimension n and order $\leq k$, set

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}}.$$

Theorem 5.9 (Taylor's theorem) *Let U be an open subset of $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ be a function of the class $C^k(U)$, where $k \geq 0$. Then, for any $x \in U$,*

$$f(x+h) = \sum_{|\alpha| \leq k} \frac{D^\alpha f(x)}{\alpha!} h^\alpha + o\left(\|h\|^k\right) \text{ as } h \to 0. \tag{5.9}$$

Here α is a multiindex of dimension n. Note that if $\|h\|$ is small enough then $x+h \in U$ by the openness of U. Hence, $f(x+h)$ is defined for sufficiently small $\|h\|$.

The function

$$\sum_{|\alpha| \leq k} \frac{D^\alpha f(x)}{\alpha!} h^\alpha$$

is called the *Taylor polynomial* of the order k of function f at point x. This is obviously a polynomial in the variables $h_1, ..., h_n$.

If U is an interval in $\mathbb{R}$, that is, $n = 1$, then we obtain the Taylor formula from Analysis I:

$$f(x+h) = \sum_{\alpha=0}^{k} \frac{f^{(\alpha)}(x)}{\alpha!} h^\alpha + o\left(|h|^k\right)$$

where α is a non-negative integer.

For an arbitrary n, let us expand the terms in (5.9) with $|\alpha| \leq 2$. If $|\alpha| = 0$ then $\alpha = (0, ..., 0)$ and

$$\frac{D^\alpha f(x)}{\alpha!} h^\alpha = f(x).$$

If $|\alpha| = 1$ then α has the form $(0, ..., 0, 1, 0, ..., 0)$ with one term 1 and all others 0. If the term 1 is at position i then

$$\frac{D^\alpha f(x)}{\alpha!} h^\alpha = \partial_i f(x) h_i$$

and

$$\sum_{|\alpha|=1} \frac{D^\alpha f(x)}{\alpha!} h^\alpha = \sum_{i=1}^{n} \partial_i f(x) h_i = \partial_1 f(x) h_1 + ... + \partial_n f(x) h_n.$$

If $|\alpha| = 2$ then there are two possibilities: either α has the form $(0, ...0, 2, 0, ..., 0)$ with the term 2 at some position i or α has two terms 1 at positions $i < j$, and all other 0. In the first case,

$$\frac{D^\alpha f(x)}{\alpha!} h^\alpha = \frac{\partial_{ii} f(x)}{2} h_i^2,$$

and in the second case

$$\frac{D^\alpha f(x)}{\alpha!} h^\alpha = \partial_{ij} f(x) h_i h_j.$$

Therefore,

$$\sum_{|\alpha|=2} \frac{D^\alpha f(x)}{\alpha!} h^\alpha = \sum_{i=1}^{n} \frac{\partial_{ii} f(x)}{2} h_i^2 + \sum_{i<j} \partial_{ij} f(x) h_i h_j = \frac{1}{2} \sum_{i,j=1}^{n} \partial_{ij} f(x) h_i h_j$$

Hence, the Taylor formula for $k = 2$ can be written in the form

$$f(x+h) = f(x) + \sum_{i=1}^{n} \partial_i f(x) h_i + \frac{1}{2} \sum_{i,j=1}^{n} \partial_{ij} f(x) h_i h_j + o\left(\|h\|^2\right). \tag{5.10}$$

5.4.3 Local extrema

Postponing the proof of the Taylor formula, let us first show that, similarly to Analysis I, the Taylor formula can be used to investigate local extrema of functions.

Definition. We say that a function $f : U \to \mathbb{R}$ has a *local maximum* at a point $x \in U$ if there is a ball $B(x,r) \subset U$ such that

$$f(x) \geq f(y) \text{ for any } y \in B(x,r),$$

that is, $f(x)$ is the maximal value of f in $B(x,r)$. A local maximum is *strict* if

$$f(x) > f(y) \text{ for any } y \in B(x,r) \setminus \{x\}.$$

Similarly one defined a *local minimum* and a *strict local minimum* A local *extremum* is either a local maximum or a local minimum.

In order to state the conditions for the local extrema of a function, we will use the notions of the *gradient* and the *Hessian* of a function. For a function $f : U \to \mathbb{R}$, the gradient $\operatorname{grad} f(x)$ is defined for any $x \in U$ as a vector in $\mathbb{R}^n$ with components

$$\operatorname{grad} f(x) = (\partial_1 f(x), \partial_2 f(x), ..., \partial_n f(x)),$$

provided the corresponding partial derivatives exist. The gradient $\operatorname{grad} f(x)$ looks identical to the Jacobian matrix $J_f(x)$ but the difference is that $\operatorname{grad} f$ is a *vector* in $\mathbb{R}^n$ while $J_f(x)$ is a *matrix* of dimension $1 \times n$ although with the same components as the gradient.

Assuming that all second order partial derivatives of f exists in U, we can consider at any point $x \in U$ the matrix called the *Hessian* of f :

$$\operatorname{Hess} f(x) = (\partial_{ij} f(x))_{i,j=1}^n = \begin{pmatrix} \partial_{11} f(x) & \partial_{12} f(x) & ... & \partial_{1n} f(x) \\ \partial_{21} f(x) & \partial_{22} f(x) & ... & \partial_{2n} f(x) \\ ... & ... & ... & ... \\ \partial_{n1} f(x) & \partial_{n2} f(x) & ... & \partial_{nn} f(x) \end{pmatrix}.$$

It follows from Theorem 5.8 that if all the second order derivatives are continuous then $\operatorname{Hess} f(x)$ is a symmetric $n \times n$ matrix.

Recall that with any $n \times n$ matrix $A = (a_{ij})$ one associates the *quadratic form*

$$Q(h) = \sum_{i,j=1}^n a_{ij} h_i h_j,$$

which is defined as a function of $h = (h_1, ..., h_n) \in \mathbb{R}^n$. The matrix A (and the form Q) is called

- *non-negative definite* if $Q(h) \geq 0$ for all $h \in \mathbb{R}^n$; in this case, write $A \geq 0$;
- *non-positive definite* if $Q(h) \leq 0$ for all $h \in \mathbb{R}^n$; in this case, write $A \leq 0$;
- *positive definite* if $Q(h) > 0$ for all $h \neq 0$; write $A > 0$;

- *negative definite* if $Q(h) < 0$ for all $h \neq 0$; write $A < 0$;
- *indefinite* if $Q(h)$ takes both positive and negative values.

For example, if $A = \mathrm{id}$ then $Q(h) = h_1^2 + ... + h_n^2$ is positive definite. Hence, $\mathrm{id} \geq 0$. If $n = 2$ and $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ then

$$Q(h) = a_{11}h_1^2 + a_{12}h_1h_2 + a_{21}h_2h_1 + a_{22}h_2^2 = 2h_1h_2.$$

Since $Q(h)$ can be both positive and negative, the matrix A is indefinite.

Theorem 5.10 *Let $U \subset \mathbb{R}^n$ be an open set and $f : U \to \mathbb{R}$ be a function on U.*

(a) (Necessary condition for a local extremum) *Let $x \in U$ be a local extremum of f. If f is differentiable at x then* $\operatorname{grad}(x) = 0$. *If $f \in C^2(U)$ and x is a point of local maximum then* $\operatorname{Hess} f(x) \leq 0$*; if x is a point of a local minimum then* $\operatorname{Hess} f(x) \geq 0$.

(b) (Sufficient condition for a local extremum) *Let $f \in C^2(U)$. If, for some $x \in U$,* $\operatorname{grad} f(x) = 0$ *and* $\operatorname{Hess} f(x) > 0$ *then x is a point of a strict local minimum of f. If* $\operatorname{grad} f(x) = 0$ *and* $\operatorname{Hess} f(x) < 0$ *then x is a point of a strict local maximum of f.*

Proof. *(a)* We need to prove that $\partial_i f(x) = 0$ for all $i = 1, ..., n$. Consider function $f(x_1, ..., x_n)$ as a function of x_i only, with fixed x_j for $j \neq i$. Then this function as a function of x_i still has a local extremum at the given point x. Using the necessary condition for a local extremum from Analysis I, we conclude that $\partial_i f(x) = 0$.

Without using Analysis I, we can argue as follows. Assume that x is a point of a local minimum. By the definition of the differentiability,

$$f(x+h) - f(x) = f'(x)\,h + o(h) = c_1h_1 + ... + c_nh_n + o(h) \text{ as } h \to 0,$$

where $c_i = \partial_i f(x)$. We need to prove that all $c_i = 0$. If some $c_i \neq 0$ then set $h = (0, ..., 0, t, 0, ..., 0)$ where t is at position i. For this h, we obtain

$$f(x+h) - f(x) = c_it + o(t) \text{ as } t \to 0.$$

If f has a local minimum at x then $f(x+h) - f(x) \geq 0$ for sufficiently small $\|h\|$, that is, $c_it + o(t) \geq 0$ for small enough $|t|$. For small enough $|t|$, the term $o(t)$ is smaller than $|c_it|$ (because $c_i \neq 0$). Therefore, the sign of $c_it + o(t)$ is the same as the sign of c_it. But we cannot have $c_it \geq 0$ for all small enough $|t|$ because changing t to $-t$ we change the sign of c_it. This contradiction shows that all c_i must be 0.

Assume now that $f \in C^2(U)$. We have by the Taylor formula (5.10)

$$\begin{aligned} f(x+h) - f(x) &= \sum_{i=1}^{n} \partial_i f(x)\, h_i + \frac{1}{2}\sum_{i,j=1}^{n} \partial_{ij} f(x)\, h_ih_j + o\left(\|h\|^2\right) \\ &= \frac{1}{2}Q(h) + o\left(\|h\|^2\right), \end{aligned} \tag{5.11}$$

where we have used that $\partial_i f(x) = 0$ and $Q(h)$ is the quadratic form associated with $\operatorname{Hess} f(x)$. Let x be a point of a local minimum and let us prove that $\operatorname{Hess} f(x) \geq 0$, that

is, $Q(h) \geq 0$ for all $h \in \mathbb{R}^n$. Indeed, if $Q(h) < 0$ for some h then replace h by th where $t \in [0,1]$. Then we obtain

$$f(x+th) - f(x) = \frac{1}{2}Q(h)\,t^2 + o\left(t^2\right),$$

and the right hand side is negative if t is sufficiently small. However, at a point of minimum the left hand side is non-negative. This proves that $Q(h) \geq 0$ and, hence, $\operatorname{Hess} f(x) \geq 0$. The case of a local maximum is treated similarly.

(*b*) If $\operatorname{grad}(x) = 0$ then we have the expansion (5.11). Assume that $\operatorname{Hess} f(x) > 0$, that is, the function $Q(h)$ is positive for all $h \in \mathbb{R}^n \setminus \{0\}$. We need to prove that

$$\frac{1}{2}Q(h) + o\left(\|h\|^2\right) > 0$$

provided $\|h\|$ is small enough but $h \neq 0$. Function $Q(h)$ is a continuous function on $\mathbb{R}^n$ being a linear combination of continuous functions $h_i h_j$. Consider a set

$$S = \{h \in \mathbb{R}^n : \|h\| = 1\}$$

where $\|h\| = N(h)$ is any norm in $\mathbb{R}^n$. Then S is bounded (obviously, it is contained is any ball $B(0,r)$ with $r > 1$) and closed (because $S = N^{-1}(\{1\})$ and N is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$). Hence, by the minimal value theorem, function $Q(h)$ has the minimum on S, let $\min_S Q = m$. Since $Q|_S > 0$, we have $m > 0$. For any $h \neq 0$, we have $\frac{h}{\|h\|} \in S$ whence

$$Q\left(\frac{h}{\|h\|}\right) \geq m,$$

and

$$Q(h) \geq m\|h\|^2.$$

On the other hand, if $\|h\|$ is sufficiently small, then $o\left(\|h\|^2\right) < \varepsilon\|h\|^2$ for any given $\varepsilon > 0$. Therefore, we obtain that, for sufficiently small $\|h\| \neq 0$

$$f(x+h) - f(x) = \frac{1}{2}Q(h) + o\left(\|h\|^2\right) \geq \left(\frac{1}{2}m - \varepsilon\right)\|h\|^2 > 0$$

provided ε is chosen small than $\frac{1}{2}m$. This means that $f(x+h) > f(x)$ in a small ball around x unless $h = 0$, that is, x is a point of a strict local minimum of f.

The case $\operatorname{Hess} f(x) < 0$ is similar. ■

Example. Let us find local extrema of the function

$$f(x,y) = xy \ln\left(x^2 + y^2\right).$$

Firstly, let us find all the *critical points*, that is, the points where $\operatorname{grad} f = 0$, which in this case amounts to the equations $\partial_x f = \partial_y f = 0$. Since

$$\begin{aligned} \partial_x f &= y \ln\left(x^2+y^2\right) + \frac{2x^2 y}{x^2+y^2} \\ \partial_y f &= x \ln\left(x^2+y^2\right) + \frac{2xy^2}{x^2+y^2} \end{aligned}$$

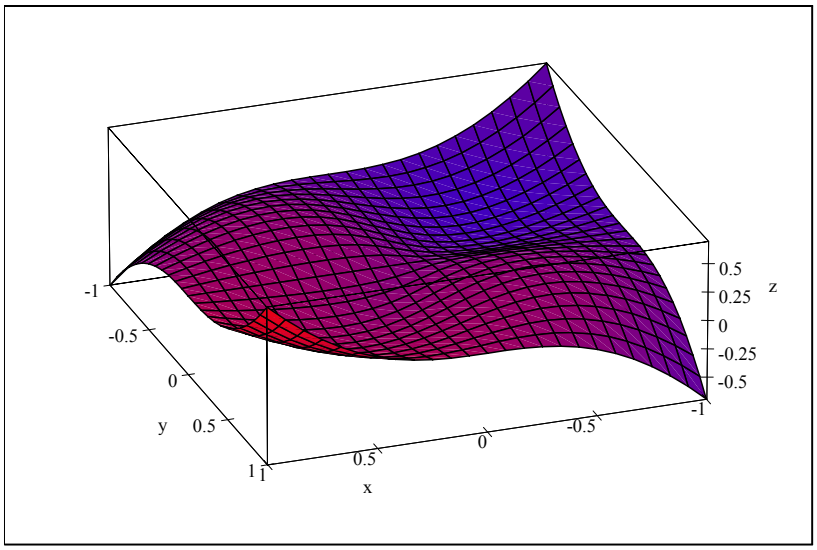


Figure 1: Function $f(x,y)$ in the domain $|x| \le 1$, $|y| \le 1$.

solving the equations $\partial_x f = \partial_y f = 0$ we obtain the following roots:

$$(0, \pm 1), (\pm 1, 0), \left(\pm \frac{1}{\sqrt{2e}}, \pm \frac{1}{\sqrt{2e}} \right), \left(\mp \frac{1}{\sqrt{2e}}, \pm \frac{1}{\sqrt{2e}} \right).$$

The points $(0, \pm 1)$, $(\pm 1, 0)$ cannot be points of local extrema because $f(x,y)$ is odd with respect to x and y separately, and an odd function cannot take a local extremum at 0. Finding the second derivatives, we obtain

$$\begin{aligned} \partial_{xx} f &= \frac{6xy}{x^2+y^2} - \frac{4x^3 y}{\left(x^2+y^2\right)^2} \\ \partial_{xy} f &= \ln\left(x^2+y^2\right) + 2 - \frac{4x^2y^2}{\left(x^2+y^2\right)^2} \\ \partial_{yy} f &= \frac{6xy}{x^2+y^2} - \frac{4xy^3}{\left(x^2+y^2\right)^2}. \end{aligned}$$

At point $\left(\pm \frac{1}{\sqrt{2e}}, \pm \frac{1}{\sqrt{2e}} \right)$ we have

$$\operatorname{Hess} f = \begin{pmatrix} \partial_{xx} f & \partial_{xy} f \\ \partial_{xy} f & \partial_{yy} f \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

and this matrix is obviously positive definite, whence it follows that these two points are the points of a strict local minimum. Similarly, $\left(\mp \frac{1}{\sqrt{2e}}, \pm \frac{1}{\sqrt{2e}} \right)$ are points of a strict local maximum.

Remark. In order to decide whether a given symmetric matrix A is positive definite one can apply one of the following approaches, known from Linear Algebra:

1. $A > 0$ if and only if, for any $1 \le k \le n$,

$$\det \left(a_{ij}\right)_{ij=1}^{k} > 0.$$

2. $A > 0$ if the corresponding quadratic form $Q(x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j$ can be represented in the form $y_1^2 + ... + y_n^2$ for some linear change of the variables.

3. $A > 0$ if all the eigenvalues (*Eigenwerte*) of A are positive.

5.4.4 Proof of Taylor's formula

Proof of Theorem 5.9. Let U be an open subset of $\mathbb{R}^n$ and $f \in C^k(U)$. Denote by $T_k(h)$ the Taylor polynomial of function f at a point $x \in U$ of the order k, that is,

$$T_k(h) = \sum_{|\alpha| \le k} \frac{D^\alpha f(x)}{\alpha!} h^\alpha. \tag{5.12}$$

We need to prove that

$$R_k(h) := f(x+h) - T_k(h) = o\left(\|h\|^k\right) \text{ as } h \to 0. \tag{5.13}$$

Use induction in k. If $k = 0$ then (5.13) becomes

$$f(x+h) - f(x) = o(1),$$

which is true by the continuity of f. Let us prove the inductive step from $k-1$ to k assuming $k \ge 1$. Fix some index $i = 1, ..., n$ and consider the derivative $\partial_i f$ which is of class $C^{k-1}(U)$.

Claim. *The Taylor polynomial of the function $\partial_i f$ of the order $k-1$ is equal to $\partial_i T_k(h)$.*

We have from (5.12)

$$\partial_i T_k(h) = \sum_{|\alpha| \le k} \frac{D^\alpha f(x)}{\alpha!} \partial_i h^\alpha.$$

If $\alpha_i = 0$ then h^α does not depend on h_i and $\partial_i h^\alpha = 0$. Hence, all α with $\alpha_i = 0$ can be omitted in the above sum; in other words, we can restrict the summation to those α with $\alpha_i \ge 1$. Denote $\beta = (0, ..., 1, ..., 0)$ where the only 1 is at position i. Then $\alpha - \beta$ is a multiindex of order $|\alpha| - 1$ and the following identities take place:

$$\partial_i h^\alpha = \alpha_i \left(h_1^{\alpha_1} ... h_i^{\alpha_i - 1} ... h_n^{\alpha_n}\right) = \alpha_i h^{\alpha - \beta},$$

$$\alpha! = \alpha_1! ... \alpha_i! ... \alpha_n! = \alpha_i \left(\alpha_1! ... (\alpha_i - 1)! ... \alpha_n!\right) = \alpha_i (\alpha - \beta)!,$$

$$D^\alpha f = D^{\alpha - \beta} D^\beta f = D^{\alpha - \beta} \partial_i f.$$

Therefore,

$$\partial_i T_k(h) = \sum_{\substack{|\alpha| \le k \\ \alpha_i \ge 1}} \frac{D^{\alpha - \beta} \partial_i f(x)}{\alpha_i (\alpha - \beta)!} \alpha_i h^{\alpha - \beta} = \sum_{|\gamma| \le k-1} \frac{D^\gamma (\partial_i f)(x)}{\gamma!} h^\gamma$$

where α_i has cancelled out and we have changed $\gamma = \alpha - \beta$. This identity proves the claim because the right hand side is the Taylor polynomial of the function $\partial_i f$ of the order $k-1$.

Applying the inductive hypothesis to the function $\partial_i f \in C^{k-1}(U)$, we obtain

$$\partial_i f(x+h) = \partial_i T_k(h) + o\left(\|h\|^{k-1}\right) \text{ as } h \to 0.$$

Using the function $R_k(h)$ defined by (5.13), we can rewrite this as

$$\partial_i R_k(h) = o\left(\|h\|^{k-1}\right) \text{ as } h \to 0. \tag{5.14}$$

Since $k \geq 1$, the function $R_k(h)$ belongs to C^1 in a small ball around 0 and, hence, is differentiable in this ball. Applying the mean value theorem to this function and using $R_k(0) = 0$, we obtain

$$R_k(h) = R_k(h) - R_k(0) = R_k'(\xi)\, h = \sum_{i=1}^{n} \partial_i R_k(\xi)\, h_i$$

where ξ is some point in $[0, h]$. Then $\|\xi\| \leq \|h\|$ and we obtain from (5.14)

$$|R_k(h)| \leq \sum_{i=1}^{n} |\partial_i R_k(\xi)|\, \|h\|_\infty = o\left(\|\xi\|^{k-1}\right) \|h\|_\infty = o\left(\|h\|^k\right),$$

which was to be proved. ■

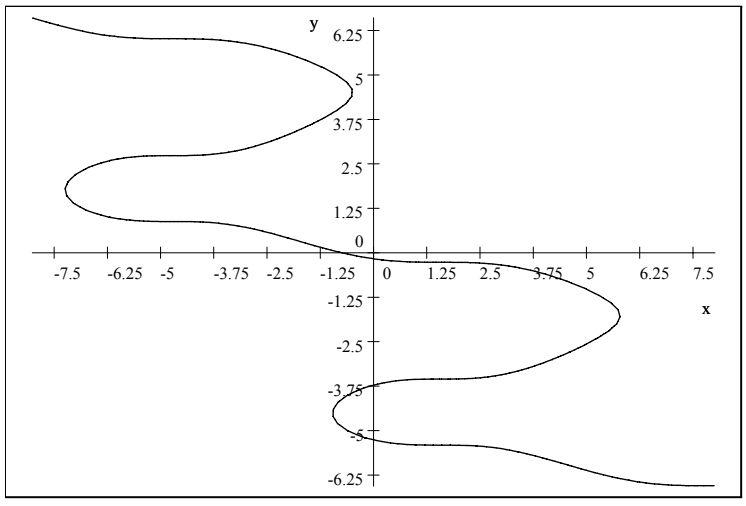
5.5 Implicit function theorem

Consider the following problem: how to find y as a function of x if its known that they are related by some identity $F(x, y) = 0$? Here x, y are so far real variables and $F(x, y)$ is a function in an open subset $W \subset \mathbb{R}^2$. Function $y = f(x)$ defined in this way is called an *implicit function*.

For example, if $x^2 + y^2 = 1$ then one can explicitly solve this to find two continuous functions $y = \pm\sqrt{1 - x^2}$. Consider another example without explicit solution:

$$x + \cos x + y + 5 \sin y = 0. \tag{5.15}$$

Here y cannot be explicitly expressed via x, but nevertheless one may hope to prove the existence of the function $y = f(x)$ that satisfies this relation. Here is the plot of the set of points (x, y) that satisfies (5.15):



This curve is not a graph of a function but it can be split into a number of graphs if we restrict the domain of x to the intervals between the *turning* points.

Consider now a more general situation when x is a point in $\mathbb{R}^n$, y is a point in $\mathbb{R}^m$ and the couple (x, y) is considered as a point in $\mathbb{R}^{n+m}$ with components

$$(x_1, ..., x_n, y_1, ..., y_m).$$

Let W be an open set in $\mathbb{R}^{n+m}$ where a mapping $F : W \to \mathbb{R}^m$ is defined, and we would like to solve the equation $F(x,y) = 0$ with respect to y in order to obtain a function $y = f(x)$. In the coordinate form, using the components F_j of F, we obtain the system of m equations

$$\begin{cases} F_1(x_1,...,x_n,y_1,...,y_m) = 0 \\ F_2(x_1,...,x_n,y_1,...,y_m) = 0 \\ ... \\ F_m(x_1,...,x_n,y_1,...,y_m) = 0 \end{cases}$$

which need to be solved with respect to m unknown $y_1,...,y_m$ considering $x_1,...,x_n$ as given parameters.

To get some flavor of what can happen consider a simple case when F is linear functions of y:

$$F(x,y) = A(x) + B(x)\,y,$$

where $A(x) \in \mathbb{R}^m$ and $B(x)$ is a $m \times m$ matrix, both depending only on x. Then the equation $F(x,y) = 0$ becomes

$$A(x) + B(x)\,y = 0,$$

which is solvable provided the matrix $B(x)$ is invertible, and we obtain

$$y = -B^{-1}(x)\,A(x).$$

Consider an arbitrary differentiable mapping $F : W \to \mathbb{R}^m$ where $W \subset \mathbb{R}^{n+m}$ is open. Denote by $\partial_x F$ the Jacobian matrix of $F(x,y)$ considered as a function of $x \in \mathbb{R}^n$ only, that is,

$$\partial_x F = (\partial_{x_i} F_j),\ \ i = 1,...,n,\ \ j = 1,...,m$$

that is, $\partial_x F$ is an $m \times n$ matrix (j is the index of rows and i is the index of columns). Similarly, the Jacobian matrix in y

$$\partial_y F = (\partial_{y_i} F_j)_{i,j=1}^m$$

is an $m \times m$ matrix. The full Jacobian matrix of $F(x,y)$ is $m \times (n+m)$ matrix that can be presented in the form

$$J_F = (\partial_x F \mid \partial_y F). \tag{5.16}$$

In the case $F = A(x) + B(x)\,y$, we have $\partial_y F(x,y) = B(x)$. Hence, the solvability of the equation $F(x,y) = 0$ in the linear case is equivalent to the fact that the matrix $\partial_y F$ **is invertible**. It turns out that the same condition will be used in the general case.

In order to state the result, we need some terminology.

Definition. Let U be an open set in some $\mathbb{R}^k$ and $f : U \to \mathbb{R}^m$. We say that function f belongs to class C^l if, for any component f_j, all the partial derivatives $D^\alpha f_j$ of the order $|\alpha| \leq l$ exist and are continuous functions in U. If $l = 0$ then f belongs to C^0 just means that f is continuous.

Notation: $f \in C^l$ or $f \in C^l(U, \mathbb{R}^m)$. The latter notation indicates also the domain of f and the target space.

Lemma 5.11 *The arithmetic operations on C^l functions result in C^l functions. Composition of C^l functions is a C^l function.*

Hint for the proof. For the case $l = 0$, this was proved in Corollaries to Theorem 4.4. For the general case, one uses induction in l and the chain rule (see Exercises). ■

The next theorem is one of the main results of multivariable Differential Calculus.

Theorem 5.12 (The implicit function theorem) *Let W be an open set in $\mathbb{R}^{n+m}$ and $F \in C^l(W, \mathbb{R}^m)$, $l \geq 1$. Assume that, for some $(a, b) \in W$,*

$$F(a, b) = 0 \quad \text{and} \quad \partial_y F(a, b) \quad \text{is invertible.}$$

Then there are open sets $U \subset \mathbb{R}^n$ and $V \in \mathbb{R}^m$ such that $a \in U$, $b \in V$, $U \times V \subset W$, and a function $f : U \to V$ of the class C^l such that

$$F(x, y) = 0 \;\Leftrightarrow\; y = f(x) \quad \text{for } x \in U, y \in V.$$

Moreover, for any $x \in U$,

$$f'(x) = -(\partial_y F)^{-1} \partial_x F. \tag{5.17}$$

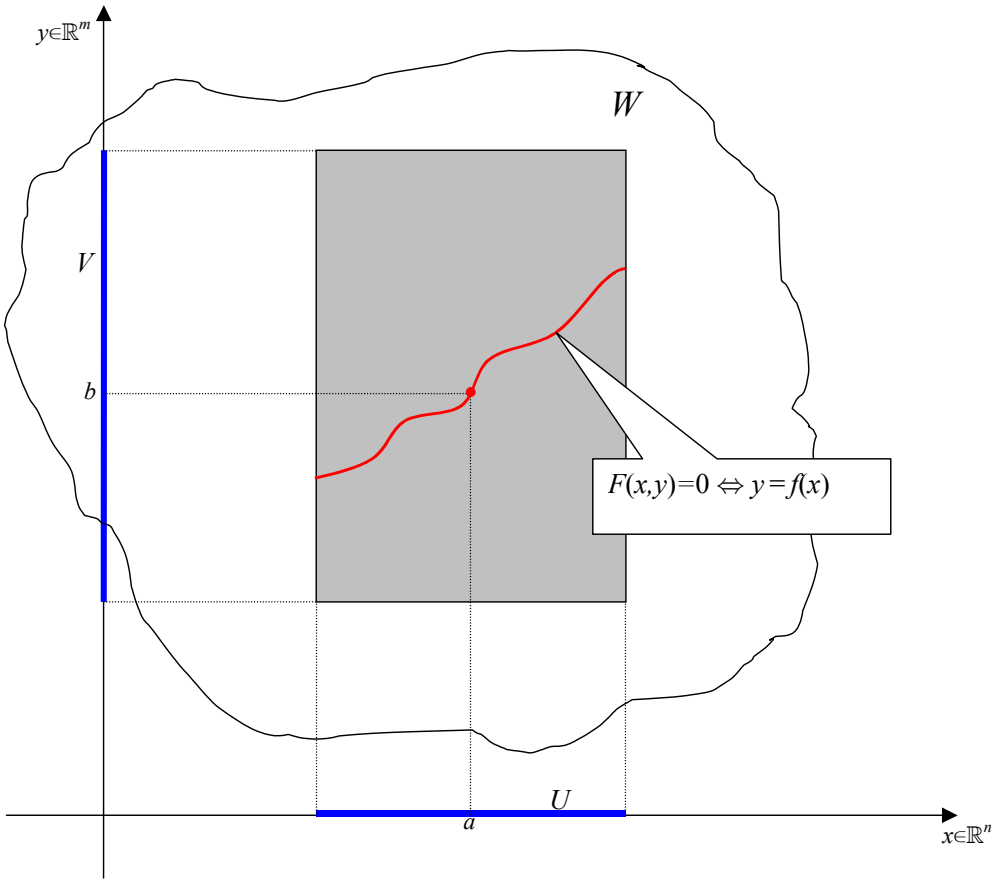


Figure 2: Illustration to Theorem 5.12

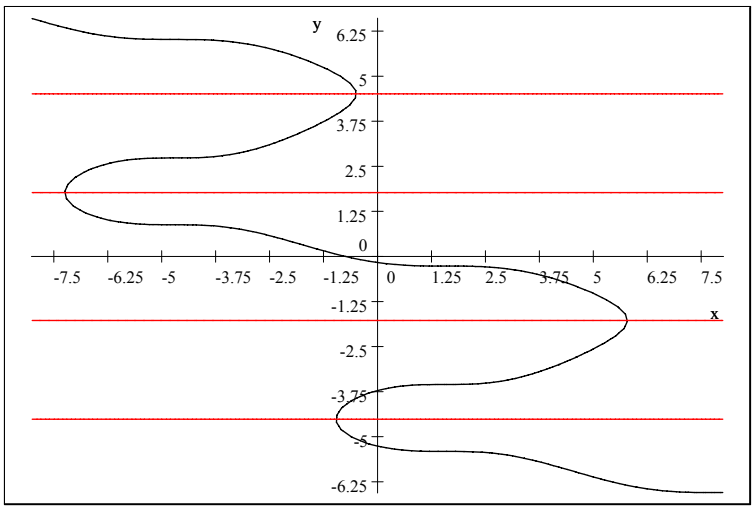
Example. Consider again the function (5.15), that is,

$$F(x, y) = x + \cos x + y + 5 \sin y$$

where $x, y \in \mathbb{R}$. We have

$$F_y = 1 + 5\cos y$$

and this is non-zero if $\cos y \neq -\frac{1}{5}$. Hence, for any point $(a, b) \in \mathbb{R}^2$ such that $F(a, b) = 0$ and $\cos b \neq -\frac{1}{5}$, the equation $F(x, y) = 0$ can be solved in a neighborhood of (a, b) and y can be represented as a function of x. In other words, in a neighborhood of such a point (a, b), the set $\{F(x, y) = 0\}$ is a graph of a function.



At the plot above, we have two sets in $\mathbb{R}^2$: $\{F(x, y) = 0\}$ and

$$\{\partial_y F(x, y) = 0\} = \left\{ y = \pm \arccos\left(-\frac{1}{5}\right) + 2\pi k,\ k \in \mathbb{Z} \right\}, \tag{5.18}$$

the latter being a collection of the horizontal lines. It is easy to see that the turning points of the curve $\{F = 0\}$ are contained in the set (5.18).

Finally, let $y = f(x)$ be a function that satisfies $F(x, y) = 0$. Then it follows from (5.17) that

$$f'(x) = -\frac{\partial_x F}{\partial_y F} = -\frac{1 - \sin x}{1 + 5\cos y} = -\frac{1 - \sin x}{1 + 5\cos f(x)}.$$

Although the function $f(x)$ is not known explicitly, (5.17) allows to compute its derivative.

Remark. In general, formula (5.17) can be memorized as follows. If f is a function as in the statement of Theorem 5.12 then, for any $x \in U$, we have

$$F(x, f(x)) \equiv 0.$$

Differentiating the function $g(x) = F(x, f(x))$ in x and using the chain rule, we obtain

$$0 = g'(x) = \partial_x F + (\partial_y F) f'(x),$$

whence (5.17) follows. This argument also proves (5.17) provided the differentiability of f is already known. However, in the actual proof, the differentiability of f is comes in a bundle with (5.17).

Proof of Theorem 5.12. For simplicity of notation, assume that $a = 0$ in $\mathbb{R}^n$ and $b = 0$ in $\mathbb{R}^m$. Also, choose in $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^{n+m}$ the ∞-norm.

Set $A = \partial_x F(0)$ and $B = \partial_y F(0)$. The we have the identity

$$F(x,y) = Ax + By + \varphi(x,y) \tag{5.19}$$

where φ is a function from W to $\mathbb{R}^m$ such that $\varphi(x,y) = o(\|(x,y)\|)$ as $(x,y) \to 0$ (indeed, by (5.16), $Ax + By$ is the differential of F at 0). Identity (5.19) can be considered as the definition of φ. We need the following properties of φ:

$$\varphi \in C^l(W, \mathbb{R}^m), \ \varphi(0) = 0, \quad \partial_y \varphi(0) = 0.$$

The first two are obvious. To obtain the third one, let us differentiate (5.19) in y so that

$$\partial_y F = B + \partial_y \varphi.$$

It follows that

$$\partial_y \varphi(0) = \partial_y F(0) - B = 0.$$

In the same way one shows $\partial_x \varphi(0) = 0$ but we do not need this.

The equation $F(x,y) = 0$ can be written as

$$Ax + By + \varphi(x,y) = 0$$

or, since the matrix B is invertible,

$$y = -B^{-1}(Ax + \varphi(x,y)) =: G(x,y).$$

Hence, the equation $F(x,y) = 0$ is equivalent to

$$y = G(x,y).$$

The idea of the proof is to show that, for any fixed x close enough to 0, $G(x,\cdot)$ can be considered as a contraction mapping in some neighborhood of 0 in $\mathbb{R}^m$ so that it has a unique fixed point y, which then satisfies the equation $y = G(x,y)$, that is, $F(x,y) = 0$.

Note that the function $G(x,y)$ has the following properties:

$$G \in C^l(W, \mathbb{R}^m),\ G(0) = 0,\quad \partial_y G(0) = 0,$$

which trivially follow from those of φ.

Set U and V to be the following balls

$$U = \{x \in \mathbb{R}^n : \|x\| < \delta\},\ V = \{y \in \mathbb{R}^m : \|y\| < \varepsilon\},$$

where positive constants ε and δ will be chosen to satisfied the following conditions:

1. Since the partial derivatives $\partial_{y_i} G_j$ are continuous and vanish at 0, for any $c > 0$ (to be specified later), there exists $\varepsilon > 0$ so that

$$\|x\| \le \varepsilon, \|y\| \le \varepsilon \implies (x,y) \in W \text{ and } |\partial_{y_i} G_j(x,y)| \le c. \tag{5.20}$$

2. Since the invertibility of the matrix $\partial_y F(0)$ means that $\det \partial_y F(0) \neq 0$ and $\det \partial_y F$ is obviously a continuous function of (x,y), ε can be chosen so small that in addition to (5.20)

$$\|x\| \le \varepsilon, \|y\| \le \varepsilon \implies \partial_y F(x,y) \text{ is invertible.} \tag{5.21}$$

3. Since function $G(x,0)$ is continuous in x, for any $\varepsilon > 0$ (in particular, for the ε defined above) there is $\delta \in (0, \varepsilon]$ such that

$$\|x\| \le \delta \implies \|G(x,0)\| < \frac{1}{2}\varepsilon. \tag{5.22}$$

Note that the condition $\delta \le \varepsilon$ ensures that in the above line $(x,0) \in W$.

Consider also the closed balls:

$$\overline{U} = \{x \in \mathbb{R}^n : \|x\| \le \delta\},\ \overline{V} = \{y \in \mathbb{R}^m : \|y\| \le \varepsilon\}.$$

It follows from the choice of ε and δ that $\overline{U} \times \overline{V} \subset W$.

Claim 1. *For an appropriate choice of c, we have for all $x \in \overline{U}$ and $y, y' \in \overline{V}$,*

$$\|G(x,y) - G(x,y')\| \le \frac{1}{2}\|y - y'\|. \tag{5.23}$$

Indeed, by the mean value theorem, for any component G_j of G, we have, for some $\xi \in [y, y']$,

$$|G_j(x,y) - G_j(x,y')| = |\partial_y G_j(x,\xi)(y - y')| = \left|\sum_{i=1}^m \partial_{y_i} G_j(x,\xi)(y_i - y_i')\right| \le cm\|y - y'\|_\infty,$$

where we have used (5.20). Hence, (5.23) follows if we choose $c = \frac{1}{2m}$.

Claim 2. *If $x \in \overline{U}$ and $y \in \overline{V}$ then $G(x,y) \in V$.*

Indeed, we have by (5.23) and (5.22)

$$\|G(x,y)\| \le \|G(x,y) - G(x,0)\| + \|G(x,0)\| < \frac{1}{2}\|y\| + \frac{1}{2}\varepsilon \le \varepsilon.$$

Hence, for any fixed $x \in \overline{U}$, we can consider the mapping $G(x,\cdot) : \overline{V} \to \overline{V}$. The set $\overline{V}$ is a closed subset of $\mathbb{R}^m$ and hence is complete as a metric space. By (5.23), this mapping is a contraction. Hence, by Theorem 4.6, there is a unique point $y \in \overline{V}$ that $G(x,y) = y$. Since such y exists for any $x \in \overline{U}$, we can define a function $f(x) = y$ that maps $\overline{U}$ to $\overline{V}$. It follows from the construction of f that $y = f(x)$ is equivalent to $G(x,y) = y$ and, hence, to $F(x,y) = 0$ (assuming that $x \in \overline{U}$ and $y \in \overline{V}$).

Let us show that function f is continuous in $\overline{U}$. Moreover, we prove the following stronger statement.

Claim 3. *There exists constant C such that*

$$\|f(x) - f(x')\| \le C\|x - x'\| \text{ for all } x, x' \in \overline{U}. \tag{5.24}$$

Indeed, we have

$$\begin{aligned} \|f(x) - f(x')\| &= \|G(x, f(x)) - G(x', f(x'))\| \\ &\le \|G(x, f(x)) - G(x', f(x))\| + \|G(x', f(x)) - G(x', f(x'))\| \end{aligned} \tag{5.25}$$

The second term in (5.23) is estimated by (5.23):

$$\|G(x', f(x)) - G(x', f(x'))\| \le \frac{1}{2}\|f(x) - f(x')\|.$$

To estimate the first term, let us consider each component G_j separately and use the mean value theorem: there is $\xi \in [x, x']$ such that

$$G_j(x, f(x)) - G_j(x', f(x)) = \partial_x G_j(\xi, f(x))(x - x') = \sum_{i=1}^{n} \partial_{x_i} G_j\left(\xi_j, f(x)\right)(x_i - x_i').$$

Since $\partial_{x_i} G_j(x,y)$ is a continuous function on a bounded closed set $\overline{U} \times \overline{V}$, by the maximal value theorem it is bounded on this set, say, by a constant C. Therefore, $\left|\partial_{x_i} G_j\left(\xi_j, f(x)\right)\right| \le C$ and we obtain

$$|G_j(x, f(x)) - G_j(x', f(x))| \le Cn\|x - x'\|$$

whence

$$\|G(x', f(x)) - G(x', f(x'))\| \le Cn\|x - x'\|.$$

Hence, it follows from (5.25) that

$$\|f(x) - f(x')\| \le Cn\|x - x'\| + \frac{1}{2}\|f(x) - f(x')\|.$$

Moving the term $\frac{1}{2}\|f(x) - f(x')\|$ to the left hand side and renaming the constant C, we obtain (5.24).

Claim 4. *The function $f : U \to \mathbb{R}^m$ is differentiable in U and*

$$f'(x) = -\left(\partial_y F\right)^{-1} \partial_x F, \tag{5.26}$$

where the right hand side is evaluated at the point $(x, f(x))$.

For any point $(x_0, y_0) \in W$, we have by the differentiability of F,

$$F(x,y) - F(x_0,y_0) = A(x-x_0) + B(y-y_0) + \varphi(x,y) \tag{5.27}$$

where $A = \partial_x F(x_0, y_0)$, $B = \partial_y F(x_0, y_0)$, and

$$\varphi(x,y) = o(\|x-x_0\| + \|y-y_0\|) \text{ as } x \to x_0 \text{ and } y \to y_0. \tag{5.28}$$

Fix $x_0 \in U$ and prove that $f(x)$ is differentiable at x_0. Set $y_0 = f(x_0)$, $y = f(x)$ for $x \in U$ and observe that

$$F(x,y) = F(x_0,y_0) = 0.$$

Recall that by (5.21), the matrix B is invertible. Hence, (5.27) yields

$$f(x) - f(x_0) = -B^{-1}A(x-x_0) - B^{-1}\varphi(x, f(x)).$$

We are left to prove that

$$B^{-1}\varphi(x, f(x)) = o(\|x-x_0\|) \text{ as } x \to x_0.$$

Since the matrix B^{-1} has a finite norm, it suffices to prove that

$$\varphi(x, f(x)) = o(\|x-x_0\|) \text{ as } x \to x_0. \tag{5.29}$$

Using (5.28), we obtain

$$\varphi(x, f(x)) = o(\|x-x_0\| + \|f(x) - f(x_0)\|) \text{ as } x \to x_0,$$

whence (5.29) follows because by (5.24)

$$\|f(x) - f(x_0)\| \le C\|x-x_0\|.$$

Claim 5. *Function f belongs to $C^l(U, \mathbb{R}^m)$.*

Induction in l. Inductive basis for $l = 1$. By (5.26), we have

$$(\partial_i f_j(x))_{i,j=1}^m = -(\partial_y F)^{-1} \partial_x F(x, f(x)). \tag{5.30}$$

The partial derivatives of F are continuous and so is $f(x)$. Hence, all the components in the right hand side of (5.30) are continuous, whence the continuity of the partial derivatives $\partial_i f_j$ follows. Hence, $f \in C^1$.

Inductive step from $l-1$ to l. Assuming that $F \in C^l$ let us prove that $f \in C^l$. By the inductive hypothesis, we have $f \in C^{l-1}$. Since the partial derivatives $\partial_x F$ and $\partial_y F$ are of the class C^{l-1}, it follows by Lemma 5.11 that the right hand side of (5.30) is in C^{l-1}, that is, any partial derivative $\partial_{x_i} f_j$ is in $C^{l-1}(U)$. This implies that $f \in C^l(U)$, which was to be proved. ■

Theorem 5.13 (The inverse function theorem) *Let Ω be open subset of $\mathbb{R}^n$ and $f : \Omega \to \mathbb{R}^n$ be a mapping of class C^l, $l \ge 1$. Assume that $f'(a)$ is invertible for some $a \in \Omega$, as an $n \times n$ matrix. Then there are open sets U and V in $\mathbb{R}^n$ such that $a \in U \subset \Omega$, $f(a) \in V$, and $f|_U$ is a bijection from U onto V. Moreover, the inverse mapping $f^{-1} : V \to U$ belongs to class C^l and*

$$\left(f^{-1}\right)'(y) = (f'(x))^{-1}, \tag{5.31}$$

for all $y \in V$, where $x = f^{-1}(y)$.

Remark. Let $f : I \to \mathbb{R}$ be a C^1 function on an open interval $I \subset \mathbb{R}$ such that $f'(x) \neq 0$ for all $x \in \mathbb{R}$. Theorem 5.12 ensures the invertibility of the function $f(x)$ *locally*, that is, in a neighborhood of any point in its image. However, in this case a stronger statement is true. Indeed, since $f'(x)$ is a continuous function, by the intermediate value theorem we have either $f'(x) > 0$ for all $x \in I$ or $f'(x) < 0$ for all $x \in I$. Then Theorem 4.12 from Analysis I yields that the inverse f^{-1} exists *globally*, that is, on the entire image $f(I)$. In higher dimensions, the existence of the global inverse require additional conditions, which we do not consider here.

Proof. The equation $y = f(x)$ is equivalent to the equation

$$F(x, y) := y - f(x) = 0.$$

Note that function $F(x, y)$ is defined for $x \in \Omega$ and $y \in \mathbb{R}^n$ and takes values in $\mathbb{R}^n$, that is, $F : \Omega \times \mathbb{R}^n \to \mathbb{R}^n$. Obviously, $F \in C^l$. Let us apply Theorem 5.12 with respect to the variable x. For that, check condition $\partial_x F \neq 0$. Indeed, $\partial_x F = -f'(x)$ and by hypothesis this is non-zero at the point $(a, f(a))$. Therefore, by Theorem 5.12, there are open neighborhoods U and V of the points a and $f(a)$, respectively, and a function $g : V \to U$ of class C^l such that

$$F(x, y) = 0 \Leftrightarrow x = g(y) \text{ for } x \in U \text{ and } y \in V,$$

that is,

$$y = f(x) \Leftrightarrow x = g(y) \text{ for } x \in U \text{ and } y \in V. \tag{5.32}$$

However, this does not mean yet that f maps U to V (although g maps V to U). To achieve that, we need to reduce U. Consider the set

$$U_0 = f^{-1}(V) \cap U$$

which is open by the continuity of f. Clearly, $a \in U_0$.

Let us show that f maps U_0 into V and g maps V into U_0. If $x \in U_0$ then $f(x) \in V$ by definition of U_0; hence, f maps U_0 into V. If $y \in V$ then $x = g(y) \in U$ whence by (5.32) $y = f(x)$ and $x \in f^{-1}(V)$. It follows that $x \in U_0$; hence, g maps V into U_0.

The both compositions $f \circ g$ and $g \circ f$ are defined by (5.32) are the identity mappings. Hence, $f : U_0 \to V$ and $g : V \to U_0$ are mutually inverse. We are left to rename U_0 to U to match the notation of the statement of the theorem.

To prove (5.31), differentiate the identity $g \circ f = I$ where $I : U \to U$ is the identity mapping. Using the chain rule, we obtain

$$\mathrm{id} = (g \circ f)' = g'(y) f'(x)$$

for all $y = f(x)$, whence $g'(y) = f'(x)^{-1}$. ■

Example. Consider the mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f(x, y) = \left(x^2 - y^2, 2xy\right).$$

The full derivative is given by the Jacobian matrix

$$f'(x, y) = J_f(x, y) = \begin{pmatrix} \partial_x f_1 & \partial_y f_1 \\ \partial_x f_2 & \partial_y f_2 \end{pmatrix} = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix},$$

and this is invertible if and only $\det J_f = 4(x^2 + y^2) \neq 0$ that is, when $(x, y) \neq 0$. By Theorem 5.12, the mapping f is invertible in a neighborhood of any non-zero point in its image.

Using complex numbers $z = x + iy$, observe that $f(z) = z^2$. This makes it clear that the full image $f(\mathbb{R}^2)$ is $\mathbb{R}^2$ because any complex number w has a square root z, that is, $z = w^2$. Furthermore, if $w \neq 0$ then there are two distinct roots z_1, z_2 of this equation such that $z_2 = -z_1$. This means that there is no global inverse f^{-1} even in $\mathbb{R}^2 \setminus \{0\}$ where the derivative f' is invertible at any point.

5.6 Surfaces in $\mathbb{R}^n$

5.6.1 Linear subspaces

Consider the following two ways of describing a subspace of $\mathbb{R}^n$.

1. For any linear mapping $C : \mathbb{R}^n \to \mathbb{R}^k$, the kernel

$$\ker C = \{v \in \mathbb{R}^n : Cv = 0\}$$

is a subspace of $\mathbb{R}^n$. Hence, for a given subspace S, one may try to find a linear mapping C so that $S = \ker C$. In this case one can also say that S is given by the equation $Cv = 0$.

2. For any linear mapping $A : \mathbb{R}^m \to \mathbb{R}^n$, the image

$$\operatorname{image} A = \{Au : u \in \mathbb{R}^m\}$$

is a subspace of $\mathbb{R}^n$. For a given subspace S, one may try to find A so that $S = \operatorname{image} A$. In this case, we obtain the *parametric equation* of S: $v = Au$, meaning that every vector $v \in S$ is represented in the form $v = Au$ where $u \in \mathbb{R}^m$ is a *parameter*.

Our aim here is to develop two similar approaches for describing *surfaces* in $\mathbb{R}^n$.

Some remarks about $\dim S$. By definition, the dimension of $\operatorname{image} A$ is the *rank* of A. Equivalently, $\operatorname{rank} A$ the maximal number of linearly independent rows of A and the maximal number of linearly independent columns of A (see Linear Algebra). Hence, if S is given in the parametric form $S = \operatorname{image} A$, then

$$\dim S = \operatorname{rank} A.$$

In the case when S is given by the equation $S = \ker C$ we have by a theorem from Linear Algebra,

$$\dim S = \dim \ker C = n - \operatorname{rank} C.$$

5.6.2 Parametric equation of a surface

Definition. A *parametric surface* in $\mathbb{R}^n$ is any continuous mapping $f : U \to \mathbb{R}^n$ where U is an open set in $\mathbb{R}^m$, $m \leq n$. The image of f, that is, the set $S = f(U)$ is called a *surface*. That is, S is defined by the parametric equation $x = f(u)$, $u \in U$.

The surface S is said to be differentiable if f is differentiable, and S is of class C^l if $f \in C^l$. A surface S of class C^1 is said to be *immersed* if the $n \times m$ matrix $f'(x)$ has the maximal rank m at any point $x \in U$. The number m is called the *dimension* of the immersed surface.

The set U is called the set of parameters of the given surface. Denoting by $u_1, ..., u_m$ the coordinates in $\mathbb{R}^m$ and by $x_1, ..., x_n$ the coordinates in $\mathbb{R}^n$, we obtain the parametric equations of the surface S in the coordinate form:

$$\begin{cases} x_1 = f_1(u_1, .., u_m) \\ ... \\ x_n = f_n(u_1, ..., u_m). \end{cases}$$

Example. *(Planes)* Let $U = \mathbb{R}^m$ and $f(u) = Au$ for some $n \times m$ matrix A. Then the linear subspace $S = f(U)$ is surface, and if $\operatorname{rank} A = m$ then this surface is immersed. Let f be an *affine* mapping, that is, $f(u) = Au + b$ where A is as above and $b \in \mathbb{R}^n$. Then the surface $S = f(\mathbb{R}^m)$ is called a *plane*. If $\operatorname{rank} A = m$ then the plane is immersed of dimension m.

Example. *(Curves)* Let U be an interval in $\mathbb{R}$ and consider a continuous mapping $f : U \to \mathbb{R}^n$. By definition, this is a parametric surface, but in the case $m = 1$ it is also called a *path*, and its image is called a *curve*. Denoting the parameter by t (instead of u) we obtain the equations of the path in the form $x_i = f_i(t)$. The parameter t can be interpreted as a time and the equation $x = f(t)$ describes the movement of a point in time in the space $\mathbb{R}^n$. The derivative $f'(t)$ has the meaning of the velocity of the point at time t.

Example. *(Graphs)* Let U be an open subset of $\mathbb{R}^m$ and $f : U \to \mathbb{R}^k$ be an arbitrary function. Consider the *graph* of f that is, the set

$$S = \{(u, w) \in \mathbb{R}^n : w = f(u), u \in U\}.$$

Here $n = m + k$, and the couple (u, w) with $u \in \mathbb{R}^m$ and $w \in \mathbb{R}^k$ is considered as a point in $\mathbb{R}^n$. Of course, this is a generalization of a familiar notion of the graph of a function $f : I \to \mathbb{R}$ where I is an interval in $\mathbb{R}$.

Since $S = \{(u, f(u)) : u \in U\}$, the graph S can be regarded as the image $\widetilde{f}(U)$ of the mapping $\widetilde{f} : U \to \mathbb{R}^n$ defined by

$$\widetilde{f}(u) = (u, f(u)).$$

Lemma 5.14 *Under the above notation, if the mapping f is of class C^l, $l \geq 1$, then the graph S of f is an immersed surface of dimension m of class C^l.*

Proof. Since $\widetilde{f} \in C^l$, the set $S = \widetilde{f}(U)$ is a surface of class C^l by definition. The derivative $\widetilde{f}'$ is the $n \times m$ matrix

$$\widetilde{f}'(u) = \left(\begin{array}{|c|}\hline \text{id} \\ \hline f'(u) \\ \hline \end{array}\right), \tag{5.33}$$

where id is the unit matrix $m \times m$ and $f'(u)$ is the $k \times m$ matrix. Since the first m rows of this matrix are linearly independent, it follows that $\operatorname{rank} \widetilde{f}'(u) = m$. Hence, the surface S is immersed of dimension m. ■

5.6.3 Tangent plane

Definition. (A tangent plane) Let $f : U \to \mathbb{R}^n$ be a parametric surface and $S = f(U)$. If f is differentiable at a point $u \in U$ then the *tangent plane* at point u to S is the plane in $\mathbb{R}^n$ given in parametric form by

$$T(h) = f(u) + f'(u)\, h,$$

where $h \in \mathbb{R}^m$ is a parameter.

Note that $T : \mathbb{R}^m \to \mathbb{R}^n$ is an affine mapping and, hence, determines a plane. Note that $T(h)$ is the first Taylor polynomial of f which is the best affine approximation to $f(u + h)$.

The tangent plane is denoted by $T_u S$ or, by some abuse of notation, by $T_x S$ where $x = f(u)$. It follows from the definition that

$$T_x S = x + \operatorname{image} f'(u).$$

In particular, $\dim T_x S = \operatorname{rank} f'(u)$. It follows that if S is an immersed surface of dimension m then $\dim T_x S = m$ for any $x \in S$.

For applications, it is convenient to write the parametric equation of the tangent plane in the form

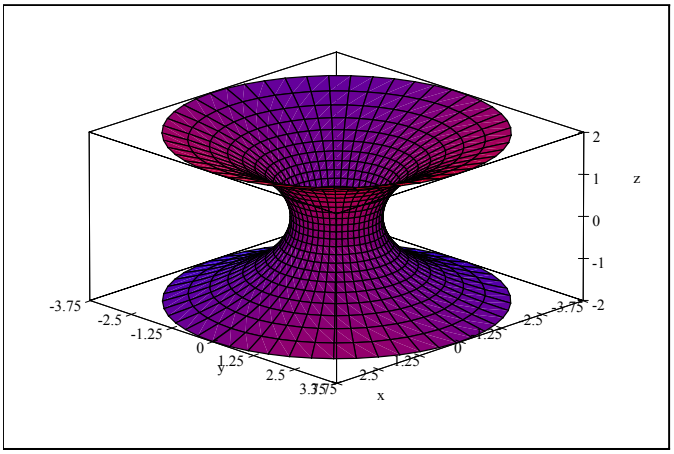
$$T(h) = f(u) + h_1 \partial_{u_1} f(u) + h_2 \partial_{u_2} f(u) + ... + h_m \partial_{u_m} f(u),$$

where the parameter $h_1, ..., h_m$ take all real values. Note that $\partial_{u_i} f$, $i = 1, ..., m$, are vectors in $\mathbb{R}^n$. If rank $f(u) = m$ then these vectors form a basis in image $f'(u)$.

Example. Consider the mapping $f : \mathbb{R}^2 \to \mathbb{R}^3$ given by in the parametric form by

$$f(u, v) = (\cosh u \cos v, \cosh u \sin v, u)$$

where $u \in \mathbb{R}$ and $v \in [0, 2\pi]$. This surface is called the *catenoid* and its is shown on the picture:



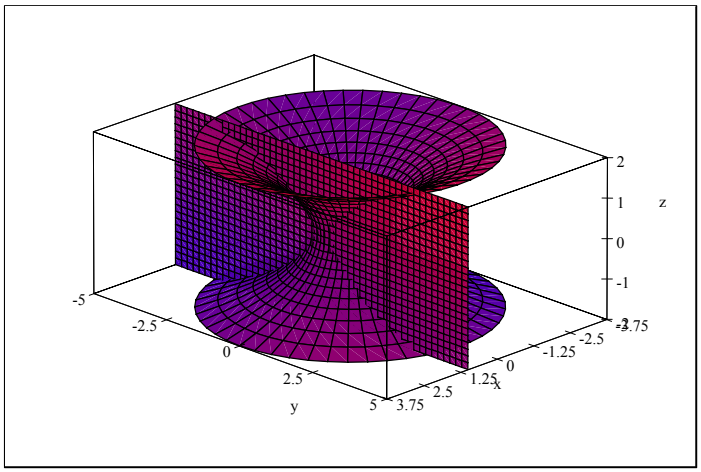
The parametric equation of the tangent plane at (u, v) is

$$\begin{aligned} T(t, s) &= f(u, v) + t\partial_u f + s\partial_v f \\ &= f(u, v) + t(\sinh u \cos v, \sinh u \sin v, 1) + s(-\cosh u \sin v, \cosh u \cos v, 0), \end{aligned}$$

where t and s take all real values. For example, at the point $u = v = 0$ we have $f(u, v) = (1, 0, 0)$, and

$$T(t, s) = (1, 0, 0) + t(0, 0, 1) + s(0, 1, 0) = (1, t, s).$$

This tangent plane is shown on the plot:



5.6.4 Surfaces given by the equation $F(x) = 0$

Theorem 5.15 *Let Ω be an open set in $\mathbb{R}^n$ and $F : \Omega \to \mathbb{R}^k$, $k \leq n$, be a function of class C^l, $l \geq 1$. If $F'(a)$ has the rank k at some point $a \in \Omega$ then there is an open set $V \subset \Omega$ containing a such that the set*

$$S = \{x \in V : F(x) = 0\}$$

is an immersed surface of the class C^l of dimension $m = n-k$. In particular, $\dim T_x S = m$ for all $x \in S$.

Furthermore, we have

$$T_x S = x + \ker F'(x).$$

In other words, $X \in T_x S$ is equivalent to $X - x \in \ker F'(x)$, that is

$$F'(x)(X - x) = 0.$$

This is the equation of the tangent plane $T_x S$.

Proof. Note that $F'(x)$ is an $k \times n$ matrix. The fact that $\operatorname{rank} F'(a) = k$ means that there is k columns of the matrix $F'(a)$ that form an invertible $k \times k$ matrix. Without loss of generality, assume that these are the last k columns. Denote

$$u = (x_1, ..., x_m) \in \mathbb{R}^m \text{ and } w = (x_{m+1}, ..., x_n) \in \mathbb{R}^k.$$

Then equation $F(x) = 0$ can be written in the form $F(u, w) = 0$ with the condition that the matrix $\partial_w F(a)$ is invertible. By the implicit function theorem, there are open sets $U \subset \mathbb{R}^m$, $W \subset \mathbb{R}^l$ such that $a \in U \times W \subset \Omega$, and a C^l function $f : U \to W$ such that

$$F(u, w) = 0 \Leftrightarrow w = f(u) \text{ for } u \in U \text{ and } w \in W. \tag{5.34}$$

Set $V = U \times W$. Then the set

$$S = \{x \in V : F(x) = 0\} = \{(u, w) \in U \times W : F(u, w) = 0\}$$

is the graph of the function $w = f(u)$, $u \in U$. By Lemma 5.14, S is an immersed surface of class C^l of dimension m, and S is given in the parametric form by $x = \widetilde{f}(u)$ where $\widetilde{f}(u) = (u, f(u))$ and $u \in U$.

Since the tangent plane at point $x = \widetilde{f}(u)$ is given by

$$T_x S = x + \operatorname{image} \widetilde{f}'(u),$$

it remains to show that

$$\operatorname{image} \widetilde{f}'(u) = \ker F'(x).$$

By (5.34), we have $F(u, f(u)) = 0$ for all $u \in U$, that is, $F \circ \widetilde{f} \equiv 0$. Differentiating this identity and using the chain rule, we obtain

$$F'(x)\widetilde{f}'(u) = 0,$$

which implies

$$\operatorname{image} \widetilde{f}'(u) \subset \ker F'(x). \tag{5.35}$$

Since $\operatorname{rank} F'(x) = k$ (we can assume that choosing U and W small enough), we obtain

$$\dim \ker F'(x) = n - \operatorname{rank} F'(x) = m,$$

while by Lemma 5.14

$$\dim \operatorname{image} \widetilde{f}'(u) = \operatorname{rank} \widetilde{f}'(u) = m.$$

Hence, the subspaces $\ker F'(x)$ and $\operatorname{image} \widetilde{f}'(u)$ have the same dimension, and the inclusion (5.35) implies that they are identical. ■

Example. Consider the equation $x_1^2 + ... + x_n^2 = 1$ that defines a unit sphere in the 2-norm. Then $F(x) = x_1^2 + ... + x_n^2 - 1$ and

$$F'(x) = (\partial_{x_1} F, ..., \partial_{x_n} F) = 2(x_1, ..., x_n)$$

whence the equation of the tangent plane is

$$\sum_{i=1}^{n} x_i (X_i - x_i) = 0.$$

In other words, $X - x$ is orthogonal to x as one expects.