Mathematik für Physiker II

Michael Dreher
Fachbereich für Mathematik und Statistik
Universität Konstanz

Studienjahr 2011/12

To those who do not know mathematics it is difficult to get across a real
feeling as to the beauty, the deepest beauty, of nature . . . .
If you want to learn about nature, to appreciate nature,
it is necessary to understand the language that she speaks in.

Richard Feynman, 1918 – 1988

Contents

Chapter 1

Differentiation in $\mathbb{R}^n$

1.1 Definitions of the Derivatives

Similarly as we have discussed the rules of differentiation in $\mathbb{R}^1$, we will now consider derivatives of functions going from $\mathbb{R}^m$ to $\mathbb{R}^n$. However, there are now several types of derivatives:

- derivatives (in the general sense of the word), also known as JACOBI matrices (Def. 1.1),
- partial derivatives (Def. 1.3),
- derivatives in a certain fixed direction (Def. 1.7).

These three types of derivatives coincide in case of $n = m = 1$.

Definition 1.1 (**Derivative, JACOBI[1] matrix**). *Let $G \subset \mathbb{R}^m$ be an open set, and $f\colon G \to \mathbb{R}^n$ be a function. We say that this function f is* differentiable[2] *at a point $x_0 \in G$ if a matrix $A \in \mathbb{R}^{n\times m}$ exists with the property that for x in a neighbourhood of x_0 we can write*

$$f(x) = f(x_0) + A(x - x_0) + R(x, x_0),$$

where the remainder term R is $\mathfrak{o}(\|x - x_0\|)$ for $x \to x_0$.

The matrix A is called derivative *or* Jacobi matrix[3].

The set of all functions $f\colon G \to \mathbb{R}^n$ that are continuously differentiable everywhere in G is denoted by $C^1(G \to \mathbb{R}^n)$. In this case, the derivative $A = A(x)$ depends continuously on $x \in G$.

Lemma 1.2. *The derivative is unique.*

Proof. Exercise: assume that there were another one, $\tilde{A}$. Subtract both defining equations, etc. □

Definition 1.3 (**Partial derivative**). *Let $G \subset \mathbb{R}^m$ be an open set and $f\colon G \to \mathbb{R}^n$ be an arbitrary function. Write f in the form $f = (f_1, \dots, f_n)^\top$. Fix a point $x_0 = (x_{0,1}, \dots, x_{0,m})^\top \in G$ and indices i, j with $1 \le i \le m$, $1 \le j \le n$. If the limit*

$$\lim_{h\to 0} \frac{1}{h} \left(f_j(x_{0,1}, \dots, x_{0,i-1}, x_{0,i} + h, x_{0,i+1}, \dots, x_n) - f_j(x_{0,1}, \dots, x_{0,i-1}, x_{0,i}, x_{0,i+1}, \dots, x_n) \right)$$

exists, then we say that the jth component of f has a partial derivative[4] *with respect to x_i, and this limit is denoted by*

$$\frac{\partial f_j}{\partial x_i}(x_0).$$

[1] CARL GUSTAV JAKOB JACOBI, 1804 – 1851

[2] differenzierbar

[3] Ableitung, Jacobi–Matrix

[4] partielle Ableitung

Proposition 1.4. *If a function f has a derivative $A = f'(x_0)$ at a point x_0, then all partial derivatives $\frac{\partial f_j}{\partial x_i}$ exist, and it holds*

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_m} \end{pmatrix}. \tag{1.1}$$

Proof. You should be able to do it yourselves. □

Warning: *The converse of this proposition is **wrong**, see below.*

Example: *In case of $n = 1$, the derivative $A = f'$ of a function $f\colon G \to \mathbb{R}^1$ is called the* gradient of f,

$$\operatorname{grad} f = \nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_m} \right).$$

Pay attention to the fact that $\operatorname{grad} f$ *is a row vector, not a column vector.*

Lemma 1.5. *If a function is differentiable at a point, then it is continuous at that point.*

Proof. The proof is exactly the same as in the one-dimensional case, which we have studied in the last semester. Just replace the modulus bars $|\cdot|$ with norm bars $\|\cdot\|$ everywhere. □

The mere existence of all partial derivatives does **not** imply the continuity of a function. There are examples of functions f, whose partial derivatives $\frac{\partial f_j}{\partial x_i}$ exist everywhere, yet the function f is discontinuous.

However, we can prove the equivalence of both types of derivatives if we sharpen the assumptions a bit:

Proposition 1.6. *Suppose that a function $f\colon G \to \mathbb{R}^n$ has partial derivatives everywhere in G, and that these partial derivatives are* continuous*.*

Then the function f is differentiable everywhere in G, and relation (1.1) holds.

Proof. Suppose, for simplicity, that $m = 2$ and $n = 1$. The general case can be proved similarly. Fix $x_0 = (x_{0,1}, x_{0,2})^\top \in \mathbb{R}^2$, and write $x = (x_1, x_2)^\top \in \mathbb{R}^2$. We want to show that

$$f(x) = f(x_0) + \frac{\partial f}{\partial x_1}(x_0) \cdot (x_1 - x_{0,1}) + \frac{\partial f}{\partial x_2}(x_0) \cdot (x_2 - x_{0,2}) + R(x, x_0), \tag{1.2}$$

with $R(x, x_0) = \mathfrak{o}(\|x - x_0\|)$ for $x \to x_0$. By the mean value theorem (of 1D calculus), we deduce that

$$\begin{aligned} f(x) = f(x_1, x_2) &= f(x_{0,1}, x_{0,2}) + (f(x_1, x_2) - f(x_{0,1}, x_2)) + (f(x_{0,1}, x_2) - f(x_{0,1}, x_{0,2})) \\ &= f(x_0) + \frac{\partial f}{\partial x_1}(\xi_1, x_2) \cdot (x_1 - x_{0,1}) + \frac{\partial f}{\partial x_2}(x_{0,1}, \xi_2) \cdot (x_2 - x_{0,2}), \end{aligned}$$

where ξ_1 is between x_1 and $x_{0,1}$; and ξ_2 is between x_2 and $x_{0,2}$. Now the continuity of the derivatives comes into play:

$$\begin{aligned} \frac{\partial f}{\partial x_1}(\xi_1, x_2) &= \frac{\partial f}{\partial x_1}(x_{0,1}, x_{0,2}) + \tilde{R}_1(x, x_0, \xi_1), \\ \frac{\partial f}{\partial x_2}(x_{0,1}, \xi_2) &= \frac{\partial f}{\partial x_2}(x_{0,1}, x_{0,2}) + \tilde{R}_2(x, x_0, \xi_2), \end{aligned}$$

where $\lim_{x \to x_0} \tilde{R}_j(x, x_0, \xi_j) = 0$. This gives us (1.2). □

Finally, the derivative of a function in a certain direction can be defined in a very similar way as the partial derivative.

Definition 1.7 (**Directional derivative**[5]). *Let $G \subset \mathbb{R}^m$ be an open set and $f\colon G \to \mathbb{R}^n$ be an arbitrary function. Choose a unit vector $e \in \mathbb{R}^m$, $\|e\| = 1$. If the limit*

$$\lim_{h\to 0} \frac{1}{h} \left(f(x_0 + he) - f(x_0) \right)$$

exists, then we say that the function f has a derivative at the point $x_0 \in G$ in direction e, and this limit is denoted by

$$\frac{\partial f}{\partial e}(x_0).$$

The partial derivatives are simply directional derivatives in the directions given by the vectors $(1, 0, \dots, 0)^\top$, $(0, 1, 0, \dots, 0)^\top$, $\dots$, $(0, \dots, 0, 1)^\top$.

Proposition 1.8 (**Directional derivative**). *Let $f\colon G \to \mathbb{R}^n$ be a continuously differentiable function, $x_0 \in G$, and $e \in \mathbb{R}^m$ a unit vector. Then the derivative of f at x_0 in direction e can be computed by*

$$\frac{\partial f}{\partial e}(x_0) = f'(x_0)e,$$

where the last multiplication is of the form "matrix times vector".

Proof. The proof requires the so-called chain rule, and therefore we postpone it. □

Proposition 1.9. *The gradient of $f \in C^1(G \to \mathbb{R}^1)$ points into the direction of steepest ascent.*

Proof. Fix $x_0 \in G$, and let $x \in G$ be close to x_0. We know that

$$f(x) - f(x_0) = \operatorname{grad} f(x_0) \cdot (x - x_0) + \mathfrak{o}(\|x - x_0\|),$$

and the remainder term becomes negligible for $x \to x_0$. By the Cauchy–Schwarz inequality, we have

$$|\operatorname{grad} f(x_0) \cdot (x - x_0)| \le \|\operatorname{grad} f(x_0)\| \, \|x - x_0\|$$

with equality if the vectors $\operatorname{grad} f(x_0)$ and $x - x_0$ are parallel. □

Proposition 1.10. *The gradient of a function is perpendicular to its level sets.*

Proof. Exercise. □

Examples:

- *If $x(t) = (x_1(t), x_2(t), x_3(t))^\top$ denotes the position of a particle at time t, then*

  $$\dot{x}(t) = \begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix}$$

  denotes the velocity of that particle at time t.

- *If $\theta = \theta(x) = \theta(x_1, x_2, x_3)^\top$ denotes the temperature at the point $x = (x_1, x_2, x_3)^\top$, then*

  $$\nabla\theta = \operatorname{grad} \theta = \left(\frac{\partial\theta}{\partial x_1}, \frac{\partial\theta}{\partial x_2}, \frac{\partial\theta}{\partial x_3} \right)$$

  is the temperature gradient. This row vector describes "how the temperature changes from one point to the next" via the relation

  $$\Delta\theta = \theta(x) - \theta(x_0) \approx \mathrm{d}\theta = (\operatorname{grad} \theta) \cdot (x - x_0).$$

  Note that Δ (capital Delta) is not the Laplacian $\triangle$.

[5] Richtungsableitung

- *If $u = (u_1(x), u_2(x), u_3(x))^\top$ denotes the velocity of a fluid at position $x = (x_1, x_2, x_3)^\top$, then*

$$\nabla u = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{pmatrix}$$

describes "how u changes from one point to the next" by

$$\Delta u = u(x) - u(x_0) \approx \mathrm{d}u = (\nabla u) \cdot \mathrm{d}x.$$

Definition 1.11 (**Total differential**). *Set $\Delta x = \mathrm{d}x = x - x_0$ and $\Delta f = f(x) - f(x_0)$, $\mathrm{d}f = f'(x_0) \cdot \mathrm{d}x$. The (column) vector $\mathrm{d}f$ is the* (total) differential of f at x_0[6].

Differentiability means the following: if $\| \mathrm{d}x\|$ is small enough, then (in general)

$$\|\Delta f - \mathrm{d}f\| \ll \| \mathrm{d}f\| ,$$

where $\ll$ means "much smaller than".

This holds, of course, only in the general case, which is $\| \mathrm{d}f\| \neq 0$.

1.2 Calculation Rules

How do the above defined derivatives interact with the usual arithmetical operations, that are

- addition of functions and multiplication with scalars,
- multiplication of functions,
- composition of functions ?

The addition and multiplication with scalars are easy:

Proposition 1.12. *The mapping that maps a function $f \in C^1(G \to \mathbb{R}^n)$ to its derivative $f' \in C(G \to \mathbb{R}^{n\times m})$ is a homomorphism.*

Proof. We only have to show that

$$(f + g)' = f' + g',$$
$$(cf)' = c \cdot f'.$$

The proof can be obtained by copying from the one-dimensional case. □

Concerning the multiplication of functions, we have to be careful. We cannot copy the old proof blindly, since the multiplication of matrices is in general not commutative.

Proposition 1.13 (**Product rule**). *Let $G \subset \mathbb{R}^m$ be an open set, and $u, v \in C^1(G \to \mathbb{R}^n)$ be continuously differentiable functions. Define a function $f\colon G \to \mathbb{R}^1$ by the formula*

$$f(x) = u(x)^\top v(x) = v(x)^\top u(x), \qquad x \in G.$$

Then f is continuously differentiable in G, $f \in C^1(G \to \mathbb{R}^1)$, and its gradient is given by

$$\operatorname{grad} f(x_0) = \nabla f(x_0) = f'(x_0) = u(x_0)^\top v'(x_0) + v(x_0)^\top u'(x_0).$$

[6] totales Differential von f in x_0

Proof. We start with

$$\begin{aligned} u(x) &= u(x_0) + u'(x_0) \cdot (x - x_0) + R_u(x, x_0), \\ v(x) &= v(x_0) + v'(x_0) \cdot (x - x_0) + R_v(x, x_0). \end{aligned}$$

Question: Which format do u, v, u' and v' have ?

We want to write down a similar expansion for $f(x)$; the factor in front of $(x - x_0)$ will then be the desired derivative. Remember that $u^\top v = v^\top u$. Here we go:

$$\begin{aligned} f(x) &= f(x_0) + (f(x) - f(x_0)) \\ &= f(x_0) + u(x)^\top (v(x) - v(x_0)) + v(x_0)^\top (u(x) - u(x_0)) \\ &= f(x_0) + u(x)^\top \left(v'(x_0) \cdot (x - x_0) + R_v(x, x_0)\right) + v(x_0)^\top \left(u'(x_0) \cdot (x - x_0) + R_u(x, x_0)\right) \\ &= f(x_0) + \left(u(x)^\top v'(x_0) + v(x_0)^\top u'(x_0)\right) \cdot (x - x_0) + \mathfrak{o}(\|x - x_0\|) \\ &= f(x_0) + \left(u(x_0)^\top v'(x_0) + v(x_0)^\top u'(x_0)\right) \cdot (x - x_0) + \mathfrak{o}(\|x - x_0\|). \end{aligned}$$

Here we have used in the last step that $u(x) = u(x_0) + \mathfrak{O}(\|x - x_0\|)$. □

Proposition 1.14 (**Chain rule**). *Let $G \subset \mathbb{R}^l$ and $H \subset \mathbb{R}^m$ be open sets, and consider 2 functions $u \in C^1(G \to \mathbb{R}^m)$, $v \in C^1(H \to \mathbb{R}^n)$ with $W_u \subset D_v = H$. Then the composed function $f = f(x) = (v \circ u)(x) = v(u(x))$ is differentiable, $f \in C^1(G \to \mathbb{R}^n)$, and its derivative is given by*

$$f'(x) = (v'(u(x))) \cdot u'(x), \qquad x \in G.$$

Proof. The proof can be copied from the 1D situation, almost word-by-word. Be careful to not divide by vectors. Divide by norms of vectors instead. □

Question: Which format do the terms $f'(x)$, $v'(u(x))$ and $u'(x)$ have ?

Example: *If $f \in C^1(\mathbb{R}^n \to \mathbb{R}^1)$ is scalar and $x = x(t) \in C^1(\mathbb{R}^1 \to \mathbb{R}^n)$ is a vector, then $g = g(t) = f(x(t)) \in C^1(\mathbb{R}^1 \to \mathbb{R}^1)$ with the derivative*

$$\dot{g}(t) = (\operatorname{grad} f)(x(t)) \cdot \dot{x}(t) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x(t)) \cdot \frac{\partial x_j}{\partial t}(t).$$

Example: *The position of a moving particle in the plane is given by*

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}_{cart.}$$

in Cartesian coordinates. The velocity vector is then $(\dot{x}(t), \dot{y}(t))^\top_{cart.}$. If you express this in polar coordinates, you have

$$\begin{aligned} x(t) &= r(t) \cos \varphi(t) = x(r(t), \varphi(t)), \\ y(t) &= r(t) \sin \varphi(t) = y(r(t), \varphi(t)), \\ \dot{x} &= \frac{\partial x}{\partial r} \cdot \frac{\partial r}{\partial t} + \frac{\partial x}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial t} = \cos(\varphi)\dot{r} - r \sin(\varphi)\dot{\varphi}, \\ \dot{y} &= \frac{\partial y}{\partial r} \cdot \frac{\partial r}{\partial t} + \frac{\partial y}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial t} = \sin(\varphi)\dot{r} + r \cos(\varphi)\dot{\varphi}, \\ \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix}_{cart.} &= \begin{pmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{pmatrix} \begin{pmatrix} \dot{r} \\ \dot{\varphi} \end{pmatrix} =: \frac{\partial(x, y)}{\partial(r, \varphi)} \begin{pmatrix} \dot{r} \\ \dot{\varphi} \end{pmatrix}. \end{aligned}$$

The matrix $\frac{\partial(x,y)}{\partial(r,\varphi)}$ is also known as fundamental matrix. *It is simply the derivative (Jacobi matrix) of that function, which maps $(r, \varphi)^\top$ to $(x, y)^\top$.*

Corollary 1.15 (**Proof of Proposition 1.8**). *Define a function $l = l(h) = x_0 + h \cdot e$, where $e \in \mathbb{R}^m$, $\|e\| = 1$, and $h \in \mathbb{R}^1$. In other words, the function l maps $\mathbb{R}^1$ into $\mathbb{R}^m$. Then the directional derivative can by computed by*

$$\frac{\partial f}{\partial e}(x_0) = \frac{\partial}{\partial h} f(l(h))\big|_{h=0} = f'(l(0)) \cdot l'(0) = \operatorname{grad} f(x_0) \cdot e.$$

Straight lines in $\mathbb{R}^m$ are a useful tool and let us play with them a bit longer. Consider two points $x, y \in G$. Then the straight line connecting them is the set

$$l(x,y) = \{z \in \mathbb{R}^m \colon z = x + t(y-x),\ 0 \le t \le 1\}.$$

The set G is said to be *convex*[7] if, for each pair (x, y) of points of G, the connecting line $l(x, y)$ belongs completely to G.

Proposition 1.16 (Mean value theorem in $\mathbb{R}^m$). *Let G be a convex open set in $\mathbb{R}^m$, and let $f \in C^1(G \to \mathbb{R}^1)$.*

Then: for each pair $(x, y) \in G^2$, there is a point $\xi \in G$ on the straight line connecting x and y, such that

$$f(y) - f(x) = \operatorname{grad} f(\xi) \cdot (y - x).$$

Proof. Define a function $l \colon [0,1] \to G$ by $l(t) = x + t(y-x)$, and put $g = g(t) = f(l(t))$. Then we have, from the 1D mean value theorem,

$$f(y) - f(x) = g(1) - g(0) = g'(\tau)(1-0),$$

for some $0 < \tau < 1$. We compute now $g'(\tau)$ by the chain rule:

$$g'(\tau) = f'(l(\tau)) \cdot l'(\tau) = \operatorname{grad} f(\xi) \cdot (y-x),$$

where we have introduced $\xi := l(\tau)$; and the proof is complete. □

The Cauchy–Schwarz inequality gives us the convenient estimate

$$\|f(y) - f(x)\| \le M \|y - x\|,$$

where we have set $M = \sup\{\|\operatorname{grad} f(\xi)\| : \xi \in l(x,y)\}$. Moreover, we can conclude that

$$\operatorname{grad} f(x) \equiv 0 \text{ in } G \Longrightarrow f \equiv \text{const. in } G$$

provided that the open set G is connected.

Warning: *In the above mean value theorem, one cannot replace $f \in C^1(G \to \mathbb{R}^1)$ by $f \in C^1(G \to \mathbb{R}^n)$. You are invited to find counter-examples yourselves. How about looking at the unit circle ?*

However, an integrated version of the mean value theorem holds in higher dimensions:

Proposition 1.17 (Integrated mean value theorem). *Let G be a convex open set in $\mathbb{R}^m$ and $f \in C^1(G \to \mathbb{R}^n)$. Then we have the following formula for each pair $(x, y) \in G^2$:*

$$f(y) - f(x) = \left(\int_{t=0}^{t=1} f'(x + t(y-x))\,\mathrm{d}t \right) \cdot (y - x).$$

Proof. Consider the first component f_1 of f. Write $g_1(t) = f_1(x + t(y-x))$. By the main theorem of calculus,

$$f_1(y) - f_1(x) = g_1(1) - g_1(0) = \int_{t=0}^{t=1} g_1'(t)\,\mathrm{d}t = \int_{t=0}^{t=1} (\operatorname{grad} f_1(x + t(y-x))) \cdot (y-x)\,\mathrm{d}t.$$

You can extract the factor $y - x$ out of the integral, and then consider the other components of f in the same way. □

If we restrict a function $f \colon G \to \mathbb{R}^1$ to a straight line connecting two points of G, then we obtain a function which only depends on a one-dimensional parameter $t \in [0,1]$. It is interesting to apply the usual 1D calculus—for instance the Taylor formula—to this restricted function. Then we will obtain a Taylor formula in higher dimensions. For this, we will need higher order derivatives, which are so important that they deserve a section of their own.

[7] konvex

1.3 Derivatives of Higher Order

Definition 1.18 (Higher order derivatives). *Let $f \in C^1(G \to \mathbb{R}^1)$ be a continuously differentiable function; and suppose that the partial derivatives of f are again continuously differentiable. Then we say that f is twice partially differentiable and write $f \in C^2(G \to \mathbb{R}^1)$. The second order partial derivatives of f are written as*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

For the mixed derivatives, the order of differentiation does not matter:

Proposition 1.19 (Theorem of Schwarz[8]**).** *Let $f \in C^2(G \to \mathbb{R})$ and $x_0 \in G$. Then*

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x_0) = \frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i} f(x_0), \qquad 1 \le i, j \le m.$$

Proof. Assume for simplicity of notation that $m = 2$ and $x_0 = 0$. We will now show that

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

Choose small numbers Δx, Δy and consider the rectangle with the corners $(0,0)$, $(\Delta x, 0)$, $(\Delta x, \Delta y)$, $(0, \Delta y)$ (draw a picture !). We define a number

$$S = f(\Delta x, \Delta y) + f(0,0) - f(\Delta x, 0) - f(0, \Delta y)$$

and represent it in two ways. On the one hand, we have

$$S = (f(\Delta x, \Delta y) - f(0, \Delta y)) - (f(\Delta x, 0) - f(0,0)) = G(\Delta y) - G(0),$$

where we have introduced $G(\eta) = f(\Delta x, \eta) - f(0, \eta)$. By the 1D mean value theorem, there is a number τ_G with $0 < \tau_G < 1$ and

$$\begin{aligned} S = G'(\tau_G \Delta y)\Delta y &= \left(\frac{\partial f}{\partial y}(\Delta x, \tau_G \Delta y) - \frac{\partial f}{\partial y}(0, \tau_G \Delta y)\right) \Delta y \\ &= \left(\frac{\partial}{\partial x}\frac{\partial}{\partial y} f\right)(\sigma_G \Delta x, \tau_G \Delta y) \cdot \Delta x \cdot \Delta y, \end{aligned}$$

where we have applied the 1D mean value theorem for the second time.

On the other hand, we have

$$S = (f(\Delta x, \Delta y) - f(\Delta x, 0)) - (f(0, \Delta y) - f(0,0)) = H(\Delta x) - H(0)$$

with $H(\xi) = f(\xi, \Delta y) - f(\xi, 0)$. By applying the mean value theorem two times more, we find that

$$\begin{aligned} S = H'(\sigma_H \Delta x)\Delta x &= \left(\frac{\partial f}{\partial x}(\sigma_H \Delta x, \Delta y) - \frac{\partial f}{\partial x}(\sigma_H \Delta x, 0)\right) \Delta x \\ &= \left(\frac{\partial}{\partial y}\frac{\partial}{\partial x} f\right)(\sigma_H \Delta x, \tau_H \Delta y) \cdot \Delta x \cdot \Delta y. \end{aligned}$$

Both representations of S together give us

$$\left(\frac{\partial}{\partial x}\frac{\partial}{\partial y} f\right)(\sigma_G \Delta x, \tau_G \Delta y) = \left(\frac{\partial}{\partial y}\frac{\partial}{\partial x} f\right)(\sigma_H \Delta x, \tau_H \Delta y).$$

Now we send Δx and Δy to 0. The continuity of the second order derivatives then completes the proof. □

[8] Hermann Amandus Schwarz, 1843 – 1921

The second order derivatives of a function $f \in C^2(G \to \mathbb{R})$ can be arranged into an $m \times m$ matrix, the so–called *Hessian*[9] [10] of f:

$$Hf(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_m}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_m^2}(x) \end{pmatrix}.$$

By the Schwarz theorem, this matrix is symmetric.

Because the derivations with respect to different directions commute, we are allowed to introduce so–called multi–indices:

Definition 1.20 (Multi–index). *A vector $\alpha = (\alpha_1, \dots, \alpha_m)$ with $\alpha_j \in \mathbb{N}_0$ is named a* multi–index[11]*. Let $h = (h_1, \dots, h_m)^\top$ be a vector of real numbers. Then we define*

$$\begin{aligned} |\alpha| &= \alpha_1 + \dots + \alpha_m, \\ \alpha! &= \alpha_1! \cdot \ldots \cdot \alpha_m!, \\ h^\alpha &= h_1^{\alpha_1} \cdot \ldots \cdot h_m^{\alpha_m}, \\ \partial_x^\alpha &= \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdot \ldots \cdot \left(\frac{\partial}{\partial x_m}\right)^{\alpha_m}. \end{aligned}$$

This notation might look a bit complicated at first. However, it enables us to write down a Taylor formula in exactly the same way as in the 1D case.

Proposition 1.21 (TAYLOR[12]**–formula).** *Let $G \subset \mathbb{R}^m$ be an open and convex set, and suppose that a function $f \in C^{N+1}(G \to \mathbb{R}^1)$ is given. Then there is, for each pair $(x_0, x) \in G^2$, a point ξ on the connecting line $l(x_0, x)$, such that*

$$\begin{aligned} f(x) &= \sum_{|\alpha| \le N} \frac{1}{\alpha!} \left(\partial_x^\alpha f\right)(x_0) \cdot (x - x_0)^\alpha + R_N(x, x_0), \\ R_N(x, x_0) &= \sum_{|\alpha| = N+1} \frac{1}{\alpha!} \left(\partial_x^\alpha f\right)(\xi) \cdot (x - x_0)^\alpha. \end{aligned}$$

Proof. Put $l = l(t) = x_0 + t(x - x_0)$ for $0 \le t \le 1$ and $g = g(t) = f(l(t))$. Then we have $g(0) = f(x_0)$ and $g(1) = f(x)$. The 1D Taylor formula gives us a number τ, $0 < \tau < 1$, such that

$$g(1) = \sum_{k=0}^N \frac{1}{k!} g^{(k)}(0) + \frac{1}{(N+1)!} g^{(N+1)}(\tau).$$

Now we compute the terms with $k = 1$ and $k = 2$:

$$\begin{aligned} g'(t) &= f'(l(t)) \cdot l'(t) = \sum_{|\alpha|=1} (\partial_x^\alpha f)(l(t))(x - x_0)^\alpha, \\ g''(t) &= \sum_{|\alpha|=1} \left(\sum_{|\beta|=1} (\partial_x^\beta \partial_x^\alpha f)(l(t))(x - x_0)^\beta \right) (x - x_0)^\alpha = \sum_{|\gamma|=2} \frac{2!}{\gamma!} (\partial_x^\gamma f)(l(t))(x - x_0)^\gamma. \end{aligned}$$

By induction, one can show that

$$g^{(k)}(t) = \sum_{|\gamma|=k} \frac{k!}{\gamma!} (\partial_x^\gamma f)(l(t))(x - x_0)^\gamma,$$

where we have omitted an explanation how the factor $\frac{k!}{\gamma!}$ appears. It is just advanced combinatorics …

The proof is complete. □

[9] Hesse–Matrix

[10] LUDWIG OTTO HESSE, 1811 – 1874, also known for the Hesse normal form of analytical geometry

[11] Multiindex

[12] BROOK TAYLOR, 1685 – 1731

Remark 1.22. *Observe that we have proved the Taylor formula only for functions $f\colon G \to \mathbb{R}^n$ with $n = 1$. This formula with the above representation of the remainder term R_N will be wrong for higher n. The reason is that the 1D Taylor formula (which we have used in the proof) needs the mean value theorem, which is not valid for $n \geq 2$. However, if we only need $R_N = \mathfrak{O}(\|x - x_0\|^{N+1})$, then any $n \in \mathbb{N}$ is admissible, as it can be seen from the integrated mean value theorem, for instance.*

Generally, one uses the Taylor formula in one of the following forms:

$$f(x) = f(x_0) + \mathfrak{O}(\|x - x_0\|), \qquad n \geq 1, \tag{1.3}$$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \mathfrak{O}(\|x - x_0\|^2), \qquad n \geq 1,$$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^\top Hf(x_0)(x - x_0) + \mathfrak{O}(\|x - x_0\|^3), \qquad n = 1, \tag{1.4}$$

where $Hf(x_0)$ is the Hessian of f at the point x_0.

From (1.4), it is clear how to find extremal values of a function f:

necessary condition If a function f has an extremum at a point x_0, then $f'(x_0) = 0$,

sufficient condition If $f'(x_0) = 0$ and the Hessian of f at x_0 is positive definite, then f has a minimum at x_0. If $f'(x_0) = 0$ and the Hessian of f is negative definite, then f has a maximum at x_0.

A real symmetric matrix A is said to be *positive definite* if $\eta^\top A\eta > 0$ for each vector $\eta \in \mathbb{R}^m \setminus \{0\}$. A real symmetric matrix A is called *negative definite* if $\eta^\top A\eta < 0$ for each vector $\eta \in \mathbb{R}^m \setminus \{0\}$. An equivalent description is: a real matrix A is positive definite if A is symmetric and all eigenvalues of A are positive. A is negative definite if A is symmetric and all eigenvalues of A are negative. If some eigenvalues of A are positive and some are negative, then A is called *indefinite*. In this case, the function f has neither a maximum nor a minimum at the point under consideration, but a so–called saddle-point. An introduction to the theory of eigenvalues of matrices will be given later, in Section 4.5.

As an example of a Taylor expansion, we wish to study the function f which maps a matrix $A \in \mathbb{R}^{m\times m}$ to its inverse A^{-1}. One can imagine that the $m \times m$ entries a_{ij} of A are written as a super-column with m^2 entries, and then f maps from some subset of $\mathbb{R}^{m^2}$ into $\mathbb{R}^{m^2}$. Of course, the big challenge is how to write down the computations without being lost in a jungle of formulas.

One can easily imagine the following: if a matrix A_0 is invertible and another matrix A is "close" to A_0, then also A should be invertible; and the inverses A_0^{-1} and A^{-1} should also be close to each other. Then natural questions are:

- what means "A is close to A_0" ?
- can we compare the distance of the inverses somehow with the distances of the original matrices ?

The key tool here is a matrix norm, which is the following. Fix a norm on $\mathbb{R}^m$, for instance $\|x\| := \sqrt{x_1^2 + \cdots + x_m^2}$. Then we define an associated matrix norm on $\mathbb{R}^{m\times m}$ via $\|A\| := \sqrt{\sum_{i,j=1}^m a_{ij}^2}$. The crucial fact is that

$$\|Ax\| \leq \|A\| \, \|x\| \,, \qquad \|AB\| \leq \|A\| \, \|B\| \,,$$

for each vector $x \in \mathbb{R}^m$ and all matrices $A, B \in \mathbb{R}^{m\times m}$. This is the reason why one calls this matrix norm *associated* to the given vector norm. In a sense, the matrix norm is *compatible* to all the operations where a matrix is involved (multiplying a matrix by a number, adding two matrices, multiplying a matrix by a vector, multiplying two matrices). Now our result is the following, and we will use it for proving the inverse function theorem.

Lemma 1.23. *Suppose that A_0 is an invertible matrix from $\mathbb{R}^{m\times m}$, and A is close to A_0 in the sense of $\left\|A_0^{-1}(A_0 - A)\right\| \leq 1/2$. Then also A is invertible, we have the estimate*

$$\left\|A^{-1} - A_0^{-1}\right\| \leq 2\left\|A_0^{-1}\right\|^2 \|A - A_0\| \,, \tag{1.5}$$

as well as the converging Taylor series

$$A^{-1} = \left(\sum_{k=0}^{\infty}(A_0^{-1}(A_0 - A))^k\right) A_0^{-1}. \tag{1.6}$$

Proof. For a start, we take a matrix B with $\|B\| \le 1/2$. Then we have $\left\|B^k\right\| \le \|B\|^k \le (1/2)^k$, and therefore the series

$$I + B + B^2 + B^3 + \dots$$

converges, even absolutely. This is the famous NEUMANN[13] series. The limit of the series is $(I - B)^{-1}$, and you can prove this limit in exactly the same way as you proved the formula $1+q+q^2+\dots = 1/(1-q)$ (for $q \in \mathbb{C}$ with $|q| < 1$) of the geometric series in school.

And you also have $\left\|(I - B)^{-1}\right\| \le \sum_{k=0}^{\infty} \left\|B^k\right\| \le \sum_{k=0}^{\infty} \|B\|^k \le \sum_{k=0}^{\infty} 2^{-k} = 2$.

Now we take the above matrices A_0 and A, and we put $B := A_0^{-1}(A_0 - A)$. Then we have $\|B\| \le \frac{1}{2}$ and

$$A = A_0 - (A_0 - A) = A_0(I - A_0^{-1}(A_0 - A)) = A_0(I - B),$$

which is the product of two invertible matrices, and consequently

$$A^{-1} = (I - B)^{-1}A_0^{-1} = \left(\sum_{k=0}^{\infty} B^k\right) A_0^{-1},$$

which is just (1.6). This is the desired Taylor expansion of that function f which maps A to A^{-1} ! The first term in this Taylor formula is $B^0 A_0^{-1} = A_0^{-1}$, and therefore

$$A^{-1} - A_0^{-1} = \left(\sum_{k=1}^{\infty} B^k\right) A_0^{-1} = B\left(\sum_{k=0}^{\infty} B^k\right) A_0^{-1} = B(I - B)^{-1}A_0^{-1},$$

which leads us to the estimate

$$\left\|A^{-1} - A_0^{-1}\right\| \le \|B\| \left\|(I - B)^{-1}\right\| \left\|A_0^{-1}\right\| \le \left\|A_0^{-1}\right\| \|A_0 - A\| \cdot 2 \cdot \left\|A_0^{-1}\right\|,$$

and this is exactly (1.5). □

1.4 Differential Operators of Vector Analysis

Definition 1.24 (Laplace–operator, divergence, rotation). *Let $\Omega \subset \mathbb{R}^n$ be an open set, and $f\colon \Omega \to \mathbb{R}^3$, $\varphi\colon \Omega \to \mathbb{R}^1$ be functions from C^1 or C^2. Then we define the operators $\triangle$ (*LAPLACE[14]*–operator),* div *(divergence–operator) and, in case $n = 3$,* rot *(rotation operator):*

$$\triangle\,\varphi(x) := \sum_{j=1}^{n} \frac{\partial^2 \varphi}{\partial x_j^2}(x),$$

$$\operatorname{div} f(x) := \sum_{j=1}^{n} \frac{\partial f_j}{\partial x_j}(x),$$

$$\operatorname{rot} f(x) := \begin{pmatrix} \frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3} \\ \frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2} \end{pmatrix}(x).$$

[13] CARL NEUMANN, 1832 – 1925, not to be confused with JOHN VON NEUMANN, renowned for his contributions to functional analysis and quantum mechanics.

[14] PIERRE–SIMON LAPLACE, 1749 – 1827

The rot–operator is sometimes also written as curl$f$. Thinking of ∇ as a vector,

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n} \right)$$

admissible to scalar product and vector product, we get the convenient notation

$$\begin{aligned}
&\triangle \varphi = \nabla^2 \varphi = \operatorname{div} \operatorname{grad} \varphi, \\
&\operatorname{div} f = \nabla \cdot f, \\
&\operatorname{rot} f = \nabla \times f \qquad\qquad \text{(only if } n = 3).
\end{aligned}$$

Next, we will list some rules for these operators. But first, we give some notation. For a moment, we do not distinguish row vectors and column vectors anymore. The Jacobi–matrix of a function $f \colon \Omega \to \mathbb{R}^n$ is denoted by Df,

$$Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

The Laplace operator can be applied to a vector-valued function component-wise:

$$\triangle f = (\triangle f_1, \triangle f_2, \dots, \triangle f_n).$$

Proposition 1.25. *Let $\Omega \subset \mathbb{R}^n$ be an open set, and $f, g \colon \Omega \to \mathbb{R}^n$ and $\varphi \colon \Omega \to \mathbb{R}^1$ be sufficiently smooth. Then the following formulae hold (if a* rot *appears, n must be equal to three):*

1. $\operatorname{rot} \operatorname{grad} \varphi = 0$,
2. $\operatorname{div} \operatorname{rot} f = 0$,
3. $\operatorname{div}(\varphi f) = \langle \operatorname{grad} \varphi, f \rangle + \varphi \operatorname{div} f$,
4. $\operatorname{rot}(\operatorname{rot} f) = \operatorname{grad} \operatorname{div} f - \triangle f$,
5. $\operatorname{div}(f \times g) = \langle \operatorname{rot} f, g \rangle - \langle f, \operatorname{rot} g \rangle$,
6. $\operatorname{rot}(\varphi f) = (\operatorname{grad} \varphi) \times f + \varphi \operatorname{rot} f$,
7. $\operatorname{rot}(f \times g) = (\operatorname{div} g) f - (\operatorname{div} f) g + (Df) g - (Dg) f$.

Proof. This is a wonderful exercise. □

1.5 Outlook: String Theory and Differential Forms

(Outlook sections are not relevant for exams.)

We play a bit with the formulas rot grad = 0 and div rot = 0, and hopefully an application of this will become visible after some time. First we make a diagram, to be read from left to right:

$$\begin{tikzcd} \boxed{C^\infty(\mathbb{R}^3 \to \mathbb{R})} \arrow[r, "\operatorname{grad}"] & \boxed{C^\infty(\mathbb{R}^3 \to \mathbb{R}^3)} \arrow[r, "\operatorname{rot}"] & \boxed{C^\infty(\mathbb{R}^3 \to \mathbb{R}^3)} \arrow[r, "\operatorname{div}"] & \boxed{C^\infty(\mathbb{R}^3 \to \mathbb{R})} \end{tikzcd}$$

The first box is the vector space of smooth scalar functions on $\mathbb{R}^3$, which are mapped by grad into the second box, which is the vector space of smooth vector fields on $\mathbb{R}^3$, which are mapped by rot again into the vector space of smooth vector fields, which are finally mapped by div into the last box, the vector space of smooth scalar fields.

For simplicity of notation, call these four vector spaces V_0, V_1, V_2, and V_3. The differential operators grad, div and rot are linear mappings from some V_j into the neighbour V_{j+1}, and then it is possible to ask for the kernel spaces and image spaces of these homomorphisms.

To this end, we look at the two vector spaces in the middle. Take V_1 first. This space contains img grad and also ker rot, and both are linear subspaces of V_1. The formula rot grad $= 0$ then simply means

$$\operatorname{img}\operatorname{grad} \subset \ker \operatorname{rot}\,.$$

Take now V_2, which contains img rot and ker div, and again both are linear subspaces of V_2. Now the formula div rot $= 0$ implies

$$\operatorname{img}\operatorname{rot} \subset \ker \operatorname{div}\,.$$

Let us formulate this in words: we have a chain of vector spaces, which are linked by linear mappings. At each vector space (neglecting the left and right end spaces), one mapping comes in from the left, and one mapping goes out to the right. And the image space of the mapping coming in from the left is contained in the kernel space of the mapping going out to the right. If you draw a picture, it will resemble a chain of fisherman's fykes[15].

Next, we wish to describe these image spaces and kernel spaces a bit closer. They are all of infinite dimension, and writing down a basis for anyone of them seems hopeless. So we settle for something less: V_1 contains img grad and ker rot, and we ask how much do img grad and ker rot differ ? So we hope to write

$$\ker \operatorname{rot} = \operatorname{img}\operatorname{grad} \oplus H_1$$

in the sense of direct sums of subspaces of V_1, and wish to know something about H_1.

Similarly, in the space V_2, we can hopefully write, with some unknown space H_2,

$$\ker \operatorname{div} = \operatorname{img}\operatorname{rot} \oplus H_2.$$

To make a long story short: Corollary 3.82 will tell us that $H_1 = \{0\}$ is a quite boring vector space, and you can compute by hand that also $H_2 = \{0\}$. (The exercise you have to solve here is the following: given a function $\vec{u}$ with div $\vec{u} = 0$, seek a function $\vec{v}$ with $\vec{u} = \operatorname{rot}\vec{v}$. If you can always find such a function $\vec{v}$, then $H_2 = \{0\}$. You will meet this exercise again in the theory of electrostatics: there are no magnetic monopoles, and therefore div $\vec{B} = 0$. Then there is a vector field $\vec{A}$ with $\vec{B} = \operatorname{rot}\vec{A}$, and $\vec{A}$ is called *vector potential* of the magnetic field $\vec{B}$.)

Now we want something less boring: the domain $\mathbb{R}^3$, where the variable x lives in, is called *universe*, for the moment. Just for the fun, let us drill a hole through the universe. That means, we remove the infinite cylinder $\{(x_1, x_2, x_3) : x_1^2 + x_2^2 \le 1\}$ from the $\mathbb{R}^3$, and we change the spaces $V_0, \dots, V_3$ accordingly. What happens with the spaces H_1 and H_2 then ? In the language of Corollary 3.82, the universe is no longer *simply connected*, and it can be shown (we will not go into the details here), that then H_1 and H_2 will be function spaces of dimension one. You can also drill some more holes, or cut the universe into pieces, or connect regions which had been far away before (think of a wormhole), and you will always have $\dim H_1 = \dim H_2$ (isn't this amazing ?).

The key idea is now: from the dimensions of H_1 and H_2 (called *Betti numbers*) you can draw some conclusions about the shape of the universe. Assume that you have two universes, and the Betti numbers of one universe are different from the Betti numbers of the other universe. Then you know that the only way to transform one universe into the other is by means of "violent action". If both universes are "topologically equivalent", then their Betti numbers are the same; but the converse need not be true.

This approach is one of the many ideas behind the **string theory**.

Literature: K.Becker, M.Becker, J.H.Schwarz: *String theory and M-Theory*

The above spaces H_1 and H_2 are closely related to something which is called DERHAM-*cohomology* (we will not go into the details of this).

And for those who do not have enough, we mention how the above spaces $V_0, \dots, V_3$ should be replaced to make everything (a bit more) precise:

- the space V_0 can remain unchanged,

[15] Reuse

- the space V_1 consists of the *one–forms.* Here a one-form is a mathematical object "that can be integrated along a one-dimensional curve in $\mathbb{R}^3$". Each one–form can be written as $f(x,y,z)\,\mathrm{d}x + g(x,y,z)\,\mathrm{d}y + h(h,y,z)\,\mathrm{d}z$. We will see these expressions again when we study *curve integrals of second kind.*

- the space V_2 consists of the *two–forms.* Here a two–form is a mathematical object "that can be integrated over a two-dimensional surface in $\mathbb{R}^3$". Each two–form can be written as $f(x,y,z)\,\mathrm{d}x \wedge \mathrm{d}y + g(x,y,z)\,\mathrm{d}y \wedge \mathrm{d}z + h(x,y,z)\,\mathrm{d}z \wedge \mathrm{d}x$, and the wedges shall remember us that commuting the two differentials next to them requires a sign change. Later we will study *surface integrals of second kind*, and they are basically the same integrals as we have here.

- the space V_3 consists of the *three–forms.* Here a three–form is a mathematical object "that can be integrated over a three-dimensional region in $\mathbb{R}^3$". Each three–form can be written as $f(x,y,z)\,\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, and the wedges shall remember us that commuting the two differentials next to them requires a sign change.

You know already (something like) a three–form: it is the usual determinant of a 3 × 3 matrix, where you interpret the columns of the matrix as three vectors. And of course you know that commuting two columns in a matrix leads to a sign change of the determinant.

One of the key advantages of the approach via differential forms is that this works in any space dimension (recall that the operator rot is only available in $\mathbb{R}^3$).

Literature: H. Goenner: *Spezielle Relativitätstheorie und die klassische Feldtheorie.* 5.2.5. Maxwellgleichungen in Differentialformenformulierung

We conclude this outlook with some mathematical mystery.

Take a convex polyhedron like a cube, or a tetrahedron, or an octahedron. Count the number V of vertices (corners), the number E of edges, and the number F of faces. Then compute the number

$$\chi = V - E + F.$$

Whatever the convex polyhedron has been, you will always get $\chi = 2$. Therefore this number χ has become famous, and its name is *Euler characteristic.* Now take a simple polyhedron like a cube, and drill a hole of prismatic shape through it, and compute χ again. Drill one more hole, and compute χ once more. What do you expect for N holes ?

And finally, we look at the angles. For each vertex of a convex polyhedron, sum up the angles which have their tip at that vertex (for instance, in case of a cube, you get $3 \times 90° = 270°$ at each corner). For each corner, compute the angle which is missing to 360° (in case of a cube, this is $360° - 270° = 90°$). Take the sum of all missing angles, for all corners.

Repeat with tetrahedron, octahedron, whatever you like. What do you observe, and what is the reason ?

Now drill a square-shaped hole through a cube (or some other polyhedron), and compute the sum of the missing angles again (attention: now some missing angles will be negative, the others positive. Respect the sign !). What will be the result if you drill one more hole ?

1.6 Inverse and Implicit Functions

In transforming polar coordinates into Cartesian coordinates, we had

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x(r,\varphi) \\ y(r,\varphi) \end{pmatrix} = \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}$$

with the derivative $\frac{\partial(x,y)}{\partial(r,\varphi)}$.

How about transforming into the other direction ?

We easily see that $r = r(x,y) = \sqrt{x^2+y^2}$. A similar formula for φ does not exist, however, we have

$$\tan\varphi = \frac{y}{x}.$$

The formula $\varphi = \arctan\frac{y}{x}$ might be obvious, but is wrong.

The partial derivatives then are

$$\begin{aligned}
&\frac{\partial r}{\partial x} = \frac{x}{\sqrt{x^2+y^2}} = \cos\varphi, \\
&\frac{\partial r}{\partial y} = \frac{y}{\sqrt{x^2+y^2}} = \sin\varphi, \\
&\frac{1}{\cos^2\varphi}\frac{\partial \varphi}{\partial x} = \frac{\partial}{\partial x}\tan\varphi = -\frac{y}{x^2} = -\frac{r\sin\varphi}{r^2\cos^2\varphi}, && \Longrightarrow \frac{\partial\varphi}{\partial x} = -\frac{\sin\varphi}{r}, \\
&\frac{1}{\cos^2\varphi}\frac{\partial \varphi}{\partial y} = \frac{\partial}{\partial y}\tan\varphi = \frac{1}{x} = \frac{1}{r\cos\varphi}, && \Longrightarrow \frac{\partial\varphi}{\partial y} = \frac{\cos\varphi}{r}.
\end{aligned}$$

This gives us the fundamental matrix

$$\frac{\partial(r,\varphi)}{\partial(x,y)} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\frac{\sin\varphi}{r} & \frac{\cos\varphi}{r} \end{pmatrix}.$$

Surprisingly, this is just the inverse matrix to

$$\frac{\partial(x,y)}{\partial(r,\varphi)} = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}.$$

We will now see that it is always like this.

Let $f \in C^1(G \to \mathbb{R}^m)$, where $G \subset \mathbb{R}^m$ is an open and convex set. Pay attention to the fact that $n = m$. Let us be given a point $x_0 \in G$, and put $y_0 = f(x_0)$. Suppose that y^* is a point close to y_0. Can we find an $x^* \in G$, such that $f(x^*) = y^*$? Is this x^* unique near x_0 ?

The answer to both questions is 'yes' provided that y^* and y_0 are close to each other, and that the Jacobian matrix $J_0 := f'(x_0)$ is invertible. This will give us an *inverse function* $x^* = g(y^*)$.

Proposition 1.26 (Inverse function theorem). *Under the above assumptions, there are positive ε and δ, with the property that for each y^* with $\|y^* - y_0\| < \varepsilon$, there is a unique $x^* \in G$ with $\|x^* - x_0\| \leq \delta$ and $f(x^*) = y^*$. The mapping $y^* \mapsto g(y^*) = x^*$ is differentiable, and its derivative satisfies*

$$g'(y) = (f'(x))^{-1}, \qquad y = f(x), \qquad \|y - y_0\| < \varepsilon.$$

The proof is quite long, but you can learn from it how bigger results can be shown if you have the proper tools. Our tools are now:

- a modified Newton iteration scheme (note that the Newton iteration which you learned at the end of the first semester works also for functions $f\colon \mathbb{R}^m \to \mathbb{R}^m$),
- the Banach fixed point theorem,
- matrix norms.

To make everything easier, we cheat a bit and assume that even the second derivatives of f exist and are continuous. As an added bonus, the proof will teach us some tricks how to handle Taylor expansions.

Proof. **Step 0: making a todo-list:** given are f, x_0, y_0, $f'(x_0) =: J_0$ and its inverse J_0^{-1}, and y^* "near" y_0.

We have to find x^* "near" x_0 with $f(x^*) = y^*$. We have to explain (twice) what "near" means. We have to show that the map $g\colon y^* \mapsto x^*$ is differentiable, and we have to compute the derivative $g'(y^*)$.

Step 1: setting up an iteration scheme: We have $y_0 = f(x_0)$, with given x_0 and y_0. Moreover, there is a given point y^* which is very close to y_0. We are looking for all x^* with $f(x^*) = y^*$. It is natural to search the x^* by means of a Newton scheme,

$$\begin{aligned}
&x_0 \text{ given}, \\
&x_k := x_{k-1} - (f'(x_{k-1}))^{-1}(f(x_{k-1}) - y^*), \qquad k = 1, 2, 3, \ldots.
\end{aligned}$$

(Draw a picture !) The proof will become easier if we modify this scheme a bit: put $J_0 := f'(x_0)$ and

$$\begin{aligned}&x_0 \text{ given,}\\ &x_k := x_{k-1} - J_0^{-1}(f(x_{k-1}) - y^*), \qquad k = 1, 2, 3, \dots .\end{aligned}$$

We will show convergence of this sequence $(x_k)_{k\in\mathbb{N}}$ to some point x^*, using Banach's fixed point theorem. This x^* is then the solution to $f(x^*) = y^*$. If y_0 and y^* are close together, this solution x^* is unique near x_0.

Step 2: preparing the Banach fixed point theorem: Write the iteration scheme in the form $x_k = T(x_{k-1})$. The fixed point theorem requires you to check two assumptions:

- the mapping T maps a closed set M into itself;
- the mapping T is contractive on M. This means $\|T(x) - T(\tilde{x})\| \le \gamma \|x - \tilde{x}\|$ for some constant $\gamma < 1$ and all $x, \tilde{x} \in M$. Let us choose $\gamma := \frac{1}{4}$.

It is reasonable to take a ball for the closed set M:

$$M := \{x \in G \colon \|x - x_0\| \le \delta\},$$

with some positive radius δ which we promise to select later.

And to show the two •, we need to know f very precisely. To this end, we write down its Taylor expansion,

$$\begin{aligned}f(x) &= f(x_0) + f'(x_0) \cdot (x - x_0) + R(x)\\ &= y_0 + J_0 \cdot (x - x_0) + R(x),\end{aligned}$$

and the remainder R is quadratically small for $x \to x_0$, since f is C^2, hence $R(x) = \mathfrak{O}(\|x - x_0\|^2)$. To make this precise: we have a positive constant C_1 with

$$\|R(x)\| \le C_1 \|x - x_0\|^2 \qquad \text{if } \|x - x_0\| \le 1 \text{ and } x \in G.$$

Let us differentiate the Taylor expansion of f: then

$$f'(x) = J_0 \cdot I + R'(x),$$

hence $R'(x) = f'(x) - J_0$, hence $R'(x_0) = 0$.

Next we discuss the mapping T and bring it into a different formula:

$$\begin{aligned}T(x) &:= x - J_0^{-1}(f(x) - y^*)\\ &= x - J_0^{-1}(y_0 + J_0(x - x_0) + R(x) - y^*)\\ &= x_0 + J_0^{-1}(y^* - y_0 - R(x)).\end{aligned}$$

This representation of T has the advantage that it contains many terms which we know very well (namely all except $R(x)$).

Step 3: the first condition in the Banach fixed point theorem: to prove that T maps M into M, we assume $x \in M$ and intent to show that also $T(x) \in M$. So, let us suppose $\|x - x_0\| \le \delta$ for our small δ. Then we have (under the reasonable assumption $\|x - x_0\| \le 1$)

$$\begin{aligned}\|T(x) - x_0\| &= \left\|J_0^{-1}(y^* - y_0 - R(x))\right\| \le \left\|J_0^{-1}\right\| \|y^* - y_0 - R(x)\|\\ &\le \left\|J_0^{-1}\right\| \cdot (\|y^* - y_0\| + \|R(x)\|)\\ &\le \left\|J_0^{-1}\right\| \cdot \left(\varepsilon + C_1 \|x - x_0\|^2\right)\\ &\le \left\|J_0^{-1}\right\| \cdot \left(\varepsilon + C_1\delta^2\right).\end{aligned}$$

We wish this to be smaller than δ, and this can be arranged as follows. First we choose δ so small that $\delta \le 1$ and

$$\left\|J_0^{-1}\right\| \cdot C_1\delta^2 \le \frac{1}{2}\delta,$$

and then we choose ε so small that

$$\left\|J_0^{-1}\right\| \cdot \varepsilon \leq \frac{1}{2}\delta.$$

Step 4: the second condition in the Banach fixed point theorem: to prove that T is contractive on M, we wish to prove that

$$\|T(x) - T(\tilde{x})\| \leq \frac{1}{4}\|x - \tilde{x}\|$$

whenever x, $\tilde{x} \in M$. We know $T(x) = x_0 + J_0^{-1}(y^* - y_0 - R(x))$, and we have a corresponding formula for $T(\tilde{x})$. Then we have

$$\|T(x) - T(\tilde{x})\| = \left\|J_0^{-1}(R(x) - R(\tilde{x}))\right\| \leq \left\|J_0^{-1}\right\| \cdot \|R(x) - R(\tilde{x})\|,$$

and this shall be smaller than $\frac{1}{4}\|x - \tilde{x}\|$.

Step 5: we need more information on R: Suppose x, $\tilde{x} \in M$, hence $\|x - x_0\| \leq \delta$ and $\|\tilde{x} - x_0\| \leq \delta$. Then also each point on on the connecting line between x and $\tilde{x}$ is in M, and we can write, by the integrated mean value theorem,

$$R(x) - R(\tilde{x}) = \left(\int_{t=0}^{1} R'(\tilde{x} + t(x - \tilde{x}))\,\mathrm{d}t\right) \cdot (x - \tilde{x}).$$

Plugging in the representation $R' = f' - J_0$ from Step 2, we then have

$$R(x) - R(\tilde{x}) = \left(\int_{t=0}^{1} f'(\tilde{x} + t(x - \tilde{x})) - f'(x_0)\,\mathrm{d}t\right) \cdot (x - \tilde{x}). \tag{1.7}$$

Now we apply the integrated mean value theorem once more, but now to the difference $f'(\dots) - f'(x_0)$ in the integrand (compare (1.3)):

$$\|f'(\tilde{x} + t(x - \tilde{x})) - f'(x_0)\| \leq C_2\|\tilde{x} + t(x - \tilde{x}) - x_0\| \leq C_2\delta,$$

for some constant C_2 which is basically computable (for the purpose of our proof it is enough to know that C_2 exists). We insert this inequality into (1.7) and obtain the nice estimate

$$\|R(x) - R(\tilde{x})\| \leq C_2\delta\|x - \tilde{x}\|.$$

Step 6: back to the second condition in the Banach fixed point theorem: we continue where we stopped in Step 4:

$$\|T(x) - T(\tilde{x})\| \leq \left\|J_0^{-1}\right\| \cdot \|R(x) - R(\tilde{x})\| \leq \left\|J_0^{-1}\right\| \cdot C_2\delta\|x - \tilde{x}\|,$$

and now we need $\left\|J_0^{-1}\right\| \cdot C_2\delta \leq \frac{1}{4}$.

Step 7: choosing δ and ε: first we select a positive δ with

$$\delta \leq 1, \qquad \left\|J_0^{-1}\right\| \cdot C_1\delta \leq \frac{1}{2}, \qquad \left\|J_0^{-1}\right\| \cdot C_2\delta \leq \frac{1}{4}.$$

Then we select a positive ε with

$$\left\|J_0^{-1}\right\| \cdot \varepsilon \leq \frac{1}{2}\delta.$$

Then the Banach fixed point theorem guarantees that there is exactly one fixed point $x^* \in M$ of the map T; $T(x^*) = x^*$. This is equivalent to $f(x^*) = y^*$.

Call the mapping $y^* \mapsto x^*$ from now on g.

Step 8: find the derivative of g: Pick two points x_*, x in M, and set $y_* = f(x_*)$, $y = f(x)$. The integrated version of the mean value theorem reads

$$f(x) - f(x_*) = \left(\int_{t=0}^{t=1} f'(x_* + t(x - x_*))\,\mathrm{d}t \right) \cdot (x - x_*) =: J_{x,x_*} \cdot (x - x_*).$$

In (1.7) we have shown that $\|J_{x,x_*} - J_0\| \le C_2\delta$, and then we have $\left\|J_0^{-1}(J_{x,x_*} - J_0)\right\| \le 1/4$, by our choice of δ in Step 7. Then Lemma 1.23 implies that also J_{x,x_*} is invertible, whence

$$g(y) - g(y_*) = x - x_* = J_{x,x_*}^{-1}(f(x) - f(x_*)) = J_{x,x_*}^{-1}(y - y_*),$$
$$g(y) = g(y_*) + J_{x,x_*}^{-1}(y - y_*).$$

This is the beginning of a Taylor expansion of g. But J_{x,x_*}^{-1} still depends on y, which is not allowed. We need a dependence on y_* only.

Put $J_* = f'(x_*)$, which is independent of y (and invertible, again by Lemma 1.23). Then we have

$$g(y) = g(y_*) + J_*^{-1}(y - y_*) + (J_{x,x_*}^{-1} - J_*^{-1})(y - y_*).$$

It suffices to show that the last item is $\mathfrak{O}(\|y - y_*\|^2)$, which will then imply that $g'(y_*) = J_*^{-1} = (f'(x_*))^{-1}$. But (1.5) makes this easy, since

$$\left\|J_{x,x_*}^{-1} - J_*^{-1}\right\| \le C\,\|J_{x,x_*} - J_*\| \le C\,\|x - x_*\| = C\left\|J_{x,x_*}^{-1}(y - y_*)\right\| \le C\,\|y - y_*\|\,,$$

where C denote highly boring computable constants, which do not depend on x, x_*, y, y_* (different occurrences of C can have different values).

This finishes the proof (which probably was the hardest proof in the second semester). □

Now we play with some functions. Let us be **given** two functions

$$y\colon \mathbb{R}^m \to \mathbb{R}^n, \qquad f\colon \mathbb{R}^{m+n} \to \mathbb{R}^n,$$

both continuously differentiable, and assume that

$$f(x, y(x)) = 0 \qquad \forall\, x \in \mathbb{R}^m.$$

We wish to differentiate this equation with respect to x. To this end, we define a function $h\colon \mathbb{R}^m \to \mathbb{R}^n$ by $h(x) := f(x, y(x))$, and another function $g\colon \mathbb{R}^m \to \mathbb{R}^{n+m}$ by

$$g(x) := \begin{pmatrix} x \\ y(x) \end{pmatrix}.$$

Then we clearly have $0 = h(x) = f(g(x))$ for all $x \in \mathbb{R}^m$, and now the chain rule gives us

$$h'(x) = f'(g(x)) \cdot g'(x),$$

with h', f', and g' as the Jacobi matrices. The Jacobi matrices of f' and g' have the block matrix form

$$f'(g) = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} \in \mathbb{R}^{n\times(m+n)}, \qquad g'(x) = \begin{pmatrix} I_m \\ y'(x) \end{pmatrix} \in \mathbb{R}^{(m+n)\times m},$$

and therefore the Jacobi matrix of h becomes

$$h'(x) = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} \begin{pmatrix} I_m \\ y'(x) \end{pmatrix} = \frac{\partial f}{\partial x}(x, y(x)) + \frac{\partial f}{\partial y}(x, y(x)) \cdot \frac{\partial y}{\partial x}(x).$$

Since $h(x) = 0$ for each $x \in \mathbb{R}^m$, we then also have $h'(x) = 0$ for each $x \in \mathbb{R}^m$. Observe that the matrix $\frac{\partial f}{\partial y}$ has size $n \times n$, so it could be an invertible matrix (also called a regular matrix). Assuming that x is a point for which $\frac{\partial f}{\partial y}(x, y(x))$ is invertible, we then have a nice representation of the Jacobi matrix $y'(x)$:

$$y'(x) = \frac{\partial y}{\partial x}(x) = -\left(\frac{\partial f}{\partial y}(x, y(x)) \right)^{-1} \cdot \frac{\partial f}{\partial x}(x, y(x)).$$

This is nice because on the right-hand side, there are only derivatives of f, but no derivative of y.

Now we take a different point of view. We now longer assume that a differentiable function y exists. Instead, we start with an equation $f(x, y) = 0$ in the space $\mathbb{R}^n$, which means that f takes values in the $\mathbb{R}^n$, and we wish to solve this equation for the vector y, which also shall be in $\mathbb{R}^n$ ("n equations for n unknowns should be a solvable system"). If such a vector y can be found, it will certainly depend on x, giving us a function $y = y(x)$. Since we have no nice formula for this function (we don't even know whether $y = y(x)$ exists at all), it is called an *implicit function.*

Let us simplify notation and write $f_y = \partial_y f$ for $\frac{\partial f}{\partial y}$, similarly for $f_x = \partial_x f$, and so on.

Proposition 1.27 (Implicit function theorem). *Let $G \subset \mathbb{R}^{m+n}$ be an open set, $(x_0, y_0) \in G$, where $x_0 \in \mathbb{R}^m$ and $y_0 \in \mathbb{R}^n$. Let $f \in C^1(G \to \mathbb{R}^n)$ be a function satisfying the following two conditions:*

- *$f(x_0, y_0) = 0 \in \mathbb{R}^n$,*
- *$(\partial_y f)(x_0, y_0)$ is a regular $n \times n$ matrix.*

Then there is a neighbourhood $U_m \subset \mathbb{R}^m$ of x_0, such that for each $x_* \in U_m$ there is a $y_* = y_*(x_*)$, with the property that

- *$(x_*, y_*(x_*)) \in G$,*
- *$f(x_*, y_*(x_*)) = 0$.*

The derivative of this function y_* is given by

$$y_*'(x_*) = -\left(\partial_y f(x_*, y_*(x_*))\right)^{-1} (\partial_x f)(x_*, y_*(x_*)). \tag{1.8}$$

Proof. This is an easy consequence of the inverse function theorem, but only if we approach it from the right angle. Namely:

Define a function $F \in C^1(G \to \mathbb{R}^{m+n})$ by

$$F\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x \\ f(x, y) \end{pmatrix}.$$

Put

$$z := \begin{pmatrix} x \\ y \end{pmatrix}, \qquad z_0 := \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \qquad w := F(z), \qquad w_0 := F(z_0) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}.$$

Then F has the Jacobi matrix (written in block matrix form)

$$F'(z) = \begin{pmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} \\ \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix}.$$

We see that $F'(z_0)$ is an invertible matrix. By the inverse function theorem, the function F has an inverse function $G = F^{-1}(w)$ in a neighbourhood of w_0. This means that for each w_* near w_0, there is a (locally unique) $z_* = G(w_*)$ with $F(z_*) = w_*$. Since $w_0 = (x_0, 0)^\top$, we are allowed to choose special w_*, namely $w_* = (x_*, 0)^\top$ with x_* near x_0. Write $z_* = G(w_*)$ as $z_* = (x_*, y_*)^\top$. Then we obtain y_* as a function of w_*, hence as a function of x_*. This vector y_* is exactly what we are looking for, namely a function $y_* = y_*(x_*)$ that solves $f(x_*, y_*(x_*)) = 0$, because

$$\begin{pmatrix} x_* \\ 0 \end{pmatrix} = w_* = F(z_*) = F\begin{pmatrix} x_* \\ y_* \end{pmatrix} = \begin{pmatrix} x_* \\ f(x_*, y_*) \end{pmatrix}.$$

The differentiability of this function y_* follows from the inverse function theorem; and the value of the derivatives as given in (1.8) have already be obtained by the chain rule. □

Example: *Let $m = n = 1$, $f = f(x, y) = x^2 + y^2 - 100$, and $(x_0, y_0) = (8, -6)$. We have $f(x_0, y_0) = 0$ and $(\partial_y f)(x_0, y_0) = 2y_0 = -12 \neq 0$. In a neighbourhood of $x_0 = 8$, we can write y as a function of x, namely $y = -\sqrt{100 - x^2}$.*

If you choose $(x_0, y_0) = (-10, 0)$ instead, you still have $f(x_0, y_0) = 0$, but you cannot write y as a function of x in a neighbourhood of x_0.

1.7 Extrema Under Side Conditions

Example: *Suppose we are given a point $P = (p_1, p_2, p_3)^\top$ in $\mathbb{R}^3$, and an ellipsoid*

$$E = \left\{ (z_1, z_2, z_3)^\top \in \mathbb{R}^3 \colon \frac{z_1^2}{a_1^2} + \frac{z_2^2}{a_2^2} + \frac{z_3^2}{a_3^2} = 1 \right\}.$$

We want to know the distance from P to E. That is, we look for a minimum of the function

$$f = f(z_1, z_2, z_3) := (z_1 - p_1)^2 + (z_2 - p_2)^2 + (z_3 - p_3)^2.$$

But $z = (z_1, z_2, z_3)^\top$ cannot be an arbitrary point in $\mathbb{R}^3$: it must satisfy the side condition $z \in E$, which means

$$g(z_1, z_2, z_3) := \frac{z_1^2}{a_1^2} + \frac{z_2^2}{a_2^2} + \frac{z_3^2}{a_3^2} - 1 = 0.$$

Now we consider a more general case.

We look for extrema of a function $f = f(z)$, where $f \in C^1(G \to \mathbb{R}^1)$ and $G \subset \mathbb{R}^{m+n}$. And we have n side conditions $g_1(z) = 0, \ldots, g_n(z) = 0$, which we can write as $g(z) = 0$ where $g = (g_1, \ldots, g_n)^\top$ is a column vector as always. The function g maps from $\mathbb{R}^{m+n}$ into $\mathbb{R}^n$. Let us hope for a moment that the system $g(z) = 0$ can be resolved in the following sense. Split the variables like

$$z = (z_1, \ldots, z_{m+n})^\top = (x_1, \ldots, x_m, y_1, \ldots, y_n)^\top$$

and assume that $g(x, y) = 0$ can be written as $y = y(x)$ via the implicit function theorem.

Looking for extrema of $f = f(x, y)$ under the system of side conditions $g(x, y) = 0$ is then equivalent to looking for extrema of

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

(without side conditions), where the function h maps from a subset of $\mathbb{R}^m$ into $\mathbb{R}^1$.

If the function h has an extremum at a point x, then

$$\nabla h(x) = h'(x) = 0 \in \mathbb{R}^m,$$

where ∇h is a row vector as always. But h' can be computed as

$$0 = h'(x) = f_x(x, y(x)) + f_y(x, y(x)) \cdot y'(x),$$

by the chain rule. And the derivative $y'(x)$ is given by

$$y'(x) = -(g_y(x, y(x)))^{-1} g_x(x, y(x)).$$

Therefore, looking for extrema of f under the side condition g means solving the system

$$0 = f_x(x, y) - f_y(x, y)(g_y(x, y))^{-1} g_x(x, y).$$

The factor f_y is an n–row, and $(g_y)^{-1}$ is an $n \times n$ matrix. Therefore, the product $-f_y(g_y)^{-1}$ is an n row, write it as $\lambda = (\lambda_1, \ldots, \lambda_n)$. Then we have the $m + n + n$ unknowns x, y, and λ, and the equations

$$\begin{aligned} f_x(x, y) + \lambda g_x(x, y) &= 0_m, \\ \lambda &= -f_y(x, y)(g_y(x, y))^{-1}, \\ g(x, y) &= 0_n, \end{aligned}$$

which we rewrite as

$$\begin{aligned} f_x(x, y) + \lambda g_x(x, y) &= 0_m, \\ f_y(x, y) + \lambda g_y(x, y) &= 0_n, \\ g(x, y) &= 0_n. \end{aligned}$$

These are $m+2n$ equations for the $m+2n$ unknowns, so there is some hope to find a solution.

Now we should undo the quite artificial splitting of the vector z into x and y. It is custom to introduce a function

$$L = L(z,\lambda) = f(z) + \lambda g(z),$$

and solve the system

$$\begin{aligned} &L_z(z,\lambda) = 0_{m+n}, \\ &L_\lambda(z,\lambda) = 0_n. \end{aligned}$$

These numbers λ_j are also known as the LAGRANGE MULTIPLIERS[16]. Remember that we have only considered necessary conditions. Sufficient conditions are much harder.

Example: *Coming back to our example with the ellipsoid, we find:*

$$\begin{aligned} &g = g(x) = \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \frac{x_3^2}{a_3^2} - 1, \\ &f = f(x) = (x_1-p_1)^2 + (x_2-p_2)^2 + (x_3-p_3)^2, \\ &L = f + \lambda g, \qquad \lambda \in \mathbb{R}, \\ &\frac{\partial L}{\partial x_j} = 2(x_j - p_j) + \lambda \frac{2x_j}{a_j^2} \overset{!}{=} 0, \\ &\frac{\partial L}{\partial \lambda} = g(x) \overset{!}{=} 0. \end{aligned}$$

From the first condition, we find that

$$x_j = \frac{p_j a_j^2}{a_j^2 + \lambda},$$

tacitly assuming that the denominator is not 0. Plugging this into the second condition, we then get

$$1 = \sum_{j=1}^3 \left(\frac{p_j a_j}{a_j^2 + \lambda} \right)^2.$$

It seems quite hard to solve this equation with respect to λ analytically. However, we may plot the right–hand side as a function of λ. Choosing, as an example, $(a_1,a_2,a_3) = (1,2,3)$ and $(p_1,p_2,p_3) = (7,4,5)$, this plot tells us that there are exactly two values $\lambda = \lambda_\pm$ for which the right–hand side becomes 1. For the mentioned parameter values, such a plot is given in Figure 1.1. The exact location of these $\lambda_\pm$ can be found by Newton's algorithm, for instance. Then we can compute

$$x_j = x_{j,\pm} = \frac{p_j a_j^2}{a_j^2 + \lambda_\pm}, \qquad j = 1,2,3.$$

1.8 Some Remarks Concerning Complex Differentiation

Let $G \subset \mathbb{C}$ be an open set in the complex plane, and $f \in C^1(G \to \mathbb{C})$ be a differentiable function. This means that for each $z_0 \in G$ the following limit exists:

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

It is important to note that the point z can approach the limit z_0 in an arbitrary way. It can run along a straight line, for instance, or along a wildly oscillating curve, or it can jump around heavily. The only restriction is that z converges to z_0. And for each imaginable path of z approaching z_0, the above limit must exist. Moreover, all these limits must coincide.

This is a restriction with deep consequences, as the following result shows.

[16] JOSEPH LOUIS LAGRANGE, 1736 – 1813

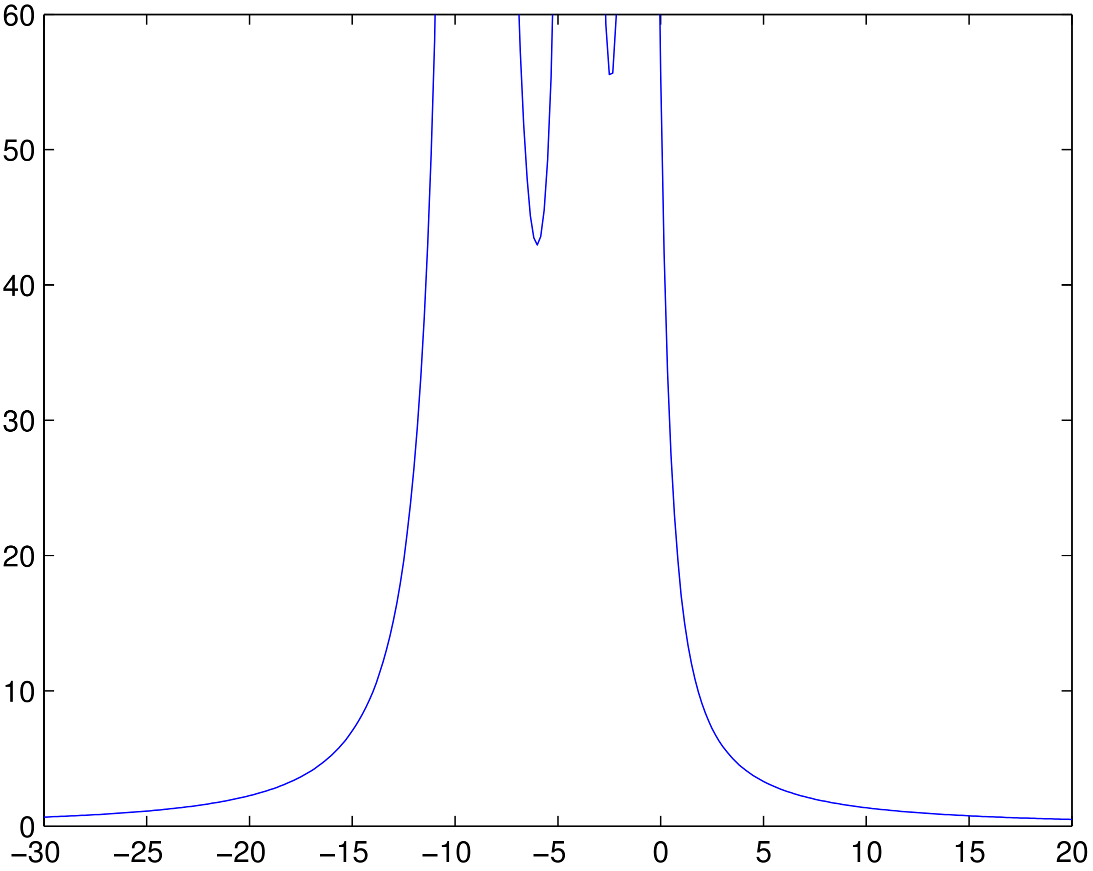


Figure 1.1: A plot of the function $\lambda \mapsto \sum_{j=1}^{3}(\frac{p_j a_j}{a_j^2+\lambda})^2$. We clearly see its poles for $\lambda = -9$, $\lambda = -4$, and $\lambda = -1$.

Proposition 1.28. *Let $f \in C^1(G \to \mathbb{C})$ be a complex continuously differentiable function. Put $f = u + \mathrm{i}v$ and $z = x + \mathrm{i}y$, where u, v, x and y are real-valued.*

Then the functions u and v solve the following partial differential equations at each point of G:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{1.9}$$

The derivative f' can be computed from the formula

$$f'(z_0) = \frac{1}{2}\left(\frac{\partial}{\partial x} - \mathrm{i}\frac{\partial}{\partial y}\right) f(z_0). \tag{1.10}$$

Proof. Let z run to z_0 vertically and horizontally; and then compare both limits. You obtain two expressions for $f'(z_0)$, which must coincide. Their arithmetic mean is the value in (1.10). □

Remark 1.29. *The equations (1.9) are the famous* CAUCHY[17]–RIEMANN[18] differential equations*. A complex differentiable function is also called* holomorphic *or* complex analytic*.*

The converse of Proposition 1.28 is also true: if a function f with real and imaginary parts u and v solves the Cauchy–Riemann differential equations everywhere, then it must be holomorphic.

[17] AUGUSTIN LOUIS CAUCHY, 1789 – 1857

[18] BERNHARD RIEMANN, 1826 – 1866

Cauchy–Riemann equations might look like academic exercises, but their importance cannot be underestimated. Let us list some conclusions of the CR equations:

- each differentiable function has all derivatives of any order; and its Taylor series always converge;
- if a function is defined on $\mathbb{C}$, differentiable everywhere, and bounded, then it must be a constant;
- if you know the values of a differentiable function on the boundary of a disk, then you can compute the values of that function inside the disk immediately;
- a lot of integrals over the real line, which are impossible to compute if you only know the real numbers, become quite easy to calculate, after an extension to the complex plane.

Details of holomorphic functions will be presented mainly during the third semester; but we will encounter some of the above conclusions in the second semester, when it is appropriate.

1.9 Outlook: the Legendre Transform

Literature: Greiner, Neise and Stöcker: *Thermodynamik und Statistische Mechanik.* Chapter 4.3: Die Legendre–Transformation

1.9.1 Mathematical Background

Let $G \subset \mathbb{R}^n$ be an open domain with smooth boundary ∂G. Let us be given a function

$$f\colon G \to \mathbb{R}$$

which is (at least) twice continuously differentiable. Our key assumption is that the mapping

$$\varphi\colon z \mapsto \varphi(z) := f'(z) \in \mathbb{R}^n, \qquad z \in G \subset \mathbb{R}^n,$$

maps G onto an open domain $G_* \subset \mathbb{R}^n$ and is invertible. By the inverse function theorem, this mapping φ is invertible if the Jacobi matrix $\varphi'(z)$ is an invertible matrix from $\mathbb{R}^{n\times n}$, for all $z \in G$. Note that the Jacobi matrix $\varphi'(z)$ is equal to the Hessian matrix of the second derivatives of f; and therefore we are on the safe side if we assume that f is strictly convex (or strictly concave), which means that the Hessian of f is always a positive definite (or negative definite) matrix.

Now we define a variable ζ, which is "conjugate" to z:

$$\zeta := \varphi(z) = f'(z), \qquad z \in G. \tag{1.11}$$

Then $\zeta \in G_*$. By the assumption of invertibility of the mapping $\varphi\colon G \to G_*$, there is an inverse mapping of φ, let us call it ψ:

$$\psi\colon G_* \to G, \qquad \psi(\zeta) = z.$$

And again by the inverse function theorem, we have

$$\varphi'(z) \cdot \psi'(\zeta) = I_n \quad \text{if} \quad z = \psi(\zeta).$$

Definition 1.30 (**Legendre transform**). *Let a function f with the above properties be given, and fix $\varphi := f'$ and $\psi := \varphi^{-1}$ as above. Then the* Legendre transform *f_* of f is defined as*

$$f_*(\zeta) := \langle \zeta, z\rangle - f(z), \qquad \zeta \in G_* \quad \text{if} \quad z = \psi(\zeta). \tag{1.12}$$

Here $\langle \zeta, z\rangle = \sum_{j=1}^n \zeta_j z_j$ is the usual scalar product on $\mathbb{R}^n$. The Legendre transform f_* maps from G_* into $\mathbb{R}$.

Our interpretation is that we start from a function f, and then we exchange its argument z against an argument ζ in a quite peculiar way and obtain a new function f_*. Sometimes also this exchange procedure is called *Legendre transform.*

Proposition 1.31. *The Legendre transform has the following properties:*

$$f_*'(\zeta) = \psi(\zeta) = z, \qquad \forall \zeta \in G_*, \tag{1.13}$$

$$f \in C^s(G), \qquad s \geq 2 \quad \Longrightarrow \quad f_* \in C^s(G_*), \tag{1.14}$$

$$(f_*)_* = f. \tag{1.15}$$

Proof. The identities (1.11) and (1.13) are valid if we treat all vectors the same and do no longer distinguish between rows and columns. For this proof however, we should exercise a little more care and consider the vectors z, ζ, ψ, φ as columns, except the gradients, f' and f_*', which are rows. And φ' as well as ψ' are the usual matrices. Then the product rule and the chain rule give, together with $z = \psi(\zeta)$,

$$\begin{aligned} f_*'(\zeta) &= \operatorname{grad}_\zeta \left(\zeta^\top \cdot \psi(\zeta) - f(\psi(\zeta))\right) \\ &= \zeta^\top \cdot \psi'(\zeta) + (\psi(\zeta))^\top \cdot I_n - f'(\psi(\zeta)) \cdot \psi'(\zeta) \\ &= \zeta^\top \cdot \psi'(\zeta) + (\psi(\zeta))^\top - f'(z) \cdot \psi'(\zeta) \\ &= \zeta^\top \cdot \psi'(\zeta) + (\psi(\zeta))^\top - \zeta^\top \cdot \psi'(\zeta) = (\psi(\zeta))^\top = z^\top. \end{aligned}$$

This proves (1.13). Now assume $f \in C^s(G)$. Then obviously $\varphi \in C^{s-1}(G)$, since $\varphi = f'$. Looking at the proof of the inverse function theorem, we then find $\psi \in C^{s-1}(G_*)$. And because the first derivative of f_* is exactly ψ, the regularity (smoothness) of f_* must be one order better than the regularity of ψ. This proves (1.14).

Now we show (1.15). The Legendre transform of f_* is defined in the same manner, replacing ζ by y:

$$y := f_*'(\zeta), \qquad (f_*)_*(y) := \langle y, \zeta\rangle - f_*(\zeta) \qquad \text{if} \quad \zeta = (f_*')^{-1}(y).$$

However, from (1.13) we know already that $y = f_*'(\zeta) = \psi(\zeta) = z$, and consequently

$$(f_*)_*(y) = \langle z, \zeta\rangle - f_*(\zeta) = \langle z, \zeta\rangle - (\langle \zeta, z\rangle - f(z)) = 0 + f(z) = f(y),$$

because of $z = y$. □

We can regard f_* as "dual function" associated to f. This is a reasonable view because transforming twice gives the original function f back.

There is a nice interpretation of f_* which makes the concept of Legendre transforms interesting for questions of optimization:

Lemma 1.32. *Assume that G is convex and that $f \in C^2(G)$ is strictly convex. Then*

$$f_*(\zeta) = \max_{z \in G} \{\langle \zeta, z\rangle - f(z)\}$$

for all $\zeta \in G_*$.

Proof. We fix some $\zeta \in G_*$, set $g(z) := \langle \zeta, z\rangle - f(z)$, and search for maxima of g. The maximal value of g is attained either at an interior point $z_0 \in G$ with $g'(z_0) = 0$, or at a point on the boundary ∂G. Note that the Hessians of f and g satisfy

$$(Hg)(z) = -(Hf)(z),$$

and therefore Hg is always negative definite, and g is a concave function. The gradient of g is

$$g'(z) = \zeta^\top - f'(z),$$

and indeed, for $z = z_0 := \psi(\zeta)$ we have $g'(z_0) = 0$, and this is the only point in G where g' vanishes, because f' has an inverse function, namely ψ. Therefore, we have found a point $z_0 \in G$ which is a candidate for the maximum of g. This point z_0 is in the interior of G, because ζ is in the interior of G_*, since G_* is open.

Pick some arbitrary $z \in \overline{G}$. Then we perform a Taylor expansion of g at z_0:

$$\begin{aligned} g(z) &= g(z_0) + g'(z_0) \cdot (z - z_0) + \frac{1}{2}(z - z_0)^\top (Hg)(\tilde{z}) \cdot (z - z_0) \\ &= g(z_0) + 0 + \frac{1}{2}(z - z_0)^\top (Hg)(\tilde{z}) \cdot (z - z_0), \end{aligned}$$

where $\tilde{z}$ is an unknown point on the connecting line between z and z_0. This connecting line runs inside G, since G is a convex domain. Therefore we find $g(z) < g(z_0)$ if $z \neq z_0$, and it turns out that the maximal value of g is attained at z_0, and never on the boundary ∂G.

On the other hand, we have $g(z_0) = \langle \zeta, z_0 \rangle - f(z_0)$ with $\zeta = \varphi(z_0)$, hence $g(z_0) = f_*(\zeta)$. □

Corollary 1.33. *Under the assumptions of the previous Lemma, we have*

$$\langle \zeta, z \rangle \leq f(z) + f_*(\zeta),$$

for arbitrary $z \in G$ and $\zeta \in G_*$.

Now we are slowly approaching applications in physics. We start with some function L depending on variables (t, x, v), and only v is getting exchanged. The variables (t, x) are just parameters.

Suppose that a function $L = L(t, x, v)$ is given on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$, we call it *Lagrange function*, and the values of L are real numbers. Our key assumption is that the mapping

$$\Phi\colon (t, x, v) \mapsto \Phi(t, x, v) := (t, x, L_v(t, x, v))$$

is an invertible mapping of $\mathbb{R}_t \times \mathbb{R}^n_x \times \mathbb{R}^n_v$ onto $\mathbb{R}_t \times \mathbb{R}^n_x \times \mathbb{R}^n_p$, where, following the spirit of (1.11),

$$p := L_v(t, x, v), \tag{1.16}$$

and L_v denotes the partial derivative. The inverse mapping of Φ is called Ψ, and we have

$$(t, x, v) = \Psi(t, x, p) \quad \text{if} \quad (t, x, p) = \Phi(t, x, v).$$

We set $z := v$, $f(z) := L(t, x, z)$ and $f_* =: H$. Then we have $\zeta = L_v = p$, and (1.12) turns into

$$H(t, x, p) := \langle p, v \rangle - L(t, x, v) \quad \text{if} \quad (t, x, v) = \Psi(t, x, p). \tag{1.17}$$

Proposition 1.34. *If $(t, x, v) = \Psi(t, x, p)$, then we have:*

$$\begin{aligned}
L(t, x, v) + H(t, x, p) &= \langle p, v \rangle\,, && (1.18)\\
L_t(t, x, v) + H_t(t, x, p) &= 0, && (1.19)\\
L_x(t, x, v) + H_x(t, x, p) &= 0, && (1.20)\\
p &= L_v(t, x, v), && (1.21)\\
v &= H_p(t, x, p). && (1.22)
\end{aligned}$$

Proof. The identity (1.18) is just the definition (1.17) of H. Differentiating (1.18) with respect to t gives (1.19), differentiating with respect to x implies (1.20). And (1.21) is only a repetition of the definition (1.16). Finally, (1.22) is a reformulation of (1.13). □

1.9.2 The Legendre Transform in Classical Mechanics

Consider a physical system, whose state at time t is described by a vector of real numbers $x(t) = (x_1(t), \dots, x_n(t))$, which can be understood as something like "position variables", not necessarily of Cartesian type.

We also have $\dot{x}(t) = (\dot{x}_1(t), \dots, \dot{x}_n(t))$, to be understood as something like "velocity variables".

Next, we select a *Lagrange function* $L = L(t, x, \dot{x})$. Physical principles recommend to choose $L = E_{\text{kin}} - E_{\text{pot}}$, the difference between kinetic and potential energies.

Choose an arbitrary time interval $[t_0, t_1]$ and consider the so-called *action integral*

$$W(x) := \int_{t=t_0}^{t=t_1} L(t, x(t), \dot{x}(t))\, \mathrm{d}t.$$

This mapping W takes a function x (which describes the state of the system during the time interval $[t_0, t_1]$) and maps this function x to a real number $W(x)$. The vector space of all admissible functions x is infinite-dimensional, because function vector spaces are almost always infinite-dimensional.

The Hamilton principle demands that the variation of W must vanish, which is a physicist's formulation of the condition that all directional derivatives (in the sense of Definition 1.7, but now for x coming from a space of infinite dimension) must be zero.

In the physics lectures, it is shown that this Hamilton principle then implies that the function x is obliged to satisfy the Euler–Lagrange equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} L_v(t, x(t), \dot{x}(t)) - L_x(t, x(t), \dot{x}(t)) = 0,$$

where we have introduced $v = \dot{x}$. This is a nonlinear system of differential equations of second order. It can be transformed to the equivalent first order system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v, \qquad \frac{\mathrm{d}}{\mathrm{d}t} L_v(t, x, v) = L_x(t, x, v).$$

However, this system is not so beautiful: the second equation looks quite different from the first, it is hard to see what the conserved quantities are, and it seems impossible to find a quantum mechanical equivalent of that approach.

The good news is that the Legendre transform improves the situation in all three aspects: we go from the variables (t, x, v) to the variables (t, x, p), where $p := L_v(t, x, v)$ are called the *canonical momentum* variables. Next we fix the Legendre transform of L,

$$H(t, x, p) := \langle p, v \rangle - L(t, x, v)$$

as usual, and this function H is called the *Hamiltonian of the system.*

Lemma 1.35. *The functions $x = x(t)$ and $p = p(t)$ are solutions to the system*

$$\frac{\mathrm{d}x}{\mathrm{d}t} = H_p(t, x, p), \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -H_x(t, x, p). \tag{1.23}$$

Proof. First we exploit (1.22):

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v = H_p(t, x, p).$$

Second we make use of (1.21) and (1.20):

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} p = \frac{\mathrm{d}}{\mathrm{d}t} L_v(t, x, v) = L_x(t, x, v) = -H_x(t, x, p).$$

□

We observe that the Legendre transform converts the Euler-Lagrange system into the *Hamiltonian system* (1.23), which looks much more symmetrical.

And it keeps getting better:

Lemma 1.36. *If the Hamiltonian H does itself not depend on the time t, then the function $t \mapsto H(x(t), p(t))$ is constant.*

Physicists call this phenomenon *conservation of energy.*

Proof. This is an exercise of the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t} H(x(t), p(t)) = H_x \dot{x} + H_p \dot{p} = H_x \cdot H_p + H_p \cdot (-H_x) = 0.$$

□

There is one more conserved quantity, albeit not so easy to recognize—the phase space volume. The phase space is the domain of the x and p variables together. Take a domain $\Omega_0 \subset \mathbb{R}^n_x \times \mathbb{R}^n_p$ which is open and bounded. The temporal evolution over a time interval $[0, t]$ transports $(x(0), p(0)) \in \Omega_0$ to a point $(x(t), p(t)) \in \Omega_t \subset \mathbb{R}^n_x \times \mathbb{R}^n_p$, and the key result is:

The subsets Ω_0 and Ω_t of the phase space have the same volume.

This conservation of the phase space volume is only visible in the Hamiltonian picture. In theoretical mechanics, this is known as *Liouville's Theorem.*

As an example, we consider a pendulum: there is a point mass m hanging at a massless wire of length l, and it is swinging frictionless in a plane. Our first step is to choose a state variable x, which will be the angle of the wire versus the vertical axis:

$$x(t) = \theta(t).$$

Attention: this is not a Cartesian coordinate. Then the "velocity" is

$$v(t) = \dot{x}(t) = \dot{\theta}(t).$$

Attention again: this is not the usual velocity which would have a unit of "meters per second".

The kinetic energy is

$$T = T(v) = \frac{m}{2} l^2 v^2,$$

and the potential energy is

$$V = V(x) = mgl(1 - \cos\theta) = mgl(1 - \cos x).$$

The Lagrangian then is

$$L = L(x, v) = T(v) - V(x) = \frac{m}{2} l^2 v^2 - mgl + mgl\cos x.$$

We have the derivatives

$$L_v = ml^2 v, \qquad L_x = -mgl\sin x,$$

and the Euler–Lagrange equations turn into

$$ml^2\dot{v} + mgl\sin x = 0,$$

which can be written as $\ddot{x} + \frac{g}{l}\sin x = 0$. This was the Lagrangian approach.

Now we come to the Hamiltonian approach. The canonical momentum variable is $p = L_v$, hence

$$p = ml^2 v, \qquad v = \frac{p}{ml^2}.$$

The (time-independent) Hamiltonian H then is the Legendre transform of L,

$$\begin{aligned} H(t, x, p) &= pv - L(x, v) = pv - (T(v) - V(x)) = p \cdot \frac{p}{ml^2} - \left(\frac{m}{2} l^2 v^2 - mgl + mgl\cos x\right) \\ &= \frac{p^2}{ml^2} - \left(\frac{m}{2} l^2 \cdot \frac{p^2}{m^2 l^4} - mgl + mgl\cos x\right) \\ &= \frac{p^2}{2ml^2} + mgl(1 - \cos x). \end{aligned}$$

This is exactly the mechanical energy of the system, and the Hamiltonian system (1.23) becomes

$$\dot{x} = H_p = \frac{p}{ml^2}, \qquad \dot{p} = -H_x = -mgl\sin x.$$

Finally, we have a (very deep, admittedly) look at the geometrical meanings of the Lagrangian and Hamiltonian approaches.

The position of the pendulum has been characterized by one real variable, the angle θ. But we can also take Cartesian coordinates $(x_1, x_2) \in \mathbb{R}^2$ which have their origin at the anchor point of the pendulum. Of course not all positions $(x_1, x_2) \in \mathbb{R}^2$ are admissible: only those with $x_1^2 + x_2^2 = l^2$ can be positions of the pendulum. This arc of a circle with radius l is a one-dimensional *manifold*[19] contained in the two-dimensional space $\mathbb{R}^2$. The space variable x lives on M, and now we tell where the velocity variable v and the generalized momentum p can be found. At each point $x \in M$, we attach the tangent line, which can be understood as a one-dimensional vector space with origin at the contact point x. This vector space is denoted by T_xM, it is called the *tangential space* at $x \in M$, and the length of its elements is measured in a certain unit. If the pendulum is at position x and has velocity v, then this velocity can be found in the space T_xM.

Now we let x vary through M and build the union of all the sets $\{x\} \times T_xM$. This union is a two-dimensional manifold TM, the so–called *tangential bundle*[20] of M. This is the domain where the variables (x, v) live, and the Lagrangian approach consists of differential equations on the tangential bundle.

Another possibility is to attach at the point $x \in M$ the dual space to the vector space T_xM. This dual space is denoted by T_x^*M, called the *cotangential space* at $x \in M$, it is also one-dimensional, and the length of its elements is measured in the reciprocal unit, compared to T_xM. Constructing the union of all the $\{x\} \times T_x^*M$ for x running through M gives the *cotangential bundle* T^*M, which is a two-dimensional manifold, where the variables (x, p) live. Then the Hamiltonian differential equations are differential equations on the cotangent bundle.

As a summary: the Legendre transform translates between the tangential bundle TM and the cotangential bundle T^*M.

1.9.3 The Legendre Transform in Thermodynamics

Consider a system with n kinds of particles. For air you might choose $n = 3$ and consider nitrogen, oxygen and water vapour. The state of this system is described by the variables

particle numbers: these are $N = (N_1, \dots, N_n) \in \mathbb{R}^n$. Here we are cheating a bit, because these numbers should be natural, but ...

absolute temperature: this is T, measured in Kelvin,

volume: this is V,

pressure: p,

entropy: S,

chemical potentials: these are numbers μ_j with $j = 1, \dots, n$.

As base variables we could choose T, V and N, and then all the other variables p, S, μ would depend on (T, V, N).

It is custom to distinguish between *extensive* and *intensive* variables. The intensive ones do not change if you “double the system”, but the extensive variables will change. In our setting, the extensive variables are S, V, N, and the intensive variables are T, p, μ.

The total energy of the system is called *inner energy*. This energy measures all the energy in the system, it is denoted by U, and its natural variables are (S, V, N), by definition. Note that these are exactly the extensive variables. In the physics lectures, the following identities will be proved/mentioned:

$$\frac{\partial U(S,V,N)}{\partial S} = T, \qquad \frac{\partial U(S,V,N)}{\partial V} = -p, \qquad \frac{\partial U(S,V,N)}{\partial N_j} = \mu_j.$$

[19] Mannigfaltigkeit
[20] Tangentialbündel

Observe that the right-hand sides are intensive variables. These identities can be summarized in the following formula for the total differential of U:

$$\mathrm{d}U = T\,\mathrm{d}S - p\,\mathrm{d}V + \sum_{j=1}^{n} \mu_j\,\mathrm{d}N_j.$$

This identity is called *Gibbs relation*, and it is no overstatement to claim that this is one of THE key relations of thermodynamics. For instance, the *first law of thermodynamics* (the conservation of energy) can be deduced from the Gibbs relation.

In the sequel, we will replace extensive variables by intensive variables via the Legendre transform, leading to a bunch of new thermodynamic potentials. Each thermodynamic potential has its own set of *natural variables*. These natural variables are fixed by definition; and if you choose the wrong ones you can produce an arbitrary amount of incorrect formulas.

the free energy: we replace $(S, V, N) \mapsto (T, V, N)$. Set $z = S$ and $f(z) := U(z, V, N)$. Then we have

$$\zeta := \frac{\partial U}{\partial z} = \frac{\partial U}{\partial S} = T,$$

and therefore the Legendre transform of f becomes

$$f_*(\zeta) = \zeta z - f(z) = TS - U(S, V, N),$$

which suggests to set $F := U - TS$ with natural variables (T, V, N). This is called the *free energy*, and it counts how much energy of the system is accessible to us (roughly spoken). We compute the derivatives of F: by (1.13), we have

$$\frac{\partial F}{\partial T}(T, V, N) = -\frac{\partial f_*}{\partial \zeta}(\zeta, V, N) = -z = -S,$$

and the other derivatives are unchanged:

$$\frac{\partial F}{\partial V} = \frac{\partial U}{\partial V}, \qquad \frac{\partial F}{\partial N_j} = \frac{\partial U}{\partial N_j}.$$

Then we arrive at the total differential of F:

$$\mathrm{d}F = -S\,\mathrm{d}T - p\,\mathrm{d}V + \sum_{j=1}^{n} \mu_j\,\mathrm{d}N_j.$$

the enthalpy: we replace $(S, V, N) \mapsto (S, p, N)$. Set $z = V$ and $f(z) := U(S, z, N)$. Then we have

$$\zeta := \frac{\partial U}{\partial z} = \frac{\partial U}{\partial V} = -p,$$

and therefore the Legendre transform of f becomes

$$f_*(\zeta) = \zeta z - f(z) = -pV - U(S, V, N),$$

which suggests to set $H := U + pV$ with natural variables (S, p, N). This is called the *enthalpy* of the system. We compute the derivatives of H: by (1.13), we have

$$\frac{\partial H}{\partial p}(S, p, N) = -\frac{\partial}{\partial p} f_*(\zeta) = \frac{\partial}{\partial \zeta} f_*(\zeta) = z = V,$$

and the other derivatives are unchanged:

$$\frac{\partial H}{\partial S} = \frac{\partial U}{\partial S}, \qquad \frac{\partial H}{\partial N_j} = \frac{\partial U}{\partial N_j}.$$

Then we arrive at the total differential of H:

$$\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}p + \sum_{j=1}^{n} \mu_j\,\mathrm{d}N_j.$$

the free enthalpy: we replace $(T, V, N) \mapsto (T, p, N)$. Set $z = V$ and $f(z) := F(T, z, N)$. Then we have

$$\zeta := \frac{\partial F}{\partial z} = \frac{\partial F}{\partial V} = -p,$$

and therefore the Legendre transform of f becomes

$$f_*(\zeta) = \zeta z - f(z) = -pV - F(T, V, N),$$

which suggests to set $G := F + pV$ with natural variables (T, p, N). This is called the *free enthalpy* or *Gibbs potential* of the system. We compute the derivatives of G: by (1.13), we have

$$\frac{\partial G}{\partial p}(T, p, N) = -\frac{\partial}{\partial p} f_*(\zeta) = \frac{\partial}{\partial \zeta} f_*(\zeta) = z = V,$$

and the other derivatives are unchanged:

$$\frac{\partial G}{\partial T} = \frac{\partial F}{\partial T} = -S, \qquad \frac{\partial G}{\partial N_j} = \frac{\partial F}{\partial N_j}.$$

Then we arrive at the total differential of G:

$$\mathrm{d}G = -S\,\mathrm{d}T + V\,\mathrm{d}p + \sum_{j=1}^{n} \mu_j\,\mathrm{d}N_j.$$

the Landau potential: we replace $(T, V, N) \mapsto (T, V, \mu)$. Set $z = N$ and $f(z) := F(T, V, z)$. Then we have

$$\zeta := \frac{\partial F}{\partial z} = \frac{\partial F}{\partial N} = \mu,$$

and therefore the Legendre transform of f becomes

$$f_*(\zeta) = \langle \zeta, z \rangle - f(z) = \sum_{j=1}^{n} \mu_j N_j - F(T, V, N),$$

which suggests to set $\Omega := F - \sum_{j=1}^{n} \mu_j N_j$ with natural variables (T, V, μ). This is called the *Landau potential* or *Grand potential* of the system. We compute the derivatives of Ω: by (1.13), we have

$$\frac{\partial \Omega}{\partial \mu}(T, V, \mu) = -\frac{\partial}{\partial \zeta} f_*(\zeta) = -z = -N,$$

and the other derivatives are unchanged:

$$\frac{\partial \Omega}{\partial T} = \frac{\partial F}{\partial T} = -S, \qquad \frac{\partial \Omega}{\partial V} = \frac{\partial F}{\partial V} = -p.$$

Then we arrive at the total differential of Ω:

$$\mathrm{d}\Omega = -S\,\mathrm{d}T - p\,\mathrm{d}V - \sum_{j=1}^{n} N_j\,\mathrm{d}\mu_j.$$

1.10 Keywords

- derivative, Jacobi matrix,
- calculation rules,
- Theorem of Schwarz, Taylor formula,
- determination of extrema,
- implicit function theorem,
- Cauchy-Riemann differential equations.

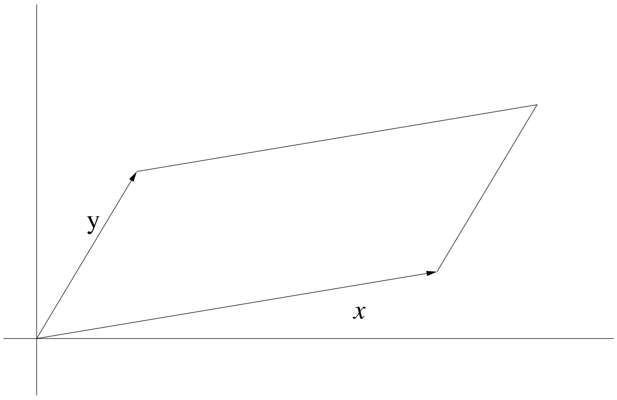
Chapter 2

Determinants

In the introductory chapter, we have considered the determinants of 3 vectors in $\mathbb{R}^3$ and their geometric properties. Now we will do the same for n vectors in $\mathbb{R}^n$, and we will do it more generally. The purpose of the determinants is multiple; they are exploited

- when investigating the inverse matrix,
- during the study of linear systems,
- in the theory of eigenvalues of matrices,
- when integrating functions of several variables.

Let us go back to the $\mathbb{R}^2$ for a moment, and consider the area of a parallelogram.



It is easy to verify that

$$A = A(x, y) = 2\left(\frac{(x_1 + y_1)(x_2 + y_2)}{2} - \frac{x_1 x_2}{2} - \frac{(x_2 + (x_2 + y_2))y_1}{2}\right) = x_1 y_2 - y_1 x_2.$$

We know a similar formula for the volume of three-dimensional parallelepipedon.

Let us look at these formulae a bit closer:

1. $A = A(x, y)$ is a linear function in x as well as y, in the sense of $A(\lambda x, y) = \lambda A(x, y)$ and $A(x + \tilde{x}, y) = A(x, y) + A(\tilde{x}, y)$; and similarly for the second variable.
2. If the vectors x and y are linearly dependent, then $A(x, y) = 0$.
3. The area of a unit square is one: $A(e_1, e_2) = 1$.

These three properties will be the foundations of our definition of *determinant functions*.

2.1 Determinant Functions

For K being a field[1], in general, $K = \mathbb{R}$ or $K = \mathbb{C}$, the vector space of vectors with n components from K is denoted by K^n. A determinant function is a function that takes n vectors of this kind and maps them to a number from K; and has to satisfy some additional conditions as listed in the following definition.

Definition 2.1 (**Determinant function**). *Let $(K^n)^n$ denote the set of n–tuples of vectors from K^n. A function $\Delta\colon (K^n)^n \to K$ is a* normalised determinant function[2] *if the following conditions hold:*

D1 *The function Δ is linear in each of its n arguments, i.e.,*

$$\Delta(\alpha_1 x_1+\beta_1 y_1, x_2, \dots, x_n) = \alpha_1\Delta(x_1, x_2, \dots, x_n)+\beta_1\Delta(y_1, x_2, \dots, x_n), \qquad x_j, y_1 \in K^n, \quad \alpha_1, \beta_1 \in K,$$

and accordingly for the other components.

D2 *If the vectors $x_1, \dots, x_n$ are linearly dependent, then $\Delta(x_1, \dots, x_n) = 0$.*

D3 *If $e_1, \dots, e_n$ denote the canonic basis vectors of K^n, then $\Delta(e_1, \dots, e_n) = 1$.*

If only D1 and D2 hold then Δ is said to be a determinant function.

Soon we will see that there is exactly one normalised determinant function on K^n.

First, we derive some computing rules.

Proposition 2.2 (**Computing rules**). *Let Δ be a determinant function, not necessarily normalised.*

- *Adding a multiple of one argument to another argument of Δ does not change the value, which means,*

$$\Delta(x_1, \dots, x_n) = \Delta(x_1, \dots, x_{i-1}, x_i + \lambda x_j, x_{i+1}, \dots, x_n), \qquad i \neq j.$$

- *Exchanging two arguments is equivalent to multiplying Δ with -1:*

$$\Delta(\dots, x_i, \dots, x_j, \dots) = -\Delta(\dots, x_j, \dots, x_i, \dots)$$

Proof. The first claim follows from D1 and D2:

$$\begin{aligned}&\Delta(x_1, \dots, x_{i-1}, x_i + \lambda x_j, x_{i+1}, \dots, x_n)\\ &\quad = \Delta(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) + \lambda\Delta(x_1, \dots, x_{i-1}, x_j, x_{i+1}, \dots, x_n)\\ &\quad = \Delta(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) + 0.\end{aligned}$$

Repeated application of the first claim gives the second as follows.

$$\begin{aligned}\Delta(x_1, x_2, \dots) &= \Delta(x_1, x_2 - x_1, \dots) = \Delta(x_1 + (x_2 - x_1), x_2 - x_1, \dots)\\ &= \Delta(x_2, x_2 - x_1, \dots) = \Delta(x_2, (x_2 - x_1) - x_2, \dots) = \Delta(x_2, -x_1, \dots)\\ &= -\Delta(x_2, x_1, \dots).\end{aligned}$$

Here we have chosen $i = 1$ and $j = 2$ to keep the notations easy. □

Remark 2.3. *Take $n = 2$. Then the identity $\Delta(x_1, x_2 + \lambda x_1) = \Delta(x_1, x_2)$ corresponds to the fact that all parallelograms with same base line and same height have the same area. And the identity $\Delta(x_2, x_1) = -\Delta(x_1, x_2)$ expresses the convention that flipping the two spanning edges of the parallelogram changes the sign of the area.*

Now we are ready to show that there is exactly one normalised determinant function on $(K^n)^n$.

Proposition 2.4 (**Uniqueness of the normalised determinant function**). *There is at most one normalised determinant function on $(K^n)^n$.*

[1] Körper

[2] normierte Determinantenfunktion

Proof. We show the following: if there is a normalised determinant function Δ, then its values can be computed using a procedure, which will be presented now.

Let us be given n vectors $x_1, \ldots, x_n$ from K^n. If they are linearly dependent, then Δ must take the value zero on them. Therefore, let these vectors be linearly independent.

We can write these vectors as row vectors, one below the other, so that they form a matrix from $K^{n\times n}$. This matrix has full rank, since the vectors are linearly independent. Then we execute the Gauss–Jordan algorithm of triangulising a matrix, repeatedly applying one of the following steps:

- adding a multiple of one row to another row (this does not change the value of Δ);
- exchanging two rows (this changes only the sign of Δ).

After several steps, we end up with a matrix of the form

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix},$$

where we have to remember how often we have exchanged two rows—an even number of times or an odd number of times. According to D1 and D3, the value of Δ is then

$$\Delta(x_1, \ldots, x_n) = \pm a_{11} \cdot \ldots \cdot a_{nn}.$$

This completes the proof. □

Up to now, the existence of a determinant function has not been established: maybe there is a contradiction which can be deduced from D1, D2 and D3; but we just haven't found it yet ?

Proposition 2.5 (Existence of a normalised determinant function). *For each $n \in \mathbb{N}_+$, there is a normalised determinant function.*

Proof. If $n = 1$ then $(K^1)^1 = K$ and a determinant function is given by

$$\begin{aligned} &\Delta_1 \colon (K^1)^1 \to K, \\ &\Delta_1 \colon (x_1) \mapsto x_{11}, \end{aligned}$$

where x_{11} is the first (and only) component of the vector x_1 from $K^1 = K$.

Let $n = 2$, and put $x_1 = (x_{11}, x_{21})^\top$, $x_2 = (x_{12}, x_{22})^\top$. Then

$$\begin{aligned} &\Delta_2 \colon (K^2)^2 \to K, \\ &\Delta_2 \colon (x_1, x_2) = \left(\begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}, \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix} \right) \mapsto x_{11}x_{22} - x_{12}x_{21} \end{aligned}$$

is a normalised determinant function.

Let now $n \in \mathbb{N}$, $n \geq 3$ be arbitrary. We define Δ_n by induction. For this, we need a notation: if $x \in K^n$ with $x = (\xi_1, \ldots, \xi_n)$, then $x^{(i)}$ is that vector from K^{n-1}, which can be obtained by crossing out the ith component of x.

Let now $1 \leq i \leq n$ be arbitrary. Then we set

$$\begin{aligned} &\Delta_n \colon (K^n)^n \to K, \\ &\Delta_n \colon (x_1, \ldots, x_n) = \left(\begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix}, \ldots, \begin{pmatrix} x_{1n} \\ \vdots \\ x_{nn} \end{pmatrix} \right) \mapsto \sum_{j=1}^{n} (-1)^{i+j} x_{ij} \Delta_{n-1}(x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}), \end{aligned}$$

where x_{ij} is the ith component of the vector x_j.

We skip the proof that this function satisfies D1, D2 and D3. □

As an example, we take $x_1 = (1,2,3)^\top$, $x_2 = (4,5,6)^\top$, $x_3 = (7,8,9)^\top$ and $i = 2$. Then

$$\Delta_3\left(\begin{pmatrix}1\\2\\3\end{pmatrix},\begin{pmatrix}4\\5\\6\end{pmatrix},\begin{pmatrix}7\\8\\9\end{pmatrix}\right) = -2\Delta_2\left(\begin{pmatrix}4\\6\end{pmatrix},\begin{pmatrix}7\\9\end{pmatrix}\right) + 5\Delta_2\left(\begin{pmatrix}1\\3\end{pmatrix},\begin{pmatrix}7\\9\end{pmatrix}\right) - 8\Delta_2\left(\begin{pmatrix}1\\3\end{pmatrix},\begin{pmatrix}4\\6\end{pmatrix}\right),$$

and the three determinants Δ_2 on the right–hand side can be evaluated in the same way as above. You could choose $i = 1$ or $i = 3$ and obtain the same value of $\Delta_3(x_1, x_2, x_3)$.

For each i, we obtain one normalised determinant function. Since there can be only one of them, they must all coincide. The following formula gives some more representations of that Δ_n:

$$\Delta_n(x_1, \dots, x_n) = \sum_{i=1}^n (-1)^{i+j} x_{ij} \Delta_{n-1}(x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_{j+1}^{(i)}, \dots, x_n^{(i)}), \qquad j = 1, \dots, n.$$

Similarly we obtain:

Corollary 2.6. *For each $\lambda \in K$, there is exactly one determinant function Δ (satisfying D1, D2) with*

$$\Delta(e_1, \dots, e_n) = \lambda.$$

This determinant function is given by $\Delta(x_1, \dots, x_n) := \lambda\Delta_{\text{norm.}}(x_1, \dots, x_n)$.

Remark 2.7. *We note that another popular way of writing is $x_1 \wedge x_2 \wedge \dots \wedge x_n = \Delta_n(x_1, \dots, x_n)$, sometimes called the* exterior (outer) product, *in contrast to the* inner product *(which is the scalar product of two vectors). This wedge symbol $\wedge$ has the same meaning as in the differential forms, by the way.*

2.2 The Determinant of a Matrix

From now on, Δ denotes the one and only normalised determinant function on $(K^n)^n$.

Definition 2.8 (Determinant). *Let $A \in K^{n\times n}$ be a matrix with columns a_1, …, a_n. Then*

$$\det A := \Delta(a_1, \dots, a_n)$$

is called determinant of the matrix A[3].

Next comes the key result concerning determinants: they are compatible to the matrix-matrix-multiplication.

Proposition 2.9. *Let A, $B \in K^{n\times n}$. Then* $\det(BA) = \det(B) \cdot \det(A)$.

Proof. We define a function

$$\Delta_B(x_1, \dots, x_n) := \Delta(Bx_1, \dots, Bx_n).$$

It is easy to check that this function fulfils D1 and D2. Then it must be a determinant function. We fix a number $\lambda \in K$ by

$$\lambda := \Delta_B(e_1, \dots, e_n).$$

By the second part of Corollary 2.6, we have $\Delta_B(x_1, \dots, x_n) = \lambda\Delta(x_1, \dots, x_n)$. Now we compute the number λ:

$$\lambda = \Delta_B(e_1, \dots, e_n) = \Delta(Be_1, \dots, Be_n) = \Delta(b_1, \dots, b_n) = \det B,$$

where b_1, …, b_n are the columns of B. Then we obtain

$$\det(BA) = \Delta(Ba_1, \dots, Ba_n) = \Delta_B(a_1, \dots, a_n) = \lambda\Delta(a_1, \dots, a_n) = \det B \cdot \det A.$$

□

[3] Determinante der Matrix A

This compatibility can be expressed as a commutative diagram,

$$\begin{tikzcd}
\boxed{(B,A)} \arrow[r] \arrow[d, "\det"'] & \boxed{BA} \arrow[d, "\det"'] \\
\boxed{(\det(B),\det(A))} \arrow[r, "\cdot"'] & \boxed{\begin{array}{c}\det(B)\cdot\det(A)\\ =\det(BA)\end{array}}
\end{tikzcd}$$

and the following proposition lists direct consequences of the compatibility identity:

Proposition 2.10. *Let $A, B \in K^{n\times n}$ and I_n the $n \times n$ unit matrix.*

1. *It holds* $\det I_n = 1$.
2. *If A is invertible, then* $\det(A^{-1}) = (\det A)^{-1}$.
3. *The rank of A is less than n if and only if* $\det A = 0$.
4. *If B is invertible, then* $\det(B^{-1}AB) = \det A$.

Proof. 1. This is D3.

2. Follows from $A^{-1} \cdot A = I_n$ for invertible A and Proposition 2.9.
3. If $\operatorname{rank} A < n$, then the columns of A are linearly dependent, and $\det A = 0$ because of D2. If $\operatorname{rank} A = n$, then A is invertible, and $\det A \neq 0$ because of part 2.
4. Follows from Proposition 2.9 and part 2.

□

The next result will give a method, which will show how to compute a determinant recursively.

Proposition 2.11 (**Expansion Theorem of** LAPLACE)**.** *Let $A \in K^{n\times n}$ be a matrix with entries a_{kl}. We write A_{ij} for that matrix from $K^{(n-1)\times(n-1)}$, which can be obtained from A by omitting row i and column j. Then we have for each $1 \le i \le n$:*

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}.$$

Moreover, we have for each $1 \le j \le n$:

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}.$$

Proof. See the end of the proof of Proposition 2.5. □

Question: What do you obtain for $n = 3$? What is the name of this object ?

Proposition 2.12 (**Computing rules**)**.** *Let $A \in K^{n\times n}$. Then we have:*

- $\det A = \det A^\top$.
- *If A has two equal rows or two equal columns, then* $\det A = 0$.
- *Adding a multiple of one row to another row preserves the determinant. Similarly for columns.*
- *Exchanging two rows multiplies the determinant by -1, so does exchanging two columns.*
- *The determinant of a triangular matrix equals the product of the entries on the diagonal.*

Proof. These are more or less direct consequences of the Laplace expansion theorem and the Gauss–Jordan algorithm for computing determinant functions. □

2.3 Applications to Linear Systems

Determinants can be utilised for inverting matrices and solving linear systems, as we will see soon.

Definition 2.13 (**Algebraic Complement**). *For $A \in K^{n\times n}$, we write A_{kl} for that matrix, which is obtained by omitting row k and column l. Then the numbers*

$$D_{ij}(A) := (-1)^{i+j} \det A_{ij}$$

are the algebraic complements[4] *of the matrix A.*

Lemma 2.14. *Take $A \in K^{n\times n}$ and a column vector $x = (x_1, \dots, x_n)^\top \in K^n$. Write $A_{j,x}$ for that matrix, which you get after replacing the jth column of A by the vector x. Then we have*

$$\det A_{j,x} = \sum_{i=1}^{n} x_i D_{ij}(A).$$

Proof. Just expand $A_{j,x}$ along the jth column, using the Laplace expansion theorem. □

Especially, we can choose the kth column of A for the vector x. Then the following result is easy:

Lemma 2.15. *Let $A \in K^{n\times n}$ with entries a_{ij}. Then the following identity holds for all j, k with $1 \le j, k \le n$:*

$$\sum_{i=1}^{n} a_{ik} D_{ij}(A) = \delta_{kj} \det A.$$

Here δ_{kj} is the Kronecker[5] *symbol.*

Note the positions of the two indices i on the left–hand side ! They do not stand next to each other.

Transposing the matrix of the D_{ij} and dividing by $\det A$ then gives you a formula for the inverse matrix:

Lemma 2.16 (**Formula for the inverse matrix**). *Let $A \in K^{n\times n}$ be an invertible matrix. Then the inverse matrix A^{-1} with entries $(A^{-1})_{ij}$ is given by*

$$(A^{-1})_{ij} = \frac{D_{ji}(A)}{\det A}, \qquad 1 \le i, j \le n.$$

Proof. Should be clear (try it with a 2×2 matrix A). □

Then you immediately get a formula for the solution of a linear system:

Proposition 2.17 (Cramer's[6] **rule**). *Let $A \in K^{n\times n}$ be invertible and $b \in K^n$. Then the solution $x = (x_1, \dots, x_n)^\top \in K^n$ to the linear system $Ax = b$ can be found by*

$$x_i = \frac{\det A_{i,b}}{\det A}, \qquad 1 \le i \le n,$$

where $A_{i,b}$ is the matrix which you get by replacing the ith column of A by b.

Proof. The assertion follows immediately from

$$x_i = \sum_{j=1}^{n} (A^{-1})_{ij} b_j = \frac{1}{\det A} \sum_{j=1}^{n} b_j D_{ji}(A).$$

□

Cramer's rule is quite handy in the case $n = 2$ and maybe for $n = 3$ as well.

[4] algebraische Komplemente

[5] Leopold Kronecker, 1823 – 1891

[6] Gabriel Cramer, 1704 – 1752

You should never use Cramer's rule for $n \geq 4$.
The effort becomes quickly huge as n increases;
and the solution formulas are not numerically stable.

Instead, you should use one of the following:

- Gauss–Jordan algorithm with pivotisation,
- LR factorisation with pivotisation,
- QR factorisation,
- special iterative algorithms for large n, say, $n \geq 10^3$.

Remark 2.18. *The QR factorisation means: let us be given a matrix $A \in \mathbb{R}^{n\times n}$. Then we search for matrices Q and R such that $QQ^\top = I_n$, R is an upper triangular matrix, and $A = QR$. Such a matrix Q is called* orthogonal, *and its columns form an orthonormal system. Applying the Gram–Schmidt orthonormalisation procedure to the columns of A is equivalent to a QR factorisation of A. Finally, we mention that decompositions $A = QR$ even exist when A has non-quadratic shape, for instance $A \in \mathbb{R}^{m\times n}$ with $m \neq n$.*

2.4 Determinants and Permutations

Literature: Greiner and Müller: *Quantenmechanik. Symmetrien.* Chapter IX: Darstellungen der Permutationsgruppe und Young–Tableaux

We have formulas for determinants in the cases $n = 2$ and $n = 3$ which involve sums of products of the matrix elements. Now we would like to generalise these formulas to higher n. For this, we need some new concepts.

Definition 2.19 (Permutation). *For $n \in \mathbb{N}_+$, denote by S_n the set of all bijective mappings*

$$\pi\colon \{1, \dots, n\} \to \{1, \dots, n\}.$$

These mappings are called permutations[7] of $\{1, \dots, n\}$.

A permutation is just a reshuffling of the (ordered) numbers from 1 to n.

Proposition 2.20. *The set S_n, together with the composition as an operation, is a group.*

Proof. You have to check that the composition of two permutations is again a permutation, that the composition is associative, that there is a unit element, and that for each permutation, there is an inverse permutation. The details are left to the student. □

This group is the so–called *symmetric group*[8].

Question: How many elements does S_n have ?

Next we connect the structure given by the group $(S_n, \circ)$ to the structure of matrices and their products.

Definition 2.21 (Permutation matrix). *Let $e_1, \dots, e_n \in K^n$ be the canonical basis vectors in K^n and $\pi \in S_n$. Then*

$$P_\pi := (e_{\pi(1)}, \dots, e_{\pi(n)}) \in K^{n\times n}$$

is the permutation matrix associated to π[9]. *The kth column of P_π is just the $\pi(k)$th basis vector.*

There is a compatibility relation between the group operation $\circ$ and the matrix product:

[7] Permutationen
[8] symmetrische Gruppe
[9] Permutationsmatrix zur Permutation π

Lemma 2.22. *For any two permutations $\sigma, \pi \in S_n$ we have*

$$P_{\sigma\circ\pi} = P_\sigma P_\pi.$$

Proof. The columns are the images of the unit basis vectors. Denoting the jth column of $P_\sigma P_\pi$ by $(P_\sigma P_\pi)_j$, we then have

$$(P_\sigma P_\pi)_j = (P_\sigma P_\pi)e_j = P_\sigma(P_\pi e_j) = P_\sigma e_{\pi(j)} = e_{(\sigma\circ\pi)(j)} = (P_{\sigma\circ\pi})_j.$$

We can do this for each column index j, which completes the proof. □

Let us express this compatibility in terms of a commutative diagram:

$$\begin{tikzcd}
\boxed{(\sigma,\pi)} \arrow[r, "\circ"] \arrow[d] & \boxed{\sigma\circ\pi} \arrow[d] \\
\boxed{(P_\sigma, P_\pi)} \arrow[r] & \boxed{\begin{array}{c} P_\sigma P_\pi \\ = P_{\sigma\circ\pi} \end{array}}
\end{tikzcd}$$

We introduce one more structure:

Definition 2.23 (Sign of a permutation). *The sign of a permutation π is defined by*

$$\operatorname{sign}(\pi) := \det P_\pi.$$

And this sign structure induces a compatibility between the group operation $\circ$ in S_n and the multiplication in $\mathbb{R}$:

Lemma 2.24. *For any two permutations $\sigma, \pi \in S_n$ we have*

$$\operatorname{sign}(\sigma\circ\pi) = \operatorname{sign}(\sigma)\cdot\operatorname{sign}(\pi).$$

Proof. This is a direct consequence of the two compatibility relations from Proposition 2.9 and Lemma 2.22:

$$\operatorname{sign}(\sigma\circ\pi) = \det(P_{\sigma\circ\pi}) = \det(P_\sigma P_\pi) = \det(P_\sigma)\cdot\det(P_\pi) = \operatorname{sign}(\sigma)\cdot\operatorname{sign}(\pi).$$

This is what we wanted to show. □

For completeness, we draw one more commutative diagram:

$$\begin{tikzcd}
\boxed{(\sigma,\pi)} \arrow[r, "\circ"] \arrow[d, "\operatorname{sign}"'] & \boxed{\sigma\circ\pi} \arrow[d, "\operatorname{sign}"] \\
\boxed{(\operatorname{sign}(\sigma), \operatorname{sign}(\pi))} \arrow[r, "\cdot"'] & \boxed{\begin{array}{c} \operatorname{sign}(\sigma)\cdot\operatorname{sign}(\pi) \\ = \operatorname{sign}(\sigma\circ\pi) \end{array}}
\end{tikzcd}$$

Now assume the following problem to solve. Given a permutation π, how to evaluate $\operatorname{sign}(\pi)$ with little effort, without handling $\det(P_\pi)$?

To this end, we consider the most simple permutations (except the identical permutation), which are those that only exchange two elements, the so–called *transpositions*[10] τ_{ij}, $i \neq j$:

$$\tau_{ij}(k) := \begin{cases} j\colon k = i, \\ i\colon k = j, \\ k\colon \text{else}. \end{cases}$$

Obviously, each transposition is its own inverse. Mappings of a set onto the same set that are their own inverses are so–called *involutions*[11].

We then quickly find:

[10] Transpositionen

[11] Involutionen

- the sign of each transposition is -1,
- each permutation from S_n is a composition of at most $n-1$ transpositions,
- if $\pi = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_K$ with each τ_j being a transposition, then $\operatorname{sign}(\pi) = (-1)^K$.

Permutations π with $\operatorname{sign}(\pi) = +1$ are called *even*, and permutations π with $\operatorname{sign}(\pi) = -1$ are called *odd*. By the way, we have shown that the composition of two even permutations is always an even permutation, and the composition of an even permutation with an odd permutation is always odd. We can even say that the subset of even permutations in S_n forms a sub-group of S_n.

Now we are in a position to give the final formula for a general determinant:

Proposition 2.25 (Leibniz[12] **formula for determinants).** *For each $A \in K^{n\times n}$ with entries $a_{i,j}$, we have*

$$\det A = \sum_{\pi\in S_n} \operatorname{sign}(\pi) a_{\pi(1),1} \cdot \ldots \cdot a_{\pi(n),n}.$$

Proof. The determinant is a linear function with respect to each column. Denote the columns of A by $a_1, \ldots, a_n$. We can decompose each column as $a_j = \sum_{i=1}^n a_{ij} e_i$. Then we have, by separate linearity in each argument of Δ,

$$\begin{aligned}\det A = \Delta(a_1, \ldots, a_n) &= \Delta\left(\sum_{i_1=1}^n a_{i_1 1} e_{i_1}, \sum_{i_2=1}^n a_{i_2 2} e_{i_2}, \ldots, \sum_{i_n=1}^n a_{i_n n} e_{i_n}\right) \\ &= \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_n=1}^n a_{i_1 1} a_{i_2 2} \cdot \ldots \cdot a_{i_n n} \Delta(e_{i_1}, e_{i_2}, \ldots, e_{i_n}).\end{aligned}$$

If two of the indices $i_1, \ldots, i_n$ coincide, then the determinant on the right–hand side vanishes, since its arguments are linearly dependent then. Therefore, the numbers i_1, ..., i_n must be the values of a certain permutation $\pi \in S_n$; $i_1 = \pi(1)$, ..., $i_n = \pi(n)$. This completes the proof. □

2.5 Outlook: Many Particle Schrödinger Functions

Consider an ensemble of N electrons. Their wave function is a function $\psi\colon \mathbb{R}^{3N} \to \mathbb{C}$, and we write it as $\psi = \psi(x_1, \ldots, x_N)$, with $x_j \in \mathbb{R}^3$ taking care of electron number j. Then $|\psi(x_1, \ldots, x_N)|^2$ describes the probability density of finding the electron ensemble at the location $(x_1, \ldots, x_N)$. Since the ensemble must be somewhere, we then have

$$\int_{x\in\mathbb{R}^{3N}} |\psi(x)|^2 \, \mathrm{d}x = 1.$$

Here we are cheating a bit and ignore the spins of the electrons.

The electrons can not be distinguished. Swapping the arguments x_2 and x_3 of ψ can be seen two ways: first, we can say that the electron 2 and electron 3 have exchanged their positions. Second, we can say that both have remained at their places, and only their names have been exchanged. In any case, a permutation of the electrons can not change the physical situation, hence we have, for each permutation $\pi \in S_N$,

$$|\psi(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)})|^2 = |\psi(x_1, x_2, \ldots, x_N)|^2.$$

Now the Pauli[13] principle comes in: there are never two electrons in the same state. In particular, two electrons can not be at the same place (we tacitly ignore the spins). This implies that we should have $\psi(x_1, \ldots, x_N) = 0$ whenever two of the x_j are equal (say $x_2 = x_3$, to have something specific).

And the Pauli principle can be guaranteed if we demand that

$$\psi(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)}) = \operatorname{sign}(\pi)\psi(x_1, x_2, \ldots, x_N),$$

for each permutation $\pi \in S_N$. This indeed implies $\psi(x_1, x_2, x_3, \ldots, x_N) = 0$ whenever $x_2 = x_3$. Just choose for the permutation π the transposition τ_{23}, for which we have $\operatorname{sign}(\tau_{23}) = -1$.

[12] Gottfried Wilhelm von Leibniz, 1646 – 1716

[13] Wolfgang Pauli, 1900 – 1958

2.6 Keywords

- calculation rules as in the Propositions 2.9, 2.10, and 2.12,
- formula for the inverse matrix,
- permutations.

Chapter 3

Integration in One Dimension, and Curves

There are several kinds of integrals in one dimension (later we will also consider integrals in higher dimensions):

Definite integrals: these are integrals over an interval of $\mathbb{R}$ (which may be unbounded); and the integral is a number from $\mathbb{R}$ or $\mathbb{C}$. This number is equal to "the area below the graph of the function". We will have to make this more precise.

Indefinite integrals or antiderivatives: a function F is the indefinite integral of a function f if $F' = f$ everywhere. We will learn how to find all such F for some given f.

Path integrals or line integrals: they are similar to the definite integrals; however, now you do not integrate over an interval of $\mathbb{R}$, but over some "path" in $\mathbb{R}^n$ or $\mathbb{C}$. We should make this more precise and study the properties of such integrals. Be careful—there are several kinds of path integrals, with sometimes completely different properties.

3.1 Definition of the Definite Integral

Before we begin, we have to make a remark on the philosophy of our approach. When we construct a theory, we have to define a lot of terms, and typically terms are defined using other terms that had been defined earlier. This way, a full genealogical tree of term definitions grows up. A naive approach to the definite integrals could be to

- first, define the term *area* "somehow",
- second, define the definite integral $\int_a^b f(x)\,\mathrm{d}x$ as the *area under the graph.*

This approach has the drawback that it is surprisingly hard to define rigorously the term *area.* Instead, our approach will go the opposite direction:

- first, we define the definite integral $\int_a^b f(x)\,\mathrm{d}x$ as a certain limit,
- second, we define the term *area* using definite integrals.

Now we start. In this section, $[a, b]$ always stands for a bounded and closed interval in $\mathbb{R}$.

Definition 3.1 (Step function). *A function $f\colon [a,b] \to \mathbb{R}$ or $f\colon [a,b] \to \mathbb{C}$ is a* step function[1] *if there are points x_0, x_1, ..., x_n with*

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

[1] Treppenfunktion

with the property that the function f is constant on the open intervals (x_j, x_{j+1}). Nothing is said about the values of f at the endpoints x_j of the sub–intervals. The points $\{x_0, \dots, x_n\}$ are called associated partition of $[a, b]$[2].

Note that we do not demand that the values of f "left of" x_j and "right of" x_j differ.

Proposition 3.2. *The set of all step functions is a linear space over the field $\mathbb{R}$ or $\mathbb{C}$, respectively.*

Proof. The only tricky part is to show that the sum of two step functions is again a step function. For this, you will have to unite the two partitions of the interval. □

It is obvious how the value of a definite integral of a step function should be defined:

If f is a step function over $[a, b]$ and $\{x_0, \dots, x_n\}$ its associated partition, then we define

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x := \sum_{k=0}^{n-1} f\left(\frac{x_k + x_{k+1}}{2}\right) \cdot (x_{k+1} - x_k).$$

Proposition 3.3. *This integral is a homomorphism from the set of all step functions to the real or complex numbers.*

Proof. Exercise. □

Integrals over step functions are very easy to define; but the step functions are ugly. The functions which we need the most are not step functions. Consequently, we should extend the definition of integral to some more reasonable functions.

For this purpose, we first need some norms.

$$f \text{ step function:} \qquad \|f\|_{L^1(a,b)} := \int_{x=a}^{x=b} |f(x)|\,\mathrm{d}x,$$

$$f \text{ bounded function:} \qquad \|f\|_{L^\infty(a,b)} := \sup_{x \in [a,b]} |f(x)|.$$

Each step function is bounded. The step function space becomes a normed space if we endow it with the L^1–norm or the L^∞–norm; but it will be not become a Banach space that way.

The two norms are related via

$$\|f\|_{L^1(a,b)} \le |b - a| \, \|f\|_{L^\infty(a,b)} \tag{3.1}$$

for each step function, but you cannot reverse this inequality.

We will extend the integrals to the following class of functions:

Definition 3.4 (Tame function). *We call a function $f\colon [a, b] \to \mathbb{R}$ or $f\colon [a, b] \to \mathbb{C}$* tame[3] *if it satisfies the following conditions:*

- *it is bounded,*
- *there is a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of step functions which converges to f in the L^∞–norm:*

$$\lim_{n\to\infty} \|\varphi_n - f\|_{L^\infty(a,b)} = 0.$$

It should be obvious to you that the set of tame functions is a vector space (take the sub-vector-space criterion to verify this).

Then it is almost clear how to define the definite integral for tame functions:

If f is a tame function which can be approximated by a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of step functions, then we define

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x := \lim_{n\to\infty} \int_{x=a}^{x=b} \varphi_n(x)\,\mathrm{d}x.$$

[2] zugehörige Zerlegung von $[a, b]$

[3] wörtlich: zahm, inhaltlich: Regelfunktion

The limit on the right–hand side exists, because the sequence of numbers $\int_a^b \varphi_n(x)\,\mathrm{d}x$ is a Cauchy sequence in $\mathbb{R}$ (or $\mathbb{C}$), and this can be seen as follows:

$$\begin{aligned}\left|\int_{x=a}^{x=b} \varphi_n(x)\,\mathrm{d}x - \int_{x=a}^{x=b} \varphi_m(x)\,\mathrm{d}x\right| &= \left|\int_{x=a}^{x=b} \varphi_n(x) - \varphi_m(x)\,\mathrm{d}x\right| \\ &\le \int_{x=a}^{x=b} |\varphi_n(x) - \varphi_m(x)|\,\mathrm{d}x = \|\varphi_n - \varphi_m\|_{L^1(a,b)} \le |b-a| \cdot \|\varphi_n - \varphi_m\|_{L^\infty(a,b)} \\ &\le |b-a| \cdot \|\varphi_n - f\|_{L^\infty(a,b)} + |b-a| \cdot \|f - \varphi_m\|_{L^\infty(a,b)}\,,\end{aligned}$$

by (3.1) and the triangle inequality. Therefore: if a tame function is approximated by a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of step functions then this sequence $(\varphi_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in the space $L^1(a,b)$.

Question: Let $(\psi_n)_{n\in\mathbb{N}}$ be another sequence of step functions that approximates f. Can you achieve another value for $\int_{x=a}^{x=b} f(x)\,\mathrm{d}x$ by replacing the φ_n with ψ_n ?

The class of tame functions is now sufficiently large:

Proposition 3.5 (Criterion for tame functions). *A function is tame if and only if at every point the limit of the function from the left exists, as well as the limit from the right (with obvious modifications for the endpoints of the interval).*

We should drop the proof, since it is rather nasty. We prove something weaker instead:

Proposition 3.6. *Continuous functions are tame.*

The proof is a bit long, but insightful for later purposes, as well.

Recall that a function is continuous at a point x_* if for each $\varepsilon > 0$, there is a $\delta = \delta(\varepsilon, x_*)$, such that $|x - x_*| < \delta$ implies $|f(x) - f(x_*)| < \varepsilon$.

Definition 3.7 (Uniformly continuous). *A function is* uniformly continuous[4] *if it is continuous, and the above δ depends only on ε, but not on x_*. This means: for each positive ε, there is a positive $\delta = \delta(\varepsilon)$ such that for all x and x_* in the domain of f with $|x - x_*| < \delta$, we have $|f(x) - f(x_*)| < \varepsilon$.*

We will now prove two things:

- uniformly continuous functions are tame,
- continuous functions over a compact interval are uniformly continuous.

Proof that uniformly continuous functions are tame. Fix $\varepsilon = \frac{1}{n}$. Then there is a number $\delta > 0$, such that $|x - x_*| < \delta$ and $x, x_* \in [a,b]$ imply $|f(x) - f(x_*)| < \frac{1}{n}$. Partition the interval $[a,b]$ into sub–intervals of length $\le \delta$. Define a step function φ_n by

$$\varphi_n(x) := f\left(\frac{x_k + x_{k+1}}{2}\right), \qquad a \le x_k < x < x_{k+1} \le b.$$

Then we have $\|f - \varphi_n\|_{L^\infty(a,b)} \le \frac{1}{n}$. Therefore, these step functions converge to f, measured in the L^∞–norm. □

Proposition 3.8. *A continuous function f on a compact set M is uniformly continuous.*

Proof. Assume the opposite, the function f is not uniformly continuous. Then there is an exceptional positive ε_0 with the following property:

For every positive δ, you can find x_δ, $x'_\delta \in M$ with $\|x_\delta - x'_\delta\| < \delta$ but $\|f(x_\delta) - f(x'_\delta)\| \ge \varepsilon_0$.

Put $\delta = \frac{1}{n}$, and let n tend to infinity. Then you have two sequences $(x_\delta)_{\delta\to 0}$ and $(x'_\delta)_{\delta\to 0}$ with the property that $\|x_\delta - x'_\delta\| < \delta$, however $\|f(x_\delta) - f(x'_\delta)\| \ge \varepsilon_0$. Each of them must have a converging subsequence, because of the compactness of the set M. Call the (common !) limit of such a converging subsequence x^*. By continuity of f, the sequences $(f(x_\delta))_{\delta\to 0}$ and $(f(x'_\delta))_{\delta\to 0}$ must converge to the same limit $f(x^*)$, since $\lim_{\delta\to 0} f(x_\delta) = f(\lim_{\delta\to 0} x_\delta) = f(x^*)$ and $\lim_{\delta\to 0} f(x'_\delta) = f(\lim_{\delta\to 0} x'_\delta) = f(x^*)$. Consequently, the differences $f(x_\delta) - f(x'_\delta)$ must become small. On the other hand, these differences must have norm greater than or equal to ε_0. This is impossible. □

[4] gleichmäßig stetig

Let us list some properties of integrable (tame) functions:

Proposition 3.9 (Properties of tame functions).

- *if f and g are tame functions over $[a, b]$ and $f \leq g$ everywhere, then $\int_{x=a}^{x=b} f(x)\,\mathrm{d}x \leq \int_{x=a}^{x=b} g(x)\,\mathrm{d}x$;*
- *if f is tame, then so is $|f|$;*
- *if f and g are tame, then also $f \cdot g$ is tame;*
- *a complex–valued function is tame if and only if its real part is tame and its imaginary part is also tame.*

We omit the (short) proof and only mention the general strategy: first you show similar properties for step functions (which is really easy), and second you show that these properties survive the limit procedure which defines tame functions from sequences of step functions. From now on we will use the terms "tame" and "integrable" as synonyms.

Sometimes it is useful to consider integrals, where the "upper end" of the integral is smaller than the "lower end" of the integral:

Definition 3.10. *Let a function f be tame on the interval $[a, b] \subset \mathbb{R}$ with $a < b$. Then we define*

$$\int_{x=b}^{x=a} f(x)\,\mathrm{d}x := -\int_{x=a}^{x=b} f(x)\,\mathrm{d}x.$$

In a similar spirit, we define $\int_{x=a}^{x=a} f(x)\,\mathrm{d}x = 0$.

We conclude this section with some estimates of integrals.

Proposition 3.11 (Properties of the integral). *Let $a < b$ and $f\colon [a, b] \to \mathbb{R}$ be tame. Then we have:*

1. $$\left|\int_{x=a}^{x=b} f(x)\,\mathrm{d}x\right| \leq \int_{x=a}^{x=b} |f(x)|\,\mathrm{d}x.$$

2. *For $M = \sup\{f(x) : a \leq x \leq b\}$ and $m = \inf\{f(x) : a \leq x \leq b\}$ we have*

$$m(b-a) \leq \int_{x=a}^{x=b} f(x)\,\mathrm{d}x \leq M(b-a).$$

3. ***(Mean value theorem of integration)***
If f and $g \geq 0$ are continuous, then there is a point $\xi \in (a, b)$ with

$$\int_{x=a}^{x=b} f(x)g(x)\,\mathrm{d}x = f(\xi)\int_{x=a}^{x=b} g(x)\,\mathrm{d}x.$$

In particular, choosing $g(x) \equiv 1$ we get

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = f(\xi)(b-a)$$

for a certain $\xi \in (a, b)$.

The first assertion also holds for complex–valued f.

Proof. The first claim follows from $-|f(x)| \leq f(x) \leq |f(x)|$ and Proposition 3.9, first • . The second claim is deduced from $m \leq f(x) \leq M$ in a very similar way.

For the third part, define m and M as above. Then we have (because the function g is never negative)

$$mg(x) \leq f(x)g(x) \leq Mg(x),$$

from which we conclude, employing Proposition 3.9 once more, that

$$m\int_{x=a}^{x=b} g(x)\,\mathrm{d}x \le \int_{x=a}^{x=b} f(x)g(x)\,\mathrm{d}x \le M\int_{x=a}^{x=b} g(x)\,\mathrm{d}x.$$

The integral over g cannot be negative. Therefore, a number $\mu \in [m, M]$ exists with the property that

$$\int_{x=a}^{x=b} f(x)g(x)\,\mathrm{d}x = \mu\int_{x=a}^{x=b} g(x)\,\mathrm{d}x.$$

According to the intermediate value theorem for continuous functions, this number μ can be written as $\mu = f(\xi)$ for some $\xi \in (a, b)$. □

Exercise: draw pictures.

3.2 The Indefinite Integral or Antiderivative

In the previous section, all functions were real–valued or complex–valued. Now complex–valued functions are forbidden. We will come back to them later.

Definition 3.12 (Antiderivative). *A function $F \in C^1([a, b] \to \mathbb{R})$ is called* indefinite integral *or* antiderivative *or* primitive function *of a function f*[5] *if $F'(x) = f(x)$ for all $x \in [a, b]$.*

We need a little result:

Lemma 3.13. *Let $F \in C^1([a, b] \to \mathbb{R})$ be a function with $F'(x) = 0$ for all $x \in [a, b]$. Then F is a constant.*

Proof. Pick two points x_1, $x_2 \in [a, b]$. Then, by the mean value theorem of differentiation, a number ξ exists (between x_1 and x_2), with $F(x_1) - F(x_2) = F'(\xi) \cdot (x_1 - x_2)$, hence $F(x_1) = F(x_2)$. Since x_1 and x_2 had been chosen arbitrarily, we get $F \equiv \text{const.}$. □

Then the antiderivatives are unique up to constants:

Proposition 3.14. *If F_1 and F_2 are indefinite integrals of a function f, then the difference $F_1(x)-F_2(x)$ is constant on $[a, b]$.*

Proof. We have $(F_1 - F_2)'(x) = f(x) - f(x) = 0$ for all $x \in [a, b]$. Now apply the previous lemma. □

The Fundamental Theorem of Calculus tells us that definite integrals with varying right endpoint are antiderivatives:

Theorem 3.15 (Fundamental Theorem of Calculus[6]**).** *If $f : [a, b] \to \mathbb{R}$ is continuous, then it has an antiderivative F which is given by*

$$F(x) := \int_{t=a}^{t=x} f(t)\,\mathrm{d}t, \qquad a \le x \le b.$$

Proof. The existence of the integral is clear because f is continuous, and continuous functions are tame (integrable). The mean value theorem of integration yields

$$\frac{F(x) - F(x_0)}{x - x_0} = \frac{1}{x - x_0}\int_{t=x_0}^{t=x} f(t)\,\mathrm{d}t = \frac{1}{x - x_0} f(\xi)(x - x_0) = f(\xi),$$

with some ξ between x and x_0. If x converges to x_0, then ξ also must converge to x_0, due to the sandwich principle. The continuity of f then shows

$$\lim_{x\to x_0} \frac{F(x) - F(x_0)}{x - x_0} = f(x_0).$$

□

[5] unbestimmtes Integral oder Stammfunktion einer Funktion f

[6] Hauptsatz der Differential– und Integralrechnung

For another choice of the lower endpoint a of the definite integral, you will get another antiderivative. However, it can happen that not each antiderivative can be written as such an "integral function". For instance, the function $x \mapsto 47 + \sin(x)$ is an antiderivative of the Cosine function. But there is no $a \in \mathbb{R}$ with $47 + \sin(x) = \int_{t=a}^{x} \cos(t)\,\mathrm{d}t$ for all $x \in \mathbb{R}$.

Corollary 3.16. *If $f\colon [a,b] \to \mathbb{R}$ is continuous and F is any antiderivative of f, then*

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = F(b) - F(a).$$

Proof. Theorem 3.15 gives us a special antiderivative F_0 of f, namely $F_0(x) = \int_{t=a}^{t=x} f(t)\,\mathrm{d}t$. Then Proposition 3.14 tells us that there must be a constant C, such that $F_0(x) = F(x) + C$ for every x. Finally, we have (because of $F_0(a) = 0$)

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = F_0(b) = F_0(b) - F_0(a) = (F(b) + C) - (F(a) + C) = F(b) - F(a).$$

□

It is custom to denote the antiderivatives of a function f by $\int f(x)\,\mathrm{d}x$; but you should be sure which one of the many antiderivatives you mean, and in which interval the variable x is running.

In the older literature, you may also find notations like

$$\int_a^x f(x)\,\mathrm{d}x \quad \text{or} \quad \int^x f(x)\,\mathrm{d}x,$$

but you should not use such expressions in this millennium anymore, for obvious reasons.

3.2.1 Antiderivatives of Elementary Functions

If you have to differentiate a given function, you can follow a fixed algorithm. You have at hand a list of derivatives of the elementary functions, as well as a set of rules how to differentiate functions which are composed of those elementary functions. Going this way, you are able to differentiate any function which is composed of known functions.

This is no longer true for the integration. There is no standard algorithm, only some heuristic techniques which are sometimes helpful, and sometimes not. As an example, one can prove that it is impossible to represent the indefinite integrals

$$\int e^{-x^2}\,\mathrm{d}x \quad \text{or} \quad \int \frac{\sin(x)}{x}\,\mathrm{d}x$$

by means of a finite number of terms built from elementary functions (polynomials, roots, logarithms, exponential functions, trigonometric functions).

However, quite a lot of important integrals can be evaluated.

For a start, the fundamental theorem of calculus gives us several antiderivatives right away. The proofs are skipped.

Lemma 3.17. *Let $\alpha \in \mathbb{R}$. The antiderivatives of $f = f(x) = x^\alpha$ are*

$$F = F(x) = C + \begin{cases} \frac{x^{\alpha+1}}{\alpha+1} & : \alpha \neq -1, \\ \ln|x| & : \alpha = -1. \end{cases}$$

If $\alpha \in \mathbb{N}_0$, then arbitrary $x \in \mathbb{R}$ are allowed. If $\alpha \in \mathbb{Z} \setminus \mathbb{N}_0$, then x can be from either $\mathbb{R}_-$ or $\mathbb{R}_+$, but not both. If α is not an integer, then x must be from $\mathbb{R}_+$.

Lemma 3.18. *Let $a > 0$ with $a \neq 1$ and $f = f(x) = a^x$, $x \in \mathbb{R}$. Then the antiderivatives of f are*

$$F = F(x) = \frac{1}{\ln a} a^x + C.$$

Further antiderivatives can be found in the following list:

$$\begin{aligned}
&\int \cos(x)\,\mathrm{d}x = \sin(x) + C, && x \in \mathbb{R},\\
&\int \sin(x)\,\mathrm{d}x = -\cos(x) + C, && x \in \mathbb{R},\\
&\int \frac{\mathrm{d}x}{\cos^2 x} = \tan(x) + C, && x \in \left(-\frac{\pi}{2} + k\pi, \frac{\pi}{2} + k\pi\right), \qquad k \in \mathbb{Z},\\
&\int \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \arcsin(x) + C, && x \in (-1,1),\\
&\int \frac{\mathrm{d}x}{1+x^2} = \arctan(x) + C, && x \in \mathbb{R},\\
&\int \sinh(x)\,\mathrm{d}x = \cosh(x) + C, && x \in \mathbb{R},\\
&\int \cosh(x)\,\mathrm{d}x = \sinh(x) + C, && x \in \mathbb{R},\\
&\int \frac{\mathrm{d}x}{\cosh^2(x)} = \tanh(x) + C, && x \in \mathbb{R},\\
&\int \frac{\mathrm{d}x}{\sqrt{1+x^2}} = \mathrm{Arsinh}(x) + C, && x \in \mathbb{R},\\
&\int \frac{\mathrm{d}x}{\sqrt{x^2-1}} = \mathrm{Arcosh}(x) + C, && |x| > 1,\\
&\int \frac{\mathrm{d}x}{1-x^2} = \mathrm{Artanh}(x) + C, && |x| < 1.
\end{aligned}$$

3.2.2 The Partial Integration

Partial integration is another name for integrating the product rule.

Proposition 3.19 (**Partial integration**). *If f and g belong to $C^1([a,b] \to \mathbb{R})$, then*

$$\begin{aligned}
&\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x, \qquad x \in [a,b],\\
&\int_{x=a}^{x=b} f(x)g'(x)\,\mathrm{d}x = f(x)g(x)\Big|_{x=a}^{x=b} - \int_{x=a}^{x=b} f'(x)g(x)\,\mathrm{d}x.
\end{aligned}$$

Proof. Product rule of differentiation and fundamental theorem of calculus. □

Example: *For certain x (which ?) the following holds:*

$$\begin{aligned}
&\int \ln(x)\,\mathrm{d}x = x\ln(x) - x + C,\\
&\int xe^x\,\mathrm{d}x = \dots,\\
&\int x^2e^x\,\mathrm{d}x = \dots,\\
&\int x\sin(x)\,\mathrm{d}x = \dots\ .
\end{aligned}$$

Fill in the blanks yourselves and figure out which values of x are allowed.

We will need the following result later, when we will study the Fourier series:

Proposition 3.20. *For all $n, m \in \mathbb{N}$ the following identities are valid:*

$$\begin{aligned}
&\int_{x=0}^{x=2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x = \int_{x=0}^{=2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x = \pi\delta_{nm}, \qquad (m,n) \neq (0,0),\\
&\int_{x=0}^{x=2\pi} \cos(nx)\sin(mx)\,\mathrm{d}x = 0.
\end{aligned}$$

We say that the functions $\sin(n\cdot)$ and $\cos(m\cdot)$ are orthogonal in $L^2([0, 2\pi])$.

Proof. Partial integration twice. As an example, we show $\int_{x=0}^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x = 0$ for $m \neq n \neq 0$. We have

$$\begin{aligned}
\int_{x=0}^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x &= \int_{x=0}^{2\pi} \left(\frac{1}{n}\sin(nx)\right)' \cos(mx)\,\mathrm{d}x \\
&= \frac{1}{n}\sin(nx)\cos(mx)\Big|_{x=0}^{x=2\pi} - \int_{x=0}^{2\pi} \frac{1}{n}\sin(nx)\cdot(-m)\sin(mx)\,\mathrm{d}x \\
&= 0 + \frac{m}{n}\int_{x=0}^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x \\
&= \frac{m}{n}\int_{x=0}^{2\pi} \left(\frac{-1}{n}\cos(nx)\right)' \sin(mx)\,\mathrm{d}x \\
&= \frac{m}{n}\cdot\frac{-1}{n}\cos(nx)\sin(mx)\Big|_{x=0}^{2\pi} - \frac{m}{n}\cdot\int_{x=0}^{2\pi} \frac{-1}{n}\cos(nx)\cdot m\sin(mx)\,\mathrm{d}x \\
&= 0 + \frac{m^2}{n^2}\int_{x=0}^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x,
\end{aligned}$$

which is the same integral as before, but now with a factor $\frac{m^2}{n^2}$, which is not equal to one. □

You really should remember this method of proof ! By a very similar idea, we will see (in the near and middle future) that

- eigenvectors of a self-adjoint matrix to different eigenvalues are orthogonal to each other (second semester),
- wave functions to different energy levels of a quantum mechanical system are orthogonal to each other (fourth semester).

The connection between both • comes from the fact that the wave functions are eigenfunctions to the Hamilton operator of that mentioned quantum mechanical system, and the Hamilton operator (which is a differential operator acting in the Hilbert space $L^2(\mathbb{R}^3 \to \mathbb{C})$) is self-adjoint.

Finally, we show how clever use of the partial integration helps in proving the Taylor expansion theorem once again. The cleverness lies in the choice of the integration constants ($-x$ and $-\frac{x^2}{2}$) in the definition of the functions g_1 and g_2 below. Let $u \in C^3([a, b] \to \mathbb{R})$ and x, $x_0 \in [a, b]$. Then we have

$$\begin{aligned}
u(x) &= u(x_0) + \int_{t=x_0}^{x} u'(t)\,\mathrm{d}t = u(x_0) + \int_{t=x_0}^{x} u'(t)\cdot 1\,\mathrm{d}t \qquad \Big| \quad f_1(t) := u'(t), \quad g_1(t) := t - x \\
&= u(x_0) + u'(t)\cdot(t-x)\Big|_{t=x_0}^{t=x} - \int_{t=x_0}^{x} u''(t)(t-x)\,\mathrm{d}t \\
&= u(x_0) + 0 - u'(x_0)(x_0 - x) + \int_{t=x_0}^{x} u''(t)(x-t)\,\mathrm{d}t \qquad \Big| \quad f_2(t) := u''(t), \quad g_2(t) := xt - \frac{t^2}{2} - \frac{x^2}{2} \\
&= u(x_0) + u'(x_0)(x - x_0) + u''(t)\frac{(x-t)^2}{-2}\Big|_{t=x_0}^{t=x} \\
&\qquad - \int_{t=x_0}^{x} u'''(t)\frac{(x-t)^2}{-2}\,\mathrm{d}t \\
&= \sum_{k=0}^{2} \frac{1}{k!}u^{(k)}(x_0)(x-x_0)^k + \frac{1}{2}\int_{t=x_0}^{x} u'''(t)(x-t)^2\,\mathrm{d}t,
\end{aligned}$$

and this prodecure could be continued several times. To handle the last integral on the right, we exploit the mean value theorem of integration (Proposition 3.11), now with $g(t) = (x-t)^2$ and $f(t) = u'''(t)$, hence there is a number ξ between x and x_0 with

$$\frac{1}{2}\int_{t=x_0}^{x} u'''(t)(x-t)^2 \,\mathrm{d}t = \frac{1}{2}u'''(\xi)\int_{t=x_0}^{x} (t-x)^2 \,\mathrm{d}t = \frac{1}{2}u'''(\xi)\int_{t=x_0-x}^{0} t^2 \,\mathrm{d}t = \frac{1}{2}u'''(\xi)\cdot\frac{-1}{3}(x_0-x)^3,$$

which equals $\frac{1}{3!}u'''(\xi)(x-x_0)^3$, and this is the well-known Lagrange form of the remainder term in the Taylor expansion. Other versions (like the Cauchy form or even the Schlömilch form) of the remainder term can be proved similarly.

3.2.3 The Substitution Rule

Proposition 3.21 (**Substitution**). *Let $f\colon [\alpha,\beta] \to \mathbb{R}$ be continuous with primitive function F, and $\varphi\colon [a,b] \to [\alpha,\beta]$ continuously differentiable. Then we have*

$$\int f(\varphi(x))\varphi'(x)\,\mathrm{d}x = F(\varphi(x)) + C, \qquad x \in [a,b],$$

$$\int_{x=a}^{x=b} f(\varphi(x))\varphi'(x)\,\mathrm{d}x = \int_{t=\varphi(a)}^{t=\varphi(b)} f(t)\,\mathrm{d}t = F(\varphi(b)) - F(\varphi(a)).$$

Proof. The chain rule implies

$$(F\circ\varphi)'(x) = F'(\varphi(x))\varphi'(x) = f(\varphi(x))\varphi'(x).$$

Integrating this identity and the fundamental theorem of calculus conclude the proof. □

When you change the variable of integration, be sure to change it everywhere:

- in the integrand,
- in the differential,
- at the endpoints of the integration interval.

Example: *Let F be a primitive function to f. Then you have*

$$\int_{x=a}^{x=b} f(x+c)\,\mathrm{d}x = \int_{t=a+c}^{t=b+c} f(t)\,\mathrm{d}t = F(b+c) - F(a+c),$$

$$\int_{x=a}^{x=b} f(cx)\,\mathrm{d}x = \frac{1}{c}\int_{t=ca}^{t=cb} f(t)\,\mathrm{d}t = \frac{1}{c}(F(cb) - F(ca)), \qquad c \neq 0,$$

$$\int_{x=a}^{x=b} x^{n-1} f(x^n)\,\mathrm{d}x = \dots,$$

$$\int_{x=a}^{x=b} x\exp(-x^2)\,\mathrm{d}x = \dots,$$

$$\int_{x=a}^{x=b} \frac{f'(x)}{f(x)}\,\mathrm{d}x = \dots, \qquad f(x) \neq 0 \text{ on } [a,b].$$

Fill in the blanks yourselves and figure out the admissible values of the variables and parameters.

3.2.4 Partial Fractions

Partial fractions[7] are the standard tool for integrating rational functions.

The fundamental theorem of algebra (to be proved later) shows us that a polynomial q of degree n has exactly n zeroes (if you count multiple zeroes according to their multiplicity). Suppose that the

[7] Partialbrüche

coefficients of this polynomial q are all real, and the highest coefficient is one. If q has a zero at a point $c \in \mathbb{R}$, then you can divide q by the factor $(x - c)$ and obtain a polynomial of degree $n - 1$. If q has a zero at the complex number $a + b\mathrm{i}$ with $b \neq 0$, then also $a - b\mathrm{i}$ is a zero of q (why is that so ?). In this case, you can divide q by the factor $(x - (a + b\mathrm{i}))(x - (a - b\mathrm{i})) = (x - a)^2 + b^2$, and get a polynomial of degree $n - 2$. Continuing in this fashion, we can write q as a product of linear or quadratic polynomials of the above structure:

$$q(x) = \prod_{i=1}^{r}(x - c_i) \prod_{j=1}^{s}((x - a_j)^2 + b_j^2).$$

Let us be given additionally a polynomial p and consider the rational function $\frac{p(x)}{q(x)}$ under the following two (non–essential) assumptions:

- the degree of p is strictly less than the degree of q,
- q has only zeroes of multiplicity 1.

Then the quotient $\frac{p(x)}{q(x)}$ can be decomposed like this:

$$\frac{p(x)}{q(x)} = \sum_{i=1}^{r} \frac{\gamma_i}{x - c_i} + \sum_{j=1}^{s} \frac{\alpha_j x + \beta_j}{(x - a_j)^2 + b_j^2}.$$

This formula holds—which we will not prove—for all points $x \in \mathbb{C}$ except the zeroes of q.

The terms on the right–hand side can be integrated easily:

$$\begin{aligned}
&\int \frac{\gamma}{x - c}\,\mathrm{d}x = \gamma \ln|x - c| + C,\\
&\frac{\alpha x + \beta}{(x - a)^2 + b^2} = \frac{\alpha}{2}\frac{2(x - a)}{(x - a)^2 + b^2} + \frac{\beta + \alpha a}{(x - a)^2 + b^2},\\
&\int \frac{2(x - a)}{(x - a)^2 + b^2}\,\mathrm{d}x = \ln|(x - a)^2 + b^2| + C,\\
&\int \frac{1}{(x - a)^2 + b^2}\,\mathrm{d}x = \frac{1}{b}\arctan\left(\frac{x - a}{b}\right) + C.
\end{aligned}$$

Here we have set $t = \frac{x-a}{b}$ and have applied the substitution rule in the last step.

Example: *The fraction* $\frac{4}{1-x^4}$ *can be decomposed into*

$$\frac{4}{1 - x^4} = \frac{1}{1 - x} + \frac{1}{1 + x} + \frac{2}{1 + x^2},$$

giving us the antiderivative

$$\begin{aligned}
\int \frac{4}{1 - x^4}\,\mathrm{d}x &= \int \frac{\mathrm{d}x}{1 - x} + \int \frac{\mathrm{d}x}{1 + x} + \int \frac{2\,\mathrm{d}x}{1 + x^2}\\
&= -\ln|1 - x| + \ln|1 + x| + 2\arctan(x) + C.
\end{aligned}$$

Example: *How to do the decomposition ? Take the function*

$$f = f(x) = \frac{3x + 7}{x(x - 1)^2(x + 2)}$$

as an example. We then make the ansatz

$$\frac{3x + 7}{x(x - 1)^2(x + 2)} = \frac{A}{x} + \frac{B}{x - 1} + \frac{C}{(x - 1)^2} + \frac{D}{x + 2}, \quad x \neq 0, \quad x \neq 1, \quad x \neq -2$$

The terms with B and C are both needed, since the left-hand side has a double pole at $x = 1$. We multiply both sides by x and get

$$\frac{3x + 7}{(x - 1)^2(x + 2)} = A + \frac{Bx}{x - 1} + \frac{Cx}{(x - 1)^2} + \frac{Dx}{x + 2}, \quad x \neq 0, \quad x \neq 1, \quad x \neq -2.$$

Now we perform the limit $x \to 0$ and find

$$\frac{7}{2} = A + B \cdot 0 + C \cdot 0 + D \cdot 0,$$

hence $A = \frac{7}{2}$. We multiply both sides of our ansatz by $(x-1)^2$ and get

$$\frac{3x+7}{x(x+2)} = \frac{A(x-1)^2}{x} + B(x-1) + C + \frac{D(x-1)^2}{x+2}, \quad x \neq 0, \quad x \neq 1, \quad x \neq -2$$

and now the limit $x \to +1$ gives us $\frac{10}{3} = C$. And we multiply both sides of our ansatz by $(x+2)$, hence

$$\frac{3x+7}{x(x-1)^2} = \frac{A(x+2)}{x} + \frac{B(x+2)}{x-1} + \frac{C(x+2)}{(x-1)^2} + D, \quad x \neq 0, \quad x \neq 1, \quad x \neq -2,$$

and sending x to -2 implies $-\frac{1}{18} = D$. Hence, only B is still to be found. One approach is: we multiply our ansatz by $x(x-1)^2(x+2)$ and compare equal powers of x on both sides. Another approach is to subtract the known terms:

$$\begin{aligned}
\frac{B}{x-1} &= \frac{3x+7}{x(x-1)^2(x+2)} - \frac{\frac{7}{2}}{x} - \frac{\frac{10}{3}}{(x-1)^2} + \frac{\frac{1}{18}}{x+2} \\
&= \frac{3x+7-\frac{7}{2}(x-1)^2(x+2) - \frac{10}{3}x(x+2) + \frac{1}{18}x(x-1)^2}{x(x-1)^2(x+2)} \\
&= \frac{3x+7-\frac{7}{2}(x^2-2x+1)(x+2) - \frac{10}{3}(x^2+2x) + \frac{1}{18}(x^3-2x^2+x)}{x(x-1)^2(x+2)} \\
&= \frac{\frac{1}{18}x^3 - \frac{31}{9}x^2 - \frac{65}{18}x + 7 - \frac{7}{2}(x^3-3x+2)}{x(x-1)^2(x+2)} \\
&= \frac{-\frac{62}{18}x^3 - \frac{31}{9}x^2 + \frac{124}{18}x}{x(x-1)^2(x+2)} = \frac{-31}{9} \cdot \frac{x^2+x-2}{(x-1)^2(x+2)} = \frac{-31}{9} \cdot \frac{(x-1)(x+2)}{(x-1)^2(x+2)},
\end{aligned}$$

and therefore $B = \frac{-31}{9}$.

3.2.5 The Half Angle Method

The half angle method enables us to integrate any rational function of $\sin(x)$ and $\cos(x)$, provided that we can find the zeroes of the denominator of a certain rational function which arises in that process. We put

$$u = \tan\frac{x}{2},$$

from which it follows that

$$\begin{aligned}
\mathrm{d}u &= \frac{\mathrm{d}u}{\mathrm{d}x}\,\mathrm{d}x = \frac{1}{2}\left(1 + \left(\tan\frac{x}{2}\right)^2\right)\mathrm{d}x = \frac{1}{2}(1+u^2)\,\mathrm{d}x, \\
\sin x &= 2\sin\frac{x}{2}\cos\frac{x}{2} = 2\tan\frac{x}{2}\cos^2\frac{x}{2} = \frac{2\tan\frac{x}{2}}{1+\tan^2\frac{x}{2}} = \frac{2u}{1+u^2}, \\
\cos x &= \cos^2\frac{x}{2} - \sin^2\frac{x}{2} = \cos^2\frac{x}{2}\left(1 - \tan^2\frac{x}{2}\right) = \frac{1-u^2}{1+u^2}.
\end{aligned}$$

Example: *Let $0 < a < b < \pi$ and $A = \tan\frac{a}{2}$, $B = \tan\frac{b}{2}$. Then we have*

$$\int_{x=a}^{x=b} \frac{\mathrm{d}x}{\sin x} = \int_{u=A}^{u=B} \frac{1+u^2}{2u} \cdot \frac{2\,\mathrm{d}u}{1+u^2} = \int_{u=A}^{u=B} \frac{\mathrm{d}u}{u} = \ln\left|\frac{\tan\frac{a}{2}}{\tan\frac{b}{2}}\right|.$$

3.2.6 Numerical Methods

You will not always succeed in finding an explicit formula for the antiderivative. It can also happen that you do not have a formula for the integrand, but just some values at some points. Then the following numerical formulae can be helpful.

The main idea is as follows: you want to compute approximately the definite integral $\int_{x=a}^{x=b} f(x)\,\mathrm{d}x$.

First, you split the interval into parts of equal length. This gives you points $x_j = a + jh$ with $h = \frac{b-a}{n}$ and $0 \le j \le n$. You take one of these sub-intervals (or maybe several adjacent of them), and on these union of sub-intervals, you take a polynomial that approximates f. Instead of f, you integrate this polynomial. Then you sum up over all possible unions of subintervals.

One of the most simple rules is obtained by piecewise linear approximation on each sub-interval.

Proposition 3.22 (**Trapezoidal rule**). *Suppose that $f \in C^2([a,b] \to \mathbb{R})$ with $|f''(x)| \le M$ for $x \in [a,b]$. Then the following approximation holds:*

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = \left(\frac{1}{2}f(a) + \sum_{j=1}^{n-1} f(a+jh) + \frac{1}{2}f(b)\right)h + R, \qquad h = \frac{b-a}{n},$$

where $|R| \le \frac{1}{12}(b-a)Mh^2$.

Proof. We consider a sub-interval $[x_j, x_{j+1}]$. On this sub-interval, we define the auxiliary function $\varphi(x) = \frac{1}{2}(x-x_j)(x_{j+1}-x)$. We observe that

$$\begin{aligned}
&\varphi(x_j) = \varphi(x_{j+1}) = 0,\\
&\varphi(x) \ge 0, \qquad x \in [x_j, x_{j+1}],\\
&\varphi'(x) = \frac{h}{2} - (x - x_j),\\
&\varphi''(x) = -1.
\end{aligned}$$

Performing partial integration twice then gives us

$$\begin{aligned}
\int_{x=x_j}^{x=x_{j+1}} f(x)\,\mathrm{d}x &= -\int_{x=x_j}^{x=x_{j+1}} \varphi''(x)f(x)\,\mathrm{d}x\\
&= -\varphi'(x)f(x)\Big|_{x=x_j}^{x=x_{j+1}} + \int_{x=x_j}^{x=x_{j+1}} \varphi'(x)f'(x)\,\mathrm{d}x\\
&= \frac{h}{2}(f(x_j) + f(x_{j+1})) + \int_{x=x_j}^{x=x_{j+1}} \varphi'(x)f'(x)\,\mathrm{d}x\\
&= \frac{h}{2}(f(x_j) + f(x_{j+1})) + \varphi(x)f'(x)\Big|_{x=x_j}^{x=x_{j+1}} - \int_{x=x_j}^{x=x_{j+1}} \varphi(x)f''(x)\,\mathrm{d}x\\
&= \frac{h}{2}(f(x_j) + f(x_{j+1})) - f''(\xi_j)\int_{x=x_j}^{x=x_{j+1}} \varphi(x)\,\mathrm{d}x\\
&= \frac{h}{2}(f(x_j) + f(x_{j+1})) - \frac{h^3}{12}f''(\xi_j),
\end{aligned}$$

where we have used the mean value theorem of integration, giving us an (unknown) point $\xi_j \in (x_j, x_{j+1})$. Summing over $j = 0, \dots, n-1$ completes the proof. □

If the function f that has to be integrated is a polynomial of degree ≤ 1, then the trapezoidal rule will give the exact value.

If one can find other rules which can integrate polynomials of higher degree than 1 exactly and which have a better estimate of the error term R (the exponent of h should be higher), then one can get a numerical approximation of the value of the integral with the same precision but requiring less effort. One of such rules is the famous KEPLER[8] rule.

[8] JOHANNES KEPLER, 1571 – 1630

Proposition 3.23 (**Kepler's barrel rule**[9]). *Suppose $f \in C^4([a,b] \to \mathbb{R})$ with $|f''''(x)| \leq M$. Let $n \in \mathbb{N}$ be* even, *and set*

$$h = \frac{b-a}{n}, \qquad x_j = a + jh, \quad j = 0, 1, 2, \dots, n.$$

Then the following approximation holds:

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = \frac{h}{3}\left(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \dots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right) + R,$$

where $|R| \leq \frac{1}{180}(b-a)Mh^4$.

We omit the proof and only remark that the function f is piecewise approximated by quadratic polynomials, which are integrated instead of f. Therefore, it is natural to expect that Kepler's rule can integrate quadratic functions without error. Surprisingly, even cubic functions can be integrated exactly, and also the bound on R has one power of h more than expected. Consequently, the Kepler's rule usually achieves the same precision as the trapezoidal rule, but with fewer evaluations of the integrand f. This is an important advantage if the evaluation of f is costly, e.g., if f itself is given as a definite integral.

The trapezoidal rule is based on piecewise *linear* interpolation of the integrand, and Kepler's rule is based on piecewise *quadratic* interpolation. You can continue in this manner, interpolating the function f piecewise by polynomials of higher and higher degree. The formulas which you will obtain are the so–called Newton[10]–Cotes[11] quadrature formulas. The highest useful polynomial degree is 6. For higher degrees, the interpolating polynomials oscillate heavily near the ends of the interval, giving a lot of trouble. Furthermore, some coefficients in the formula will become negative, leading to numerical instabilities because of cancellation effects.

There is another approach to numerical integration. Let us be given a function f on the interval $[-1, 1]$. We would like to find an approximate value of the integral $\int_{x=-1}^{x=1} f(x)\,\mathrm{d}x$ like this:

$$\int_{x=-1}^{x=1} f(x)\,\mathrm{d}x \approx \sum_{j=1}^{n} w_j f(x_j),$$

where the x_j are certain points in the interval $[-1, 1]$ (not necessarily equidistant), and the w_j are so–called weights.

Is it possible to choose the x_j and w_j in such a way that all polynomials up to a certain degree are integrated exactly ? How high can that degree be ?

We have $2n$ free parameters (namely the x_j and the w_j) available, so we hope to integrate all polynomials of degree less than or equal to $2n-1$ exactly. This leads us to the Gauss quadrature formulas. The x_j must be chosen as the zeroes of the Legendre[12] polynomial $P_n = P_n(x)$, which is defined as being the (only) polynomial solution to the differential equation

$$(1-x^2)y''(x) - 2xy'(x) + n(n+1)y(x) = 0$$

with $P_n(1) = 1$. The functions P_0, P_1, P_2, ..., P_n form an $L^2(-1,1)$–*orthogonal* basis of the space of all polynomials of degree at most n. All the n zeroes of P_n are in the interval $[-1, 1]$. And the weights w_k must be chosen as

$$w_k = \int_{x=-1}^{x=1} \left(\prod_{j=1, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \right) \mathrm{d}x = \int_{x=-1}^{x=1} \left(\prod_{j=1, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \right)^2 \mathrm{d}x.$$

For the convenience of the reader, we list the data for the case $n = 7$:

More parameters for the Gaussian quadrature (up to $n = 96$) can be found in [1].

[9] Keplersche Faßregel
[10] Sir Isaac Newton, 1642 – 1727
[11] Roger Cotes, 1682 – 1716
[12] Adrien–Marie Legendre, 1752 – 1833

i	x_i	w_i
1	-0.949107912342759	0.129484966168870
2	-0.741531185599384	0.279705391489277
3	-0.405845151377397	0.381830050505119
4	0	0.417959183673469
5	$+0.405845151377397$	0.381830050505119
6	$+0.741531185599384$	0.279705391489277
7	$+0.949107912342759$	0.129484966168870

As an example, we use this to evaluate $\ln 5 = \int_{x=1}^{x=5} \frac{1}{x}\,\mathrm{d}x$. The exact value is

$$\ln 5 \approx 1.6094379124341002818\dots.$$

First, we shift the interval $[1, 5]$ to $[-1, 1]$:

$$t = \frac{x-3}{2}, \qquad x = 2t+3, \qquad \ln 5 = \int_{x=1}^{x=5} \frac{\mathrm{d}x}{x} = \int_{t=-1}^{t=1} \frac{2\,\mathrm{d}t}{2t+3}.$$

Evaluating this last integral with Gaussian quadrature with $n = 7$ and with one or two sub-divisions, as well as with the trapezoidal rule ($n = 6$) and Kepler's rule ($n = 6$), we obtain the following numbers:

sub-division	value	error
0	1.6094346840305430703	$3.228e-06$
1	1.6094378965041162519	$1.592e-08$
2	1.6094379124141617306	$1.993e-11$

Table 3.1: Gauss quadrature

sub-division	value	error
0	1.6436008436008436009	$3.416e-02$
1	1.6182289932289932289	$8.791e-03$
2	1.611653797587057737	$2.215e-03$

Table 3.2: Trapezoidal rule

sub-division	value	error
0	1.6131128131128131129	$3.675e-03$
1	1.6097717097717097716	$3.338e-04$
2	1.6094620657064125734	$2.415e-05$

Table 3.3: Kepler's rule

More on Legendre Polyomials can be found here:

Literature: Greiner: *Klassische Elektrodynamik.* Chapter I.3: Entwicklung beliebiger Funktionen in vollständige Funktionssysteme

3.2.7 Improper Integrals

Our definite integral as defined in Section 3.1 suffers from several restrictions. Two of them are that:

- the interval of integration must be bounded,
- the function to be integrated (the integrand) must be bounded.

Now we will overcome these restrictions, as much as possible.

Unbounded Interval of Integration

Definition 3.24. *Let $f\colon [a,\infty) \to \infty$ be a function that is integrable on every interval $[a, R]$ with $a < R$. If the limit*

$$\lim_{R\to\infty} \int_{x=a}^{x=R} f(x)\,\mathrm{d}x$$

exists, then this limit is denoted by $\int_{x=a}^{x=\infty} f(x)\,\mathrm{d}x$.

Example 3.25. *For which values of α does the integral*

$$\int_{x=1}^{x=\infty} \frac{1}{x^\alpha}\,\mathrm{d}x$$

exist ? Compare with series of real numbers.

After defining integrals $(-\infty, b]$ in a very similar way, we then define

$$\int_{x=-\infty}^{x=\infty} f(x)\,\mathrm{d}x := \int_{x=-\infty}^{x=0} f(x)\,\mathrm{d}x + \int_{x=0}^{x=\infty} f(x)\,\mathrm{d}x,$$

under the assumption that the right–hand side exists. Or, equivalently,

$$\int_{x=-\infty}^{x=\infty} f(x)\,\mathrm{d}x := \lim_{R_-\to-\infty, R_+\to+\infty} \int_{x=R_-}^{x=R_+} f(x)\,\mathrm{d}x.$$

The important thing to note here is that R_- and R_+ do not depend on each other. Each one can approach its limit at its own preferred pace.

Question: How about the integrals $\int_{x=-\infty}^{x=\infty} \sin(x)\,\mathrm{d}x$ and $\int_{x=-\infty}^{x=\infty} \cos(x)\,\mathrm{d}x$?

Unbounded Integrand

Definition 3.26. *Let $f\colon [a,b] \to \mathbb{R}$ be a function which is integrable on every interval $[a+\varepsilon, b]$ with $\varepsilon > 0$. If the limit*

$$\lim_{\varepsilon\to+0} \int_{x=a+\varepsilon}^{x=b} f(x)\,\mathrm{d}x$$

exists, then this limit is denoted by $\int_{x=a}^{x=b} f(x)\,\mathrm{d}x$.

Example: *For which value of α does the integral*

$$\int_{x=0}^{x=1} \frac{1}{x^\alpha}\,\mathrm{d}x$$

exist ? Compare with Example 3.25.

Integrals over the interval $[a, b]$ with an unbounded integrand at the right endpoint b are defined similarly. Next we will define integrals with an integrand having the one and only pole at a point c inside the interval (a, b) by splitting the interval;

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x := \lim_{\varepsilon_1\to+0} \int_{x=a}^{x=c-\varepsilon_1} f(x)\,\mathrm{d}x + \lim_{\varepsilon_2\to+0} \int_{x=c+\varepsilon_2}^{x=b} f(x)\,\mathrm{d}x$$

provided that the right–hand side exists.

Note that the ε_1 and ε_2 do not depend on each other.

Sometimes this restriction is considered too severe; and one wants to take advantage from cancellation properties. In this case, one could use the *Cauchy principal value*[13]:

$$\text{p.v.} \int_{x=a}^{x=b} f(x)\,\mathrm{d}x = \lim_{\varepsilon\to+0} \left(\int_{x=a}^{x=c-\varepsilon} f(x)\,\mathrm{d}x + \int_{x=c+\varepsilon}^{x=b} f(x)\,\mathrm{d}x \right).$$

[13] Cauchy–Hauptwert

Unfortunately, often the “p.v.” is omitted, leading to two different meanings of the same symbol.

Question: How about the integral $\int_{x=-2}^{x=1} \frac{1}{x}\,\mathrm{d}x$? Consider the usual integral and the Cauchy principal value.

There is a nice criterion, connecting the convergence of a series of real numbers and the existence of an improper integral.

Proposition 3.27. *Let $f\colon [1,\infty) \to \mathbb{R}$ be a monotonically decreasing function, taking only non–negative values. Then the series $\sum_{n=1}^{\infty} f(n)$ converges if and only if the improper integral $\int_{x=1}^{x=\infty} f(x)\,\mathrm{d}x$ exists.*

Proof. Due to the monotonicity of f, we deduce for all n that

$$f(n) \le \int_{x=n-1}^{x=n} f(x)\,\mathrm{d}x \le f(n-1),$$

hence

$$\sum_{n=2}^{m} f(n) \le \int_{x=1}^{x=m} f(x)\,\mathrm{d}x \le \sum_{n=1}^{m} f(n).$$

Then the convergence of the series implies the convergence of the improper integral, and vice versa. □

The proof yields the following estimates from above and below:

$$\int_{x=1}^{x=\infty} f(x)\,\mathrm{d}x \le \sum_{n=1}^{\infty} f(n) \le \int_{x=1}^{x=\infty} f(x)\,\mathrm{d}x + f(1).$$

For instance, $\sum_{n=1}^{\infty} n^{-2}$ can be estimated by

$$1 \le \sum_{n=1}^{\infty} \frac{1}{n^2} \le 2.$$

The theory of Fourier series will give us as a by–product that $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$.

3.3 Commuting Limit Processes

Now we have a sequence of functions $(f_n)_{n\in\mathbb{N}}$ that converges (in whichever sense of the word) to a limit function f. We would like to know whether

$$\begin{aligned} &\int_{x=a}^{x=b} \lim_{n\to\infty} f_n(x)\,\mathrm{d}x \overset{?}{=} \lim_{n\to\infty} \int_{x=a}^{x=b} f_n(x)\,\mathrm{d}x, \\ &\frac{\mathrm{d}}{\mathrm{d}x} \lim_{n\to\infty} f_n(x) \overset{?}{=} \lim_{n\to\infty} \frac{\mathrm{d}}{\mathrm{d}x} f_n(x), \qquad a \le x \le b. \end{aligned}$$

Just as a warning example, consider the functions $f_n\colon [0,1] \to \mathbb{R}$ defined by

$$f_n(x) := \begin{cases} 0 & : x = 0, \\ n & : 0 < x < \frac{1}{n}, \\ 0 & : \frac{1}{n} \le x \le 1. \end{cases}$$

We see that $\lim_{n\to\infty} f_n(x) = 0$ for each $x \in [0,1]$, but

$$1 = \lim_{n\to\infty} 1 = \lim_{n\to\infty} \int_{x=0}^{x=1} f_n(x)\,\mathrm{d}x \ne \int_{x=0}^{x=1} \lim_{n\to\infty} f_n(x)\,\mathrm{d}x = \int_{x=0}^{x=1} 0\,\mathrm{d}x = 0.$$

Or choose functions $g_n\colon \mathbb{R} \to \mathbb{R}$ with

$$g_n(x) := \frac{1}{n}\sin(nx),$$

which go to zero for $n \to \infty$, but the derivatives $g_n'(x) = \cos(nx)$ do not converge for $n \to \infty$ to the zero function.

These two examples teach us that we need more conditions than just convergence for each point x.

Remark 3.28. *Remember that the integrals and derivatives are defined as limits, namely limits over a sequence of step functions, and limits of quotients of differences, respectively. From examples we learn that limits, in general, cannot be commuted, e.g.,*

$$\lim_{n\to\infty}\lim_{m\to\infty} a_{n,m} \neq \lim_{m\to\infty}\lim_{n\to\infty} a_{n,m}, \qquad a_{n,m} = \frac{n}{n+m}.$$

Seen from this point of view, the above two examples do not come as a surprise. As another example (showing that the limit function of a sequence of continuous functions might be discontinuous),

$$\lim_{n\to\infty}\lim_{x\to 1-0} x^n \neq \lim_{x\to 1-0}\lim_{n\to\infty} x^n.$$

The situation becomes even more complicated if you ask whether

$$\int_{x=a}^{x=\infty} \lim_{n\to\infty} f_n(x)\,\mathrm{d}x \overset{?}{=} \lim_{n\to\infty}\int_{x=a}^{x=\infty} f_n(x)\,\mathrm{d}x,$$

because now three limit symbols appear on each side.

We should distinguish several types of convergence of a sequence of functions to a limit function.

Definition 3.29 (Pointwise convergence). *We say that a sequence $(f_n)_{n\in\mathbb{N}}$ of functions mapping $[a,b]$ into $\mathbb{R}$ converges* pointwise[14] *to a function $f\colon [a,b]\to\mathbb{R}$ if, for each $x\in[a,b]$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ of real numbers converges to $f_n(x)$. Written symbolically:*

$$\forall x\in[a,b]\ \forall\varepsilon>0 :\ \exists N_0(x,\varepsilon) :\ \forall n\geq N_0(x,\varepsilon) :\ |f_n(x)-f(x)|<\varepsilon.$$

Definition 3.30 (Uniform convergence). *We say that a sequence $(f_n)_{n\in\mathbb{N}}$ of functions mapping $[a,b]$ into $\mathbb{R}$ converges* uniformly[15] *to a function $f\colon [a,b]\to\mathbb{R}$ if it converges pointwise and the above–mentioned N_0 can be chosen independent of $x\in[a,b]$. Written symbolically:*

$$\forall\varepsilon>0 :\ \exists N_0(\varepsilon) :\ \forall x\in[a,b]\ \forall n\geq N_0(\varepsilon) :\ |f_n(x)-f(x)|<\varepsilon.$$

The uniform convergence is the same as the convergence in the L^∞–norm or sup–norm.

As we have seen in the above examples, pointwise convergence does not enable us to commute the integral symbol and the limit symbol. However, the uniform convergence does.

Now we show that:

- uniform convergence preserves continuity,
- uniform convergence preserves integrability, and you can commute $\int$ and $\lim_n$,
- uniform convergence of the sequence of *derivatives* $(f_n')_{n\in\mathbb{N}}$ and pointwise convergence of the sequence $(f_n)_{n\in\mathbb{N}}$ together allow to commute $\frac{\mathrm{d}}{\mathrm{d}x}$ and $\lim_n$.

Proposition 3.31. *Let $(f_n)_{n\in\mathbb{N}}\subset C([a,b]\to\mathbb{R})$ be a sequence of continuous functions, uniformly converging to a function $f\colon[a,b]\to\mathbb{R}$. Then f is continuous.*

By obvious changes of the notation, you can prove much more. Let U,V be Banach spaces and $M\subset U$ be a closed set. If a sequence $(f_n)_{n\in\mathbb{N}}\subset C(M\to V)$ converges to a function $f\colon M\to V$, then f is continuous. This is the missing proof to Satz 5.15 from the first term.

Proof. Fix $x_0\in[a,b]$ and choose a positive ε. Then there is an $n=n(\varepsilon)\in\mathbb{N}$, such that

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

for all $x\in[a,b]$. Keep this n. Since f_n is continuous, there is a $\delta_0=\delta_0(\varepsilon,x_0,n)$, such that for all $x\in[a,b]$ with $|x-x_0|<\delta_0$, we have

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

[14] punktweise

[15] gleichmäßig

For such n and x, we then can write

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

□

Proposition 3.32. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of integrable functions, mapping $[a,b]$ to $\mathbb{R}$ and converging uniformly to a function $f\colon [a,b] \to \mathbb{R}$.*

Then the function f is integrable, and

$$\lim_{n\to\infty} \int_{x=a}^{x=b} f_n(x)\,\mathrm{d}x = \int_{x=a}^{x=b} \lim_{n\to\infty} f_n(x)\,\mathrm{d}x = \int_{x=a}^{x=b} f(x)\,\mathrm{d}x.$$

Proof. By definition of integrability, the functions f_n are tame. That means: for each $n \in \mathbb{N}$, we have a sequence $(\varphi_{n,m})_{m\in\mathbb{N}}$ of step functions with

$$\lim_{m\to\infty} \|\varphi_{n,m} - f_n\|_{L^\infty(a,b)} = 0,$$

or, equivalently, the sequence $(\varphi_{n,m})_{m\in\mathbb{N}}$ converges uniformly to f_n.

Choose a positive ε. Then there is an f_n with

$$\|f_n - f\|_{L^\infty(a,b)} < \frac{\varepsilon}{2},$$

by the uniform convergence of the f_n to f. For this f_n, we find a step function $\varphi_{n,m}$ with

$$\|\varphi_{n,m} - f_n\|_{L^\infty(a,b)} < \frac{\varepsilon}{2}.$$

Hence we have found a step function $\varphi_{n,m}$ with

$$\|\varphi_{n,m} - f\|_{L^\infty(a,b)} < \varepsilon.$$

Consequently, the function f is tame. The second claim follows from

$$\begin{aligned}\left|\int_{x=a}^{x=b} f_n(x)\,\mathrm{d}x - \int_{x=a}^{x=b} f(x)\,\mathrm{d}x\right| &\leq \int_{x=a}^{x=b} |f_n(x) - f(x)|\,\mathrm{d}x \\ &= \|f_n - f\|_{L^1(a,b)} \leq |b-a|\,\|f_n - f\|_{L^\infty(a,b)} \to 0 \quad \text{for} \quad n\to\infty.\end{aligned}$$

□

The situation is more delicate for the derivatives. Uniform convergence of a sequence of functions $(f_n)_{n\in\mathbb{N}}$ **does not** ensure differentiability of the limit function. Take $f_n = \frac{1}{n}\sin(nx)$, for instance. You need uniform convergence of the sequence of the **derivatives** $(f_n')_{n\in\mathbb{N}}$ instead, and additionally convergence of the sequence $(f_n)_{n\in\mathbb{N}}$, at least in one point.

Question: Why is this additional condition necessary ?

Proposition 3.33. *Let $(f_n)_{n\in\mathbb{N}} \subset C^1([a,b] \to \mathbb{R})$ be a sequence of differentiable functions with the following properties:*

- *$(f_n')_{n\in\mathbb{N}}$ converges uniformly to a function $g\colon [a,b] \to \mathbb{R}$,*
- *$\lim_{n\to\infty} f_n(c)$ exists, for at least one $c \in [a,b]$.*

Then the limit $\lim_{n\to\infty} f_n(x) =: f(x)$ exists for each $x \in [a,b]$, this convergence is uniform, the limit function f is continuously differentiable, and

$$f'(x) = g(x), \qquad x \in [a,b].$$

Proof. The functions f_n' are continuous, and the sequence $(f_n')_{n\in\mathbb{N}}$ converges uniformly, hence g is continuous, too. For each $x \in [a,b]$, we have

$$f_n(x) = f_n(c) + \int_{t=c}^{t=x} f_n'(t)\,\mathrm{d}t.$$

According to Proposition 3.32, the limit of the right–hand side exists and has the value $f(c)+\int_{t=c}^{t=x} g(t)\,\mathrm{d}t$, which is a continuously differentiable function on $[a,b]$ (why ?).

Then the limit of the left–hand side must exist, too; and it must be a continuously differentiable function. Call it $f(x)$. □

Let us now apply the above results, starting with power series.

We consider a power series $f = f(x) = \sum_{j=0}^{\infty} a_j(x-x_0)^j$. It is known that this series converges in a ball of the complex plane with centre x_0. It especially converges in an interval $(x_0 - R, x_0 + R)$ of the real axis, and the convergence is uniform in any compactly contained interval.

Question: Why ?

We put $f_n(x) = \sum_{j=0}^{n} a_j(x-x_0)^j$, which is obviously continuously differentiable and converges to $f(x)$ for $n \to \infty$.

Question: Check that the sequence $(f_n')_{n\in\mathbb{N}}$ converges uniformly in any interval $(x_0 - R', x_0 + R')$ with $R' < R$.

Since the sequence $(f_n(x_0))_{n\in\mathbb{N}}$ trivially converges to $f(x_0)$, we find that the limit function f is differentiable, and the derivative is

$$f'(x) = \sum_{j=0}^{\infty} a_j j(x-x_0)^{j-1}, \qquad x \in (x_0 - R, x_0 + R).$$

Question: Why can we write R instead of R' ?

Power series can be differentiated term-wise,
and the power series of the derivative has the same radius of convergence.

Examples:

- *Prove that the power series of* $\ln(1+x)$ *from last term converges for* $-1 < x \le 1$.
- *Give power series for* arctan *and* arcsin. *For which values of the argument do they converge ?*

We conclude this section with some more results on commuting limit processes. We do not possess the tools to prove them and refer the reader to [4, Vol. 2, Nr. 127].

Proposition 3.34 (Theorem of Arzela[16]). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of tame functions over a bounded interval $[a,b]$ that converges* pointwise *to a function f. We assume:*

- *the limit f is tame,*
- *the f_n are* uniformly bounded*: there is a constant M, such that for all n and all $x \in [a,b]$, the inequality $|f_n(x)| \le M$ holds.*

Then $\lim_{n\to\infty} \int_{x=a}^{x=b} f_n(x)\,\mathrm{d}x = \int_{x=a}^{x=b} f(x)\,\mathrm{d}x$.

Proposition 3.35. *Let I be an interval of the form $[a,+\infty)$ or $(-\infty,a]$, and let us be given a sequence of functions $(f_n)_{n\in\mathbb{N}}$. We assume:*

- *each function f_n is continuous on I and improperly integrable on I,*
- *on each compact sub-interval of I, the sequence $(f_n)_{n\in\mathbb{N}}$ converges* uniformly *to a limit function f,*

[16] Cesare Arzela, 1847 – 1912

- *there is a continuous function* $g\colon I \to \mathbb{R}$*, such that* $|f_n(x)| \le g(x)$ *for all* n *and all* $x \in I$*,*
- *this majorising function* g *is improperly integrable on* I*.*

Then the limit function f *is improperly integrable on* I*, and we can commute:*

$$\lim_{n\to\infty} \int_I f_n(x)\,\mathrm{d}x = \int_I f(x)\,\mathrm{d}x.$$

Note that in the first proposition, we had to assume that the limit function is integrable; however, this was not necessary in the second proposition.

The next result permits us to "differentiate under the integral".

Proposition 3.36 (Differentiation with respect to parameters). *Let* $\Lambda \subset \mathbb{R}$ *be a compact interval and* $-\infty < a < b < +\infty$*. Let* $\alpha = \alpha(\lambda)$*,* $\beta = \beta(\lambda)$ *be* C^1 *functions from* Λ *into* (a, b)*, and consider a continuous function* $f\colon [a, b] \times \Lambda \to \mathbb{R}$ *with continuous derivative* $\frac{\partial f}{\partial \lambda}$*. Then the function*

$$g = g(\lambda) = \int_{x=\alpha(\lambda)}^{x=\beta(\lambda)} f(x, \lambda)\,\mathrm{d}x$$

maps Λ *into* $\mathbb{R}$*, is continuously differentiable there, and has derivative*

$$g'(\lambda) = f(\beta(\lambda), \lambda) \cdot \beta'(\lambda) - f(\alpha(\lambda), \lambda) \cdot \alpha'(\lambda) + \int_{x=\alpha(\lambda)}^{x=\beta(\lambda)} \frac{\partial f}{\partial \lambda}(x, \lambda)\,\mathrm{d}x.$$

Differentiation under the integral sign is tricky if the interval of integration is unbounded and the integral is improper. You will need an integrable majorant for the derivative $\partial_\lambda f$.

3.4 Fourier Series

Literature: Greiner: *Klassische Elektrodynamik.* Chaper I.3: Fourierreihen

Literature: Greiner: *Klassische Mechanik II.* Chapter III: Schwingende Systeme

Imagine a physical system that can "oscillate" at various frequencies simultaneously. For instance

- an electrical current in a metal wire and carrying a sound signal,
- a vibrating violine string,
- a quantum mechanical system that can take on various (discrete) states simultaneously, with certain (complex) probabilities.

At least for the first two examples it is known that the system has a fundamental frequency[17] and a large number of *harmonics*[18] which are integer multiples of the fundamental frequency. Writing a periodic function as a Fourier series means to decompose it into a collection of oscillations, each being a *pure tone*[19] with its own frequency.

We say that a function $f\colon \mathbb{R} \to \mathbb{C}$ is *periodic*[20] with period $\omega \in \mathbb{R}$ if for all $t \in \mathbb{R}$ the identity

$$f(t + \omega) = f(t)$$

holds. Typical examples of periodic functions are the trigonometric functions and their linear combinations. The goal of this section is to show that *every* periodic function (with reasonable smoothness) can be written (in a certain sense) as a series of trigonometric functions, so-called FOURIER[21] series. By

[17] Grundfrequenz
[18] Oberschwingungen, Obertöne
[19] Sinuston
[20] periodisch
[21] JEAN BAPTISTE JOSEPH FOURIER, 1768 – 1830

scaling the variable t, we may assume that $\omega = 2\pi$. Then we would like to show that for a 2π–periodic function f, there are constants a_n, b_n, such that for all $x \in \mathbb{R}$, we have

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(nx) + b_n \sin(nx)\right).$$

Another possibility to write f could be

$$f(x) = \sum_{n=-\infty}^{+\infty} c_n \exp(\mathrm{i}nx).$$

Unfortunately, it turns out that this is too much to ask from the poor function f. There are many important 2π–periodic functions for which these identities are not always true, i.e., not for all x.

We have to be more careful and find answers to the following questions:

- do the series on the right–hand sides converge ?
- if yes, do they converge to $f(x)$?
- is this convergence pointwise, or uniform, or something else ?

The most beautiful convergence result refers to the L^2 norm, as we will see later.

We start with some basic result which avoids the delicate issue of the convergence of the series since we add up only a finite number of terms:

Proposition 3.37 (**Formulas for the Fourier coefficients**). *Let the function f be a so–called* trigonometric polynomial*, i.e., a function of the form*

$$f(x) = \sum_{n=-N}^{n=N} c_n \exp(\mathrm{i}nx), \qquad x \in \mathbb{R}.$$

Then f can be written as

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos(nx) + b_n \sin(nx)\right), \qquad x \in \mathbb{R},$$

where the two sets of coefficients can be converted into each other via

$$a_n = c_n + c_{-n}, \quad b_n = \mathrm{i}(c_n - c_{-n}), \quad c_n = \frac{1}{2}(a_n - \mathrm{i}b_n), \quad c_{-n} = \frac{1}{2}(a_n + \mathrm{i}b_n), \qquad n \geq 0, \qquad b_0 := 0.$$

Moreover, these coefficients can be computed from the function f as follows:

$$\begin{aligned}
a_n &= \frac{1}{\pi} \int_{x=0}^{x=2\pi} f(x) \cos(nx) \,\mathrm{d}x, && n \geq 0, \\
b_n &= \frac{1}{\pi} \int_{x=0}^{x=2\pi} f(x) \sin(nx) \,\mathrm{d}x, && n \geq 1, \\
c_n &= \frac{1}{2\pi} \int_{x=0}^{x=2\pi} f(x) \exp(-\mathrm{i}nx) \,\mathrm{d}x, && n \in \mathbb{Z}.
\end{aligned}$$

Proof. Multiply f by the appropriate trigonometric function and integrate over $[0, 2\pi]$, exploiting Proposition 3.20. □

Question: Consider an abstract unitary vector space U of dimension k with orthonormal basis $(u_1, \dots, u_k)$. How can you determine the coefficients of a vector $u \in U$ with respect to that basis ?

From now on, all functions can be complex-valued, and we will formulate most of the results in terms of the exponential functions

$$e_n = e_n(x) := \exp(\mathrm{i}nx), \qquad n \in \mathbb{Z}, \quad x \in \mathbb{R}.$$

You can rewrite all these results in terms of the Sine and Cosine functions yourselves.

The above integrals make sense not only for trigonometric polynomials f, but also for general 2π–periodic tame functions. Consequently, we set

$$\hat{f}_n := \frac{1}{2\pi}\int_{x=0}^{x=2\pi} f(x)\exp(-\mathrm{i}nx)\,\mathrm{d}x, \qquad n \in \mathbb{Z}, \tag{3.2}$$

commonly known as the nth *Fourier coefficient*[22], and write

$$f(x) \sim \sum_{n=-\infty}^{\infty} \hat{f}_n e_n(x).$$

Nothing has been said about the convergence of the series on the right; for this reason we write $\sim$ instead of $=$. Additionally, we introduce the notation

$$(S_N f)(x) := \sum_{n=-N}^{N} \hat{f}_n e_n(x),$$

and we would like to know under which conditions and in which sense we have $f = \lim_{N\to\infty} S_N f$.

This partial sum $S_N f$ is closely related to approximation problems, which we have studied in the last term. Let us define a scalar product on the space of 2π–periodic tame functions:

$$\langle f, g\rangle_{L^2} := \int_{x=0}^{x=2\pi} f(x)\overline{g(x)}\,\mathrm{d}x.$$

Then we have a nice formula for the Fourier coefficients $\hat{f}_n$ for the function f:

$$\hat{f}_n = \frac{\langle f, e_n\rangle}{\langle e_n, e_n\rangle}.$$

As usual, we define a norm $\|f\|_{L^2} := \sqrt{\langle f, f\rangle}$. The space of tame functions, together with this norm, becomes a normed space (which is, unfortunately, not a Banach space).

Denote $V_N := \operatorname{span}(e_{-N}, e_{-N+1}, \dots, e_N)$, which is a vector space over $\mathbb{C}$ with dimension $2N+1$. The elements of V_N are called *trigonometric polynomials*[23] of degree less than or equal to N. If f is a tame function, then clearly $S_N f \in V_N$.

Proposition 3.38 (**Best approximation**). *Let f be a 2π–periodic tame function and $N \in \mathbb{N}_0$. Define $S_N f \in V_N$ as above; and let $g_N \neq S_N f$ be an arbitrary function from V_N. Then*

$$\|f - S_N f\|_{L^2} < \|f - g_N\|_{L^2}, \qquad\qquad \|f\|_{L^2}^2 = \|S_N f\|_{L^2}^2 + \|f - S_N f\|_{L^2}^2.$$

If you have a complicated function f and are looking for an approximation to f in the space V_N, your best choice is $S_N f$.

Proof. This is Satz 2.31 from the first term in disguise. See Figure 3.1. □

3.4.1 General Approximation Results

The proof of $\lim_{N\to\infty} S_N f = f$ requires some more tools. We start with the representation

$$(S_N f)(x) = \frac{1}{2\pi}\int_{t=0}^{t=2\pi} f(t)\left(\sum_{n=-N}^{n=N} \exp(\mathrm{i}n(x-t))\right)\mathrm{d}t.$$

[22] Fourierkoeffizient

[23] trigonometrische Polynome

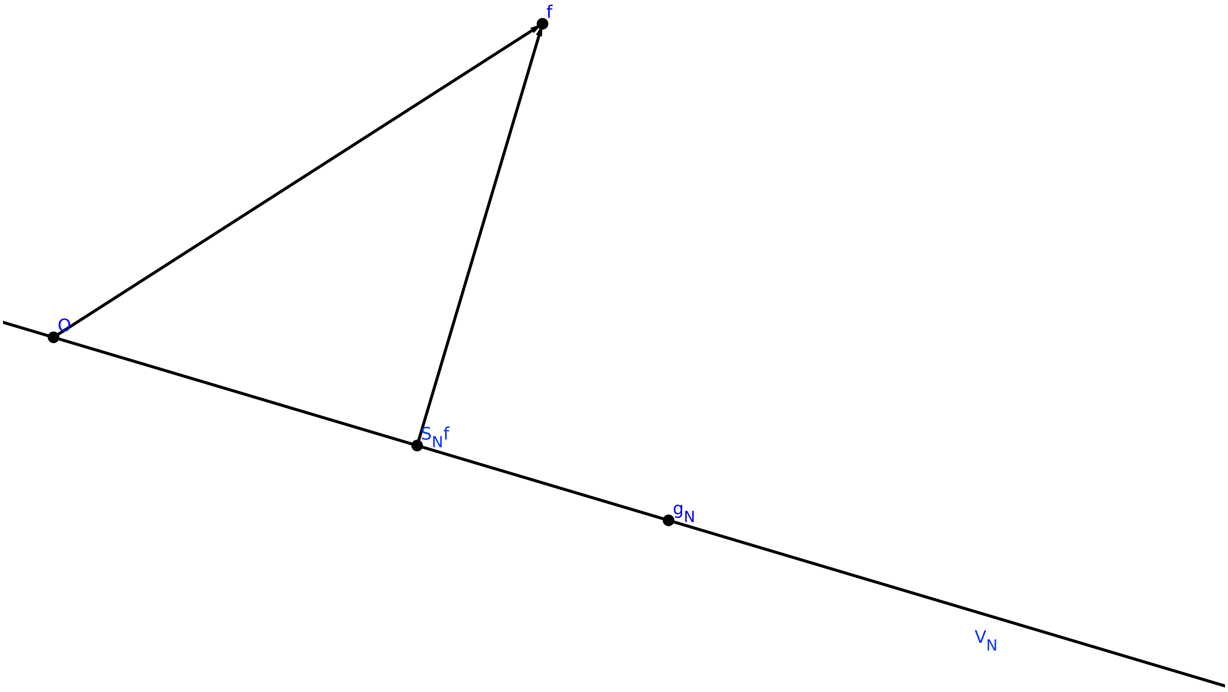


Figure 3.1: In the sub-space V_N, $S_N f$ is the best approximation to the given f. Any other element $g_N \in V_N$ is a worse approximation to f.

Introducing the Dirichlet[24] kernel

$$D_N(x) := \sum_{n=-N}^{n=N} \exp(inx),$$

we then can also write

$$(S_N f)(x) = \frac{1}{2\pi} \int_{t=0}^{t=2\pi} f(t) D_N(x-t) \, \mathrm{d}t,$$

which is a typical example of a convolution.

Definition 3.39 (**Convolution on $\mathbb{R}^1$**). *Let f and g be tame functions from $\mathbb{R}^1$ to $\mathbb{C}$, one of them identically vanishing for large values of the argument. Then the* convolution[25] $f * g = (f * g)(x)$ *is the function mapping $\mathbb{R}^1$ to $\mathbb{C}$ defined by*

$$(f * g)(x) := \int_{t=-\infty}^{t=+\infty} f(t) g(x-t) \, \mathrm{d}t.$$

Definition 3.40 (**Convolution of periodic functions**). *Let f and g be 2π–periodic tame functions from $\mathbb{R}$ to $\mathbb{C}$. Then the* convolution $f * g = (f * g)(x)$ *is the 2π–periodic function mapping $\mathbb{R}^1$ to $\mathbb{C}$ defined by*

$$(f * g)(x) := \frac{1}{2\pi} \int_{t=0}^{t=2\pi} f(t) g(x-t) \, \mathrm{d}t.$$

[24] Peter Gustav Lejeune Dirichlet, 1805 – 1859

[25] Faltung

Question: Instead of the interval $[0, 2\pi]$, you can take any other interval of length 2π as interval of integration, and will always get the same value of $(f * g)(x)$. Why ?

Question: Show the following:

- the convolution is commutative: $f * g = g * f$,
- the convolution is linear in the first factor: $(\alpha_1 f_1 + \alpha_2 f_2) * g = \alpha_1(f_1 * g) + \alpha_2(f_2 * g)$.

Clearly, $(S_N f)(x) = (f * D_N)(x)$. We want this to be an approximation of $f(x)$. To prove this, we introduce so–called DIRAC[26] sequences, proposed by the physicist PAUL DIRAC.

Definition 3.41 (Dirac sequence). *A sequence $(\delta_n)_{n\in\mathbb{N}}$ of real-valued tame functions is a* Dirac sequence *if the following conditions are fulfilled:*

- $\delta_n(x) \geq 0$ *for all* $x \in \mathbb{R}$ *and all* $n \in \mathbb{N}$,
- $\int_{t=-\infty}^{t=\infty} \delta_n(t)\,\mathrm{d}t = 1$ *for all* $n \in \mathbb{N}$,
- *For every $\varepsilon > 0$ and every $r > 0$, there is an $N_0(\varepsilon, r)$, such that for all $n \geq N_0(\varepsilon, r)$, we have*
$$\int_{\mathbb{R}\setminus[-r,r]} \delta_n(t)\,\mathrm{d}t < \varepsilon.$$

The functions δ_n have a peak at $t = 0$, which gets higher and higher as n grows.

Under the assumption that the function f behaves not too bad, the sequence of convolutions $(\delta_n * f)_{n\in\mathbb{N}}$ converges to f. We consider only the $\mathbb{R}^1$–version of the convolution; the periodic case can be proved similarly.

Proposition 3.42 (General approximation result). *Let $f\colon \mathbb{R} \to \mathbb{C}$ be a tame function, vanishing identically for large arguments; and $(\delta_n)_{n\in\mathbb{N}}$ be a Dirac sequence. Put $f_n := f * \delta_n$. Then the following holds:*

1. *if f is continuous at a point x_0, then the sequence $(f_n(x_0))_{n\in\mathbb{N}}$ converges to $f(x_0)$;*
2. *if f is uniformly continuous on $\mathbb{R}$, then the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly on $\mathbb{R}$,*
3. *if the functions δ_n are even, then the sequence $(f_n(x_0))_{n\in\mathbb{N}}$ converges to $\frac{1}{2}(f(x_0 - 0) + f(x_0 + 0))$, the arithmetic mean of the left and right limits, for all $x_0 \in \mathbb{R}$.*

Proof. 1. Fix a positive ε. Since the function f is continuous at x_0, there is a positive r, such that $|f(x_0) - f(x_0 - t)| < \varepsilon$ for $|t| < r$. Let $n \geq N_0(\varepsilon, r)$ be arbitrary. Then we can write

$$\begin{aligned}
|f(x_0) - f_n(x_0)| &= \left| f(x_0)\cdot 1 - \int_{t=-\infty}^{t=\infty} f(x_0 - t)\delta_n(t)\,\mathrm{d}t \right| \\
&= \left| f(x_0)\cdot \int_{t=-\infty}^{\infty} \delta_n(t)\,\mathrm{d}t - \int_{t=-\infty}^{t=\infty} f(x_0 - t)\delta_n(t)\,\mathrm{d}t \right| = \left| \int_{t=-\infty}^{t=\infty} (f(x_0) - f(x_0 - t))\delta_n(t)\,\mathrm{d}t \right| \\
&\leq \int_{t=-\infty}^{t=\infty} |(f(x_0) - f(x_0 - t))\cdot\delta_n(t)|\,\mathrm{d}t = \int_{t=-\infty}^{t=\infty} |(f(x_0) - f(x_0 - t))|\cdot\delta_n(t)\,\mathrm{d}t \\
&= \int_{t=-r}^{t=r} |f(x_0) - f(x_0 - t)|\cdot\delta_n(t)\,\mathrm{d}t + \int_{\mathbb{R}\setminus[-r,r]} |f(x_0) - f(x_0 - t)|\cdot\delta_n(t)\,\mathrm{d}t \\
&\leq \int_{t=-r}^{t=r} \varepsilon\cdot\delta_n(t)\,\mathrm{d}t + \int_{\mathbb{R}\setminus[-r,r]} 2\,\|f\|_{L^\infty(\mathbb{R})}\cdot\delta_n(t)\,\mathrm{d}t \\
&\leq \varepsilon\cdot 1 + 2\,\|f\|_{L^\infty(\mathbb{R})}\cdot\varepsilon \\
&= \varepsilon(1 + 2\,\|f\|_{L^\infty(\mathbb{R})}).
\end{aligned}$$

Here we have exploited all the three defining properties of the Dirac sequences. Therefore, the difference $|f(x_0) - f_n(x_0)|$ can be made arbitrarily small.

[26] PAUL ADRIEN MAURICE DIRAC, 1902 – 1984

2. If f is uniformly continuous, you can choose the above number r independent of x_0. Then also the number $N_0(\varepsilon, r)$ does not depend on x_0.

3. For even δ_n, we have $\int_{t=0}^{t=\infty} \delta_n(t)\,\mathrm{d}t = \int_{t=-\infty}^{t=0} \delta_n(t)\,\mathrm{d}t = \frac{1}{2}$. Then we can write

$$\left|\frac{f(x_0-0)+f(x_0+0)}{2} - f_n(x_0)\right|$$
$$\leq \int_{t=-\infty}^{t=0} |f(x_0-0) - f(x_0-t)|\delta_n(t)\,\mathrm{d}t + \int_{t=0}^{t=\infty} |f(x_0+0) - f(x_0-t)|\delta_n(t)\,\mathrm{d}t,$$

and continue in a similar way as in part 1.

□

Corollary 3.43 (WEIERSTRASS[27] **Approximation Theorem).** *For each continuous function on a compact interval, there is a sequence of polynomials converging uniformly to that continuous function.*

Sketch of proof. Choose

$$\delta_n(t) := \begin{cases} 0 & : \quad t \leq -1, \\ c_n(1-t^2)^n & : \quad -1 < t < 1, \\ 0 & : \quad t \geq 1, \end{cases}$$

where the constant c_n is determined by the condition $\int_{t=-\infty}^{t=+\infty} \delta_n(t)\,\mathrm{d}t = 1$. Dive into some details and then apply Proposition 3.42. □

Corollary 3.44. *For each 2π–periodic continuous function, there is a sequence of trigonometric polynomials converging uniformly to that periodic continuous function.*

Sketch of proof. A trigonometric polynomial is a linear combination of terms $\exp(\mathrm{i}nt)$, hence the nth power of $z := \exp(\mathrm{i}t)$. All that remains is to apply a variant of the Weierstrass approximation theorem.

□

3.4.2 Pointwise Convergence

Now we determine the pointwise limit of the Fourier approximations, $\lim_{N\to\infty} S_N f$.

We have $(S_N f)(x) = (f * D_N)(x)$, so our first try is take advantage from Proposition 3.42. However, this does not work, because D_N can become negative, due to

$$D_N(x) = \frac{\sin\left(N+\frac{1}{2}\right)x}{\sin\frac{1}{2}x}, \qquad x \notin 2\pi\mathbb{Z}. \tag{3.3}$$

Question: Show (3.3).

In this situation, the following brilliant lemma saves us.

Lemma 3.45. *Let $(a_n)_{n\in\mathbb{N}_+}$ be an arbitrary sequence, and put*

$$\alpha_n = \frac{1}{n}(a_1 + a_2 + \cdots + a_{n-1} + a_n).$$

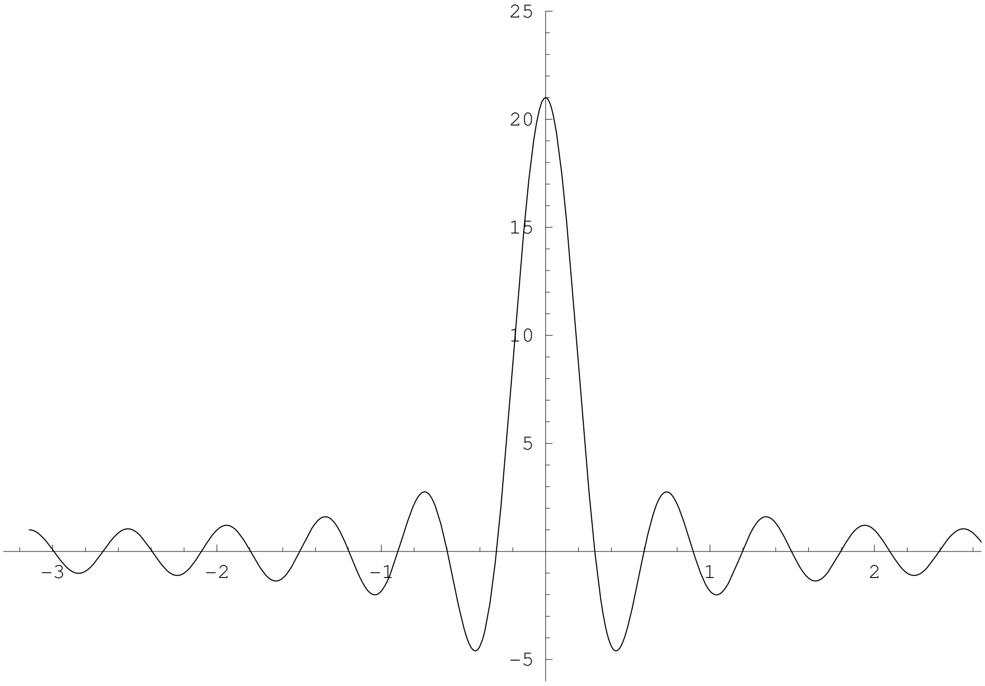
If the sequence $(a_n)_{n\in\mathbb{N}_+}$ converges to a limit A, then also the sequence $(\alpha_n)_{n\in\mathbb{N}_+}$ converges to the same limit A.

Proof. We know that $\lim_{n\to\infty} a_n = A$, which means

$$\forall\, \varepsilon > 0 \quad \exists\, N_{0,a}(\varepsilon)\colon \forall\, n \geq N_{0,a}(\varepsilon)\colon \ |a_n - A| < \varepsilon,$$

[27] KARL THEODOR WILHELM WEIERSTRASS, 1815 – 1897

```
In[1]:= Plot[Sin[10.5 x] / Sin[0.5 x], {x, -Pi, Pi}, PlotRange → {-6, 25}]
```



```
Out[1]= - Graphics -
```

Figure 3.2: The Dirichlet kernel for $N = 10$

and we wish to show that

$$\forall\, \varepsilon > 0 \quad \exists\, N_{0,\alpha}(\varepsilon)\colon \forall\, n \geq N_{0,\alpha}(\varepsilon)\colon\ |\alpha_n - A| < \varepsilon.$$

If we succeed in constructing $N_{0,\alpha}(\varepsilon)$, then we are done. The following $N_{0,\alpha}(\varepsilon)$ will give us what we need:

$$N_{0,\alpha}(\varepsilon) := \left\lceil \frac{N_{0,a}(\varepsilon/3)}{\varepsilon/3} \cdot (C + |A|) \right\rceil, \qquad C := \sup_{n \in \mathbb{N}} |a_n|,$$

with $\lceil \dots \rceil$ denoting the rounding *up*. This works, because, for $n \geq N_{0,\alpha}(\varepsilon)$, we have

$$\begin{aligned}
|\alpha_n - A| &\leq \sum_{j=1}^{N_{0,a}(\varepsilon/3)} \frac{1}{n}|a_n - A| + \sum_{j=N_{0,a}(\varepsilon/3)+1}^{n} \frac{1}{n}|a_n - A| \\
&\leq \sum_{j=1}^{N_{0,a}(\varepsilon/3)} \frac{C + |A|}{n} + \sum_{j=N_{0,a}(\varepsilon/3)+1}^{n} \frac{\varepsilon/3}{n} \\
&= \frac{N_{0,a}(\varepsilon/3)}{n} \cdot (C + |A|) + \frac{n - N_{0,a}(\varepsilon/3)}{n} \cdot \frac{\varepsilon}{3} \\
&\leq \frac{N_{0,a}(\varepsilon/3)}{N_{0,\alpha}(\varepsilon)} \cdot (C + |A|) + \frac{\varepsilon}{3} \\
&= \left(\frac{N_{0,a}(\varepsilon/3)}{\varepsilon/3} \cdot (C + |A|) \right) \cdot \frac{\varepsilon/3}{N_{0,\alpha}(\varepsilon)} + \frac{\varepsilon}{3} \\
&\leq N_{0,\alpha}(\varepsilon) \cdot \frac{\varepsilon/3}{N_{0,\alpha}(\varepsilon)} + \frac{\varepsilon}{3} = \frac{2}{3}\varepsilon.
\end{aligned}$$

This was our goal. □

In the spirit of this lemma, we define

$$(\sigma_n f)(x) := \frac{1}{n}\left((S_0 f)(x) + (S_1 f)(x) + \cdots + (S_{n-1} f)(x)\right), \qquad n \in \mathbb{N}_+.$$

Clearly, we have the integral representation (by the linearity of the convolution and the identity $2\sin\alpha\sin\beta = \cos(\beta-\alpha) - \cos(\beta+\alpha)$)

$$\begin{aligned}
(\sigma_n f)(x) &= (F_n * f)(x), \\
F_n(x) := \frac{1}{n}(D_0(x) + D_1(x) + \cdots + D_{n-1}(x)) &= \frac{1}{n\sin\frac{1}{2}x}\sum_{j=0}^{n-1}\sin\left(j+\frac{1}{2}\right)x \\
&= \frac{1}{n(\sin\frac{1}{2}x)^2}\sum_{j=0}^{n-1}\sin\frac{1}{2}x\ \sin\left(j+\frac{1}{2}\right)x \\
&= \frac{1}{2n(\sin\frac{1}{2}x)^2}\sum_{j=0}^{n-1}(\cos(jx) - \cos((j+1)x)) = \frac{1}{2n(\sin\frac{1}{2}x)^2}(1-\cos(nx)) \\
&= \frac{1}{n}\left(\frac{\sin\frac{n}{2}x}{\sin\frac{1}{2}x}\right)^2,
\end{aligned}$$

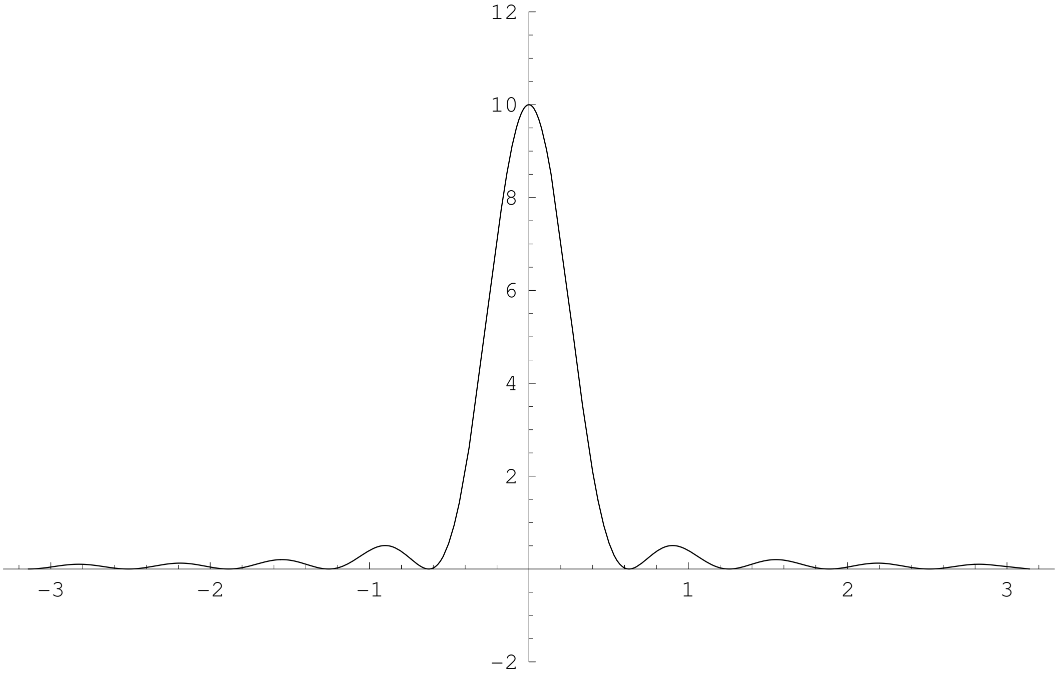
from which we easily get that the F_n form a Dirac sequence in the following sense:

Lemma 3.46. *The* FEJER[28] *kernels F_n are even functions; and additionally,*

- $F_n(x) \geq 0$ *for all x and $n \in \mathbb{N}_+$,*
- $\frac{1}{2\pi}\int_{t=-\pi}^{t=\pi} F_n(t)\,\mathrm{d}t = 1$,
- *For every $\varepsilon > 0$ and $r > 0$, there is an $N_0(\varepsilon, r)$, such that for all $n \geq N_0(\varepsilon, r)$, we have*

$$\int_{[-\pi,\pi]\setminus[-r,r]} F_n(t)\,\mathrm{d}t < \varepsilon.$$

```
In[2]:= Plot[0.1 (Sin[5 x] / Sin[0.5 x])^2, {x, -Pi, Pi}, PlotRange → {-2, 12}]
```



```
Out[2]= - Graphics -
```

Figure 3.3: The Fejer kernel for $n = 10$

Then the periodic version of Proposition 3.42 immediately yields the following result:

[28] LIPOT FEJER, 1880 – 1959

Proposition 3.47. *If f is a 2π–periodic tame function, then:*

1. *if f is continuous at x_0, then the sequence $((\sigma_n f)(x_0))_{n\in\mathbb{N}}$ converges to $f(x_0)$;*
2. *if f is continuous on $\mathbb{R}$, then the sequence $(\sigma_n f)_{n\in\mathbb{N}}$ converges to f uniformly on $[0, 2\pi]$,*
3. *the sequence $((\sigma_n f)(x_0))_{n\in\mathbb{N}}$ always converges to $\frac{1}{2}(f(x_0-0)+f(x_0+0))$.*

Going back to the original Fourier polynomials $S_n f$, we then find, by the aid of Lemma 3.45:

Proposition 3.48. *Let f be a 2π–periodic tame function. If the sequence $((S_n f)(x_0))_{n\in\mathbb{N}}$ converges, then it must converge to $\frac{1}{2}(f(x_0-0)+f(x_0-0))$. If, additionally, f is continuous at x_0, then $((S_n f)(x_0))_{n\in\mathbb{N}}$ can only converge to $f(x_0)$.*

This is the best result we can get, for there are continuous 2π–periodic functions f, whose sequence of Fourier approximations $(S_n f)_{n\in\mathbb{N}}$ does not converge everywhere.

3.4.3 Convergence in L^2

Our above result on pointwise convergence is not completely satisfactory. We did not succeed in showing that the Fourier series of a tame function (or continuous function) converges everywhere.

The situation is better if we ask for a weaker convergence, namely the convergence in the L^2–norm. As an added bonus: the L^2 convergence is physically highly meaningful.

Proposition 3.49 (Convergence in L^2). *Let f be a 2π–periodic tame function. Then the Fourier series converges in the L^2–norm:*

$$\lim_{n\to\infty} \|f - S_n f\|_{L^2(0,2\pi)} = 0.$$

Proof. **Step 1: strengthen the assumption; let f be continuous on $\mathbb{R}$.**

By Proposition 3.47, part 2, the sequence $(\sigma_n f)_{n\to\infty}$ converges uniformly to f:

$$\lim_{n\to\infty} \|f - \sigma_n f\|_{L^\infty(0,2\pi)} = 0.$$

Note that each function $u \in L^\infty(0, 2\pi)$ satisfies

$$\begin{aligned}\|u\|_{L^2(0,2\pi)} &= \sqrt{\int_{x=0}^{2\pi} |u(x)|^2 \,\mathrm{d}x} \le \sqrt{\int_{x=0}^{2\pi} \|u\|^2_{L^\infty(0,2\pi)} \,\mathrm{d}x} = \|u\|_{L^\infty(0,2\pi)} \cdot \sqrt{\int_{x=0}^{2\pi} 1 \cdot \mathrm{d}x} \\ &= \sqrt{2\pi}\, \|u\|_{L^\infty(0,2\pi)} \,.\end{aligned}$$

Therefore, we have, by Proposition 3.38,

$$\|f - S_n f\|_{L^2(0,2\pi)} \le \|f - \sigma_n f\|_{L^2(0,2\pi)} \le \sqrt{2\pi}\, \|f - \sigma_n f\|_{L^\infty(0,2\pi)} \,,$$

and this implies $\lim_{n\to\infty} \|f - S_n f\|_{L^2(0,2\pi)} = 0$.

Step 2: going back to the original assumption; f is just tame.

For each positive ε, you can find a continuous function f_ε with $\|f - f_\varepsilon\|_{L^2} < \varepsilon$.

Question: Build such an f_ε, drawing a picture if necessary.

We can always write, by the triangle inequality,

$$\|f - S_n f\|_{L^2(0,2\pi)} \le \|f - f_\varepsilon\|_{L^2(0,2\pi)} + \|f_\varepsilon - S_n f_\varepsilon\|_{L^2(0,2\pi)} + \|S_n(f_\varepsilon - f)\|_{L^2(0,2\pi)} \,.$$

Note that $\|S_n u\|_{L^2(0,2\pi)} \le \|u\|_{L^2(0,2\pi)}$, which is plausible from a picture like Figure 3.1, and a rigorous proof is in Proposition 3.38. Then we get

$$\begin{aligned}\|f - S_n f\|_{L^2(0,2\pi)} &\le \|f - f_\varepsilon\|_{L^2(0,2\pi)} + \|f_\varepsilon - S_n f_\varepsilon\|_{L^2(0,2\pi)} + \|f_\varepsilon - f\|_{L^2(0,2\pi)} \\ &\le \varepsilon + \|f_\varepsilon - S_n f_\varepsilon\|_{L^2(0,2\pi)} + \varepsilon.\end{aligned}$$

Note that the second item can be made arbitrarily small for large n; this follows from Step 1 of this proof. All together we obtain $\|f - S_n f\|_{L^2(0,2\pi)} < 3\varepsilon$ for $n \ge N_0(\varepsilon, f)$.

□

This beautiful result can be even improved: for the L^2–convergence of the Fourier series, it is not necessary to assume that the function be tame. Recall what tame means: the function has to be bounded, and at every point, the limit from the left and the limit from the right must exist. We do not need this for the L^2–convergence of the Fourier series. It suffices to assume that the function to be approximated belongs to the Lebesgue–space $L^2(0, 2\pi)$.

Definition 3.50 (L^p **space**)**.** *A function f belongs to the Lebesgue space $L^p(0, 2\pi)$ for $1 \le p < \infty$ if there is a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of step functions with the following properties:*

- *the sequence $(\varphi_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in the $L^p(0, 2\pi)$–norm:*

$$\|g\|_{L^p(0,2\pi)} = \left(\int_{t=0}^{t=2\pi} |g(t)|^p \, \mathrm{d}t\right)^{1/p} ;$$

- *the sequence $(\varphi_n(x))_{n\in\mathbb{N}}$ converges almost everywhere to $f(x)$.*

"Almost everywhere" means "everywhere with the exception of a set of Lebesgue measure[29] zero". A subset A of $\mathbb{R}$ has the Lebesgue measure zero if, for every $\varepsilon > 0$, you can find a countable[30] collection of intervals that cover A, and whose sum of lengths is at most ε.

A function from such a Lebesgue space may be continuous nowhere; and its limits from left or right may exist nowhere; and it can have (soft) poles. A function belongs to $L^p(0, 2\pi)$ if its L^p–norm is finite.

Proposition 3.51 (**Convergence in** L^2)**.** *The Fourier series of a 2π–periodic function from $L^2(0, 2\pi)$ converges in the L^2–norm.*

The proof of this nice result is very similar to the proof of Proposition 3.49.

You can generalise this result a bit more:

Proposition 3.52 (**Convergence in** L^p)**.** *The Fourier series of a 2π–periodic function from $L^p(0, 2\pi)$ with $1 < p < \infty$ converges in the L^p–norm.*

The convergence in the L^p–norm for $p > 2$ is stronger than the L^2–convergence.

The proof is insightful and amazing; we could learn a lot from it. Regrettably, the shortness of this term (only 14 weeks) and the length of the proof (about 20 pages) force us to omit it.

We conclude the section on Fourier series with a famous result, which you can consider as "splitting of energy" if you would like.

Proposition 3.53 (Bessel's[31] **Identity**)**.** *Let $f \in L^2(0, 2\pi)$ be a 2π–periodic function, and let $\hat{f}_n$ be its Fourier coefficients as defined in (3.2). Then we have*

$$\|f\|^2_{L^2(0,2\pi)} = \int_{t=0}^{t=2\pi} |f(t)|^2 \, \mathrm{d}t = 2\pi \sum_{n=-\infty}^{+\infty} |\hat{f}_n|^2.$$

Proof. We know that

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}_n e_n(x),$$

where we understand the equality sign as "convergence in the L^2–norm". Rewrite this as $f = \sum_{n=-N}^{N} \hat{f}_n e_n + R_N$. The functions e_n are orthogonal to each other (why ?) and to R_N. Then the Pythagoras theorem yields

$$\|f\|^2_{L^2} = \sum_{n=-N}^{N} |\hat{f}_n|^2 \, \|e_n\|^2_{L^2} + \|R_N\|^2_{L^2} = 2\pi \sum_{n=-N}^{N} |\hat{f}_n|^2 + \|R_N\|^2_{L^2} .$$

The last term goes to zero for $N \to \infty$. □

[29] Lebesgue–Maß

[30] abzählbar

[31] Friedrich Wilhelm Bessel, 1784 – 1846

We should explain why Bessel's identity can be seen as a splitting of the energy. Typically, kinetic energies are quadratic functionals. Suppose that a particle of mass m performs at the same time two linear motions with the velocity vectors $\vec{v}_1$ and $\vec{v}_2$. Then the kinetic energy associated to the motion $\vec{v}_1$ and $\vec{v}_2$, respectively, are

$$E_1 = \frac{m}{2} \|\vec{v}_1\|^2, \qquad E_2 = \frac{m}{2} \|\vec{v}_2\|^2.$$

And the total energy is

$$E_{\text{total}} = \frac{m}{2} \|\vec{v}_1 + \vec{v}_2\|^2.$$

Since $\|\vec{v}_1 + \vec{v}_2\|^2 = \|\vec{v}_1\|^2 + 2\langle \vec{v}_1, \vec{v}_2\rangle + \|\vec{v}_2\|^2$, we only have $E_{\text{total}} = E_1 + E_2$ if $\vec{v}_1 \perp \vec{v}_2$.

The kinetic energy functional will only respect the splitting into two sub-motions if those sub-motions are perpendicular to each other !

In the case of the Fourier series, we are lucky, since $e_n \perp e_m$ for $n \neq m$, hence the electrical energy of the sound signal in the wire to your earphones really can be split into the various frequency components.

3.4.4 Excursion: Dirac's Delta–Distribution

The delta distribution is an indispensable tool for physicists, so you should know it.

Never ever say in a maths lecture that the delta function takes the value $+\infty$ at the origin, everywhere else it takes the value zero, but its integral over the real line equals one.

That is preposterous nonsense.[32]

Every physicist should have seen a rigorous definition of the delta distribution at least once; and since this definition is quite similar to that of Dirac sequences (Definition 3.41), it is no big issue to discuss this matter right now. An added bonus is that you will better understand where the computing rules of the delta distribution come from.

To this end, we need some preparations:

Definition 3.54 (Test function space). *The set $C_0^\infty(\mathbb{R})$ consists of all functions $\varphi\colon \mathbb{R} \to \mathbb{R}$ which are differentiable an infinite number of times, and which vanish for large arguments: for $\varphi \in C_0^\infty(\mathbb{R})$, there is a number $M > 0$ such that $\varphi(x) = 0$ for all $|x| \geq M$. The members of that space are called* test functions. *We also write $\mathcal{D}$ instead of $C_0^\infty(\mathbb{R})$.*

Definition 3.55 (Distribution). *A distribution T is a* linear *and* continuous *map from $\mathcal{D} = C_0^\infty(\mathbb{R})$ to $\mathbb{R}$. The set of all distributions is written as $\mathcal{D}'$.*

We should explain this a bit:

linear: if $T \in \mathcal{D}'$ and $\varphi_1, \varphi_2 \in \mathcal{D}$, and $\alpha_1, \alpha_2 \in \mathbb{R}$, then $T(\alpha_1\varphi_1 + \alpha_2\varphi_2) = \alpha_1 T(\varphi_1) + \alpha_2 T(\varphi_2)$.

continuous: if a sequence $(\varphi_k)_{k\in\mathbb{N}}$ of test functions converges to a test function φ_* in $\mathcal{D}$, then $\lim_{k\to\infty} T(\varphi_k) = T(\varphi_*)$.

convergence in $\mathcal{D}$: we say that a sequence $(\varphi_k)_{k\in\mathbb{N}}$ converges to a test function φ_* in $\mathcal{D}$ if there is a number $M > 0$ such that $\varphi_k(x) = 0$ for all k and all $|x| \geq M$, and if $\|\partial_x^m(\varphi_k - \varphi_*)\|_{L^\infty} \to 0$ for $k \to \infty$ and all derivative orders $m \in \mathbb{N}$.

[32] And it is also strategically not really clever. If you deal with mathematical objects like functions or vector fields or tensors or coordinates of a point, it is advantageous to have an eye on how they are transformed when the coordinate system changes. Each object can transform in covariant style or contravariant style. And it turns out that distributions always transform in the opposite style: if the underlying test functions transform covariantly, then the associated distributions transform contravariantly, and vice versa.

It turns out (and is quite easy to prove) that $\mathcal{D}'$ is a vector space over the field $\mathbb{R}$. Its dimension is infinite, and it is *not* a normed space in the sense that one norm is not enough to describe its topology.

The two key examples of distributions are the following:

delta distribution: define a mapping $\delta_0 \colon \mathcal{D} \to \mathbb{R}$ by $\delta_0(\varphi) = \varphi(x = 0)$, for $\varphi \in \mathcal{D}$.

regular function: let $f \colon \mathbb{R} \to \mathbb{R}$ be a tame function. Then we set $I_f(\varphi) = \int_{x=-\infty}^{\infty} f(x)\varphi(x)\,\mathrm{d}x$, for $\varphi \in \mathcal{D}$. The I stands for "integration".

In that sense, every integrable function can be interpreted as a distribution.

A crucial step is the following general representation (whose proof is beyond our reach, unfortunately):

Proposition 3.56. *Every distribution T can be approximated via smooth regular functions in the following sense:*

For each $T \in \mathcal{D}'$, there is a sequence $(f_k)_{k\in\mathbb{N}}$ with $f_k \in \mathcal{D}$ for all k, with the property that

$$T(\varphi) = \lim_{k\to\infty} I_{f_k}(\varphi), \qquad \forall \varphi \in \mathcal{D}.$$

For instance, the regular functions f_k giving an approximation of the delta distribution δ_0 can be chosen very similarly to the Dirac sequences from Definition 3.41.

Now we are in a position to understand some computing rules of the delta distribution:

$$\begin{aligned}
&(\delta_0(ax))(\varphi(x)) = \frac{1}{a}\delta_0(\varphi) = \frac{1}{a}\varphi(x = 0), \qquad a > 0,\\
&\delta_0'(\varphi) = -\varphi'(x = 0).
\end{aligned}$$

To show them, we proceed as follows. Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of functions approximating δ_0 in the above sense. Then we have

$$\begin{aligned}
&(\delta_0(ax))(\varphi(x)) = \lim_{k\to\infty}\int_{x=-\infty}^{\infty} f_k(ax)\varphi(x)\,\mathrm{d}x = \lim_{k\to\infty}\int_{y=-\infty}^{\infty} f_k(y)\varphi(y/a)\frac{1}{a}\,\mathrm{d}y = \frac{1}{a}\varphi(y = 0),\\
&\delta_0'(\varphi) = \lim_{k\to\infty}\int_{x=-\infty}^{\infty} f_k'(x)\varphi(x)\,\mathrm{d}x = -\lim_{k\to\infty}\int_{x=-\infty}^{\infty} f_k(x)\varphi'(x)\,\mathrm{d}x = -\varphi'(x = 0).
\end{aligned}$$

3.5 Curves

3.5.1 General Properties

Definition 3.57 (**Curve, Image**). *A* curve[33] *is a continuous function γ which maps a compact interval into $\mathbb{R}^n$:*

$$\begin{aligned}
&\gamma \colon [a, b] \to \mathbb{R}^n,\\
&\gamma \colon t \mapsto \gamma(t) = x(t) = (x_1(t), \dots, x_n(t))^\top.
\end{aligned}$$

The set $\Gamma = \{\gamma(t) \colon a \le t \le b\}$ is called image of the curve.

Different curves may have the same image. Here are two curves for the upper semi-circle, run in counterclockwise orientation:

$$\begin{aligned}
&\gamma \colon [0, \pi] \to \mathbb{R}^2, &&\gamma(\varphi) = (\cos\varphi, \sin\varphi)^\top,\\
&\gamma \colon [-1, 1] \to \mathbb{R}^2, &&\gamma(t) = \left(\frac{-2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right)^\top.
\end{aligned}$$

Always distinguish a curve γ from its image Γ !

[33] Kurve

Definition 3.58. *A curve $\gamma\colon [a,b] \to \mathbb{R}^n$ is called* simple *if $\gamma(t) = \gamma(s)$ implies $s = t$ or $\{s,t\} = \{a,b\}$.*

A curve $\gamma\colon [a,b] \to \mathbb{R}^n$ is said to be a loop[34] *or* closed *if $\gamma(a) = \gamma(b)$.*

A simple closed curve is called a Jordan curve[35] [36].

You may think of a curve γ as a trajectory of a moving particle; and the image Γ can be visualised like a condensation trail of a jet flying in the sky. The image Γ is the object we are really interested in, and the curve γ is mostly an analytical tool for the investigation of Γ. Be advised that the denotation in the literature is not uniform. Sometimes the function γ is called *parametrisation*, and then the image Γ is called *curve.*

Be warned that the curves can be scary creatures. Naively, one should expect the image of a curve to be an one-dimensional object. Nevertheless, in 1890, Peano[37] constructed a (non-simple) curve whose image fills the whole square $(0,1) \times (0,1) \subset \mathbb{R}^2$ and is, consequently, a two-dimensional object. More information can be found in the Wikipedia, under the entries *Peano curve* and *Hilbert curve.*

But if we consider only *Jordan* curves, such strange things cannot happen:

Proposition 3.59 (Jordan Curve Theorem). *A Jordan curve in the plane divides the plane into three parts: a part "inside the image", a part "outside the image", and the image itself. The image is the boundary of the "inner part", and is the boundary of the "outer part".*

Looking at the Peano example, it should be no surprise to you that the proof of the Jordan curve theorem is longer than we have the time for.

For a curve $\gamma\colon [a,b] \to \mathbb{R}^n$, consider a subdivision of the interval $[a,b]$:

$$a = t_0 < t_1 < t_2 < \dots < t_{m-1} < t_m = b.$$

Then you can connect $\gamma(t_{i-1})$ and $\gamma(t_i)$ by a straight line. The union of all these straight lines should give us an approximation of the image of the curve by a polygon, which clearly has a *length.* And intuitively it is clear that the length of the curve γ can never be shorter than the length of the polygon.

Definition 3.60 (Length). *The length of a curve $\gamma\colon [a,b] \to \mathbb{R}^n$ is defined as*

$$\operatorname{length}(\gamma) := \sup \left\{ \sum_{j=1}^{m} \|\gamma(t_j) - \gamma(t_{j-1})\| \ : \ a = t_0 < t_1 < \dots < t_{m-1} < t_m = b \right\}.$$

A curve γ is called rectifiable[38] *if its length is finite.*

A deeper look at this definition will convince you that two curves γ, $\tilde{\gamma}$ with the same image Γ have the same length. We should speak in terms of lengths of *images of curves*, not just lengths of curves.

Definition 3.61 (Tangent vector). *A curve $\gamma\colon [a,b] \to \mathbb{R}^n$ is said to be* differentiable *if the function γ is differentiable. The vector $\dot{\gamma}(t)$ is called* tangent vector *to the curve at the point $\gamma(t)$, for $t \in [a,b]$.*

Warning: *The differentiability is the property of the* curve*, not the property of the image. It can happen that the curve γ is differentiable (even differentiable an infinite number of times), but the image Γ of that curve has a "corner" like this: $\ulcorner$. Moreover, if the curve is not a simple curve, but "intersects" itself like the symbol ∞, the intersection point of the image can have more than one tangent.*

Corners of the image will be impossible if we demand the curve to be regular:

Definition 3.62 (Regular curve). *A curve is said to be* regular *if it is continuously differentiable and the tangent vector is never the null vector.*

[34] Schleife

[35] Jordankurve

[36] Marie Ennemond Camille Jordan, 1838 – 1922, also explorer of the Jordan normal form, not to be confused with Wilhelm Jordan, 1842 – 1899, famous for the Gauss–Jordan method

[37] Giuseppe Peano, 1858 – 1932

[38] rektifizierbar

Proposition 3.63 (Length). *The length of a continuously differentiable curve $\gamma\colon [a,b] \to \mathbb{R}^n$ can be calculated as follows:*

$$\operatorname{length}(\gamma) = \int_{t=a}^{t=b} \|\dot\gamma(t)\| \,\mathrm{d}t,$$

where $\|\dot\gamma(t)\| = \sqrt{\dot x_1^2(t) + \dots + \dot x_n^2(t)}$. Different curves with the same image have the same length.

Proof. Geometrically, it should be plausible. The details are dropped. □

Corollary 3.64. *Consider a differentiable function $y = f(x)$ mapping the interval $[a,b]$ into $\mathbb{R}^1$. The graph of this function is the image Γ of a curve γ, whose length is*

$$\operatorname{length}(\gamma) = \int_{x=a}^{x=b} \sqrt{1 + (f'(x))^2}\,\mathrm{d}x.$$

Proof. Without loss of generality, we can write γ as $\gamma\colon [a,b] \to \mathbb{R}^2$, $\gamma(x) = (x, f(x))^\top$. □

Example: *Consider the interval $[a,b] = [-1,1]$ and the function $y = f(x) = \sqrt{1-x^2}$, and compute the length of the curve described by the graph of f.*

The curve γ is also called a *parametrisation*[39] of the image Γ. If you have a parametrisation, you can always replace it by another parametrisation giving it the same image. It suffices to choose an arbitrary strictly monotone increasing function $t = t(s)$ which maps an interval $[c,d]$ onto $[a,b]$, and consider

$$(\gamma \circ t)(s) = \gamma(t(s)).$$

If the function $t = t(s)$ is strictly monotone *de*creasing instead, mapping the interval $[c,d]$ onto $[a,b]$, then the re-parametrisation $\gamma \circ t$ induces an "inversion of the direction": the "particle" is now travelling along Γ, but backwards. Sometimes we will write $-\Gamma$ instead of Γ to make it clear that the direction has been inverted.

Definition 3.65. *A parametrisation γ of an image Γ is said to be* natural *or* unit speed *or* a parametrisation by arc-length[40] *if $\|\dot\gamma(t)\| = 1$ for all t.*

Then the tangent vector is always normalised, and the length of the curve $\gamma\colon [a,b] \to \mathbb{R}^n$ is simply $b-a$.

Finally, we consider the important case of *planar curves*[41]. These are curves whose image is contained in the two–dimensional Euclidean plane.

Proposition 3.66. *Let us be given a planar curve,*

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix}, \qquad a \le t \le b,$$

where the functions f and g are continuously differentiable. If $\dot f(t)$ is always positive, then the image of this curve can be written in the form $y = y(x)$ with derivative

$$y'(x) = \frac{\dot g(t)}{\dot f(t)}, \qquad x = f(t).$$

If, additionally, f and g are twice continuously differentiable, then the second derivative of the function $y = y(x)$ is

$$y''(x) = \frac{\ddot g(t)}{(\dot f(t))^2} - \frac{\dot g(t) \ddot f(t)}{(\dot f(t))^3}, \qquad x = f(t).$$

[39] Parametrisierung

[40] Bogenlängenparametrisierung

[41] ebene Kurven

Proof. The function $x = f(t)$ is invertible, because $\dot{f}$ never vanishes. Then we have an inverse function $t = \tau(x)$ with $x = f(\tau(x))$; and $y = y(x)$ is $y = g(\tau(x))$. Differentiating this twice with respect to x gives

$$\begin{aligned}
&1 = \dot{f}(\tau(x))\tau'(x),\\
&0 = \ddot{f}(\tau(x))(\tau'(x))^2 + \dot{f}(\tau(x))\tau''(x),\\
&y'(x) = \dot{g}(\tau(x))\tau'(x),\\
&y''(x) = \ddot{g}(\tau(x))(\tau'(x))^2 + \dot{g}(\tau(x))\tau''(x).
\end{aligned}$$

All that remains is to plug these equations into each other. □

A fundamental property of a planar curve is the *curvature*[42]. Roughly, it tells you how much you have to turn the steering wheel at a point if you are driving along the image of the curve with a car. The curvature is positive or negative if you are driving at a left curve or a right curve, respectively.

More precisely, the curvature at a point (x_0, y_0) of the image of the curve is defined as follows: suppose without loss of generality that we are given a unit speed parametrisation $(x, y)^\top = \gamma(s) = (\gamma_1(s), \gamma_2(s))^\top$ of the curve. Denote the tangential vector (velocity) by $T(s) = \gamma'(s)$. The unit normal vector $N(s)$ is obtained by rotating $T(s)$ to the left 90 degrees,

$$N(s) = \begin{pmatrix} \cos 90^\circ & -\sin 90^\circ \\ \sin 90^\circ & \cos 90^\circ \end{pmatrix} \gamma'(s) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1'(s) \\ \gamma_2'(s) \end{pmatrix} = \begin{pmatrix} -\gamma_2'(s) \\ \gamma_1'(s) \end{pmatrix}.$$

The vectors $N(s)$ and $T'(s)$ are parallel (this is a key advantage of the unit speed parametrisation).

Question: Why ?

Finally, the curvature $\kappa(s)$ is defined by $T'(s) = \kappa(s)N(s)$. The modulus of the curvature can be easily computed as $|\kappa(s)| = \|T'(s)\| = \|\gamma''(s)\|$.

Another approach to the curvature is the following: Suppose for simplicity reasons that in a neighbourhood of the point under consideration, the image of the curve can be written as a twice continuously differentiable function $y = y(x)$. Now you determine a circle that passes through (x_0, y_0) and whose describing function $y = c(x)$ has (at the point x_0) the same first and second derivatives as $y = y(x)$. There is exactly one such circle. The modulus of the curvature of the curve is the inverse of the radius of that circle.

Proposition 3.67 (**Curvature**)**.** *For a regular C^2–curve $\gamma = (x, y)^\top$, the curvature is*

$$\kappa(t) = \frac{\dot{x}(t)\ddot{y}(t) - \ddot{x}(t)\dot{y}(t)}{(\dot{x}^2(t) + \dot{y}^2(t))^{3/2}}.$$

Especially, for a graph of a function given by $y = f(x)$, the curvature is

$$\kappa(x) = \frac{f''(x)}{(1 + (f'(x))^2)^{3/2}}.$$

And in case of the unit speed parametrisation $\gamma = (x(s), y(s))^\top$, we have

$$\kappa(s) = x'(s)y''(s) - x''(s)y'(s).$$

Proof. Nice exercise. □

We continue with some remarks about areas.

Proposition 3.68 (**Sectorial area**)**.** *Let $\gamma\colon [a, b] \to \mathbb{R}^2$ be a continuously differentiable curve. The* sectorial area[43] *between the image of this curve and the origin is defined as follows: take a polygonal line approximating the image of γ and adjoin the origin to it, which gives you a closed polygon for which you can compute the area. The sectorial area between the image of γ and the origin is then the limit that you obtain if you make the partition of the interval $[a, b]$ underlying the polygonal approximation infinitesimally fine.*

[42] Krümmung

[43] Sektorfläche

The sectorial area is equal to

$$A(\gamma) = \frac{1}{2}\int_{t=a}^{t=b} (x(t)\dot{y}(t) - y(t)\dot{x}(t))\,\mathrm{d}t.$$

Proof. The area of a triangle with the corners $(x_1, y_2)^\top$, $(x_2, y_2)^\top$, and $(x_3, y_3)^\top$ is

$$A = \frac{1}{2}\det\begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}.$$

The details are left to the student. □

3.5.2 Applications

Definition 3.69 (**Conic sections**). *Fix positive real numbers a, b, p. An* ellipse, hyperbola, parabola in normal form[44] *is the set of all points $(x, y)^\top \in \mathbb{R}^2$ that solve the equations*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad \textit{(ellipse)},$$

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \qquad \textit{(hyperbola)},$$

$$y^2 = 2px, \qquad \textit{(parabola)}.$$

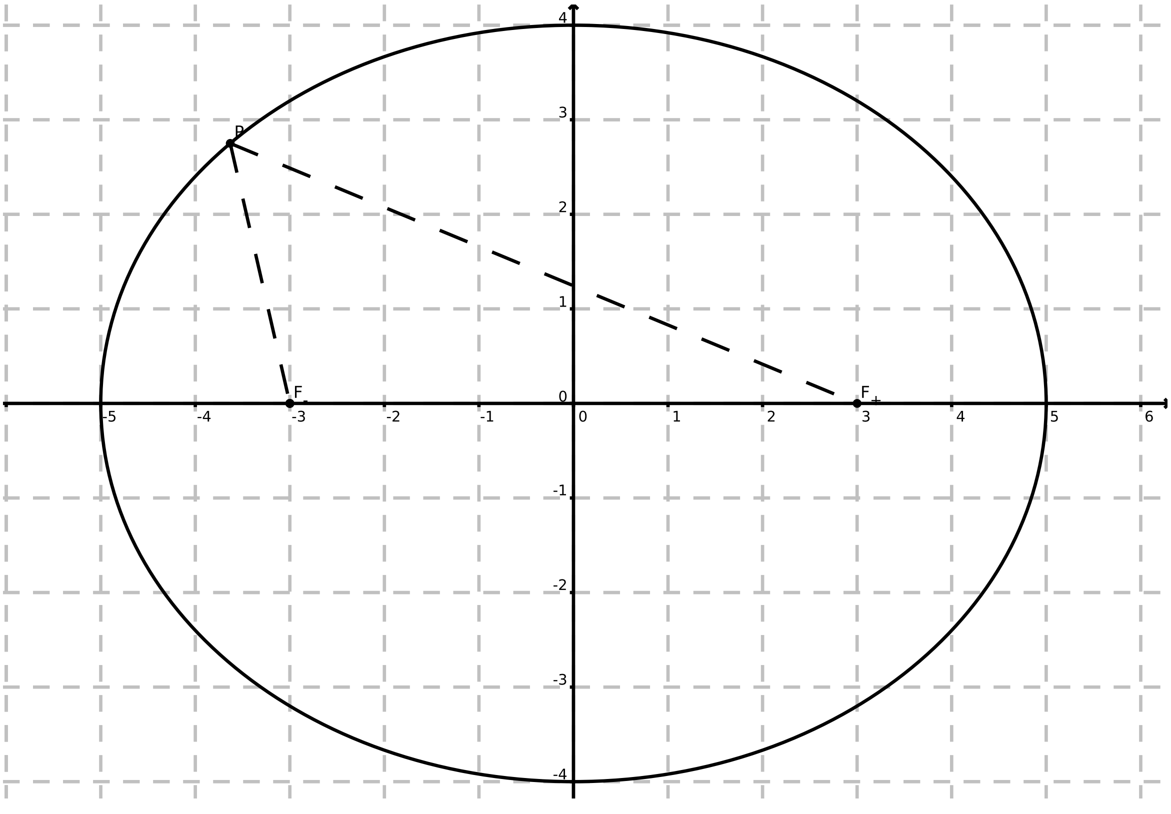


Figure 3.4: An ellipse with $a = 5$ and $b = 4$. The two focal points F_- and F_+ are at $(\pm m, 0)$ with $m = \sqrt{a^2 - b^2} = 3$. The two dashed lines, connecting the two foci to the generic point P on the ellipse, add up to $2a = 10$. A ray of light that starts at F_- and gets reflected at the ellipse obeying the well-known reflection rule, passes through F_+. The numerical eccentricity $\varepsilon := \frac{m}{a} < 1$ measures how far the ellipse differs from a circle. The shifted polar coordinates have their center in F_-.

[44] Ellipse, Hyperbel, Parabel in Normalform

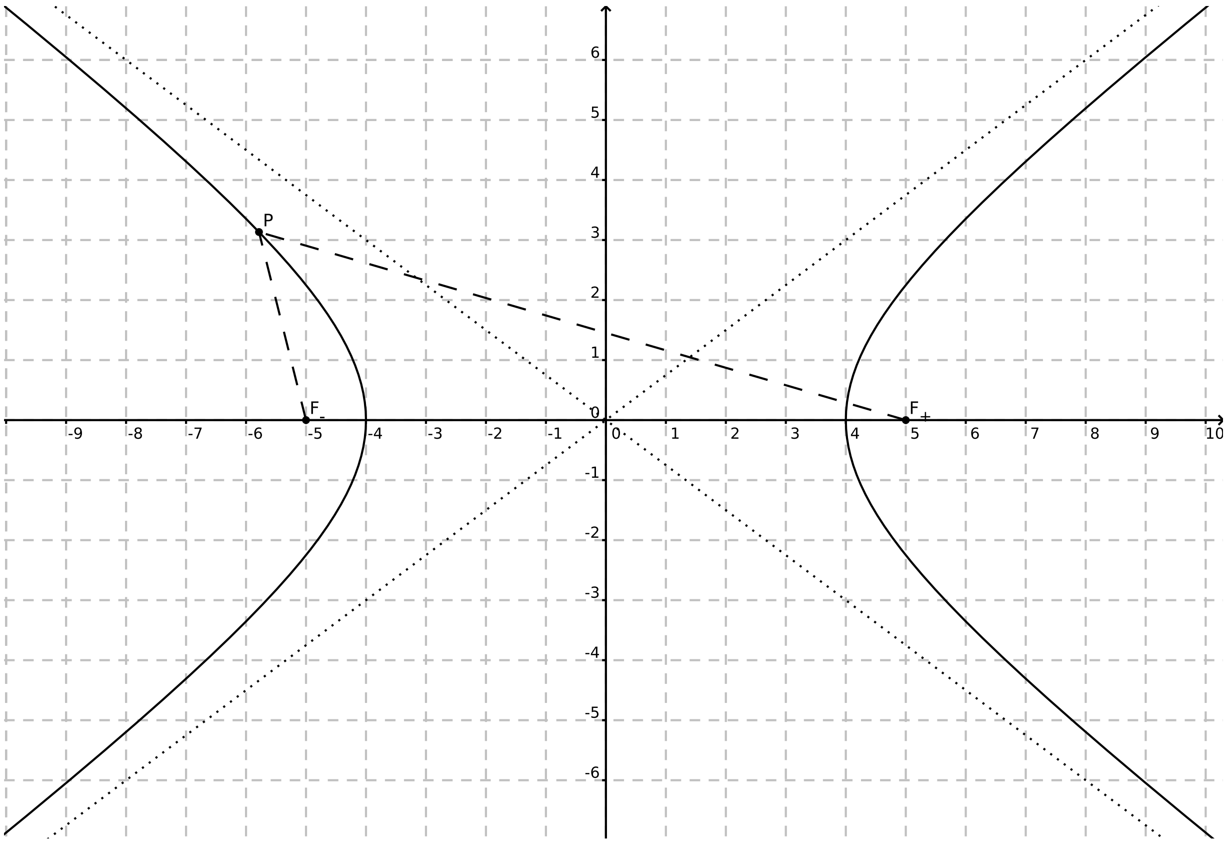


Figure 3.5: A hyperbola with $a = 4$ and $b = 3$. The asymptotic lines (dotted) are given by $|\frac{x}{a}| = |\frac{y}{b}|$. The two focal points F_- and F_+ are at $(\pm m, 0)$ with $m = \sqrt{a^2 + b^2} = 5$. The two dashed lines, connecting the two foci to the generic point P on the hyperbola, have lengths whose absolute difference is $2a = 8$. The numerical eccentricity $\varepsilon := \frac{m}{a}$ is now greater than 1. The shifted polar coordinates have their center now in the right focal point F_+, and the formula $r(\varphi) = \frac{p\varepsilon}{1-\varepsilon\cos\varphi}$ only describes the right branch of the hyperbola, not the left one.

You can always rewrite these equations in **shifted** polar coordinates,

$$r = r(\varphi) = \frac{p\varepsilon}{1 - \varepsilon\cos\varphi},$$

where $\varepsilon < 1$, $\varepsilon > 1$ and $\varepsilon = 1$ in case of the ellipse, hyperbola, parabola, respectively. The parameters (a, b) and (p, ε) are connected via

$$a = \frac{\varepsilon p}{|1 - \varepsilon^2|}, \qquad b = \frac{\varepsilon p}{\sqrt{|1 - \varepsilon^2|}}.$$

Question: Show that, for given a and b, you can always find such ε and p. Prove the above polar coordinate formula for these *conic sections*[45].

Then we can consider the motion of a *celestial body*[46] around the sun.

Put the origin in the centre of the sun, and call the masses of the moving body and the sun m and M, respectively. The gravitation law and NEWTON's third axiom give

$$m\ddot{x} = -\gamma M m \frac{x}{\|x\|^3}.$$

[45] Kegelschnitte

[46] Himmelskörper

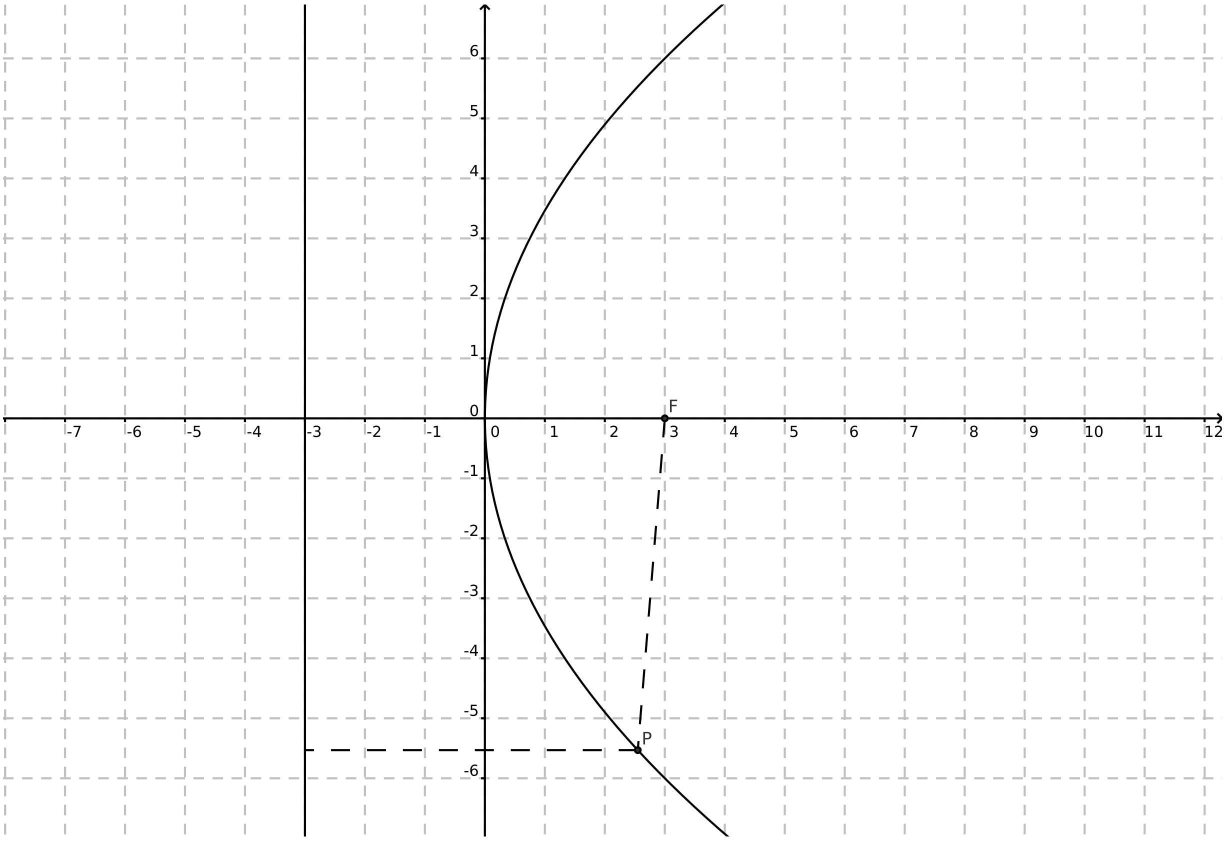


Figure 3.6: A parabola with equation $y^2 = 2px$, where $p = 6$. The focus F is at $(\frac{p}{2}, 0)$, and the *directrix* (in German: Leitlinie) is the line given by the equation $x = -\frac{p}{2}$. The two dashed lines have equal length. A ray of light that travels horizontally leftward and gets reflected at the parabola obeying the well-known reflection rule, passes through F. This is the reason why your satellite dish antenna works ! The numerical eccentricity ε is now exactly equal to 1. The shifted polar coordinates have their center in the focal point F, and the parabola can be written as $r(\varphi) = \frac{p}{1-\cos\varphi}$ in these polar coordinates.

We define the *vector of angular momentum*[47] and the *axial vector*[48] as

$$J = x \times m\dot{x},$$
$$A = \frac{1}{\gamma Mm} J \times \dot{x} + \frac{x}{\|x\|}.$$

Lemma 3.70. *The vectors J and A are constant in time, and perpendicular to each other.*

Proof. Show that the time derivatives vanish. Be careful to not turn your calculations into a big mess. □

Therefore, the three-dimensional vector x must live in a two-dimensional plane. Introduce a system of polar coordinates (r, φ) as usual in this plane, where you count φ starting from A. Put $\varepsilon = \|A\|$. Then we have two expressions for the scalar product $\langle A, x\rangle$:

$$\begin{aligned}\langle A, x(t)\rangle &= \varepsilon \, \|x(t)\| \cos\varphi(t),\\ \langle A, x(t)\rangle &= \frac{1}{\gamma Mm} \langle J \times \dot{x}(t), x(t)\rangle + \|x(t)\| = \frac{1}{\gamma Mm} \det(J, \dot{x}(t), x(t)) + \|x(t)\|\\ &= \frac{1}{\gamma M} \det(m\dot{x}(t), x(t), J) + \|x(t)\| = \frac{1}{\gamma M} \langle m\dot{x}(t) \times x(t), J\rangle + \|x(t)\| = -\frac{J^2}{\gamma Mm^2} + \|x(t)\|.\end{aligned}$$

Here we have used that the determinant function (written as det) in $\mathbb{R}^3$ does not change under cyclic permutations of its three arguments.

[47] Drehimpulsvektor

[48] Achsenvektor

In case of $\|A\| = 0$, we see that x is running around in circles. Otherwise we have

$$r(t) = \|x(t)\| = \frac{\varepsilon p}{1 - \varepsilon \cos \varphi(t)}, \qquad p = \frac{J^2}{\gamma M m^2 \|A\|}.$$

This proves that the celestial body is moving along a conic section.

In case of a comet, you may have either a hyperbola or an ellipse. In case of a planet, it is an ellipse.

KEPLER's **First Law:** The planets move along an ellipse with the sun as one of its *focal points*[49].

Now define Cartesian coordinates with basis vectors e_1, e_2, e_3 which are given by the relations $e_1 \parallel A$ and $e_3 \parallel J$. Then $x_3(t) \equiv 0$, and

$$\frac{1}{m} J = x \times \dot{x} = \begin{pmatrix} 0 \\ 0 \\ x_1 \dot{x}_2 - \dot{x}_1 x_2 \end{pmatrix}.$$

Then Proposition 3.68 tells you that the sectorial area between the position at time t_1 and the position at time t_2 (and the origin) is $\pm \frac{1}{2m} \|J\| \, |t_2 - t_1|$, which depends only on the *difference* of the times.

KEPLER's **Second Law:** The position vector covers equal areas in equal time intervals.

Calling the time needed for one revolution T and the area of the ellipse A_0, we find that $A_0 = \frac{1}{2m} \|J\| \, T$. On the other hand, the area of an ellipse with parameters (a, b) is $A_0 = \pi ab = \pi a^2 \sqrt{1 - \varepsilon^2}$. Together with $a = \frac{\varepsilon p}{1 - \varepsilon^2}$ and the above formula for p, you then compute like this:

$$\begin{aligned} T^2 &= \left(\frac{2mA_0}{\|J\|}\right)^2 = \frac{4m^2}{J^2} \cdot (\pi a^2 \sqrt{1 - \varepsilon^2})^2 = \frac{4m^2\pi^2}{J^2} a^3 \cdot a(1 - \varepsilon^2) = \frac{4m^2\pi^2}{J^2} a^3 \cdot \varepsilon \cdot p \\ &= \frac{4m^2\pi^2}{J^2} a^3 \cdot \|A\| \cdot \frac{J^2}{\gamma M m^2 \|A\|}, \end{aligned}$$

which brings you to the famous formula

$$T^2 = \frac{4\pi^2}{\gamma M} a^3.$$

KEPLER's **Third Law:** The squares of the periods of revolution are proportional to the cubes of the long axes.

3.6 Curve Integrals

We know integrals $\int_{t=a}^{t=b} f(t) \, \mathrm{d}t$ over an interval $[a, b]$, which can be construed as a straight line in $\mathbb{R}^n$.

In this section, we will replace this straight line by a curve in $\mathbb{R}^n$. There are (at least) two choices for the replacement of the integrand f and the differential $\mathrm{d}t$: scalar-valued or vector-valued.

Curve integrals of first kind: the integrand f and the differential $\mathrm{d}t$ are scalars. Think of f as a density of charges distributed along a curve, and $\mathrm{d}t = \mathrm{d}s$ is the differential of arc-length. Then the value of the integral is the total charge.

Curve integrals of second kind: the integrand f and the differential $\mathrm{d}t$ are n–vectors; and the product $f \, \mathrm{d}t$ is to be understood as scalar product. Think of f as being a force vector and $\mathrm{d}t$ as the unit tangent vector along the curve. Then the value of the integral can be seen as the amount of energy you have to spend in order to drag an object against the force along the curve.

From these interpretations, it should be clear that the direction of integration along the curve does not matter for integrals of the first kind, in contrast to integrals of the second kind.

The energy interpretation of curve integrals of the second kind predicts that the value of such an integral sometimes depends only on the location of the endpoints of the curve, not the curve itself. We say that the integral is *path-independent*[50] and will clarify the conditions for the path-independence below. Path-independent curve integrals are sometimes called *path integrals*[51].

[49] Brennpunkte

[50] wegunabhängig

[51] do not mix them up with the FEYNMAN path integrals; they are completely different objects

3.6.1 Curve Integrals of First Kind

Definition 3.71 (Curve integral of first kind). *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a continuous function and $\gamma\colon [a,b] \to \mathbb{R}^n$ be a regular C^1–curve. Then we define a* curve integral of first kind[52] *as follows:*

$$\int_\gamma f \,\mathrm{d}t := \int_{t=a}^{t=b} f(\gamma(t))\,\|\dot\gamma(t)\|\,\mathrm{d}t.$$

Proposition 3.72 (Re-parametrisation). *The value of a curve integral of first kind does not change under a C^1 re-parametrisation of the curve.*

Proof. Substitution rule. Be careful when the re-parametrisation changes the direction of the curve. □

Consequently, the value of the curve integral does not depend on the curve γ, but only on the *image* Γ of the curve. In the sequel, we will write $\int_\Gamma f\,\mathrm{d}t$ instead of $\int_\gamma f\,\mathrm{d}t$.

Curve integrals for curves that are only *piecewise*[53] C^1 can be defined in a natural way: take a curve γ which maps the interval $[a,b]$ into $\mathbb{R}^n$ continuously, and is C^1 for $t \notin \{t_1, t_2, \dots, t_m\}$ with $a = t_0 < t_1 < t_2 < \dots < t_m < t_{m+1} = b$. Then the curve integral $\int_\Gamma f\,\mathrm{d}t$ is defined as

$$\int_\Gamma f\,\mathrm{d}t := \sum_{j=0}^{m} \int_{t_j}^{t_{j+1}} f(\gamma(t))\,\|\dot\gamma(t)\|\,\mathrm{d}t.$$

If we have two images Γ and Δ of two regular piecewise C^1 curves, and if the union $\Gamma \cup \Delta$ can be also written as the image of a regular piecewise C^1 curve, then

$$\int_{\Gamma\cup\Delta} f\,\mathrm{d}t = \int_\Gamma f\,\mathrm{d}t + \int_\Delta f\,\mathrm{d}t.$$

It is easy to check that

$$\int_{-\Gamma} f\,\mathrm{d}t = \int_\Gamma f\,\mathrm{d}t,$$
$$\int_\Gamma (c_1 f_1 + c_2 f_2)\,\mathrm{d}t = c_1 \int_\Gamma f_1\,\mathrm{d}t + c_2 \int_\Gamma f_2\,\mathrm{d}t,$$
$$\left|\int_\Gamma f\,\mathrm{d}t\right| \le \|f\|_{L^\infty(\Gamma)}\,\mathrm{length}(\Gamma).$$

3.6.2 Curve Integrals of Second Kind

Definition 3.73 (Curve integral of second kind). *Let $f = f(x)\colon \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function,*

$$f = f(x) = \begin{pmatrix} f_1(x_1, x_2, \dots, x_n) \\ f_2(x_1, x_2, \dots, x_n) \\ \vdots \\ f_n(x_1, x_2, \dots, x_n) \end{pmatrix},$$

and $\gamma\colon [a,b] \to \mathbb{R}^n$ be a regular C^1 curve,

$$\gamma = \gamma(t) = x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}.$$

Then we define the curve integral of second kind[54] *as follows:*

$$\int_\gamma f \cdot \mathrm{d}x := \int_\gamma f_1\,\mathrm{d}x_1 + \dots + f_n\,\mathrm{d}x_n := \int_{t=a}^{t=b} \Big(f_1(\gamma(t))\dot\gamma_1(t) + \dots + f_n(\gamma(t))\dot\gamma_n(t) \Big)\,\mathrm{d}t.$$

[52] Kurvenintegral erster Art

[53] stückweise

[54] Kurvenintegral zweiter Art

Proposition 3.74 (**Re-parametrisation**). *The value of a curve integral of second kind does not change under a C^1 re-parametrisation which preserves the orientation of the curve.*

Re-parametrisations that change the orientation of the curve only change the sign of the value of the curve integral.

Proof. Substitution rule. □

You can define curve integrals over *piecewise* C^1 regular curves in the same way as you did for curve integrals of first kind. Then it is easy to check that

$$\begin{aligned}
&\int_{\Gamma\cup\Delta} f\cdot \mathrm{d}x = \int_\Gamma f\cdot \mathrm{d}x + \int_\Delta f\cdot \mathrm{d}x,\\
&\int_{-\Gamma} f\cdot \mathrm{d}x = -\int_\Gamma f\cdot \mathrm{d}x,\\
&\int_\Gamma (\kappa f + \lambda g)\cdot \mathrm{d}x = \kappa\int_\Gamma f\cdot \mathrm{d}x + \lambda\int_\Gamma g\cdot \mathrm{d}x,\\
&\left|\int_\Gamma f\cdot \mathrm{d}x\right| \le \|f\|_{L^\infty(\Gamma)}\,\mathrm{length}(\Gamma),
\end{aligned}$$

where $\|f\|_{L^\infty(\Gamma)} = \sup_{x\in\Gamma}\sqrt{f_1^2(x)+\cdots+f_n^2(x)}$.

From the energy interpretation of such curve integrals, we know that sometimes the value of the curve integral only depends on the endpoints, not on the curve between them.

Definition 3.75 (**Conservative vector field**). *A vector field f is called* conservative *or* exact[55] *if every curve integral over a loop vanishes.*

An alternative description: if γ and δ are any two regular piecewise C^1 curves with coinciding endpoints A and B, then $\int_\Gamma f\cdot \mathrm{d}x = \int_\Delta f\cdot \mathrm{d}x$. We will often write $\int_A^B f\cdot \mathrm{d}x$.

Example 3.76. *Compute the integral $\int_\Gamma f\cdot \mathrm{d}x$ with Γ being the unit circle in $\mathbb{R}^2$ and*

$$f = f(x) = \frac{1}{x_1^2+x_2^2}\begin{pmatrix}-x_2\\ x_1\end{pmatrix}, \qquad x = (x_1,x_2)^\top \neq (0,0)^\top.$$

Proposition 3.77. *A vector field $f\colon \Omega\to\mathbb{R}^n$ is conservative if and only if it is a* gradient field[56]*, i.e., if there is a scalar function $\varphi\colon\Omega\to\mathbb{R}$ with $f = \operatorname{grad}\varphi$. In this case, we have*

$$\int_A^B f\cdot \mathrm{d}x = \varphi(B)-\varphi(A), \qquad A,B\in\Omega.$$

It is custom to call φ (or $-\varphi$) a *potential*[57] to the vector field f.

Proof. First, let f be a gradient field, and let γ be a regular C^1 curve connecting A and $B\in\Omega$. Then

$$\int_\gamma f\cdot \mathrm{d}x = \int_{t=a}^{t=b} f(\gamma(t))\cdot\dot\gamma(t)\,\mathrm{d}t = \int_{t=a}^{t=b}\frac{\mathrm{d}}{\mathrm{d}t}\varphi(\gamma(t))\,\mathrm{d}t = \varphi(\gamma(b))-\varphi(\gamma(a)) = \varphi(B)-\varphi(A),$$

which depends only on the points B and A, not on the curve connecting them.

Second, let the vector field f be conservative. Pick a point $x_0\in\Omega$, and define a function $\varphi\colon\Omega\to\mathbb{R}$,

$$\varphi(x_*) := \int_{x_0}^{x_*} f\cdot \mathrm{d}x,$$

where, by assumption, the image Γ of the curve γ connecting x_0 and x_* does not matter as long as it stays in Ω.

[55] konservativ bzw. exakt
[56] Gradientenfeld
[57] Potential

We would like to know $\frac{\partial}{\partial x_j}\varphi(x)$. Since we have the freedom in choosing the curve, we can fix Γ in such a way that it arrives at the point x on a straight line parallel to the e_j–axis. In the case of $n = 2$, think of curves like ⌜ or ⌟, with x_0 at the lower left corner, and x_* at the upper right corner. Then the "terminal part" of the curve integral can be seen as an integral over an interval on the real axis, and all that remains is to exploit the fundamental theorem of calculus. □

Proposition 3.78 (**Integrability conditions**). *A continuously differentiable conservative vector field* $f\colon \Omega \to \mathbb{R}^n$ *must satisfy the* integrability conditions[58]

$$\left(\frac{\partial f_j}{\partial x_k}\right)(x) = \left(\frac{\partial f_k}{\partial x_j}\right)(x), \qquad 1 \le j, k \le n, \qquad x \in \Omega.$$

Proof. This follows from the Schwarz theorem in Proposition 1.19, applied to the potential φ. □

Warning: *The converse is not true, as Example 3.76 demonstrates.*

We need a special condition on the domain Ω.

Definition 3.79 (**Simply connected**). *We say that an open set* $\Omega \subset \mathbb{R}^n$ *is* simply connected[59] *if every loop entirely contained in* Ω *can be shrunk to a point, always staying in* Ω*.*

An open ball, an open triangle, an open rectangle in $\mathbb{R}^2$ are simply connected; an open ball in $\mathbb{R}^2$ without centre is not. A hollow cylinder in $\mathbb{R}^3$ is not simply connected, but a hollow ball in $\mathbb{R}^3$ is.

Proposition 3.80. *A* C^1 *vector field that satisfies the integrability conditions in a simply connected domain is conservative there.*

It suffices to prove the following: if you have a curve between two endpoints and "wiggle it a bit", keeping the endpoints fixed, then the value of the curve integral does not change. It even suffices to "wiggle" only a short part of the curve. However, such a short part of the curve is always contained in a certain ball that is a subset of Ω, provided that the part of the curve is chosen short enough.

Consequently, it suffices to prove the following result:

Lemma 3.81. *A* C^1 *vector field that satisfies the integrability conditions in a ball is conservative there.*

Proof. Let the ball B be centered at the origin. For each $x \in B$, we choose a curve γ_x connecting the origin with the point x, namely the straight line, which is entirely contained in B:

$$\gamma_x = \gamma_x(t) = tx, \qquad 0 \le t \le 1.$$

We have $\dot\gamma(t) = x$. Then we define a scalar function $\varphi\colon B \to \mathbb{R}$ by

$$\varphi(x) = \int_{\gamma_x} f \cdot \mathrm{d}x = \int_{t=0}^{t=1} f(\gamma(t)) \cdot \dot\gamma(t)\,\mathrm{d}t = \int_{t=0}^{t=1} \sum_{j=1}^n f_j(tx) x_j \,\mathrm{d}t = \sum_{j=1}^n x_j \int_{t=0}^{t=1} f_j(tx)\,\mathrm{d}t.$$

We want to show that $\nabla\varphi = f$. Since the partial derivatives of the f_j are uniformly continuous on B, Proposition 3.36 allows us to differentiate under the integral sign:

$$\begin{aligned}
\frac{\partial\varphi}{\partial x_k}(x) &= \int_{t=0}^{t=1} f_k(tx)\,\mathrm{d}t + \sum_{j=1}^n x_j \int_{t=0}^{t=1} \left(\frac{\partial f_j}{\partial x_k}\right)(tx) \cdot t\,\mathrm{d}t \\
&= t \cdot f_k(tx)\Big|_{t=0}^{t=1} - \int_{t=0}^{t=1} t \cdot \frac{\partial f_k(tx)}{\partial t}\,\mathrm{d}t + \sum_{j=1}^n x_j \int_{t=0}^{t=1} t \cdot \left(\frac{\partial f_j}{\partial x_k}\right)(tx)\,\mathrm{d}t \\
&= f_k(x) - \int_{t=0}^{t=1} t \cdot \sum_{j=1}^n \left(\frac{\partial f_k}{\partial x_j}\right)(tx) \cdot x_j\,\mathrm{d}t + \sum_{j=1}^n x_j \int_{t=0}^{t=1} t \cdot \left(\frac{\partial f_j}{\partial x_k}\right)(tx)\,\mathrm{d}t = f_k(x),
\end{aligned}$$

because of the integrability conditions. □

[58] Integrabilitätsbedingungen

[59] einfach zusammenhängend

Corollary 3.82. *A function $f\colon \Omega \to \mathbb{R}^3$, where $\Omega \subset \mathbb{R}^3$ is simply connected, is conservative if and only if* $\operatorname{rot} f = \vec{0}$.

Question: Take f as in Example 3.76 and Ω either the upper half-plane $\{(x_1, x_2)\colon x_2 > 0\}$ or the lower half-plane $\{(x_1, x_2)\colon x_2 < 0\}$, and construct, in either case, a potential function φ. If you choose the two potential functions in such a way that $\varphi_+(1,0) = 0$ and $\varphi_-(1,0) = 0$, what happens on the left half-axis $\{(x_1, 0)\colon x_1 < 0\}$?

3.6.3 Complex Curve Integrals

Definition 3.83 (**Complex curve integral**)**.** *Let $\gamma\colon [a,b] \to \mathbb{C}$ be a regular C^1 curve and $w\colon \Omega \to \mathbb{C}$ be a continuous function, where $\Omega \subset \mathbb{C}$ is an open set. Then a curve integral $\int_\gamma w \,\mathrm{d}z$ is defined as*

$$\int_\gamma w \,\mathrm{d}z = \int_{t=a}^{t=b} w(\gamma(t))\dot{\gamma}(t) \,\mathrm{d}t.$$

If γ is a loop, we also write $\oint_\gamma w \,\mathrm{d}z := \int_\gamma w \,\mathrm{d}z$.

We split w and $\gamma = z$ into real part and imaginary part:

$$w = u + \mathrm{i}v, \qquad z = x + \mathrm{i}y.$$

Then the curve integral can be interpreted as linear combination of two real curve integrals of second kind:

$$\begin{aligned}\int_\gamma w \,\mathrm{d}z &= \int_{t=a}^{t=b} (u(z(t)) + \mathrm{i}v(z(t))) \cdot (\dot{x}(t) + \mathrm{i}\dot{y}(t)) \,\mathrm{d}t \\ &= \int_{t=a}^{t=b} \Big(u(z(t))\dot{x}(t) - v(z(t))\dot{y}(t)\Big) \,\mathrm{d}t + \mathrm{i} \int_{t=a}^{t=b} \Big(v(z(t))\dot{x}(t) + u(z(t))\dot{y}(t)\Big) \,\mathrm{d}t \\ &= \int_\Gamma u(x,y) \,\mathrm{d}x - v(x,y) \,\mathrm{d}y + \mathrm{i} \int_\Gamma v(x,y) \,\mathrm{d}x + u(x,y) \,\mathrm{d}y.\end{aligned}$$

This representation tells us that an orientation-preserving re-parametrisation does not change the value of the complex curve integral. An orientation-changing re-parametrisation changes only the sign of the value of the complex curve integral.

The integrability conditions for the complex curve integral are
the Cauchy–Riemann differential equations.

Then Proposition 3.80 gives us the following result of eminent importance.

Proposition 3.84 (Cauchy**'s Integral Theorem**)**.** *Let the function $w\colon \Omega \to \mathbb{C}$ be holomorphic (complex differentiable) in Ω with continuous derivative w'. Assume that the domain Ω is simply connected. Then the integral*

$$\oint_\Gamma w \,\mathrm{d}z$$

vanishes for every regular C^1 loop Γ in Ω.

The supposition of the continuity of w' can be dropped, as a (completely) different proof would show.

Convention: we always choose the direction of integration along the loop Γ in such a way that the bounded domain “inside” Γ is lying “to the left”.

Remark 3.85. *The path-independence of curve integrals is equivalent to the vanishing of all loop integrals. Then it makes sense to speak about* definite *integrals*

$$\int_{z_0}^{z} w(\zeta) \,\mathrm{d}\zeta,$$

where w is a holomorphic function and the points z_0 and z are connected by an arbitrary curve in Ω. Only the endpoints z_0 and z matter for the value of the integral. In the third semester, we will learn that such an integral is an antiderivative for the function w.

Corollary 3.86 (of CAUCHY's integral theorem)**.** *Let the function $w\colon \Omega \to \mathbb{C}$ be holomorphic in Ω, and assume that the domain Ω is simply connected, and Γ is a loop in Ω. Pick a point z_0 inside this loop, and a small positive ε, such that the ball $B_\varepsilon(z_0)$ about z_0 with radius ε does not intersect Γ. Then we have*

$$\oint_\Gamma \frac{w(z)}{z-z_0}\,\mathrm{d}z = \oint_{B_\varepsilon(z_0)} \frac{w(z)}{z-z_0}\,\mathrm{d}z.$$

Proof. Connect Γ and the ball $B_\varepsilon(z_0)$ by a straight line and apply Cauchy's integral theorem to the domain "between" Γ and $B_\varepsilon(z_0)$. □

Now we choose ε extremely small and use the differentiability of w:

$$|z-z_0| = \varepsilon \quad \Longrightarrow \quad w(z) = w(z_0) + \mathfrak{O}(\varepsilon).$$

It follows that

$$\left|\oint_{B_\varepsilon(z_0)} \frac{\mathfrak{O}(\varepsilon)}{z-z_0}\,\mathrm{d}z\right| \le |\mathfrak{O}(\varepsilon)| \cdot \frac{1}{\varepsilon} \cdot \operatorname{length}(B_\varepsilon(z_0)) = \mathfrak{O}(\varepsilon),$$

$$\oint_\Gamma \frac{w(z)}{z-z_0}\,\mathrm{d}z = \oint_{B_\varepsilon(z_0)} \frac{w(z_0)}{z-z_0}\,\mathrm{d}z + \mathfrak{O}(\varepsilon) = w(z_0)\oint_{B_\varepsilon(z_0)} \frac{1}{z-z_0}\,\mathrm{d}z + \mathfrak{O}(\varepsilon).$$

In the last loop integral, we parametrise $z = z_0 + \varepsilon\exp(\mathrm{i}t)$ with $0 \le t \le 2\pi$ and $\mathrm{d}z = \varepsilon\mathrm{i}\exp(\mathrm{i}t)\,\mathrm{d}t$:

$$\oint_{B_\varepsilon(z_0)} \frac{1}{z-z_0}\,\mathrm{d}z = \int_{t=0}^{t=2\pi} \frac{1}{\varepsilon\exp(\mathrm{i}t)}\varepsilon\mathrm{i}\exp(\mathrm{i}t)\,\mathrm{d}t = 2\pi\mathrm{i}.$$

We send ε to zero, and the result is amazing:

Proposition 3.87 (CAUCHY's **Integral Formula**)**.** *Let the function $w\colon \Omega \to \mathbb{C}$ be holomorphic in Ω, and assume that the domain Ω is simply connected. Then we have the following formula for any loop Γ in Ω and* any *point z_0 inside the loop Γ:*

$$w(z_0) = \frac{1}{2\pi\mathrm{i}}\oint_\Gamma \frac{w(z)}{z-z_0}\,\mathrm{d}z.$$

The values of a holomorphic function on a loop
determine the values of that function inside the loop.

A variant of Proposition 3.36 allows us to differentiate under the integral sign as often as we want:

$$\partial_{z_0}^k w(z_0) = \frac{k!}{2\pi\mathrm{i}}\oint_\Gamma \frac{w(z)}{(z-z_0)^{k+1}}\,\mathrm{d}z,$$

which tells us that $w \in C^\infty(\Omega)$. We can do even better: if $|z-z_0| > R$ for all points $z \in \Gamma$, then

$$|\partial_{z_0}^k w(z_0)| \le \frac{k!}{2\pi}\,\|w\|_{L^\infty(\Gamma)}\,\frac{1}{R^{k+1}}\operatorname{length}(\Gamma). \tag{3.4}$$

Then the usual remainder term estimates show us that the Taylor series of the function $w(z)$ converges for z near z_0; and its limit is $w(z)$. The function w is better than just C^∞; it even has a converging Taylor series.

We go back to (3.4) once more, with $k = 1$:

$$|\partial_{z_0} w(z_0)| \le \frac{1}{2\pi}\,\|w\|_{L^\infty(\Gamma)}\,\frac{1}{R^2}\operatorname{length}(\Gamma).$$

What happens if w is holomorphic in all of $\mathbb{C}$, and, furthermore, if it is bounded ? In this case, you can choose a huge circular loop $\Gamma = B_R(z_0)$, giving you $\operatorname{length}(\Gamma) = 2\pi R$ and

$$|w'(z_0)| \le \frac{\|w\|_{L^\infty(\mathbb{C})}}{R}.$$

However, R can be any big number, hence $w'(z_0) = 0$. Since z_0 can be chosen arbitrarily in $\mathbb{C}$, the function w must be constant, because $w'(z) = 0$ everywhere in $\mathbb{C}$.

Proposition 3.88 (LIOUVILLE's[60] Theorem)**.** *If a function is holomorphic on all of $\mathbb{C}$ and bounded, then it is constant.*

We have one more famous result:

Proposition 3.89 (**Fundamental Theorem of Algebra**)**.** *Every polynomial of degree at least one has a zero in $\mathbb{C}$.*

Proof. Write the polynomial as $P = P(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$. We may assume that $a_n = 1$, divide otherwise. Assume that P has no zero in $\mathbb{C}$. It is clear that $|P(z)| \geq 1$ for large values of $|z|$, because the highest power z^n is stronger than all other powers for large $|z|$. Say that $|P(z)| \geq 1$ for $|z| \geq R$. We define

$$Q(z) = \frac{1}{P(z)}.$$

The function Q is holomorphic on $\mathbb{C}$, since we never divide by zero. The function Q is bounded for $|z| \geq R$, namely by 1. The function $|Q|$ is also bounded for $|z| \leq R$, since $|Q|$ is continuous there and $B_0(R)$ is compact. Consequently, Q is bounded on $\mathbb{C}$ and is holomorphic on $\mathbb{C}$. By the Liouville Theorem, Q must be a constant. However, this is not possible for $n \geq 1$, giving the desired contradiction. □

Try to prove the fundamental theorem of algebra without holomorphic functions !

3.7 Keywords

- norms of functions,
- fundamental theorem of calculus,
- antiderivatives of standard functions,
- rules for finding antiderivatives,
- numerical methods,
- improper integrals,
- the three results on commutation of limit processes,
- formulae for the Fourier coefficients,
- L^2 theory of Fourier series,
- definition of Dirac sequences,
- definition of the three kinds of convergence of Fourier sequences,
- curves and images,
- tangent vectors and length,
- three kinds of curve integrals,
- conservative vector fields and integrability conditions.

[60] JOSEPH LIOUVILLE, 1809 – 1882

Chapter 4

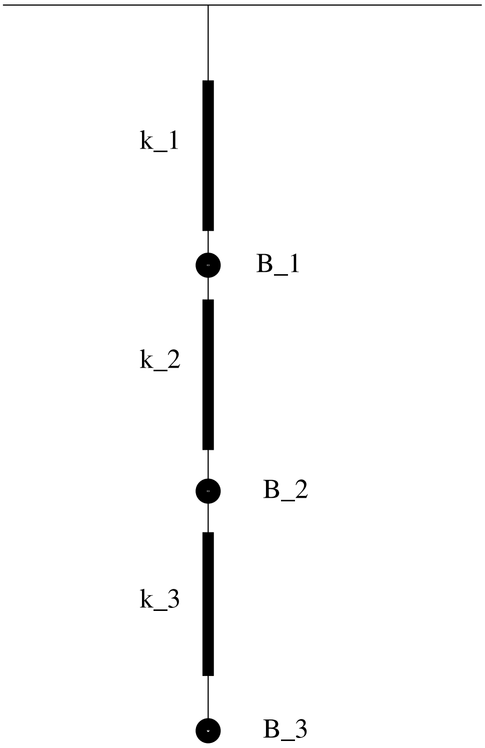
Eigenvalues and Eigenvectors

Literature: Greiner: *Quantenmechanik. Einführung.* Chapter IV.20: Eigenwerte und Eigenfunktionen

4.1 Introduction

Literature: Greiner: *Klassische Mechanik II.* Chapter III.7: Schwingungen gekoppelter Massepunkte

Consider three balls B_1, B_2, B_3 with equal mass 1, connected by springs with spring constants k_1, k_2, and k_3. At time $t = 0$, the system is in rest position, then we touch it and let it swing (only vertically). Assuming that there is no damping, how does the system evolve ?



We denote the deviation of the ball B_i from the rest position by $y_i(t)$ and get the following system of linear ordinary differential equations:

$$\begin{aligned}
-y_1'' &= k_1 y_1 - k_2(y_2 - y_1),\\
-y_2'' &= k_2(y_2 - y_1) - k_3(y_3 - y_2),\\
-y_3'' &= k_3(y_3 - y_2).
\end{aligned}$$

Experience says that oscillations should be modelled with sine functions:

$$y_1(t) = \xi_1 \sin \omega t, \quad y_2(t) = \xi_2 \sin \omega t, \quad y_3(t) = \xi_3 \sin \omega t.$$

Going with this ansatz into the ODE (ordinary differential equation) system, we get

$$\begin{aligned} \omega^2 \xi_1 \sin \omega t &= ((k_1 + k_2)\xi_1 - k_2 \xi_2) \sin \omega t, \\ \omega^2 \xi_2 \sin \omega t &= (-k_2 \xi_1 + (k_2 + k_3)\xi_2 - k_3 \xi_3) \sin \omega t, \\ \omega^2 \xi_3 \sin \omega t &= (-k_3 \xi_2 + k_3 \xi_3) \sin \omega t, \end{aligned}$$

which can be simplified to

$$\begin{pmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \omega^2 \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}.$$

We will abbreviate this to $Ax = \lambda x$ with A being this 3×3 matrix, $x = (\xi_1, \xi_2, \xi_3)^\top$, and $\lambda = \omega^2$. The matrix A maps the vector x to a multiple of itself.

We know A and want to find all λ and x with $Ax = \lambda x$. Since mass points in idle position are not interesting, the vector x should not be the null vector.

Definition 4.1 (Eigenvalue, eigenvector). *Let U be a linear space over the field K, and f be an endomorphism in U. We say that $\lambda \in K$ is an* eigenvalue[1] *for f if a vector $x \in U$, $x \neq 0$, exists with*

$$f(x) = \lambda x.$$

The vector x is called eigenvector[2] *to the eigenvalue λ.*

Recall that an endomorphism is a linear map from a vector space into itself.

In quantum mechanics or elasticity theory, U might be a space of functions, and f a differential operator. The eigenvalue λ is then related to a frequency, or the energy, or similar quantities connected to the state of the system.

In the rest of this chapter, we restrict ourselves to $U = \mathbb{R}^n$ or $U = \mathbb{C}^n$, and $f = f_A$ is the endomorphism from U to U associated to a matrix $A \in K^{n \times n}$. Then we will talk about eigenvalues and eigenvectors of this matrix A, instead of the mapping f_A.

4.2 Basic Properties

Proposition 4.2. *Let $A \in K^{n \times n}$. A number $\lambda \in K$ is an eigenvalue to the matrix A if and only if* $\det(A - \lambda I) = 0$.

Any vector $x \in \ker(A - \lambda I)$ is an eigenvector to the eigenvalue λ.

The linear space $\ker(A - \lambda I)$ is called *eigenspace*[3].

Proof. The system $Ax = \lambda x$ is equivalent to $(A - \lambda I)x = 0$, which has a solution $x \neq 0$ if and only if $\det(A - \lambda I) = 0$. Moreover, each vector from $\ker(A - \lambda I)$ is, by definition, a solution to $(A - \lambda I)x = 0$. □

Definition 4.3 (Characteristic polynomial). *The expression $\chi_A(\lambda) = \det(A - \lambda I)$ is the* characteristic polynomial[4] *of the matrix A.*

Proposition 4.4. *The characteristic polynomial of $A \in K^{n \times n}$ is a polynomial of degree n,*

$$\chi_A(\lambda) = a_n \lambda^n + a_{n-1} \lambda^{n-1} + \cdots + a_1 \lambda + a_0,$$

with $a_n = (-1)^n$, $a_{n-1} = (-1)^{n-1} \operatorname{trace}(A) = (-1)^{n-1} \sum_{j=1}^n a_{jj}$, and $a_0 = \det A$.

[1] Eigenwert
[2] Eigenvektor
[3] Eigenraum
[4] charakteristisches Polynom

Proof. You can compute a_n and a_{n-1} from the Leibniz formula in Proposition 2.25. The absolute term a_0 can be computed easily by setting $\lambda = 0$. □

Proposition 4.5. *Let U be a linear space over $\mathbb{C}$ with finite dimension, and f be a linear mapping of U into itself. Then the mapping f has at least one eigenvector.*

Proof. Given a basis for U, there is a matrix A connected to f via $Ax = f(x)$ for all $x \in U$. The characteristic polynomial to this matrix has at least one zero in $\mathbb{C}$, by the fundamental theorem of algebra (Proposition 3.89). This zero is an eigenvalue of f. □

Question: What happens if you replace $\mathbb{C}$ by $\mathbb{R}$ in Proposition 4.5 ? Consider $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$.

Proposition 4.6. *Similar matrices have the same characteristic polynomial.*

Proof. Recall that two matrices A and $\tilde{A}$ are similar if there is an invertible matrix B with $\tilde{A} = B^{-1}AB$. The assertion follows from the Proposition 2.9, 2.10, and

$$\begin{aligned} \det(\tilde{A} - \lambda I) &= \det(B^{-1}AB - B^{-1}\lambda IB) = \det(B^{-1}(A - \lambda I)B) \\ &= \det(B^{-1}) \det(A - \lambda I) \det(B) = \det(A - \lambda I). \end{aligned}$$

□

Lemma 4.7. *If $\lambda_1, \dots, \lambda_m$ are distinct eigenvalues of a matrix A, and $u_1, \dots, u_m$ are associated eigenvectors, then these vectors are linearly independent.*

Proof. Assume the opposite; the vectors $u_1, \dots, u_m$ are linearly dependent. We can renumber the vectors in such a way that u_m can be written as a linear combination of $u_1, \dots, u_{m-1}$. Moreover, we can assume that $u_1, \dots, u_{m-1}$ are linearly independent. Otherwise, we throw u_m away and continue with the remaining vectors. Therefore, we have numbers $\alpha_1, \dots, \alpha_{m-1}$ (not all of them zero) with

$$u_m = \alpha_1 u_1 + \dots + \alpha_{m-1} u_{m-1}.$$

We apply A:

$$\lambda_m u_m = \alpha_1 \lambda_1 u_1 + \dots + \alpha_{m-1} \lambda_{m-1} u_{m-1}.$$

From these two representations we then have

$$0 = \alpha_1(\lambda_1 - \lambda_m)u_1 + \dots + \alpha_{m-1}(\lambda_{m-1} - \lambda_m)u_{m-1}.$$

Not all of the coefficients can vanish, because the λ_j are mutually distinct. Consequently, the vectors $u_1, \dots, u_{m-1}$ must be linearly dependent, which is a contradiction. □

Corollary 4.8. *If a matrix $A \in K^{n \times n}$ has n mutually distinct eigenvalues, then the associated eigenvectors form a basis of K^n.*

Write these n eigenvectors column-wise next to each other and call that matrix S. Then we have

$$AS = SD,$$

where $D = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ is a diagonal matrix containing the eigenvalues (in the same order as the eigenvectors appear in S).

Definition 4.9 (Diagonalisable matrix). *We say that a matrix A is* diagonalisable[5] *if there is an invertible matrix S, such that the matrix $D = S^{-1}AS$ is diagonal.*

[5] diagonalisierbar

If a matrix A is diagonalisable, then the matrix D contains n eigenvalues of A, and S is a matrix of eigenvectors. This follows from Proposition 4.6.

What is the purpose of diagonalising a matrix ?

In practical applications, a matrix often describes a mapping in $\mathbb{R}^n$, e.g., a rotation or a reflection. Remember that the columns of the matrix are the components of the images of the basis vectors. If you choose another basis of the space $\mathbb{R}^n$, the matrix A has to be changed accordingly. Later computations will become easier if A has an easy structure. And the easiest structure you can get is that of a diagonal matrix. For this, the basis vectors will be the eigenvectors of the matrix A, *provided that A has enough (n) eigenvectors.*

An example should clarify the matter:

$$A = \begin{pmatrix} -1 & 2 & 2 \\ 2 & 2 & 2 \\ -3 & -6 & -6 \end{pmatrix}.$$

The characteristic polynomial is

$$\chi_A(\lambda) = \det(A - \lambda I) = \det \begin{pmatrix} -1-\lambda & 2 & 2 \\ 2 & 2-\lambda & 2 \\ -3 & -6 & -6-\lambda \end{pmatrix}.$$

After some calculation, you will find that

$$\chi_A(\lambda) = -\lambda(\lambda^2 + 5\lambda + 6) = -\lambda(\lambda+2)(\lambda+3)$$

with zeros $\lambda_1 = 0$, $\lambda_2 = -2$, $\lambda_3 = -3$. This is a good moment to perform a quick check:

The sum of the eigenvalues equals the trace of the matrix;
the product of the eigenvalues equals the determinant of the matrix.

Question: Why is that so ?

The eigenvectors are solutions to the systems $(A - \lambda_j I)x = 0$:

$$\lambda_1 = 0\colon \quad \begin{pmatrix} -1 & 2 & 2 \\ 2 & 2 & 2 \\ -3 & -6 & -6 \end{pmatrix} x = 0, \quad \text{with a solution} \quad x = u_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix},$$

$$\lambda_2 = -2\colon \quad \begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 2 \\ -3 & -6 & -4 \end{pmatrix} x = 0, \quad \text{with a solution} \quad x = u_2 = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix},$$

$$\lambda_3 = -3\colon \quad \begin{pmatrix} 2 & 2 & 2 \\ 2 & 5 & 2 \\ -3 & -6 & -3 \end{pmatrix} x = 0, \quad \text{with a solution} \quad x = u_3 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Write the eigenvectors as columns in a matrix:

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

Since the eigenvectors form a basis, the matrix S is invertible. Then we have

$$S^{-1}AS = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = D.$$

You can check this if you compute S^{-1}, and then multiply the three matrices S^{-1}, A, and S by hand.

4.3 The Jordan Normal Form

Things are not always that nice as presented in the last section. It can happen that a matrix $A \in K^{n\times n}$ does not have n eigenvectors, but less. Then you do not have enough vectors for a basis.

This phenomenon can only occur if some eigenvalues are multiple.

Definition 4.10 (**Algebraic, geometric multiplicity**). *The* algebraic multiplicity[6] *of an eigenvalue λ of a matrix is defined as the multiplicity of the zero λ of the characteristic polynomial.*

The geometric multiplicity[7] *of an eigenvalue λ of a matrix A is defined as the dimension of* $\ker(A-\lambda I)$.

Example 4.11. *Take $A = \left(\begin{smallmatrix} 3 & 1 \\ 0 & 3 \end{smallmatrix}\right)$. This matrix has eigenvalues $\lambda_1 = \lambda_2 = 3$, but every eigenvector is a multiple of $u_1 = (1,0)^\top$. The algebraic multiplicity is two, and the geometric multiplicity is one.*

Proposition 4.12. *The geometric multiplicity is less than or equal to the algebraic multiplicity.*

If all eigenvalues of a matrix $A \in K^{n\times n}$ have equal algebraic and geometric multiplicity, then you can find n linearly independent eigenvectors of A. Selecting these eigenvectors as a new basis for K^n will diagonalise A.

You can no longer diagonalise if there is an eigenvalue whose algebraic multiplicity is greater than the geometric multiplicity. Instead of a diagonalisation, we resort to an *almost diagonalisation* of the matrix, using a family of eigenvectors and *principal vectors*. The result will be the *Jordan normal form*.

Definition 4.13 (**Principal vectors**). *A family $(u_1, \dots, u_m)$ of vectors (none of them being the null vector) is said to be a* chain of principal vectors[8] *to the eigenvalue λ of a matrix A if u_1 is an eigenvector to the eigenvalue λ, and*

$$(A-\lambda I)u_i = u_{i-1}, \qquad 2 \le i \le m.$$

The vector u_j is called jth level principal vector[9].

Question: Determine eigenvectors and principal vectors for the following *Jordan block*:

$$A = \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \lambda & 1 & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix}. \tag{4.1}$$

Lemma 4.14. *The principal vector u_i to the eigenvalue λ of A belongs to $\ker(A-\lambda I)^i$, but not to $\ker(A-\lambda I)^{i-1}$. Moreover, a chain of principal vectors is linearly independent.*

Proof. The first claim is easy to check, after having computed $(A-\lambda I)^k u_i$ for $i \le k$ and $i > k$. The second claim then follows from the first. □

The principal vectors are exactly that vectors, which will complete the family of eigenvectors to a basis of the whole space.

Examples: *If an eigenvalue has algebraic multiplicity six and geometric multiplicity one, then there is one eigenvector and, starting from this eigenvector, a chain of five principal vectors.*

If an eigenvalue has algebraic multiplicity six and geometric multiplicity two, then there are two eigenvectors u_1, u_2, and two chains of principal vectors. These two chains start from two eigenvectors (which may not be u_1 and u_2) and have altogether four principal vectors.

Proposition 4.15 (JORDAN **normal form**). *For any matrix $A \in \mathbb{C}^{n\times n}$, you can find a regular $G \in \mathbb{C}^{n\times n}$ whose columns are eigenvectors and principal vectors of A, such that $G^{-1}AG$ is a block-diagonal matrix* $\operatorname{diag}(J_1, \dots, J_k)$. *The J_i are Jordan blocks as in (4.1). Different Jordan blocks can belong to the same eigenvalue of A. These blocks are uniquely determined, only their arrangement can change.*

[6] algebraische Vielfachheit

[7] geometrische Vielfachheit

[8] Kette von Hauptvektoren

[9] Hauptvektor j-ter Stufe

The proof is much longer than we have the time for. We only remark that you can not replace $\mathbb{C}$ by $\mathbb{R}$, and conclude with two examples.

Example: *The characteristic polynomial of the matrix*

$$A = \begin{pmatrix} 17 & 0 & -25 \\ 0 & 2 & 0 \\ 9 & 0 & -13 \end{pmatrix}$$

is $\chi_A(\lambda) = (2-\lambda)^3$, hence A has an eigenvalue $\lambda = 2$ with algebraic multiplicity three. The system $(A-\lambda I)x = 0$ has only two linearly independent solutions, namely $u_1 = (0,1,0)^\top$ and $u_2 = (5,0,3)^\top$. The geometric multiplicity is, therefore, two. We need one principal vector. The first try is to solve $(A-\lambda I)x = u_1$, which is unfortunately unsolvable. The second try is $(A-\lambda I)x = u_2$ which has many solutions, for instance $u_3 = (\frac{1}{3},0,0)^\top$. Put $G = (u_1,u_2,u_3)$, which is the new basis of the $\mathbb{C}^3$. Then the Jordan normal form of A is

$$G^{-1}AG = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Example: *Consider the matrix*

$$A = \begin{pmatrix} -4 & -9 & 12 & 1 & -967 & -671 \\ 4 & 8 & -1 & 4 & 392 & 272 \\ 0 & 0 & 8 & 4 & -194 & -135 \\ 0 & 0 & -9 & -4 & 305 & 212 \\ 0 & 0 & 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & 0 & 9 & 8 \end{pmatrix}.$$

The characteristic polynomial is, after some calculation, $\chi_A(\lambda) = (2-\lambda)^6$, giving you an eigenvalue $\lambda = 2$ of algebraic multiplicity six. Eigenvectors are (for instance)

$$u_1 = (-7,0,-4,6,0,0)^\top, \qquad u_2 = (-3,2,0,0,0,0)^\top.$$

Of course, linear combinations of u_1 and u_2 will give more eigenvectors. Consequently, we expect two chains of principal vectors, both together containing 4 principal vectors. After a much longer calculation, you will find a matrix G of eigenvectors and principal vectors:

$$G = \begin{pmatrix} -9 & 0 & 16 & 12 & -671 & 0 \\ 6 & 1 & -6 & -5 & 272 & 0 \\ 0 & 0 & 4 & 4 & -135 & 0 \\ 0 & 0 & -6 & -5 & 212 & 0 \\ 0 & 0 & 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 & 6 & 1 \end{pmatrix}$$

Then the Jordan normal form of A is:

$$G^{-1}AG = J = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

We see a chain with one principal vector (plus one eigenvector) and a chain with three principal vectors (plus one eigenvector). The eigenvectors of A can be found in the first and third columns of G. The first column of G contains the eigenvector $3u_2$, from which the chain with one principal vector originates. And the third column of G contains the eigenvector $-u_1 - 3u_2$, which is the anchor point of the second chain. These anchor points can not be chosen arbitrarily. The full algorithm is quite complicated, and we recommend to let a computer algebra system do the work, like the opensource system `maxima`. The relevant commands are `matrix`, `eigenvalues`, `eigenvectors`, `jordan`, `dispJordan`, `ModeMatrix`.

4.4 Normal Matrices and Projections

Literature: Greiner: *Quantenmechanik. Einführung.* Chapter XVII: Das formale Schema der Quantenmechanik

Our goal is to study eigenvalues and eigenvectors to certain important matrices. These matrices are the *self-adjoint* matrices and the *unitary* matrices, which are special cases of *normal* matrices.

Recall the scalar product in $\mathbb{C}^n$ and the definition of the adjoint matrix A^*:

$$\langle x, y \rangle = \sum_{j=1}^{n} \xi_j \overline{\eta_j},$$
$$A = (a_{jk}) \in \mathbb{C}^{n \times m} \quad \Longrightarrow \quad A^* = (\overline{a_{kj}}) \in \mathbb{C}^{m \times n}.$$

All vectors are column vectors, as always.

Warning: *The notation in many physics books differs from ours: there you will find the convention* $\langle x, y \rangle = \sum_{j=1}^{n} \overline{\xi_j} \eta_j$*, and then several formulas below must be changed. Moreover,* A^* *in mathematics corresponds to* $A^\dagger$ *in physics.*

Proposition 4.16 (Adjoint matrices). *As a matrix product, the scalar product is* $\langle x, y \rangle = y^* x$*. For all compatible vectors* x, y *and matrices* A*, we have*

$$\langle Ax, y \rangle = \langle x, A^* y \rangle .$$

Fix $A \in \mathbb{C}^{n \times m}$*. If a matrix* B *satisfies*

$$\langle Ax, y \rangle_{\mathbb{C}^n} = \langle x, By \rangle_{\mathbb{C}^m} \tag{4.2}$$

for all compatible x *and* y*, then* B *must be the adjoint to* A*.*

Proof. The first claim is obvious. By the usual rules for the adjoint matrix, we conclude that

$$\langle Ax, y \rangle = y^*(Ax) = (y^* A)x = (A^* y)^* x = \langle x, A^* y \rangle .$$

Testing (4.2) with unit vectors for x and y gives $B = A^*$. □

We can "shift-conjugate" the operator A from the left factor to the right factor of a scalar product, but also in the other direction:

$$\langle x, Ay \rangle = \overline{\langle Ay, x \rangle} = \overline{\langle y, A^* x \rangle} = \langle A^* x, y \rangle .$$

Definition 4.17. *Let* $A \in \mathbb{C}^{n \times n}$*. Then we define:*

A **is normal** *if* $AA^* = A^* A$*,*

A **is self-adjoint or hermitian** *if* $A = A^*$*,*

A **is symmetric** *if* A *is self-adjoint and* $A \in \mathbb{R}^{n \times n}$*,*

A **is unitary** *if* $A^{-1} = A^*$*,*

A **is orthogonal** *if* A *is unitary and* $A \in \mathbb{R}^{n \times n}$*.*

Examples are numerous:

- Matrices describing rotations or reflections in $\mathbb{R}^n$ are orthogonal, as we will see below.
- The Hessian matrix of a real-valued (twice continuously differentiable) function is symmetric.
- The matrix of the *inertial tensor*[10] is symmetric.

[10] Trägheitstensor

- Most operators from quantum mechanics are self-adjoint, although they frequently do not have a matrix form. For instance, the energy levels for the electron in the hull of the hydrogen atom are eigenvalues of the HAMILTON[11] operator, which is a partial differential operator of second order acting in the function space $L^2(\mathbb{R}^3 \to \mathbb{C})$ (whose dimension is infinite). The wave function of the electron is then the eigenvector to that eigenvalue.

Orthogonal projection operators are another example. Take the space $\mathbb{C}^n$ and a subspace $U \subset \mathbb{C}^n$. There is exactly one mapping which maps a point x from $\mathbb{C}^n$ to that element of U, which has least distance to x. This mapping is called *orthogonal projection*[12]. Let $\{u_1, \dots, u_m\}$ be an orthonormal basis for the subspace U. By Satz 2.31 from the first term, the projection operator P has the following form:

$$\begin{aligned} &P\colon \mathbb{C}^n \to U, \\ &P\colon x \mapsto Px = \sum_{j=1}^m \langle x, u_j \rangle \, u_j = \sum_{j=1}^m (u_j^* x) u_j = \sum_{j=1}^m u_j (u_j^* x) = \left(\sum_{j=1}^m u_j u_j^* \right) x =: A_P x. \end{aligned}$$

From now on, we will no longer distinguish between an orthogonal projector P (which is a mapping) and its associated matrix A_P (which is, well, a matrix). There should be no danger of misunderstandings.

Proposition 4.18 (Properties of orthogonal projectors, I). *Orthogonal projectors are self-adjoint and* idempotent[13]*, meaning $P \circ P = P$.*

The relation $P \circ P = P$ holds for all projections (not only orthogonal projections), and maybe it becomes clearer if you remember that a point directly on the ground will have no shadow from the sunbeams. Or you could draw pictures with $U \subset \mathbb{C}^n$, x, Px, PPx.

Proof. Just compute it, taking advantage from the sesquilinearity[14] of the scalar product and from the system $(u_1, \dots, u_m)$ being orthonormal:

$$\begin{aligned} \langle Px, y \rangle &= \left\langle \sum_{j=1}^m \langle x, u_j \rangle \, u_j, y \right\rangle = \sum_{j=1}^m \langle x, u_j \rangle \langle u_j, y \rangle \, , \\ \langle x, Py \rangle &= \left\langle x, \sum_{j=1}^m \langle y, u_j \rangle \, u_j \right\rangle = \sum_{j=1}^m \overline{\langle y, u_j \rangle} \langle x, u_j \rangle = \sum_{j=1}^m \langle x, u_j \rangle \langle u_j, y \rangle \, , \end{aligned}$$

$$\begin{aligned} P^2 x = PPx &= \sum_{j=1}^m \langle Px, u_j \rangle \, u_j = \sum_{j=1}^m \left\langle \sum_{l=1}^m \langle x, u_l \rangle \, u_l, u_j \right\rangle u_j \\ &= \sum_{j=1}^m \sum_{l=1}^m \langle x, u_l \rangle \langle u_l, u_j \rangle \, u_j = \sum_{j=1}^m \langle x, u_j \rangle \, u_j = Px. \end{aligned}$$

□

Proposition 4.19 (Properties of orthogonal projectors, II). *Let U and V be subspaces of $\mathbb{C}^n$ with $U \perp V$. Denote the orthogonal projectors from $\mathbb{C}^n$ onto U and V by P and Q, respectively. Then $PQ = 0$, the null operator.*

Proof. The spaces U and V have orthonormal bases $(u_1, \dots, u_k)$ and $(v_1, \dots, v_l)$. Then the projectors P and Q are given by

$$Px = \sum_{i=1}^k \langle x, u_i \rangle \, u_i, \qquad Qx = \sum_{j=1}^l \langle x, v_j \rangle \, v_j.$$

[11] WILLIAM ROWAN HAMILTON, 1805 – 1865

[12] Orthogonalprojektion

[13] idempotent

[14] this means that the scalar product is linear in each factor, but constants in the second factor must be conjugated before dragging them in front of the product

What still remains is to compute PQx for an arbitrary $x \in \mathbb{C}^n$:

$$PQx = \sum_{i=1}^{k} \langle Qx, u_i \rangle \, u_i = \sum_{i=1}^{k} \left\langle \sum_{j=1}^{l} \langle x, v_j \rangle \, v_j, u_i \right\rangle u_i = \sum_{i=1}^{k} \sum_{j=1}^{l} \langle x, v_j \rangle \langle v_j, u_i \rangle \, u_i = 0,$$

because of $\langle v_j, u_i \rangle = 0$ due to $U \perp V$. □

Now we get back to the normal matrices.

Proposition 4.20. *For any matrix $A \in \mathbb{C}^{n\times n}$, the following holds:*

1. $(\operatorname{img} A)^\perp = \ker A^*$,
2. *A is normal* $\implies \quad \ker A = \ker A^*$,
3. *A is normal* $\implies \quad \mathbb{C}^n = \operatorname{img} A \oplus \ker A$.

Proof. 1. If $y \in (\operatorname{img} A)^\perp$, then $\langle y, Ax \rangle = 0$ for each $x \in \mathbb{C}^n$. Choosing $x = A^* y$, we then have $0 = \langle y, AA^* y \rangle = \langle A^* y, A^* y \rangle$, hence $y \in \ker A^*$. This gives you the inclusion $(\operatorname{img} A)^\perp \subset \ker A^*$.

Put $r := \operatorname{rank}(A) = \operatorname{rank}(A^*)$. Then $r = \dim \operatorname{img} A$, hence $\dim(\operatorname{img} A)^\perp = n - r$, where we have used the dimension formula for sub-spaces. By the dimension formula for linear mappings, we have $\dim \ker A^* = n - \dim \operatorname{img} A^* = n - \operatorname{rank}(A^*) = n - r$. Recall $(\operatorname{img} A)^\perp \subset \ker A^*$, and both these sub-spaces of $\mathbb{C}^n$ have the same dimension $n - r$. Therefore, they are equal.

2. From $A^* A = AA^*$, we deduce that

$$\|Ax\|^2 = \langle Ax, Ax \rangle = \langle x, A^* Ax \rangle = \langle x, AA^* x \rangle = \langle A^* x, A^* x \rangle = \|A^* x\|^2 ,$$

which says that $Ax = 0$ if and only if $A^* x = 0$.

3. Satz 2.32 from the first terms yields $\mathbb{C}^n = (\operatorname{img} A) \oplus (\operatorname{img} A)^\perp$. Now apply part 1 and part 2.

□

Corollary 4.21 (Fredholm[15] **alternative**). *From* $(\operatorname{img} A)^\perp = \ker A^*$ *we get* $\operatorname{img} A = (\ker A^*)^\perp$, *and this gives a beautiful conclusion:*

If $Ax = b$ is solvable, then b must be perpendicular to $\ker A^*$.

If b is perpendicular to $\ker A^*$, *then the system $Ax = b$ is solvable.*

Question: Check the following: if A and B are normal matrices, then $A + B$ need not be a normal matrix. Hence the set of all normal matrices of $\mathbb{C}^{n\times n}$ will not form a vector space (if $n \geq 2$).

We continue with eigenvalues and eigenvectors of normal matrices:

Proposition 4.22 (Normal matrices). *Let $A \in \mathbb{C}^{n\times n}$ be a normal matrix, and $\lambda \in \mathbb{C}$. Then:*

1. $(A - \lambda I)^* = A^* - \overline{\lambda} I$,
2. $A - \lambda I$ *is normal,*
3. $Ax = \lambda x \iff A^* x = \overline{\lambda} x$*; especially, the matrices A and A^* have the same eigenvectors,*
4. *eigenvectors to different eigenvalues are perpendicular to each other.*

Proof. 1. Conjugating is an additive operation.

2. Follows from 1 by a computation like this:

$$\begin{aligned}(A - \lambda I)(A - \lambda I)^* &= (A - \lambda I)(A^* - \overline{\lambda} I) = AA^* - \lambda A^* - \overline{\lambda} A + |\lambda|^2 I \\ &= A^* A - \lambda A^* - \overline{\lambda} A + |\lambda|^2 I = (A^* - \overline{\lambda} I)(A - \lambda I) \\ &= (A - \lambda I)^*(A - \lambda I).\end{aligned}$$

[15] Erik Ivar Fredholm, 1866 – 1927

3. By Proposition 4.20 and 2., $\ker(A - \lambda I) = \ker((A - \lambda I)^*) = \ker(A^* - \overline{\lambda} I)$.

4. Take eigenvectors x and x' to eigenvalues λ and λ' with $\lambda \neq \lambda'$. Part 3 gives $A^* x' = \overline{\lambda'} x'$, hence

$$\lambda \langle x, x' \rangle = \langle \lambda x, x' \rangle = \langle Ax, x' \rangle = \langle x, A^* x' \rangle = \langle x, \overline{\lambda'} x' \rangle = \lambda' \langle x, x' \rangle .$$

This is only possible if $\langle x, x' \rangle = 0$, due to $\lambda \neq \lambda'$.

□

The next result is the highlight of this chapter, and it tells us that normal matrices are always diagonalisable because they possess enough linearly independent eigenvectors. The ugly machinery of Jordan normal forms is then not needed.

Theorem 4.23 (Spectral theorem). *The following statements are equivalent for any* $A \in \mathbb{C}^{n\times n}$*:*

1. *A is normal,*
2. *there is an orthonormal basis of* $\mathbb{C}^n$*, consisting of eigenvectors of* A*,*
3. *there are orthogonal projectors* $P_1, \ldots, P_n$ *and* $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ *with the following properties:*
   (a) $P_j P_k = 0$ *for* $j \neq k$*,*
   (b) $\sum_{j=1}^n P_j = I$*,*
   (c) $\sum_{j=1}^n \lambda_j P_j = A$*.*

The last formula is known as *spectral representation*[16] *of* A.

Proof. $1 \Longrightarrow 2$: By Proposition 4.5, the matrix A has at least one eigenvalue. Denote the eigenvalues of A by $\lambda_1, \ldots, \lambda_m$. Each eigenspace $\ker(A - \lambda_j I)$ has an orthonormal basis. Collecting these bases gives you a family of vectors $(u_1, \ldots, u_r)$ which are orthonormal, due to Proposition 4.22.

If $r = n$, we are done. Aiming for a contradiction, we suppose that $r < n$. Put $U := \operatorname{span}(u_1, \ldots, u_r)$ and $V := U^\perp$. The vectors u_j are eigenvectors to A as well as A^*, hence $AU \subset U$ and $A^* U \subset U$, by Proposition 4.22, part 3. Next we show that A maps the sub-space V into itself: $AV \subset V$. If $v \in V$ is fixed and $u \in U$ is arbitrary, then

$$\langle u, Av \rangle = \langle A^* u, v \rangle = 0,$$

since $A^* u \in U$ and $U \perp V$. This implies that $Av \in U^\perp = V$, or $AV \subset V$ as claimed. By Proposition 4.5, the matrix A must have an eigenvector in V which is impossible, from the construction of U and V.

$2 \Longrightarrow 3$: Write this orthonormal basis of $\mathbb{C}^n$ as $(u_1, \ldots, u_n)$, with u_j being an eigenvector to the eigenvalue λ_j. Define orthonormal projectors P_j as $P_j x := \langle x, u_j \rangle u_j$. Proposition 4.19 gives (a) and Satz 2.29 from the first term gives (b). Then (c) follows easily:

$$Ax = AIx = A \sum_{j=1}^n P_j x = A \sum_{j=1}^n \langle x, u_j \rangle u_j = \sum_{j=1}^n \langle x, u_j \rangle A u_j = \sum_{j=1}^n \langle x, u_j \rangle \lambda_j u_j = \sum_{j=1}^n \lambda_j P_j x.$$

$3 \Longrightarrow 1$: Formula (c) gives you a representation of A, from which it is easy to verify that $AA^* = A^* A$. Here you could use $P^* = P$ from Proposition 4.18 and $P_j P_k = \delta_{jk} P_j$. □

Question: Choose $A = 3I$ and $u_j = e_j$, the canonical basis vectors of $\mathbb{C}^n$. How do the projectors P_j look like ?

We discuss an example. Let $A \in \mathbb{C}^{n\times n}$ be normal with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $u_1, \ldots, u_n$, such that $\langle u_j, u_k \rangle = \delta_{jk}$. Given $x \in \mathbb{C}^n$, can we describe Ax in another way ? We may decompose x,

$$x = \alpha_1 u_1 + \cdots + \alpha_n u_n,$$

[16] Spektraldarstellung

since $(u_1, \dots, u_n)$ is a basis of $\mathbb{C}^n$. And the α_j can be quickly found as $\alpha_j = \langle x, u_j \rangle$, because uf $\langle u_k, u_l \rangle = \delta_{kl}$. The result then is

$$x = \sum_{j=1}^{n} \langle x, u_j \rangle \, u_j = \sum_{j=1}^{n} \Big(P_j x \Big) = \left(\sum_{j=1}^{n} P_j \right) x,$$

or part 3(b). And the action of A then is described by

$$\begin{aligned} Ax &= A \sum_{j=1}^{n} P_j x = \sum_{j=1}^{n} A P_j x = \sum_{j=1}^{n} A \, \langle x, u_j \rangle \, u_j = \sum_{j=1}^{n} \langle x, u_j \rangle \, A u_j = \sum_{j=1}^{n} \langle x, u_j \rangle \, \lambda_j u_j = \sum_{j=1}^{n} \Big(\lambda_j P_j x \Big) \\ &= \left(\sum_{j=1}^{n} \lambda_j P_j \right) x, \end{aligned}$$

hence part 3(c). The matrix A performs its action onto the vector x in three steps:

- first x is decomposed by the family of projectors $P_1, \dots, P_n$ into the components living in the eigenspaces,
- second: in each eigenspaces, the operator A behaves like a dilation[17] by the associated eigenvalue,
- third: the individual results are recombined ($\sum_{j=1}^{n}$).

Things become even nicer for self-adjoint and unitary matrices:

Proposition 4.24 (Self-adjoint matrices). *All eigenvalues of a self-adjoint matrix are real.*

Each symmetric matrix possesses an ONB of real *eigenvectors.*

Proof. If $A = A^*$ then A is normal, and we can exploit Proposition 4.22. Let now $Ax = \lambda x$, then $\lambda x = Ax = A^* x = \overline{\lambda} x$, and therefore $(\lambda - \overline{\lambda})x = 0$, which is only possible for $\lambda \in \mathbb{R}$.

The eigenvectors are elements of $\ker(A - \lambda I)$, and $A - \lambda I$ is a *real* matrix if A is symmetric and λ is real. Then you can choose a basis of $\ker(A - \lambda I)$ as real vectors. □

Example 4.25 (the Fourier series revisited). *We compare a self-adjoint matrix $A \in \mathbb{C}^{n \times n}$ and the differential operator $\mathcal{A} := \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}$ acting on the space $C^1_{(2\pi)}(\mathbb{R} \to \mathbb{C})$ of those functions that are continuously differentiable and 2π-periodic.*

$U = \mathbb{C}^n$	$U = C^1_{(2\pi)}(\mathbb{R} \to \mathbb{C})$
$\dim U = n$	$\dim U = \infty$
$\langle x, y \rangle = \sum_{j=1}^{n} \xi_j \overline{\eta_j}$	$\langle f, g \rangle = \int_{t=-\pi}^{\pi} f(t)\overline{g(t)} \, \mathrm{d}t$
$A = A^*$, $\langle Ax, y \rangle = \langle x, Ay \rangle \quad \forall \, x, y \in \mathbb{C}^n$	$\mathcal{A} = \mathcal{A}^*$: partial integration can be used to show $\left\langle \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t} f, g \right\rangle = \left\langle f, \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t} g \right\rangle \quad \forall \, f, g \in U$
eigenvalues of A are $\lambda_1, \dots, \lambda_n \in \mathbb{R}$	eigenvalues of $\mathcal{A}$ are $\dots -3, -2, -1, 0, 1, 2, \dots$
eigenvectors of A are $u_1, \dots, u_n \in \mathbb{C}^n$ $Au_k = \lambda_k u_k$	eigenfunctions of $\mathcal{A}$ are $e_k = e_k(t) = e^{-\mathrm{i}kt}$, $k \in \mathbb{Z}$ $\mathcal{A} e_k = k e_k$
$\langle u_j, u_k \rangle = \delta_{jk}$	$\langle e_j, e_k \rangle = 2\pi \delta_{jk}$
decomposition of $x \in \mathbb{C}^n$: $x = \sum_{k=1}^{n} \alpha_k u_k, \quad \alpha_k = \langle x, u_k \rangle$	Fourier series decomposition of $f \in U$: $f(t) = \sum_{k \in \mathbb{Z}} \hat{f}_k e_k(t), \quad \hat{f}_k = \frac{1}{2\pi} \langle f, e_k \rangle$
Pythagorean Theorem: $\|x\|^2 = \sum_{k=1}^{n} \lvert \alpha_k \rvert^2$	Bessel identity: $\|f\|^2 = \sum_{k \in \mathbb{Z}} 2\pi \lvert \hat{f}_k \rvert^2$

[17] Streckung, Dehnung

Proposition 4.26 (Unitary matrices). *Let either $A \in \mathbb{C}^{n\times n}$ or $A \in \mathbb{R}^{n\times n}$. Then the following statements are equivalent:*

1. *A is unitary (or orthogonal), i.e., $AA^* = I$ (or $AA^\top = I$),*
2. *the columns of A are an ONB of $\mathbb{C}^n$ (or $\mathbb{R}^n$),*
3. *if $(v_1, \dots, v_n)$ is an ONB, then so is $(Av_1, \dots, Av_n)$,*
4. *for all $x, y \in \mathbb{C}^n$ (or $\in \mathbb{R}^n$), we have $\langle x, y\rangle = \langle Ax, Ay\rangle$,*
5. *A is* isometric[18]*, i.e., $\|Ax\| = \|x\|$ for all $x \in \mathbb{C}^n$ (or $\in \mathbb{R}^n$),*
6. *A is normal and all eigenvalues have modulus 1.*

Sketch of proof. We only consider the $\mathbb{C}$–case:

$1 \iff 2$: This is just the definition of the matrix-matrix-product.

$4 \iff 5$: This is a direct consequence of

$$4\langle x, y\rangle = \|x+y\|^2 - \|x-y\|^2 + \mathrm{i}\,\|x+\mathrm{i}y\|^2 - \mathrm{i}\,\|x-\mathrm{i}y\|^2,$$
$$4\langle Ax, Ay\rangle = \|Ax+Ay\|^2 - \|Ax-Ay\|^2 + \mathrm{i}\,\|Ax+\mathrm{i}Ay\|^2 - \mathrm{i}\,\|Ax-\mathrm{i}Ay\|^2.$$

$1 \implies 4$: This is due to $\langle Ax, Ay\rangle = \langle A^*Ax, y\rangle = \langle x, y\rangle$.

$4 \implies 1$: Put $B = A^*A$. Then $\langle x, y\rangle = \langle Ax, Ay\rangle = \langle x, A^*Ay\rangle = \langle x, By\rangle$. On the other hand, $\langle x, y\rangle = \langle Ix, y\rangle$. Then Proposition 4.16 gives $B = I^* = I$.

Observe that we have shown: the four statements 1, 2, 4, 5 are logically equivalent.

$(1, 5) \implies 6$: Every unitary matrix is normal. If $Ax = \lambda x$ with $x \neq 0$ and $\|Ax\| = \|x\|$, then $|\lambda| = 1$.

$6 \implies 1$: The spectral theorem gives us n eigenvectors and projections P_j onto the fibres, generated by those eigenvectors, such that $A = \sum_{j=1}^n \lambda_j P_j$. Then

$$AA^* = \left(\sum_{j=1}^n \lambda_j P_j\right)\left(\sum_{k=1}^n \overline{\lambda_k} P_k\right) = \sum_{j=1}^n\sum_{k=1}^n \lambda_j\overline{\lambda_k}P_jP_k = \sum_{j=1}^n \lambda_j\overline{\lambda_j}P_j = \sum_{j=1}^n |\lambda_j|^2 P_j = I,$$

because of $|\lambda_j| = 1$. Therefore, A is unitary.

$3 \implies 2$: Choose $v_j = e_j$, the canonical basis vectors. They are an ONB, hence $(Ae_1, \dots, Ae_n)$ must be an ONB. But these vectors are just the columns of A.

$4 \implies 3$: If $\langle v_j, v_k\rangle = \delta_{jk}$ then $\langle Av_j, Av_k\rangle = \delta_{jk}$. □

We learn the following:

Each mapping of a (finite-dimensional) vector space into itself
that preserves the length of a vector is described by a unitary/orthogonal matrix.

Proposition 4.26 gives us enough information to completely characterise orthogonal matrices:

$n = 2$**:** The first column is a vector of length one, hence it can be written as $(\cos\varphi, \sin\varphi)^\top$, for some $\varphi \in [0, 2\pi]$. Then the second column must be also a vector of length one, and it must be perpendicular to the first column. Therefore the second column is $\pm(-\sin\varphi, \cos\varphi)^\top$, leading us to two cases:

$$A = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}, \qquad \det A = 1,$$
$$A = \begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix}, \qquad \det A = -1.$$

[18] isometrisch

The first matrix has the characteristic polynomial $\chi_A(\lambda) = (\cos\varphi - \lambda)^2 + \sin^2\varphi$, with the complex roots[19] $\lambda_{1,2} = \exp(\pm \mathrm{i}\varphi)$. The associated mapping is a rotation with angle φ.

The second matrix has the characteristic polynomial $\chi_A(\lambda) = \lambda^2 - 1$ with roots $\lambda_{1,2} = \pm 1$. The associated mapping is a reflection along the axis of the eigenvector to the eigenvalue $\lambda_1 = +1$, $(\cos(\frac{\varphi}{2}), \sin(\frac{\varphi}{2}))$.

$n = 3$**:** The characteristic polynomial $\chi_A(\lambda) = -\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0$ has real coefficients -1, c_2, c_1, c_0. We have $\lim_{\lambda\to-\infty}\chi_A(\lambda) = +\infty$ and and $\lim_{\lambda\to+\infty}\chi_A(\lambda) = -\infty$, therefore the continuous function χ_A must have a real root, which can only be $+1$ or -1. For the other two roots, there are two possibilities. Either they are both real, and each of them must be $+1$ or -1. Or they are both non-real, and each is the complex conjugate of the other. Since their modulus must equal 1, we can write them as $\exp(\pm\mathrm{i}\varphi)$, for some $\varphi \in (0, \pi)$.

In the first case, the eigenvalues are either $(\lambda_1, \lambda_2, \lambda_{33}) = (1, 1, 1)$ or $(1, 1, -1)$ or $(1, -1, -1)$ or $(-1, -1, -1)$, with associated eigenvectors (u_1, u_2, u_3). The mappings are the identity mapping, a reflection on the plane spanned by (u_1, u_2), a rotation around u_1 with angle π, or a point-reflection at the origin.

In the second case, there is a real eigenvector u_1 to the real eigenvalue $\lambda_1 = \pm 1$. The other two non-real eigenvalues are $(\lambda_2, \lambda_3) = (\exp(\mathrm{i}\varphi), \exp(-\mathrm{i}\varphi))$ with $0 < \varphi < \pi$, and these non-real eigenvalues have never a real eigenvector. Then A maps the plane $E = (\operatorname{span}(u_1))^\perp$ into itself, namely as a rotation with angle φ. Hence there is a real orthogonal matrix S, such that

$$S^{-1}AS = \begin{pmatrix} \pm 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}.$$

The mapping A is either a rotation around u_1 with angle φ (in case of $\lambda_1 = +1$), or it is a rotation followed by a plane reflection (in case of $\lambda_1 = -1$).

We can generalise this to arbitrary n.

Proposition 4.27 (**Orthogonal matrices**). *For any orthogonal matrix $A \in \mathbb{R}^{n\times n}$, an orthonormal basis $(u_1, \dots, u_n)$ of $\mathbb{R}^n$ can be found which transforms the matrix A to $\tilde{A}$,*

$$\tilde{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & R_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & R_k \end{pmatrix},$$

where the R_j are 2×2 rotation matrices.

For later use, we define some groups of matrices. The operation of the group is always the multiplication.

Definition 4.28. 1. *The group of all invertible $n \times n$ matrices is called* linear group $GL(n)$.

2. *The group of all unitary $n \times n$ matrices is called* unitary group $U(n)$.

3. *The group of all orthogonal $n \times n$ matrices is called* orthogonal group $O(n)$.

4. *The groups of all invertible, unitary, orthogonal matrices with determinant equal to $+1$ are called* special linear group $SL(n)$, special unitary group $SU(n)$, *and* special orthogonal group $SO(n)$.

5. *The group of all matrices of the form* $\exp(\mathrm{i}\varphi)I$ *with $\varphi \in \mathbb{R}$ is called* circle group.

It turns out that the discussion of quarks and anti-quarks requires a deeper understanding of the structures of $SU(3)$ and $SU(4)$, compare the already cited book of Greiner and Müller on symmetries in quantum mechanics.

[19] Wurzel bzw. Nullstelle

4.5 Definiteness, Quadratic Forms, and Quadrics

For a self-adjoint matrix A and arbitrary vector x, we have

$$\langle Ax, x\rangle = \langle x, A^*x\rangle = \langle x, Ax\rangle = \overline{\langle Ax, x\rangle},$$

hence $\langle Ax, x\rangle$ must be real. Then the following definition is reasonable:

Definition 4.29. *Let $A \in \mathbb{C}^{n\times n}$ be self-adjoint and $U \subset \mathbb{C}^n$ be a sub-space. Then we call*

- *A* positive definite on *U if $\langle Ax, x\rangle > 0$ for all $x \in U$, $x \neq 0$,*
- *A* negative definite on *U if $\langle Ax, x\rangle < 0$ for all $x \in U$, $x \neq 0$,*
- *A* positive semi-definite on *U if $\langle Ax, x\rangle \geq 0$ for all $x \in U$,*
- *A* negative semi-definite on *U if $\langle Ax, x\rangle \leq 0$ for all $x \in U$.*

If no sub-space U is mentioned, then $U = \mathbb{C}^n$ is meant.

Proposition 4.30 (**Definiteness**)**.** *A self-adjoint matrix is*

positive definite *if and only if all eigenvalues are positive,*

positive semi-definite *if and only if all eigenvalues are greater than or equal to zero,*

negative semi-definite *if and only if all eigenvalues are less than or equal to zero,*

negative definite *if and only if all eigenvalues are negative.*

Proof. Exercise. □

In particular, a definite matrix is invertible.

Proposition 4.31. *Let $f \in C^2(G \to \mathbb{R})$ be a function on an open set $G \subset \mathbb{R}^n$. Let x_0 be an interior point of G. Then the following holds:*

- *if $\nabla f(x_0) = 0$ and the Hessian matrix $(Hf)(x_0)$ is positive definite, then f has a minimum at the point x_0;*
- *if $\nabla f(x_0) = 0$ and the Hessian matrix $(Hf)(x_0)$ is negative definite, then f has a maximum at the point x_0;*
- *if $\nabla f(x_0) = 0$ and the Hessian matrix $(Hf)(x_0)$ has some negative eigenvalues and some positive ones, then f has a saddle point at x_0.*

If $\nabla f(x_0) = 0$ and one eigenvalue of $(Hf)(x_0)$ is zero, then everything can happen.

Definition 4.32 (**Signature**)**.** *Denote by U_+, U_-, and U_0 the linear spaces that are spanned by the eigenvectors to positive, negative, and zero eigenvalues of a self-adjoint matrix A. Define $k_+ = \dim U_+$, $k_- = \dim U_-$, and $k_0 = \dim U_0$. The triple (k_+, k_-, k_0) is the* signature[20] *of the matrix A. (If there are no eigenvalues of the sign $+$, $-$, or 0, choose the null space for U_+, U_-, U_0.)*

The signature only depends on the eigenvalues, which do not change under similarity transformations (since similar matrices have the same characteristic polynomial, Proposition 4.6).

This can be sharpened a bit to the following Law which we do not prove:

Proposition 4.33 (SYLVESTER's **Law of Inertia**[21] [22])**.** *If $A \in \mathbb{C}^{n\times n}$ is self-adjoint and $G \in \mathbb{C}^{n\times n}$ is invertible, then also G^*AG is self-adjoint and has the same signature as A.*

[20] Signatur

[21] SYLVESTERscher Trägheitssatz

[22] JAMES JOSEPH SYLVESTER, 1814 – 1897

Proposition 4.34 (Definiteness). *A self-adjoint matrix $A \in \mathbb{C}^{n\times n}$ is positive definite if and only if all sub-determinants of the form*

$$\det A_k := \det(a_{ij})_{1\le i,j\le k}, \qquad k = 1, \dots, n,$$

are positive. It is negative definite if and only if $(-1)^k \det A_k > 0$ for all k.

Proof. If A is positive definite, then all eigenvalues are positive, then also $\det A$ is positive, being the product of the eigenvalues. Choose $U_k = \operatorname{span}(e_1, \dots, e_k)$ and $x \in U_k$. Then

$$\langle Ax, x\rangle = \sum_{i,j=1}^{k} \overline{\xi_i} a_{ij} \xi_j > 0,$$

showing that A is positive definite on U_k. Consequently, $\det A_k > 0$.

A matrix A is negative definite if and only if $-A$ is positive definite. The other two parts of the proof are a bit tricky, and we prefer to omit them. □

To present another criterion on the definiteness which is sometimes helpful, we need an auxiliary result, which gives you a hint where the eigenvalues of a matrix are located, without computing them:

Lemma 4.35 (GERSCHGORIN's **circle theorem**). *Let $A \in \mathbb{C}^{n\times n}$ be any matrix, and for $j = 1, \dots, n$, call C_j the closed circle in the complex plane with center a_{jj} and radius $\sum_{k\ne j} |a_{jk}|$. Then the union of all these circles C_1, …, C_n contains all the eigenvalues of A.*

Be careful with the logics here ! The circle theorem *does not* say that each circle C_j contains one eigenvalue.

Proof. Take an eigenvalue λ of A, an eigenvector x to that eigenvalue, and let x_k be a component of x with biggest modulus. This means $|x_k| \ge |x_j|$ for all $j = 1, \dots, n$. We do not know the value of k, and we can not choose k, but we know that k exists. The kth row of the equation $Ax = \lambda x$ reads

$$\sum_{j\ne k} a_{kj} x_j = (\lambda - a_{kk}) x_k,$$
$$|\lambda - a_{kk}| \cdot |x_k| \le \sum_{j\ne k} |a_{kj}| \cdot |x_j| \le |x_k| \sum_{j\ne k} |a_{kj}|,$$

which implies $|\lambda - a_{kk}| \le \sum_{j\ne k} |a_{kj}|$. This is what we wanted. □

Here comes the promised criterion, which is an easy consequence of Gerschgorin's theorem:

Proposition 4.36 (Definiteness). *Any self-adjoint matrix A that is* strictly diagonal dominant*, i.e.,*

$$a_{kk} > \sum_{j\ne k} |a_{kj}|, \qquad 1 \le k \le n,$$

is positive definite.

Lemma 4.37. *If a rectangular matrix $A \in \mathbb{C}^{m\times n}$ has rank equal to n with $n \le m$, then the matrix $A^* A \in \mathbb{C}^{n\times n}$ is positive definite and invertible.*

Proof. Beautiful exercise. □

We conclude this section with a classification of the *quadric surfaces*[23] in $\mathbb{R}^n$. These consist of all points $x = (x_1, \dots, x_n)^\top$ from $\mathbb{R}^n$ with

$$\sum_{i,j=1}^{n} a_{ij} x_i x_j + 2 \sum_{j=1}^{n} b_j x_j + b_0 = 0, \tag{4.3}$$

[23] Quadriken

where the a_{ij}, and b_j, are given real numbers.

Neglecting exceptional cases of minor importance, the quadrics in $\mathbb{R}^2$ are the ellipse, the hyperbola, or the parabola. Now we present a general procedure to completely classify the quadrics in $\mathbb{R}^n$, with $n = 2, 3$.

We start with the example

$$3x_1^2 - 7x_2^2 + 10x_3^2 + 28x_1x_2 - 38x_1x_3 + 42x_2x_3 + 12x_1 - 16x_2 + 20x_3 + 25 = 0,$$

and we observe that we can rewrite it into the form

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 3 & 14 & -19 \\ 14 & -7 & 21 \\ -19 & 21 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 12 & -16 & 20 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + 25 = 0, \tag{4.4}$$

which is equivalent to

$$\begin{pmatrix} x_1 & x_2 & x_3 & 1 \end{pmatrix} \begin{pmatrix} 3 & 14 & -19 & 6 \\ 14 & -7 & 21 & -8 \\ -19 & 21 & 10 & 10 \\ 6 & -8 & 10 & 25 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = 0.$$

Note that both matrices are symmetric. It is quite hard to understand the shape of the solution set (which is some set of points of the $\mathbb{R}^3$) of this scalar equation, because the matrices are quite complicated. A new coordinate system will bring us to a simpler matrix, in the sense that the new matrix contains many zero entries.

Step 1: Eliminating the mixed entries $a_{jk}x_jx_k$ by a rotation.

Write (4.3) as $\langle Ax, x\rangle + 2\langle b, x\rangle + b_0 = 0$, which is of the form (4.4). The matrix A is real symmetric. Then there is an orthonormal system of real eigenvectors of A that form a basis of $\mathbb{R}^n$. Writing these vectors as columns one next to the other, you have an orthogonal matrix G. Put

$$y := G^\top x, \qquad x = Gy$$

as new coordinates. With the new notation $L := G^\top AG$, $c := G^\top b$, $c_0 := b_0$, you then get

$$\langle Ly, y\rangle + 2\langle c, y\rangle + c_0 = 0.$$

By choice of G, the matrix L is a diagonal matrix comprising the eigenvalues of A. By suitable ordering of the columns of G, we can achieve that

$$L = \operatorname{diag}(\lambda_1, \dots, \lambda_p, \lambda_{p+1}, \dots, \lambda_{p+m}, 0, \dots, 0),$$

with $\lambda_1, \dots, \lambda_p$ being the positive eigenvalues and $\lambda_{p+1}, \dots, \lambda_{p+m}$ being the negative eigenvalues. The signatures of A and L are the same, namely $(p, m, n-p-m)$ (by Sylvester's Law).

Step 2: Eliminating many lower order terms by a shift of the origin.

Now we shift the origin, which means to introduce new coordinates z by $z := y + \delta$, for some carefully chosen shift vector δ. We have two cases.

Case I: if all the pure squares y_j^2 are present, then all the linear terms c_jy_j can be eliminated, but the constant term will remain almost always.

Case II: only some pure squares y_j^2 are present, other pure squares are absent. Then some linear terms can be eliminated, and the constant term can be eliminated if some linear item is present whose quadratic brother is absent.

We discuss an example of Case I: consider

$$0 = 4y_1^2 - 9y_2^2 + 16y_3^2 + 8y_1 - 36y_2 + 32y_3 + 26.$$

We rewrite this equation as follows:

$$\begin{aligned} 0 &= 4(y_1^2 + 2y_1) - 9(y_2^2 + 4y_2) + 16(y_3^2 + 2y_3) + 26 \\ &= 4\Big((y_1+1)^2 - 1\Big) - 9\Big((y_2+2)^2 - 4\Big) + 16\Big((y_3+1)^2 - 1\Big) + 26 \\ &= 4(y_1+1)^2 - 9(y_2+2)^2 + 16(y_3+1)^2 + \Big(-4 + 36 - 16 + 26\Big). \end{aligned}$$

We set $z_1 := y_1 + 1$, $z_2 := y_2 + 2$, $z_3 := y_3 + 1$, $d_0 := 42$ and get $0 = 4z_1^2 - 9z_2^2 + 16z_3^2 + d_0$. A matrix formulation is

$$\begin{pmatrix} z_1 & z_2 & z_3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & -9 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 42 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ 1 \end{pmatrix} = 0. \tag{4.5}$$

Next we discuss an example of Case II: consider

$$0 = 4y_1^2 - 9y_2^2 + 8y_1 - 36y_2 + 32y_3 + 26,$$

and observe that no term with y_3^2 is available, but we have linear terms y_3^1. We rewrite the equation:

$$\begin{aligned} 0 &= 4(y_1^2 + 2y_1) - 9(y_2^2 + 4y_2) + 32y_3 + 26 \\ &= 4\Big((y_1 + 1)^2 - 1\Big) - 9\Big((y_2 + 2)^2 - 4\Big) + 32y_3 + 26 \\ &= 4(y_1 + 1)^2 - 9(y_2 + 2)^2 + 32y_3 + \Big(-4 + 36 + 26\Big) \\ &= 4(y_1 + 1)^2 - 9(y_2 + 2)^2 + 32\left(y_3 + \frac{58}{32}\right). \end{aligned}$$

Now we set $z_1 := y_1 + 1$, $z_2 := y_2 + 2$, $z_3 := y_3 + \frac{58}{32}$, and the result then is $0 = 4z_1^2 - 9z_2^2 + 32z_3$, for which we have the matrix formulation

$$\begin{pmatrix} z_1 & z_2 & z_3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & -9 & 0 & 0 \\ 0 & 0 & 0 & 16 \\ 0 & 0 & 16 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ 1 \end{pmatrix} = 0. \tag{4.6}$$

And we only mention one more example of Case II: if in our model equation also the term $32y_3$ were absent, then the zero order item 58 would have survived.

We can join Case I and Case II in the formula $0 = \langle Lz, z\rangle + 2\langle d, z\rangle + d_0$ where one of d and d_0 vanishes. After switching to block matrix notation,

$$\begin{aligned} L &= \begin{pmatrix} \Lambda & O \\ O & O \end{pmatrix} \in \mathbb{R}^{n\times n}, & \Lambda &= \operatorname{diag}(\lambda_1, \dots, \lambda_m, \lambda_{m+1}, \dots \lambda_{m+p}), \\ d &= \begin{pmatrix} O \\ d' \end{pmatrix} \in \mathbb{R}^n, & d' &= (d_{m+p+1}, \dots, d_n)^\top, \\ \hat{z} &= \begin{pmatrix} z \\ 1 \end{pmatrix} \in \mathbb{R}^{n+1}, & z &= (z_1, \dots, z_n)^\top, \end{aligned}$$

we can compress the quadric formula even more: $\left\langle \hat{A}\hat{z}, \hat{z} \right\rangle = 0$, with

$$\hat{A} = \begin{pmatrix} L & d \\ d^\top & d_0 \end{pmatrix} = \begin{pmatrix} \Lambda & O & O \\ O & O & d' \\ O & d'^\top & d_0 \end{pmatrix},$$

where at least one of d' and d_0 is zero. Compare (4.5) and (4.6) for examples.

This procedure enables us to completely classify all quadric curves for $n = 2$:

The signature of A is $(2, 0, 0)$ or $(0, 2, 0)$: Then we are in Case I, hence $d = 0$, and the curve is either an ellipse or a point or the empty set, depending on the sign of d_0.

The signature of A is $(1, 1, 0)$: Then we are in Case I, hence $d = 0$, and the curve is either a pair of straight lines or a hyperbola, depending on whether $d_0 = 0$ or $d_0 \neq 0$.

The signature of A is $(1, 0, 1)$ or $(0, 1, 1)$: Then we are in Case II. If $d' \neq 0$, then $d_0 = 0$ and we have a parabola. If $d' = 0$, then you get three exceptional cases. Figure them out yourselves.

The signature of A is $(0,0,2)$: Then A is the matrix full of zeros, and the curve is either a straight line or the empty set.

The situation is more intricate for $n = 3$. If we agree to neglect most of the exceptional cases, the following classification is found. Here $+$ and $-$ stand for positive of negative entries in $\hat{A}$; all other entries are zero. The numbers μ_j are always positive. We can always arrange that $\lambda_1 > 0$, otherwise we multiply the equation with -1.

$p + m = 3$: The vector d is always the null vector, and the sub-cases are:

$$\hat{A} = \begin{pmatrix} + & & & \\ & + & & \\ & & + & \\ & & & - \end{pmatrix} \quad \text{ellipsoid,} \quad \mu_1 x_1^2 + \mu_2 x_2^2 + \mu_3 x_3^2 - 1 = 0,$$

$$\hat{A} = \begin{pmatrix} + & & & \\ & \pm & & \\ & & - & \\ & & & \mp \end{pmatrix} \quad \text{hyperboloid of one sheet,} \quad \mu_1 x_1^2 + \mu_2 x_2^2 - \mu_3 x_3^2 - 1 = 0,$$

$$\hat{A} = \begin{pmatrix} + & & & \\ & \pm & & \\ & & - & \\ & & & \pm \end{pmatrix} \quad \text{hyperboloid of two sheets,} \quad \mu_1 x_1^2 - \mu_2 x_2^2 - \mu_3 x_3^2 - 1 = 0,$$

$$\hat{A} = \begin{pmatrix} + & & & \\ & \pm & & \\ & & - & \\ & & & 0 \end{pmatrix} \quad \text{elliptic double cone,} \quad \mu_1 x_1^2 + \mu_2 x_2^2 - \mu_3 x_3^2 = 0.$$

$p + m = 2$ and $d = 0$: The only reasonable cases are

$$\hat{A} = \begin{pmatrix} + & & & \\ & + & & \\ & & 0 & \\ & & & - \end{pmatrix} \quad \text{elliptic cylinder,} \quad \mu_1 x_1^2 + \mu_2 x_2^2 = 1,$$

$$\hat{A} = \begin{pmatrix} + & & & \\ & - & & \\ & & 0 & \\ & & & \pm \end{pmatrix} \quad \text{hyperbolic cylinder,} \quad \mu_1 x_1^2 - \mu_2 x_2^2 = 1.$$

$p + m = 2$ and $d \neq 0$: Then $d_0 = 0$; and the two sane cases are

$$\hat{A} = \begin{pmatrix} + & & & \\ & + & & \\ & & 0 & - \\ & & - & 0 \end{pmatrix} \quad \text{elliptic paraboloid,} \quad \mu_1 x_1^2 + \mu_2 x_2^2 = 2x_3,$$

$$\hat{A} = \begin{pmatrix} + & & & \\ & - & & \\ & & 0 & - \\ & & - & 0 \end{pmatrix} \quad \text{hyperbolic paraboloid,} \quad \mu_1 x_1^2 - \mu_2 x_2^2 = 2x_3.$$

$p + m = 1$ and $d \neq 0$: Then $d_0 = 0$; and the only interesting case is

$$\hat{A} = \begin{pmatrix} + & & & \\ & 0 & & - \\ & & 0 & \\ & - & & 0 \end{pmatrix} \quad \text{parabolic cylinder,} \quad \mu_1 x_1^2 = 2x_2.$$

Pictures of these surfaces can be found in [2].

4.6 Outlook: the Google PageRank Algorithm

Let us have a look at Google's PageRank algorithm, called after Larry Page, one of the founders of Google[24]. Here we only show some basic ideas; of course, the algorithm to compute the rank of a web page has been refined over the passing of the years, and naturally, some details are kept secret for understandable reasons ...

The rank of a web page shall describe its "importance", and this importance is related to the links from one web page to another. Intuitively, we make some rough assumptions:

- a web page which receives many links should have higher rank than a page which receives only a few links,
- if a web page A receives some links coming from other pages of high rank, and a web page B receives links coming from other pages of low rank, than the links pointing to A shall count more than the links pointing to B.

The page rank of a web page P shall be written as $r(P)$, which should be a positive number, and it should hold

$$r(P) = \sum_{Q\in B_P} \frac{r(Q)}{|Q|},$$

where B_P is the set of all the pages in the WWW which contain a hyperlink to the page P, and $|Q|$ is the total number of links which start on the page Q.

The unknown numbers $r(P)$ and $r(Q)$ appear on both sides of the equation, and for each web page P which is reachable in the WWW, we have one equation. Consequently, we have to solve a system of linear equations, where the matrix comes from $\mathbb{R}^{N\times N}$, with $N \approx 10^{10}$. Call these pages P_1, ..., P_N, and arrange their page ranks to a column vector $\pi = (r(P_1), \dots, r(P_N))^\top$. Then we wish to solve the system

$$\pi^\top = \pi^\top \mathbf{P},$$

where $\pi^\top$ is a row vector, and $\mathbf{P}$ is a matrix of size $N \times N$ with entries

$$p_{ij} = \begin{cases} \frac{1}{|P_i|} & : P_i \text{ links to } P_j, \\ 0 & : \text{ otherwise} \end{cases}$$

To obtain a notation more familiar, we transpose and get $\mathbf{P}^\top \pi = \pi$, which is an eigenvalue problem.

How to solve it ? How to find π ?

Due to the enormous size of the matrix, the only viable approach seems to be: guess an initial vector $\pi^{(0)}$, for instance $\pi_j^{(0)} = 1/N$ for each component, and then iterate $\pi^{(k+1)} := \mathbf{P}^\top \pi^{(k)}$, and hope for fast convergence as k goes to infinity. Natural questions are:

- is the problem solvable at all ? If 1 is *not* an eigenvalue of $\mathbf{P}^\top$, then the solution π can not exist !
- does the iteration converge ?
- does the iteration converge with reasonable speed ?

We start with the first question. All the columns of $\mathbf{P}^\top$ have sum equal to one, by definition of the p_{ij}. Then each of the columns of $(\mathbf{P}^\top - I)$ adds up to zero. Therefore, adding all the rows of the matrix $(\mathbf{P}^\top - I)$ produces the zero row vector. Then the rows of $(\mathbf{P}^\top - I)$ are linearly dependent, and the determinant of that matrix must be zero, and therefore one is an eigenvalue of $\mathbf{P}^\top$.

To consider the second question, we play with a toy model: take $N = 2$ and $\mathbf{P}^\top = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. If your initial vector is $\pi^{(0)} = (0.2, 0.8)^\top$, for instance, then the sequence of the $\pi^{(k)}$ does not converge, as can be computed by hand quickly. The reason is that $\mathbf{P}^\top$ has another eigenvalue, -1, which has the same

[24] Our representation follows the article of A.Langville, C. Meyer: *A survey of eigenvector methods for web information retrieval,* SIAM Review, Vol. 47, No. 1 (2005), 135–161.

modulus as our favourite eigenvalue +1. This is bad. We want that no eigenvalue has modulus larger than one, and +1 must be the only eigenvalue with modulus equal to one, and +1 shall be an eigenvalue of algebraic multiplicity one. (Think about why we wish all this !)

So we would like to know, where the other eigenvalues of $\mathbf{P}^\top$ are located. Here the GERSCHGORIN principle helps and tells us that all the eigenvalues are (possibly complex) numbers with modulus ≤ 1. (We apply the Gerschgorin principle to $\mathbf{P}$, but this matrix has the same eigenvalues as $\mathbf{P}^\top$). At least we have proved that eigenvalues of modulus larger than one are impossible !

Now comes the moment where we should modify the matrix $\mathbf{P}$ a bit. First: if a row of $\mathbf{P}$ contains only zeroes, then we replace each of these zeroes by $1/N$. This gives us a matrix $\overline{\mathbf{P}}$. Next, we set

$$\overline{\overline{\mathbf{P}}} := \alpha\overline{\mathbf{P}} + (1-\alpha)\mathbf{E}, \qquad 0 < \alpha < 1,$$

where $\mathbf{E} = \vec{e}\otimes\vec{v}$, and $\vec{e}$ is a column vector full of ones, and $\vec{v}$ is a special vector full of positive entries that sum up to one. This vector $\vec{v}$ can describe that a user might type in a popular URL by hand into the browser, or it can be used for manual adjustions of certain pageranks for political/commercial/whatever reasons (think of stopping link spammers). The number α is a parameter for fine tuning.

The key step is now: the matrix $\overline{\overline{\mathbf{P}}}$ has only positive entries, and then we can cite the FROBENIUS–PERRON theorem that tells us that the eigenvalue of biggest modulus is unique, its eigenvector has only positive entries, and all the other eigenvalues have smaller modulus. And it keeps getting better: if the eigenvalues of $\overline{\mathbf{P}}$ are $(1, \mu_1, \mu_2, \dots, \mu_N)$ with $|\mu_j| \leq 1$, then the eigenvalues of $\overline{\overline{\mathbf{P}}}$ are $(1, \alpha\mu_1, \alpha\mu_2, \dots, \alpha\mu_N)$. Google has chosen $\alpha \approx 0.85$, which implies that all the other eigenvalues are (in modulus) considerably smaller than one.

This is related to the question of the speed of convergence: if $\lambda_1 = 1$ is the largest eigenvalue, and $\lambda_2 \in \mathbb{C}$ is the second-largest (in modulus) eigenvalue of the iteration matrix $\overline{\overline{\mathbf{P}}}$, then the error in the vector π after k steps of iterations can be bounded by $(|\lambda_2|/\lambda_1)^k$ (times a constant). But $|\lambda_2| \leq 0.85$, by the choice of α, which makes the convergence quite fast. It is said that Google can compute the page rank vector π in just a few days, and a new version of π is computed about once per month.

Now you have an idea how to do “the world’s largest matrix computation”.

4.7 Keywords

- Eigenvalues, eigenvectors, and how to compute them,
- multiplicities of eigenvalues, diagonalisation of a matrix,
- orthogonal projectors, self-adjoint operators,
- spectral theorem,
- definiteness.

Chapter 5

Integration in Several Dimensions, and Surfaces

5.1 Integration on Cuboids

Definition 5.1 (Cuboid). *A set $Q \subset \mathbb{R}^n$ is called* cuboid[1] *if there are real numbers a_i, b_i with $-\infty < a_i < b_i < +\infty$ and*

$$Q = \left\{ x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n \colon a_i \le x_i \le b_i,\ i = 1, \dots, n \right\}.$$

The volume[2] *of this cuboid Q is defined as*

$$\operatorname{vol}(Q) := \prod_{i=1}^{n} (b_i - a_i).$$

We say that a collection of numbers

$$(x_{1,0}, x_{1,1}, \dots, x_{1,m_1}),\ (x_{2,0}, x_{2,1}, \dots, x_{2,m_2}),\ \dots, (x_{n,0}, x_{n,1}, \dots, x_{n,m_n}),$$

forms a grid-partition[3] *of the above Q if*

$$a_i = x_{i,0} < x_{i,1} < x_{i,2} < \dots < x_{i,m_{i-1}} < x_{i,m_i} = b_i, \qquad i = 1, \dots, n.$$

Given indices $1 \le j_1 \le m_1$, $1 \le j_2 \le m_2$, $\dots$, $1 \le j_n \le m_n$, we define an open sub-cuboid *$Q_{j_1 \dots j_n}$ as*

$$Q_{j_1 \dots j_n} := (x_{1,j_1-1}, x_{1,j_1}) \times (x_{2,j_2-1}, x_{2,j_2}) \times \dots \times (x_{n,j_n-1}, x_{n,j_n}).$$

The closures $\overline{Q_{j_1 \dots n_h}}$ of all sub-cuboids joined together give again Q:

$$Q = \bigcup_{j_1, \dots, j_n} \overline{Q_{j_1 \dots j_n}}.$$

Definition 5.2 (Step function). *A function f from a cuboid into the real numbers is said to be a* step function[4] *if a grid-partition of the cuboid exists, with the property that f is constant on each sub-cuboid.*

Note two things:

- we do not say anything about the values of a step-function on the borders of the sub-cuboids,
- for each step-function, you can find an infinite number of grid-partitions (just split the sub-cuboids once more).

[1] Quader

[2] Volumen

[3] Gitterzerlegung

[4] Treppenfunktion

Definition 5.3 (**Integral of step functions**). *Let $f\colon Q \to \mathbb{R}$ be a step function, taking the values $c_{j_1 j_2 \cdots j_n}$ at the sub-cuboid $Q_{j_1 j_2 \cdots j_n}$. Then the integral of f over Q is defined as*

$$\int_Q f(x)\,\mathrm{d}x := \sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\cdots\sum_{j_n=1}^{m_n} c_{j_1 j_2 \cdots j_n}\,\mathrm{vol}(Q_{j_1 j_2 \cdots j_n}).$$

Question: If different grid-partitions of Q gave different values of the integral, this definition would become absurd. Show that this cannot happen.

As in (3.1), any step function f satisfies the following estimate:

$$\left|\int_Q f(x)\,\mathrm{d}x\right| \le \|f\|_{L^\infty(Q)}\,\mathrm{vol}(Q). \tag{5.1}$$

We define *tame functions* as in one dimension:

Definition 5.4 (**Tame function**). *We call a function $f\colon Q \to \mathbb{R}$* tame[5] *if f is bounded and there is a sequence $(\varphi_m)_{m\in\mathbb{N}}$ of step functions which converges to f in the $L^\infty(Q)$–norm:*

$$\lim_{m\to\infty}\|\varphi_m - f\|_{L^\infty(Q)} = 0.$$

Definition 5.5 (**Integral of tame functions**). *Let $(\varphi_m)_{m\in\mathbb{N}}$ be a sequence of step functions which converges to a tame function in the $L^\infty(Q)$–norm. Then we define*

$$\int_Q f(x)\,\mathrm{d}x := \lim_{m\to\infty}\int_Q \varphi_m(x)\,\mathrm{d}x.$$

Question: If this limit did not exist, or if different sequences of step functions gave different limits, then this definition would become absurd, too. Show that this is impossible, taking advantage of (5.1).

The integral of tame functions over cuboids shares many properties with its one-dimensional counterpart:

Proposition 5.6.

- *The estimate (5.1) holds for tame functions, too.*
- *Continuous functions are tame.*
- *If f and g are tame functions over a cuboid Q and $f \le g$ everywhere, then $\int_Q f(x)\,\mathrm{d}x \le \int_Q g(x)\,\mathrm{d}x$.*
- *If f is tame, then so is $|f|$.*
- *If f and g are tame, then also $f\cdot g$ is tame.*
- *Each tame function f over a cuboid Q satisfies the estimate*

$$\left|\int_Q f(x)\,\mathrm{d}x\right| \le \int_Q |f(x)|\,\mathrm{d}x.$$

- *For $M = \sup\{f(x)\colon x\in Q\}$ and $m = \inf\{f(x)\colon x \in Q\}$, we have*

$$m\,\mathrm{vol}(Q) \le \int_Q f(x)\,\mathrm{d}x \le M\,\mathrm{vol}(Q).$$

- *If f and $g \ge 0$ are continuous, then there is a point $\xi \in Q$ with*

$$\int_Q f(x)g(x)\,\mathrm{d}x = f(\xi)\int_Q g(x)\,\mathrm{d}x.$$

Proof. See the one-dimensional version, Propositions 3.6, 3.9, and 3.11. □

[5] Regelfunktion

Proposition 5.7. *If a sequence $(f_m)_{m\in\mathbb{N}}$ of tame functions over a cuboid Q converges* uniformly *to a limit function f, then this limit function is tame, too, and we can commute the limit and the integration:*

$$\lim_{m\to\infty}\int_Q f_m(x)\,\mathrm{d}x = \int_Q \lim_{m\to\infty} f_m(x)\,\mathrm{d}x = \int_Q f(x)\,\mathrm{d}x.$$

Proof. The proof is exactly the same as in the 1D case, see Proposition 3.32. □

Now to the differences between integration in one dimension and integration in several dimensions:

- there are no anti-derivatives of a function over a cuboid,
- in the multi-dimensional case, *iterated integrals* can be considered.

Iterated integrals give us a means to calculate the value of the integral. Let $Q \subset \mathbb{R}^n = \mathbb{R}^n_x$ be a cuboid, and put $n = l + m$ with $1 \le l, m \le n-1$. Write

$$\begin{aligned}&\mathbb{R}^n = \mathbb{R}^l \times \mathbb{R}^m,\\ &x = (x_1,\dots,x_n) = (v_1,\dots,v_l,w_1,\dots,w_m) = (v,w),\\ &Q = O \times P,\end{aligned}$$

with a cuboid $O \subset \mathbb{R}^l$ and a cuboid $P \subset \mathbb{R}^m$.

Take a function $f \in C(Q \to \mathbb{R})$ and write it as $f = f(x) = f(v,w)$. If you freeze the variable v, you obtain a function $f_v = f_v(w)$ which depends on w only, is defined on P and continuous there. Then the integral $\int_P f_v(w)\,\mathrm{d}w$ makes sense. Note that this integral is a function of the variable v.

Proposition 5.8 (Fubini'[6]**s Theorem).** *Write a cuboid $Q \subset \mathbb{R}^n$ as $Q = O \times P$ as above, and pick a continuous function $f\colon Q \to \mathbb{R}$. For frozen $v \in O$ (and frozen $w \in P$), define a function f_v (and a function f_w) as*

$$\begin{aligned}&f_v\colon P \to \mathbb{R}, &&f_v\colon w \mapsto f(v,w),\\ &f_w\colon O \to \mathbb{R}, &&f_w\colon v \mapsto f(v,w).\end{aligned}$$

Then the following holds:

1. *The function f_v is a continuous function over P, and the function f_w is continuous over O.*
2. *The integrals $\int_P f_v(w)\,\mathrm{d}w$ and $\int_O f_w(v)\,\mathrm{d}v$ exist, for every v and w, respectively. The first integral is a continuous function over O, the second integral is a continuous function over P.*
3. *We have the identity*

$$\int_Q f(x)\,\mathrm{d}x = \int_O\left(\int_P f_v(w)\,\mathrm{d}w\right)\,\mathrm{d}v = \int_P\left(\int_O f_w(v)\,\mathrm{d}v\right)\,\mathrm{d}w. \tag{5.2}$$

Sketch of proof. Continuous functions are tame and can be approximated by step functions, for which the above claims are quite obvious. The *uniform* convergence of the step functions to tame functions allows to commute limits and integration operators (but you have to be careful in the details). □

You could try yourselves to prove the same result once more, replacing everywhere *continuous* by *tame*. Be warned that the devil is in the details, though. One such detail is that we have no information about the values of a step-function on the borders of the sub-cuboids, making claim 2 of the above proposition wrong, for some values of v and w . . .

Question: Take $Q = (0,1) \times (0,2)$. Which value has the integral $\int_Q y\sin(xy)\,\mathrm{d}(x,y)$?

[6] Guido Fubini, 1879 – 1943

5.2 Integration on Arbitrary Bounded Domains

We start with an example.

Let B be a two-dimensional disk with electrical charges on it. The charge density function is a continuous function $f = f(x)$, for $x \in B$. We expect the total charge to be

$$\int_B f(x)\,\mathrm{d}x,$$

and now we wish to make sense of this expression.

Our considerations of the previous section cannot be applied directly, since the disk is no cuboid. Another idea might be to choose a cuboid Q containing B and to define the zero extension f_0 of f,

$$f_0(x) := \begin{cases} f(x) & : x \in B, \\ 0 & : x \in Q \setminus B. \end{cases}$$

Our naive intuition suggests that $\int_B f(x)\,\mathrm{d}x$ can be defined as $\int_Q f_0(x)\,\mathrm{d}x$. But this does not work, because f_0 is not a tame function on the cuboid Q. There is no sequence of step functions on Q which converges uniformly to f_0, except in the artificial case that f vanishes on the boundary ∂B, making f_0 continuous on Q. In general, f_0 will be discontinuous on ∂B.

However, we can argue that the function f_0 is *almost tame*: if we stay a bit away from the bad part ∂B, the function is tame there. And the bad part ∂B has no two-dimensional area, because it is a one-dimensional object. This hints at how to overcome the trouble: just cut-off the boundary ∂B.

Definition 5.9 (**Sets of JORDAN–measure zero**). *We say that a set $\Gamma \subset \mathbb{R}^n$ has* n-dimensional Jordan-measure zero[7] *if, for every positive ε, you can find a finite number of cubes with total volume less than ε whose union covers Γ.*

The boundary of a two-dimensional disk is a set with two-dimensional Jordan-measure zero.

Definition 5.10 (**JORDAN-measurable**). *We say that a set $G \subset \mathbb{R}^n$ is* Jordan-measurable[8] *if it is bounded in $\mathbb{R}^n$, and its boundary ∂G has n-dimensional Jordan-measure zero.*

Definition 5.11 (**ε-boundary-cut-off**). *Let $G \subset \mathbb{R}^n$ be an open Jordan-measurable set. Take a cube $Q \subset \mathbb{R}^n$ that contains G and a grid-partition of Q with sub-cubes of equal size.*

- *The sub-cubes whose closure intersects ∂G are the* boundary sub-cubes*.*
- *The sub-cubes that are contained in G and are neighbours of boundary sub-cubes are the* intermediate sub-cubes*. We make the agreement that two sub-cubes are neighbours if they have at least one vertex[9] in common.*
- *All other sub-cubes that are contained in G are the* interior sub-cubes*.*
- *All sub-cubes whose closure does not intersect $\overline{G}$ are the* exterior sub-cubes*.*

A function $\varphi_\varepsilon \colon Q \to \mathbb{R}$ with $\varepsilon > 0$ is called an ε-boundary-cut-off *function of the domain G if the following conditions are met:*

- *There is a cube Q of $\mathbb{R}^n$ which contains G and a grid-partition of Q, such that the total volume of the boundary sub-cubes and intermediate sub-cubes is less than ε,*
- *the function φ_ε takes the value 0 on the exterior and boundary sub-cubes, the value 1 on the interior sub-cubes, and values between zero and one on the intermediate sub-cubes,*
- *the function φ_ε is continuous on Q.*

One can show (we will not do it) that, for any Jordan-measurable set, such boundary-cut-off functions can always be found. See Figure 5.1 for an example.

[7] n-dimensionales Null-Jordan-Maß

[8] Jordan-meßbar

[9] Eckpunkt

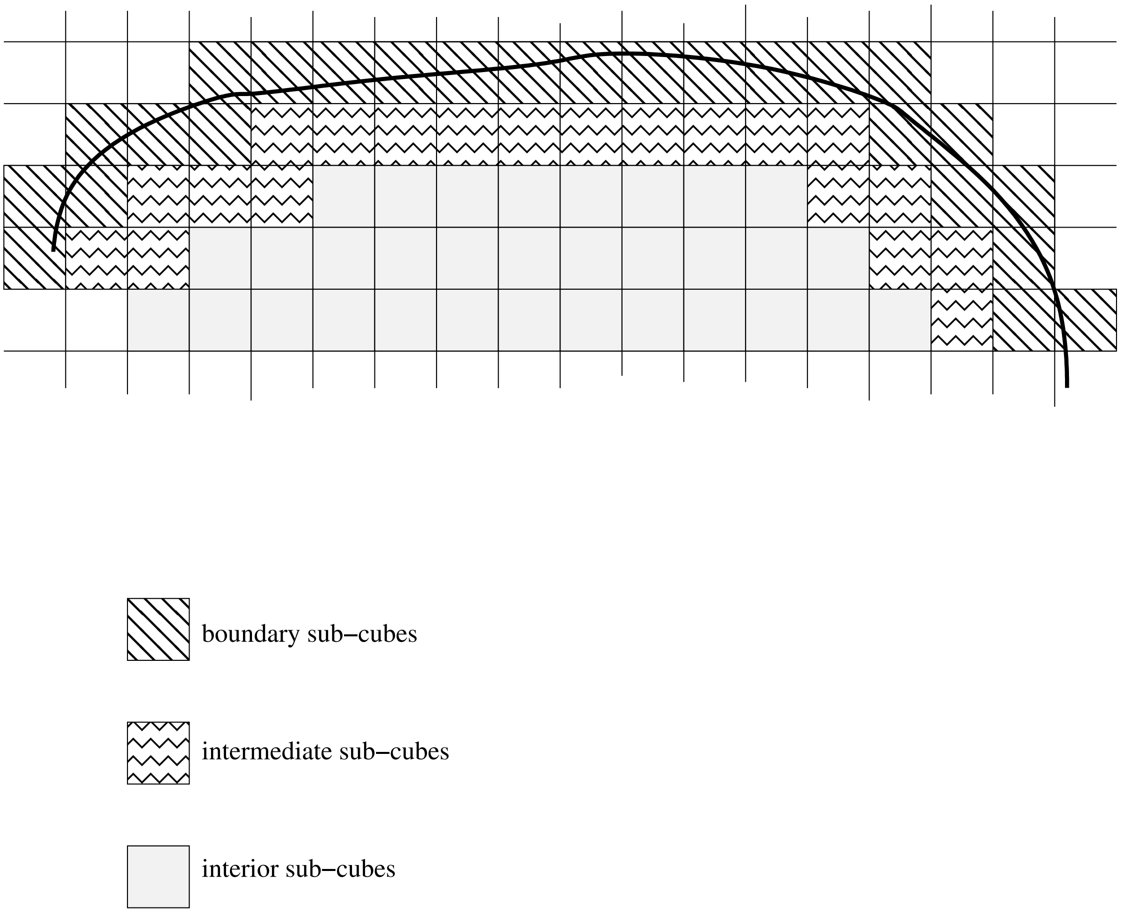


Figure 5.1: A (part of a) grid partition of $G \subset \mathbb{R}^2$. The fat curved line is ∂G, and G lies "below" it. The exterior sub-cubes are the white ones "above" the fat curved line. The cut-off-function φ_ε is zero on the boundary sub-cubes, and one on the interior sub-cubes.

Definition 5.12 (**Integral over bounded domains**). *Let G be an open Jordan-measurable set in $\mathbb{R}^n$. For any positive ε, let φ_ε be an ε-boundary-cut-off. Then the integral $\int_G f(x)\,\mathrm{d}x$ over G of a continuous function $f\colon \overline{G} \to \mathbb{R}$ is defined as*

$$\int_G f(x)\,\mathrm{d}x := \lim_{\varepsilon\to+0} \int_Q \varphi_\varepsilon(x) f(x)\,\mathrm{d}x,$$

where Q is a cube containing G, and the function f has been tacitly zero-extended to Q.

Take two boundary-cut-off functions φ_ε and $\varphi_{\varepsilon'}$. They coincide everywhere except a set with volume at most $\varepsilon + \varepsilon'$. Then the estimate

$$\left| \int_Q \varphi_\varepsilon(x) f(x)\,\mathrm{d}x - \int_Q \varphi_{\varepsilon'}(x) f(x)\,\mathrm{d}x \right| \le (\varepsilon + \varepsilon') \left\| f \right\|_{L^\infty(G)}$$

follows and convinces us that the sequence $(\int_Q \varphi_\varepsilon f\,\mathrm{d}x)_{\varepsilon\to+0}$ is a Cauchy sequence of real numbers. Therefore, the limit in Definition 5.12 exists, making it a sane definition.

Definition 5.13 (**Volume of a bounded domain**). *The volume of an open Jordan-measurable set G is defined as*

$$\mathrm{vol}(G) := \int_G 1\,\mathrm{d}x.$$

Proposition 5.14. *Let G be an open Jordan-measurable set in $\mathbb{R}^n$, and $f, g\colon \overline{G} \to \mathbb{R}$ be continuous. Then the assertions of Propositions 5.6 hold (replace Q with G everywhere, and "tame" with "continuous").*

Proof. Omitted. □

You can also consider iterated integrals. Just extend the integrand by zero outside the domain G and write down the equation (5.2).

Lemma 5.15 (Fubini). *Let $[a, b] \subset \mathbb{R}$ be a bounded interval, and $\psi_\pm \in C([a, b] \to \mathbb{R})$ two continuous functions with $\psi_-(x) < \psi_+(x)$, for $x \in (a, b)$. Define a set*

$$G = \{(x, y) \in \mathbb{R}^2 \colon a < x < b,\ \psi_-(x) < y < \psi_+(x)\}.$$

Then this set is open and Jordan-measurable in $\mathbb{R}^2$, and the integral of every continuous function f over G can be computed as

$$\int_G f(x, y)\,\mathrm{d}(x, y) = \int_{x=a}^{x=b} \left(\int_{y=\psi_-(x)}^{y=\psi_+(x)} f(x, y)\,\mathrm{d}y \right) \mathrm{d}x.$$

How to prove it should be obvious: pick a small ε; cut-off the boundary of G with a cut-off function φ_ε; then you can read the integral over G as an integral over a larger cube (by zero-extension of f); for such an integral iterated integration is permitted. Then you send ε to zero and commute $\lim_{\varepsilon \to 0}$ with the integral symbols (carefully checking that you are allowed to do that).

Of course, you can switch the roles of x and y, as well as formulate similar lemmas in higher dimensions.

Example: *Let $G = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 < R^2\}$ be a ball of radius R, and compute the integral*

$$\int_G (x^2 + y^2)\,\mathrm{d}(x, y).$$

Differentiation under the integral is possible in higher dimensions, too:

Proposition 5.16 (**Differentiation with respect to parameters**). *Let $\Lambda \subset \mathbb{R}$ be a compact interval and $G \subset \mathbb{R}^n$ be an open Jordan-measurable set. Assume that the function $f \colon \overline{G} \times \Lambda \to \mathbb{R}$ is continuously differentiable. Then the function*

$$g = g(\lambda) = \int_G f(x, \lambda)\,\mathrm{d}x$$

maps Λ into $\mathbb{R}$, is continuously differentiable there, and has derivative

$$g'(\lambda) = \int_G \frac{\partial f}{\partial \lambda}(x, \lambda)\,\mathrm{d}x.$$

We have already skipped the proof of the one-dimensional version, so we should skip the proof now, too.

Proposition 5.17 (**Substitution**). *Let H be an open set in $\mathbb{R}^n$, and let $\varphi \colon H \to \mathbb{R}^n$ be a function with the following properties:*

- *φ is C^1 on H,*
- *φ is injective,*
- *the Jacobi matrix φ' is always regular.*

Let $G \subset \mathbb{R}^n$ be an open Jordan-measurable set, with $\overline{G}$ being contained in H (then G is a strictly smaller set than H). Finally, let f be a continuous function $f \colon \varphi(\overline{G}) \to \mathbb{R}$. Then the following holds:

- *the set $\varphi(G)$ is Jordan-measurable and open,*
- *the substitution formula holds:*

$$\int_{\varphi(G)} f(y)\,\mathrm{d}y = \int_G f(\varphi(x))\,|\det \varphi'(x)|\,\mathrm{d}x.$$

The formula remains true if φ' becomes singular on a subset of G with Jordan-measure zero.

Proof. Consult [4, Vol. 2, Nr. 205] for the gory details. □

Example 5.18 (Polar coordinates). *We come back to the integral $\int_B (x^2+y^2)\,\mathrm{d}(x,y)$ where B is a ball about 0 of radius R. The function φ is $\varphi = \varphi(r,\phi) = (x,y)$ with*

$$x = r\cos\phi, \qquad y = r\sin\phi.$$

This gives $\det\varphi' = r$. The ball B is described by $\{(r,\phi) : 0 \le \phi \le 2\pi,\ 0 \le r < R\}$. We choose $f = f(x,y) = x^2+y^2$ and obtain, by Fubini's theorem,

$$\int_B (x^2+y^2)\,\mathrm{d}(x,y) = \int_{\phi=0}^{\phi=2\pi}\int_{r=0}^{r=R} r^2\cdot r\,\mathrm{d}(r,\phi) = 2\pi\frac{R^4}{4}.$$

5.3 Integration on Unbounded Domains

Now we investigate integrals $\int_G f(x)\,\mathrm{d}x$ with unbounded domains G. You can see them as multi-dimensional analogues to improper integrals.

Definition 5.19 (Integrability). *Let G be a (possibly unbounded) open set in $\mathbb{R}^n$, and Q_R the centred cube with side length $2R$:*

$$Q_R := \left\{x = (x_1,\dots,x_n)^\top \in \mathbb{R}^n\colon\ -R < x_j < R,\ j = 1,\dots,n\right\}.$$

We assume that $G\cap Q_R$ is Jordan-measurable, for every positive R, and say that a bounded function $f \in C(G\to\mathbb{R})$ is integrable over *G if the following limit exists:*

$$\lim_{R\to\infty}\int_{G\cap Q_R} |f(x)|\,\mathrm{d}x < \infty.$$

Question: If a function f is integrable over G, then also the limit

$$\lim_{R\to\infty}\int_{G\cap Q_R} f(x)\,\mathrm{d}x$$

exists. Why ?

Definition 5.20 (Integral). *Let the function f be bounded, continuous and integrable over G; then we define*

$$\int_G f(x)\,\mathrm{d}x := \lim_{R\to\infty}\int_{G\cap Q_R} f(x)\,\mathrm{d}x.$$

Proposition 5.21. *Take two functions f and g that are continuous, bounded and integrable over G.*

- *Then also the functions $\alpha f + \beta g$ are continuous, bounded and integrable over G, and we have $\int_G(\alpha f+\beta g)\,\mathrm{d}x = \alpha\int_G f\,\mathrm{d}x + \beta\int_G g\,\mathrm{d}x$.*
- *The functions $|f|$ and $f\cdot g$ are integrable over G, too.*
- *If $f \le g$ everywhere, then $\int_G f\,\mathrm{d}x \le \int_G g\,\mathrm{d}x$.*
- *In particular, we have $|\int_G f\,\mathrm{d}x| \le \int_G |f|\,\mathrm{d}x$.*

That was the easy part.

Question: Take $f = f(x,y) = \exp(-|y|(1+x^2))$. Check the integrability of f over $G = \mathbb{R}^2$, and compute the integrals $\int_{\mathbb{R}}(\int_{\mathbb{R}} f\,\mathrm{d}x)\,\mathrm{d}y$ and $\int_{\mathbb{R}}(\int_{\mathbb{R}} f\,\mathrm{d}y)\,\mathrm{d}x$ if you can.

Observe that the integral $\int_{x=-\infty}^{x=+\infty} f(x,y=0)\,\mathrm{d}x$ does not exist in this example, which should discourage you from applying Fubini's rule mindlessly.

Stronger assumptions will rescue Fubini's rule, as we will see now. For clarity of exposition, we consider only the case $G = \mathbb{R}^2$. You can write down a version for $G = \mathbb{R}^n = \mathbb{R}^l\times\mathbb{R}^m$ yourselves.

Proposition 5.22 (Fubini's **Theorem**). *Take a bounded function $f \in C(\mathbb{R}^2 \to \mathbb{R})$. Suppose that, for every compact interval $I \subset \mathbb{R}_x$, a continuous and bounded function $g_I \colon \mathbb{R}_y \to \mathbb{R}$ exists, such that*

$$|f(x,y)| \le g_I(y), \qquad (x,y) \in I \times \mathbb{R}_y, \qquad \int_{y=-\infty}^{y=+\infty} g_I(y)\,\mathrm{d}y < \infty. \tag{5.3}$$

Then the following holds:

- *The following functions exist and are continuous on $\mathbb{R}$:*

$$F(x) := \int_{y=-\infty}^{y=+\infty} f(x,y)\,\mathrm{d}y, \qquad F_{|\cdot|}(x) := \int_{y=-\infty}^{y=+\infty} |f(x,y)|\,\mathrm{d}y, \qquad x \in \mathbb{R}.$$

- If *the function $F_{|\cdot|}$ is integrable over $\mathbb{R}$, then the function F is integrable over $\mathbb{R}$, and the function f is integrable over $\mathbb{R}^2$. In this case we have the equivalence*

$$\int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}(x,y) = \int_{\mathbb{R}_x} \left(\int_{\mathbb{R}_y} f(x,y)\,\mathrm{d}y \right) \mathrm{d}x.$$

Warning: *We did not assert that you can switch the order of integration, i.e.*

$$\int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}(x,y) \overset{?}{=} \int_{\mathbb{R}_y} \left(\int_{\mathbb{R}_x} f(x,y)\,\mathrm{d}x \right) \mathrm{d}y,$$

because it would be wrong. For this equivalence, you need a counterpart of (5.3) with the roles of x and y interchanged.

You can probably guess the idea of the proof: first you replace $\mathbb{R}^2$ by a big square Q_R, for which the classical Fubini rule is valid. Then you send R to infinity and try to commute the $\lim_R$–operator with an integral symbol. This is the point where you need (5.3). The details are left to the student.

5.4 Surfaces

Literature: Greiner and Stock: *Hydrodynamik.* Chapter 17: Mathematische Ergänzung: Zur Theorie der Flächen

5.4.1 Definition and Examples

The typical example of a surface is the upper hemisphere of a ball in the usual three-dimensional space. In some sense, surfaces in $\mathbb{R}^3$ are like (images of) curves in $\mathbb{R}^2$. We recall the ingredients for a curve:

parametrisation: this is a continuous mapping $\gamma \colon [a,b] \to \mathbb{R}^2$,

parameter domain: this is the interval $[a,b]$ where the parameter t lives in,

image: this is the set $\{\gamma(t) \colon t \in [a,b]\} \subset \mathbb{R}^2$, (different parametrisations can give the same image).

Understandably, we wish to forbid such monsters as the Peano curve. To this end, we requested so-called *regularity conditions*:

- the parameter domain should be compact and connected (hence, a closed and bounded interval),
- the mapping γ should be injective, at least in the open interval (a,b) (but we want to allow $\gamma(a) = \gamma(b)$, for we can not consider loops otherwise),
- the parametrisation γ should be differentiable with continuous derivative $\dot\gamma$, and this derivative must never be the null vector.

If these conditions hold, the curves behave as expected: they have a tangent vector at every point, they have a finite length, the images have no *corner points*[10], and you can consider curve integrals.

Surfaces (more precisely, surface *patches*) are similar. The main differences are that the parameter domain is a subset of $\mathbb{R}^2$, and the image is a subset of $\mathbb{R}^3$.

Definition 5.23 (**Surface patch**)**.** *A set $S \subset \mathbb{R}^3$ is said to be a* surface patch[11] *if a set $U \subset \mathbb{R}^2$ and a continuous mapping $\Phi\colon U \to \mathbb{R}^3$ exists with $\Phi(U) = S$. The set U is called* parameter domain[12]*, and the mapping Φ is named* parametrisation. *The following conditions on U and Φ are assumed:*

- *the parameter domain U is compact and connected; the boundary ∂U has Jordan-measure zero,*
- *the mapping Φ is injective on the interior $\Omega := U \setminus \partial U$, and $\Phi(\Omega) \cap \Phi(\partial U) = \emptyset$,*
- *the derivative Φ' (Jacobi matrix) is continuous on U and has rank 2 everywhere in U.*

Question: Let $u = (u_1, u_2)^\top$ denote the parameters of $u \in U$, and write Φ as a column vector $(\Phi_1, \Phi_2, \Phi_3)^\top$. What is the geometrical meaning of the two column vectors of Φ' ?

Example 5.24 (Plane)**.** *Take three vectors a_0, a_1, $a_2 \in \mathbb{R}^3$ with a_1 and a_2 being linearly independent, and define the parametrisation*

$$\Phi(u) = a_0 + u_1 a_1 + u_2 a_2, \qquad u \in U.$$

Then $\partial_{u_1}\Phi = a_1$ and $\partial_{u_2}\Phi = a_2$.

Example: *The* lateral surface[13] *of a cylinder: choose $U = \{0 \le \varphi \le 2\pi\} \times \{0 \le z \le 1\}$ and*

$$\Phi(u) = \begin{pmatrix} \cos\varphi \\ \sin\varphi \\ z \end{pmatrix}, \qquad u \in U.$$

How about the complete surface of a cylinder or the surface of a cone ? Are they surface patches ?

Example: *Take $U = \{0 \le \varphi \le 2\pi\} \times \{0 \le \theta \le \pi\}$ and*

$$\Phi(u) = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}.$$

This describes the unit sphere, but violates the last condition of Definition 5.23.

The famous HEDGEHOG Theorem[14] [15] states that this is unavoidable—it is impossible to find a parametrisation of the whole sphere. Instead you parametrise the lower and the upper hemisphere separately (for instance), and then glue the pieces together. This is where the name *surface patch* comes from.

Example: *For a parameter domain $U \subset \mathbb{R}^2$ and a C^1 function $f\colon U \to \mathbb{R}$, put*

$$\Phi(u) = \begin{pmatrix} u_1 \\ u_2 \\ f(u_1, u_2) \end{pmatrix}.$$

This describes the surface patch "generated by the function f over the domain U". Obviously,

$$\frac{\partial\Phi}{\partial u_1} = \begin{pmatrix} 1 \\ 0 \\ f_{u_1} \end{pmatrix}, \qquad \frac{\partial\Phi}{\partial u_2} = \begin{pmatrix} 0 \\ 1 \\ f_{u_2} \end{pmatrix}.$$

[10] Knick

[11] Flächenstück

[12] Parametergebiet

[13] Mantel

[14] Satz vom Igel

[15] You cannot comb a hedgehog so that all its prickles stay flat; there will be always at least one singular point, like the head crown.

Question: Give one more parametrisation of the upper hemisphere, in the spirit of the last example.

Combining the last two examples, we now have two different parametrisations for the upper hemisphere: one via polar coordinates with parameter domain $U = \{0 \leq \varphi \leq 2\pi\} \times \{0 \leq \theta \leq \pi/2\}$, the other via cartesian coordinates with parameter domain $V = \{(x, y) \colon x^2 + y^2 \leq 1\}$. Which one you choose is up to you; and, depending on the goal you are trying to achieve, you select that one which makes your calculations easier.

It turns out that the mapping which translates between polar parameters (φ, θ) and cartesian parameters (x, y) of the upper hemisphere is invertible and differentiable, with differentiable inverse mapping:

Proposition 5.25. *If U and V are parameter domains, and $\Phi \colon U \to \mathbb{R}^3$, $\Psi \colon V \to \mathbb{R}^3$ two parametrisations of the same surface patch S, then there is a C^1* diffeomorphism[16] *$\tau \colon U \to V$ with $\Phi = \Psi \circ \tau$. This diffeomorphism τ is named* parameter transform[17] *between Φ and Ψ.*

Proof. Omitted. But it is not that hard. □

5.4.2 Tangential Vectors and Tangential Planes

Everyone has a rough understanding what a tangential plane at a point on a sphere geometrically means: it is the plane that gives the "closest approximation of the sphere near that point". However, if you want a more precise description of a tangential plane, or if you want to compute something, you will have to use parametrisations of the surface under consideration.

Note that there is a tricky point: we have already seen that the same surface can have different parametrisations. Somehow we should also describe tangential planes via parametrisations. It would be quite bad if two different parametrisations of the same surface patch would lead to differing tangential planes. We would not know which of them is the right one.

Therefore, our job is now the following: how to define tangential planes of a surface patch at a point (by means of a parametrisation), in such a way that different parametrisations of the surface patch agree on what the tangential plane is.

To start with the easy things, we consider tangential *vectors* on a surface patch first:

Consider a surface patch S with a parameter domain U and parametrisation $\Phi \colon U \to S$. Pick a point $u_0 = (u_{1,0}, u_{2,0})$ in the parameter domain U; call $x_0 = \Phi(u_0)$ the associated point on S. To give a formula for a tangential vector on S at the point x_0, we take a short curve $\gamma = \gamma(\tau) = (u_1(\tau), u_2(\tau))$ in U, where $|\tau| < \varepsilon \ll 1$ and $\gamma(\tau = 0) = u_0$. Then $\Phi \circ \gamma$ is a curve on S. The tangent vector of that curve, taken at the point $x_0 = \Phi(u_0)$, is

$$\vec{t} = \frac{\partial}{\partial \tau} \Phi(\gamma(\tau))_{|\tau=0} = \Phi'(u_0) \cdot \gamma'(0) = (\partial_{u_1} \Phi) \cdot u_1'(0) + (\partial_{u_2} \Phi) \cdot u_2'(0).$$

We obtain different tangent vectors at x_0 if we let $(u_1'(0), u_2'(0))$ vary. If for instance the curve γ runs horizontally with unit speed through the point u_0 in the parameter domain U, then $(u_1'(0), u_2'(0)) = (1, 0)$, and the tangential vector becomes $\vec{t} = \partial_{u_1} \Phi(u_0)$. Similarly for $\partial_{u_2} \Phi(u_0)$. Our requirement that the rank of Φ' be two everywhere simply means that those two tangent vectors are linearly independent.

Then we are tempted to define the tangential plane at x_0 on the surface patch S as the plane that goes through the point $x_0 = \Phi(u_0)$ and is spanned by the vectors $\partial_{u_1} \Phi(u_0)$ and $\partial_{u_2} \Phi(u_0)$.

Next we have to check the independence of this plane from the choice of parametrisation. Take another parametrisation $\Psi \colon V \to \mathbb{R}^3$ of S, with $x_0 = \Psi(v_0)$. We have to verify that the vectors $\partial_{v_1} \Psi(v_0)$, $\partial_{v_2} \Psi(v_0)$ span the same plane as $\partial_{u_1} \Phi(u_0)$, $\partial_{u_2} \Phi(u_0)$. Equivalently, we may verify that the cross products of the spanning vectors point along the same line:

Proposition 5.26 (Re-parametrisation). *Let $\Phi \colon U \to \mathbb{R}^3$ and $\Psi \colon V \to \mathbb{R}^3$ be two parametrisations of the same surface S and $\tau \colon U \to V$ a diffeomorphism with $\Phi = \Psi \circ \tau$. Then we have:*

$$(\partial_{u_1} \Phi) \times (\partial_{u_2} \Phi) = \det\left(\tau'\right) \left((\partial_{v_1} \Psi) \times (\partial_{v_2} \Psi)\right). \tag{5.4}$$

[16] Diffeomorphismus. This is a C^1 mapping that is bijective from U onto V.

[17] Parametertransformation

Proof. The chain rule of differentiation says

$$\Phi' = \Psi' \cdot \tau'. \tag{5.5}$$

In this formula, Φ' and Ψ' are matrices with 2 columns and 3 rows, and τ' is a 2×2–matrix. Since τ is a bijective mapping, the derivative τ' is always an invertible matrix. Choose an arbitrary column vector $a \in \mathbb{R}^3$. Then (5.5) can be extended to an identity of matrices from $\mathbb{R}^{3\times 3}$:

$$\left(a, \partial_{u_1}\Phi, \partial_{u_2}\Phi\right) = \left(a, \partial_{v_1}\Psi, \partial_{v_2}\Psi\right) \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & (\tau')_{11} & (\tau')_{12} \\ 0 & (\tau')_{21} & (\tau')_{22} \end{pmatrix}.$$

Taking determinants on both sides, and interpreting two of them as *parallelepipedial products*[18] imply

$$\begin{aligned} &\det\left(a, \partial_{u_1}\Phi, \partial_{u_2}\Phi\right) = \det\left(a, \partial_{v_1}\Psi, \partial_{v_2}\Psi\right) \cdot \det(\tau'), \\ &\langle a, (\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\rangle = \langle a, (\partial_{v_1}\Psi) \times (\partial_{v_2}\Psi)\rangle \cdot \det(\tau'). \end{aligned}$$

The vector a is completely arbitrary; then (5.4) follows immediately. □

Of course, the cross product of the spanning vectors of a plane is just a normal vector of that plane. This justifies the next definition, and the obtained tangent plane is independent of the parametrisation.

Definition 5.27 (**Tangential plane of a surface patch**). *Let $S \subset \mathbb{R}^3$ be a surface patch, parametrised by a mapping Φ with parameter domain U. Then the tangential plane at a point $x_0 = \Phi(u_0) \in S$ is defined as the plane that goes through x_0 and has normal vector $\partial_{u_1}\Phi(u_0) \times \partial_{u_2}\Phi(u_0)$.*

Definition 5.28 (**Unit normal vector**). *Let $S \subset \mathbb{R}^3$ be a surface patch with parametrisation Φ. Then*

$$n(x) := \frac{(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)}{\|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\|}(u), \qquad \text{with } x = \Phi(u)$$

is called unit normal vector[19].

This unit normal vector is, at the same time, a normal vector to the tangent plane and a normal vector to the surface patch. Proposition 5.26 guarantees that the unit normal vector changes at most its sign when we choose to parametrise the surface in another way.

Definition 5.29 (**Orientation**). *Two parametrisations Φ and Ψ of a surface patch $S \subset \mathbb{R}^3$ are said to have the* same orientation[20] *if the parameter transform τ with $\Phi = \Psi \circ \tau$ has a Jacobi-Matrix τ' with positive determinant. Otherwise, we say that Φ and Ψ have* opposite orientations[21].

All parametrisations form two classes. Elements in the same class have the same orientation. We pick one class and call it positive orientation. *Then we talk about an* oriented surface patch *S.*

In practice, you often have a surface that is too big to be parametrised by only one surface patch (take the sphere, for instance). In such situations, it is standard to take several surface patches, each of them parametrising only a part of the surface, and glue these patches together. Of course, you do not want the unit normal vector to flip when you go from one patch to its neighbour patch.

Proposition 5.30 (**Unit normal vector field**). *An oriented surface patch S has, for each point $x \in S$, a uniquely determined unit normal vector*

$$n(x) = \pm\frac{(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)}{\|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\|}(u), \qquad \text{with } x = \Phi(u),$$

where $+$ stands for positive orientations and $-$ for negative orientations. The mapping $x \mapsto n(x)$ is called unit normal vector field[22] *on S.*

[18] Spatprodukte
[19] Einheitsnormalenvektor
[20] gleichorientiert
[21] entgegengesetzt orientiert
[22] Einheitsnormalenfeld

Definition 5.31 (**Orientable**). *A surface is said to be* orientable[23] *if a unit normal vector field on that surface can be chosen in such a way that the normal vectors vary continuously on the surface.*

There are plenty of examples of non-orientable surfaces, the most prominent being the Möbius[24] strip, showing that a surface may have "only one side".

Example: *Consider the unit sphere:*

$$\Phi(\varphi,\theta) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)^\top, \qquad 0 \le \varphi \le 2\pi, \quad 0 \le \theta \le \pi.$$

We easily compute

$$\partial_\varphi \Phi = \begin{pmatrix} -\sin\theta\sin\varphi \\ \sin\theta\cos\varphi \\ 0 \end{pmatrix}, \qquad \partial_\theta \Phi = \begin{pmatrix} \cos\theta\cos\varphi \\ \cos\theta\sin\varphi \\ -\sin\theta \end{pmatrix},$$

$$\big((\partial_\varphi\Phi) \times (\partial_\theta\Phi)\big)(\varphi,\theta) = -\sin\theta \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix} = -\sin\theta\Phi(\varphi,\theta).$$

This normal vector points to the interior of the ball, except at the north and south poles, where it vanishes (which makes this parametrisation illegal at those points).

5.4.3 Outlook: General Relativity

Roughly spoken, Einstein's theory of relativity tells us that our world is four-dimensional, and large masses make this four-dimensional space curved. Then this space is no longer a flat space that can be described by four cartesian coordinates, but it is a four-dimensional manifold which has a curvature.

In this section we try to explain what this means.

We will follow the *Einstein summation convention*: when you see a single term or a product with one index appearing as lower index and as upper index at the same time, then you have to take the summation over that index. For example, the expressions

$$a_j^j, \qquad a^j b_j^k, \qquad a^j b_l^k c^l d_{mk}$$

are to be understood as

$$\sum_{j=1}^n a_j^j, \qquad \sum_{j=1}^n a^j b_j^k, \qquad \sum_{l=1}^n \sum_{k=1}^n a^j b_l^k c^l d_{mk},$$

where n is the space dimension.

The spatial variables always have upper indices, and in a partial derivative $\frac{\partial y^k}{\partial x^m}$, k is an upper index, and m is a lower index.

Next we consider two coordinate systems describing the same point:

$$(x^1, \dots, x^n), \qquad (x'^1, \dots, x'^n).$$

For instance, (x^1, x^2, x^3) could be the usual cartesian coordinates in $\mathbb{R}^3$, and (x'^1, x'^2, x'^3) could be the polar coordinates:

$$(x'^1, x'^2, x'^3) = (r, \theta, \varphi).$$

Now let us be given a physical quantity that depends on the space variables like a velocity field or a density field (like mass per volume), and now we will study how the *coordinate representation* of this quantity changes when we switch the coordinate system.

[23] orientierbar

[24] August Ferdinand Möbius, 1790–1868

Thinking of a moving particle, we should consider time derivatives of the coordinate functions, and they transform (due to the chain rule) like this:

$$\frac{\mathrm{d}x'^j}{\mathrm{d}t} = \frac{\partial x'^j}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}t} \qquad \text{(Einstein summation convention).}$$

And a scalar function f depending on the spatial variables has partial derivatives which transform as follows (again by the chain rule):

$$\frac{\partial f}{\partial x'^j} = \frac{\partial x^i}{\partial x'^j}\frac{\partial f}{\partial x^i}.$$

Observe that these two transformation rules are quite different, and in particular the matrix of the $\frac{\partial x'^j}{\partial x^i}$ is the inverse of the matrix of the $\frac{\partial x^i}{\partial x'^j}$.

Definition 5.32. *A* contravariant tensorfield *is a function that maps (depending on the coordinate system) a point P to numbers $a^1(P)$, ..., $a^n(P)$ which transform according to*

$$a'^j = \frac{\partial x'^j}{\partial x^i}a^i$$

when we switch from the x–coordinate system to the x'–coordinate system.

A covariant tensorfield *is a function that maps (depending on the coordinate system) a point P to numbers $a_1(P)$, ..., $a_n(P)$ which transform according to*

$$a'_j = \frac{\partial x^i}{\partial x'^j}a_i$$

when we switch from the x–coordinate system to the x'–coordinate system.

Be careful: a mathematical object with upper indices need not be a contravariant tensor field, and an object with lower indices need not be a covariant tensor field. The key to the definition is always the transformation rule !

There are also twice contravariant tensor fields a^{jk}, and they transform according to the formula

$$a'^{jk} = \frac{\partial x'^j}{\partial x^l}\frac{\partial x'^k}{\partial x^m}a^{lm}.$$

Similarly you can define tensor fields $a^{j_1\dots j_p}_{k_1\dots k_q}$ which are p–fold contravariant and q–fold covariant.

Before we come to examples, a few remarks on the notations are in order. Geometric points and vectors will be written in bold letters, and the geometric scalar product of two vectors **a** and **b** will be expressed as **ab**. The function which maps the coordinates of a geometric point to that point is written as

$$\mathbf{r} = \mathbf{r}(x^1, x^2, x^3), \qquad \mathbf{r} = \mathbf{r}(x'^1, x'^2, x'^3).$$

In $\mathbb{R}^3$, fix an orthonormal frame of base vectors **i**, **j**, **k** (these will never change neither their meaning nor their directions).

In case of polar coordinates $(x^1, x^2, x^3) = (r, \theta, \varphi)$ we then have

$$\begin{aligned}\mathbf{r} &= r\sin\theta\cos\varphi\,\mathbf{i} + r\sin\theta\sin\varphi\,\mathbf{j} + r\cos\theta\,\mathbf{k}\\ &= x^1\sin x^2\cos x^3\,\mathbf{i} + x^1\sin x^2\sin x^3\,\mathbf{j} + x^1\cos x^2\,\mathbf{k}.\end{aligned}$$

local base vectors: near a chosen point, let one of the coordinates x^j run, and keep the others fixed. You obtain a curve, and the local base vector $\mathbf{e}_j$ is just the tangential vector to that curve:

$$\mathbf{e}_j := \frac{\partial \mathbf{r}}{\partial x^j}.$$

These vectors might not have length one, and are not necessarily perpendicular to each other. In case of polar coordinates, you have

$$\mathbf{e}_1 = \sin x^2\cos x^3\,\mathbf{i} + \sin x^2\sin x^3\,\mathbf{j} + \cos x^2\,\mathbf{k}$$

(pointing away from the origin in radial direction),

$$\mathbf{e}_2 = x^1 \cos x^2 \cos x^3\,\mathbf{i} + x^1 \cos x^2 \sin x^3\,\mathbf{j} - x^1 \sin x^2\,\mathbf{k}$$

(pointing towards the south pole along a meridian),

$$\mathbf{e}_3 = -x^1 \sin x^2 \sin x^3\,\mathbf{i} + x^1 \sin x^2 \cos x^3\,\mathbf{j} + 0\,\mathbf{k}$$

(pointing from west to east).

Some people prefer normalizing the local base vectors $\mathbf{e}_j$ to length one, but this destroys the beauty of the following formulas.

arc length and metric tensor field: consider a curve with parametrization

$$\mathbf{r} = \mathbf{r}(x^1(t), \dots, x^n(t)).$$

Then the tangential vector can be calculated by the chain rule:

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \frac{\partial \mathbf{r}}{\partial x^j}\frac{\mathrm{d}x^j}{\mathrm{d}t} = \mathbf{e}_j \dot{x}^j,$$

and its squared length is

$$\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2 = \mathbf{e}_j\mathbf{e}_k\dot{x}^j\dot{x}^k = g_{jk}\dot{x}^j\dot{x}^k$$

with $g_{jk} = \mathbf{e}_j\mathbf{e}_k$, or in shorter notation,

$$\mathrm{d}s^2 = g_{jk}\,\mathrm{d}x^j\,\mathrm{d}x^k.$$

Here $\mathrm{d}s$ is called the *arc length element*, and the g_{jk} form the *metric tensor field* which is twice covariant. This tensor field is symmetric in the sense of $g_{jk} = g_{kj}$. In our derivation, the matrix of the g_{jk} is positive definite, but in general relativity, one of the variables will be the time, and the 4×4 matrix of the g_{jk} will then have one positive and three negative eigenvalues.

The inverse matrix of the g_{jk} has entries g^{jk} (by definition), and this inverse matrix is a twice contravariant tensor field.

In case of the polar coordinates in $\mathbb{R}^3$, we have $g_{jk} = 0$ for $j \neq k$, and

$$g_{11} = 1, \qquad g_{22} = r^2 = (x^1)^2, \qquad g_{33} = (r\sin\theta)^2 = (x^1 \sin x^2)^2,$$
$$g^{11} = 1, \qquad g^{22} = \frac{1}{r^2} = \frac{1}{(x^1)^2}, \qquad g^{33} = \frac{1}{(r\sin\theta)^2} = \frac{1}{(x^1\sin x^2)^2}.$$

We also set $g = \det g_{ij}$, which is $r^4 \sin^2\theta$ in case of the polar coordinates in $\mathbb{R}^3$. The square root of $|g|$ will always be the factor which appears when evaluating volume integrals !

dual local base vectors: given the local base vectors $\mathbf{e}_1$, ..., $\mathbf{e}_n$, we define

$$\mathbf{e}^j := g^{jk}\mathbf{e}_k,$$

and these are just the vectors of the dual basis, since

$$\mathbf{e}^j\mathbf{e}_k = (g^{jl}\mathbf{e}_l)\mathbf{e}_k = g^{jl}g_{lk} = \delta^j_k \qquad \text{(Kronecker symbol).}$$

components of a vector field: at each point P, we attach a vector, giving us a vector field $\mathbf{v} = \mathbf{v}(P)$. Then we can write it in terms of the $\mathbf{e}_j$ basis,

$$\mathbf{v}(P) = v^j(P)\mathbf{e}_j,$$

where the v^j can be computed using the scalar product:

$$v^j(P) = \mathbf{e}^j\mathbf{v}(P).$$

The components v^j form a contravariant tensor field.

Now we have enough knowledge to write down the typical differential operators in general coordinates:

$$\begin{aligned}\operatorname{grad} f &= \frac{\partial f}{\partial x^j}\mathbf{e}^j = \left(g^{ij}\frac{\partial f}{\partial x^i}\right)\mathbf{e}_j,\\ \operatorname{div}\mathbf{v} &= \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^i}\left(\sqrt{|g|}v^i\right),\\ \triangle f = \operatorname{div}\operatorname{grad} f &= \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^i}\left(\sqrt{|g|}g^{ij}\frac{\partial f}{\partial x^j}\right).\end{aligned}$$

Next we wish to understand how the local base vectors change when we go from one point to a neighbouring point. Then we should consider the derivatives $\frac{\partial \mathbf{e}_i}{\partial x^j}$, which are again vectors which can be decomposed in terms of the local basis or the dual basis, leading to

$$\begin{aligned}\frac{\partial \mathbf{e}_i}{\partial x^j} &= \Gamma_{ij}^k\mathbf{e}_k, & \frac{\partial \mathbf{e}_i}{\partial x^j} &= \Gamma_{ij,k}\mathbf{e}^k,\\ \Gamma_{ij}^k &= \mathbf{e}^k\frac{\partial \mathbf{e}_i}{\partial x^j}, & \Gamma_{ij,k} &= \mathbf{e}_k\frac{\partial \mathbf{e}_i}{\partial x^j}.\end{aligned}$$

The Γ_{ij}^k and $\Gamma_{ij,k}$ are the *Christoffel symbols*, and they can be computed via

$$\Gamma_{ij,m} = \frac{1}{2}\left(\frac{\partial g_{jm}}{\partial x^i} + \frac{\partial g_{im}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^m}\right), \qquad \Gamma_{ij}^k = g^{km}\Gamma_{ij,m}.$$

Now there is a nasty issue coming: spatial derivatives of a tensor field will (in general) not be tensor fields again, and then also the Christoffel symbols will not be tensor fields (because their transformation rule when exchanging the coordinate system will be different). To compensate this, we introduce a new derivative: the *covariant derivative*. For a general p–fold covariant and q–fold contravariant tensor, it is quite hard to define, but it is easier for tensor fields which are covariant or contravariant only a few number of times:

$$\begin{aligned}&\nabla_k f = \frac{\partial f}{\partial x^k} && \text{(scalar functions)},\\ &\nabla_k a^i = \frac{\partial a^i}{\partial x^k} + \Gamma_{kl}^i a^l, \quad \nabla_k a_i = \frac{\partial a_i}{\partial x^k} - \Gamma_{ki}^l a_l && \text{(once contravariant/covariant, respectively)},\\ &\nabla_k a^{ij} = \frac{\partial a^{ij}}{\partial x^k} + \Gamma_{ks}^i a^{sj} + \Gamma_{ks}^j a^{is}, \quad \nabla_k a_{ij} = \frac{\partial a_{ij}}{\partial x^k} - \Gamma_{ki}^s a_{sj} - \Gamma_{kj}^s a_{si}.\end{aligned}$$

Then the divergence of a vector field $\mathbf{v}$ can be simply written as $\operatorname{div}\mathbf{v} = \nabla_j v^j$. When checking this, note that we have the Ricci Lemma:

$$\nabla_r g_{ik} = \nabla_r g^{ik} = 0, \qquad \nabla_i g = 0.$$

And we can also write down the simple formulas

$$\operatorname{grad} f = (\nabla_k f)\mathbf{e}^k, \qquad \triangle f = g^{ij}\nabla_i\nabla_j f, \qquad \triangle\mathbf{v} = (g^{ij}\nabla_i\nabla_j v^k)\mathbf{e}_k.$$

Next we come to the curvature of a manifold. The *Riemannian curvature tensor* is a tensor R_{ikm}^j with

$$\nabla_k\nabla_m u^j - \nabla_m\nabla_k u^j = R_{ikm}^j u^i.$$

This tensor measures how much the rule $\partial_k\partial_m - \partial_m\partial_k = 0$ is violated (roughly spoken). In case of a flat metric, the Riemannian curvature tensor is everywhere zero.

Explicitely, we have the formulae

$$R_{ikm}^j = \frac{\partial \Gamma_{mi}^j}{\partial x^k} - \frac{\partial \Gamma_{ki}^j}{\partial x^m} + \Gamma_{ks}^j\Gamma_{mi}^s - \Gamma_{ms}^j\Gamma_{ki}^s.$$

Further, we define

$$R_{ijkm} := g_{js}R_{ikm}^s$$

and also the Ricci *tensor*

$$R_{ik} := R^s_{iks}.$$

Finally, the SCALAR CURVATURE is defined as

$$R := g^{rs} R_{rs}.$$

Then the *Einstein field equations* are

$$R_{ij} - \frac{1}{2} g_{ij} R = \kappa T_{ij}, \qquad i, j = 0, 1, 2, 3,$$

where x_0 denotes the time variable. Here T_{ij} is the *energy momentum tensor* which describes the distribution of the masses. The tensor of the g_{ij} is the covariant metric tensor field belonging to the four-dimensional space-time manifold (which will be deformed if masses are present, and it will be flat in case of an empty universe). These Einstein field equations, one of the key elements of Einstein's theory of general relativity, are a highly complicated system of partial differential equations of second order, and the unknown functions are the components of the metric tensor g_{ij}.

Have fun !

> Wie die spezielle Relativitätstheorie auf das Postulat gegründet ist, daß ihre Gleichungen bezüglich linearer, orthogonaler Transformationen kovariant sein sollen, so ruht die hier darzulegende Theorie auf dem Postulat der *Kovarianz aller Gleichungssysteme bezüglich Transformationen von der Substitutionsdeterminante 1.*
>
> Dem Zauber dieser Theorie wird sich kaum jemand entziehen können, der sie wirklich erfaßt hat; sie bedeutet einen wahren Triumph der durch Gauss, Riemann, Christoffel, Ricci und Levi-Civiter (sic) begründeten Methode des allgemeinen Differentialkalküls.[25]

5.5 Surface Integrals

5.5.1 Surface Integrals of First Kind

It is natural to ask for the area of a surface; and it is also natural to expect that the area (in a naive sense of the word) will emerge if you "integrate the function 1 over the surface".

Now we have to explain how to integrate over a surface (patch).

Fix a point $u_0 = (u_{0,1}, u_{0,2})^\top$ and take a small rectangle in the parameter domain:

$$R_\Delta := \{(u_1, u_2) \colon u_{0,1} \leq u_1 \leq u_{0,1} + \Delta_1, \ u_{0,2} \leq u_2 \leq u_{0,2} + \Delta_2\}, \tag{5.6}$$

where Δ_1 and Δ_2 are small positive numbers. This rectangle is mapped by the parametrisation Φ onto a quadrangle $\Phi(R_\Delta) \subset S$ (with non–straight edges) with corners

$$\Phi(u_0), \quad \Phi(u_0 + (\Delta_1, 0)), \quad \Phi(u_0 + (0, \Delta_2)), \quad \Phi(u_0 + (\Delta_1, \Delta_2)).$$

The Taylor expansion of our parametrisation reads

$$\Phi(u) = \Phi(u_0) + \Phi'(u_0) \cdot (u - u_0) + R(u, u_0)$$

with a remainder vector R much shorter than $\|\Phi'(u_0) \cdot (u - u_0)\|$ — recall that $\Phi'(u_0)$ has full rank. The proof that the remainder R indeed is negligible is beautiful, since it combines many ideas from various branches of mathematics. Enjoy how they nicely fit together:

Lemma 5.33. *Suppose that the function $\Phi \colon U \to \mathbb{R}^3$ that describes the surface patch S is twice continuously differentiable, and $u_0 \in U$. Then the matrix $(\Phi'(u_0))^\top \Phi'(u_0) \in \mathbb{R}^{2\times 2}$ is positive definite. Call its positive eigenvalues λ_1 and λ_2. Then we have, for all $u \in U$, the length estimate from below*

$$\|\Phi'(u_0) \cdot (u - u_0)\| \geq \sqrt{\min(\lambda_1, \lambda_2)}\, \|u - u_0\| ,$$

and there is a positive constant C_2 such that the remainder R satisfies the length estimate from above

$$\|R(u, u_0)\| \leq C_2 \|u - u_0\|^2 .$$

[25] A. Einstein, *Zur allgemeinen Relativitätstheorie.* Königlich Preußische Akad. Wiss. (Berlin). Sitz.ber. (1915), 778–786.

We remark that both inequalities together say $\|R(u,u_0)\| \ll \|\Phi'(u_0)\cdot(u-u_0)\|$ if $\|u-u_0\| \ll 1$.

Proof. For simplicity of notation, put $A := \Phi'(u_0)$. Then $A \in \mathbb{R}^{3\times 2}$ maps from $\mathbb{R}^2$ into $\mathbb{R}^3$, and A has full rank two, since Φ is a parametrisation of a surface patch. The dimension formula for the linear map associated to the matrix A then reads

$$\dim \mathbb{R}^2 = \dim\ker A + \dim \operatorname{img} A,$$

and therefore $\dim\ker A = 0$, or $\ker A = \{0\}$. Take any vector $\Delta u := u - u_0 \in \mathbb{R}^2$, with $\Delta u \neq 0$. Then $A\Delta u \neq 0$ because of $\ker A = \{0\}$, hence

$$0 < \|A\Delta u\|^2 = \langle A\Delta u, A\Delta u\rangle_{\mathbb{R}^3} = \left\langle A^\top A\Delta u, \Delta u\right\rangle_{\mathbb{R}^2},$$

which is just the definition that the matrix $B := A^\top A$ is positive definite (of course, B is symmetric !). By Proposition 4.30, the two eigenvalues of B must be positive. Call them λ_1 and λ_2. Since B is symmetric and real, it is self-adjoint, and therefore the spectral theorem holds. Call the associated projectors P_1 and P_2, which are matrices from $\mathbb{R}^{2\times 2}$ with the properties

$$P_1 + P_2 = I_2, \qquad P_j^2 = P_j, \qquad P_j^\top = P_j, \qquad P_1P_2 = P_2P_1 = 0 \in \mathbb{R}^{2\times 2}, \qquad B = \lambda_1 P_1 + \lambda_2 P_2.$$

Now we compute the length of $\Phi'(u_0)\cdot(u-u_0) = A\Delta u$:

$$\begin{aligned}
\|A\Delta u\|^2 &= \langle B\Delta u, \Delta u\rangle_{\mathbb{R}^2} = \langle (\lambda_1P_1+\lambda_2P_2)\Delta u, (P_1+P_2)\Delta u\rangle_{\mathbb{R}^2} \\
&= \sum_{j=1}^2\sum_{k=1}^2 \langle \lambda_j P_j\Delta u, P_k\Delta u\rangle_{\mathbb{R}^2} = \sum_{j=1}^2\sum_{k=1}^2 \lambda_j \left\langle P_k^\top P_j\Delta u, \Delta u\right\rangle_{\mathbb{R}^2} = \sum_{j=1}^2\sum_{k=1}^2 \lambda_j \langle P_kP_j\Delta u, \Delta u\rangle_{\mathbb{R}^2} \\
&= \sum_{j=1}^2 \lambda_j \langle P_jP_j\Delta u, \Delta u\rangle_{\mathbb{R}^2} = \sum_{j=1}^2 \lambda_j \left\langle P_j^\top P_j\Delta u, \Delta u\right\rangle_{\mathbb{R}^2} = \sum_{j=1}^2 \lambda_j \langle P_j\Delta u, P_j\Delta u\rangle_{\mathbb{R}^2} \\
&\geq \min(\lambda_1,\lambda_2)\sum_{j=1}^2 \langle P_j\Delta u, P_j\Delta u\rangle_{\mathbb{R}^2}.
\end{aligned}$$

We interrupt this computation for a little calculation of the length of Δu:

$$\begin{aligned}
\|\Delta u\|^2 &= \langle \Delta u, \Delta u\rangle_{\mathbb{R}^2} = \langle (P_1+P_2)\Delta u, (P_1+P_2)\Delta u\rangle_{\mathbb{R}^2} \\
&= \sum_{j=1}^2\sum_{k=1}^2 \langle P_j\Delta u, P_k\Delta u\rangle_{\mathbb{R}^2} = (\ \ldots\text{repeat the dance with the adjoints from above}\ldots) \\
&= (\ \ldots\text{but now without the } \lambda_j\ \ldots) = \sum_{j=1}^2 \langle P_j\Delta u, P_j\Delta u\rangle_{\mathbb{R}^2},
\end{aligned}$$

and therefore we have proved

$$\|A\Delta u\|^2 \geq \min(\lambda_1,\lambda_2)\,\|\Delta u\|^2.$$

This is the first estimate. And the second estimate $\|R(u,u_0)\| \leq C_2\|u-u_0\|^2$ is exactly the claim of Taylor's Theorem, compare Remark 1.22. □

We obtain a parametrisation Ψ of the tangential plane at $\Phi(u_0)$ if we drop the remainder term R:

$$\Psi(u) = \Phi(u_0) + \Phi'(u_0)\cdot(u-u_0).$$

The image of the rectangle R_Δ (defined in (5.6)) from the parameter domain U under the parametrisation Ψ of the tangential plane is just a parallelogram $\Psi(R_\Delta)$ with the corners

$$\Phi(u_0), \quad \Phi(u_0) + \partial_{u_1}\Phi(u_0)\cdot\Delta_1, \quad \Phi(u_0) + \partial_{u_2}\Phi(u_0)\cdot\Delta_2, \quad \Phi(u_0) + \partial_{u_1}\Phi(u_0)\cdot\Delta_1 + \partial_{u_2}\Phi(u_0)\cdot\Delta_2.$$

It is natural to expect that the areas of the quadrangle with non–straight edges $\Phi(R_\Delta)$ and of the parallelogram $\Psi(R_\Delta)$ should be roughly the same, and that this approximation should become better and better if the side lengths Δ_1 and Δ_2 of the rectangle R_Δ go to zero. But the area $A(\Psi(R_\Delta))$ is easy to compute, and we obtain

$$A(\Psi(R_\Delta)) = \|(\partial_{u_1}\Phi(u_0)\cdot\Delta_1)\times(\partial_{u_2}\Phi(u_0)\cdot\Delta_1)\| = \|(\partial_{u_1}\Phi)\times(\partial_{u_2}\Phi)\|\cdot|\Delta_1|\cdot|\Delta_2|.$$

This reasoning justifies the following definition of the area of a surface patch.

Definition 5.34 (Area of a surface patch). *Let $S \subset \mathbb{R}^3$ be a surface patch, parametrised by a mapping Φ with parameter domain U. Then the area of that parametrisation is defined as*

$$A(\Phi(U)) := \int_U \|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\| \, \mathrm{d}(u_1, u_2).$$

This definition is not satisfactory: we know that a surface patch may have several differing parametrisations. They should all have the same area, because the area is a *geometric* property and should only depend on the surface patch, but not on the parametrisation (which is just a tool for the *analytical* description). And indeed, now we show that the parametrisation does not matter (beware that we recycle the identifier Ψ, since we do not need it anymore as parametrisation of the tangential plane):

Proposition 5.35 (Re-parametrisation). *Let $\Phi\colon U \to \mathbb{R}^3$ and $\Psi\colon V \to \mathbb{R}^3$ be two parametrisations of the same surface S and $\tau\colon U \to V$ a diffeomorphism with $\Phi = \Psi \circ \tau$. Put*

$$g = \|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\|^2, \qquad h = \|(\partial_{v_1}\Psi) \times (\partial_{v_2}\Psi)\|^2.$$

Then the following identity is true:

$$\int_U \sqrt{g}\, \mathrm{d}(u_1, u_2) = \int_U \sqrt{h \circ \tau}\, |\det(\tau')|\, \mathrm{d}(u_1, u_2) = \int_V \sqrt{h}\, \mathrm{d}(v_1, v_2).$$

Proof. From Proposition 5.26 we know already $\sqrt{g} = |\det \tau'|\sqrt{h}$. Now apply Proposition 5.17. □

Definition 5.36 (Scalar surface element). *The expression*

$$\mathrm{d}\sigma = \|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\| \, \mathrm{d}(u_1, u_2) = \sqrt{g}\, \mathrm{d}(u_1, u_2)$$

is called scalar surface element[26]*. It is invariant under re-parametrisations.*

Then the area of the surface patch is given by

$$A(\Phi(U)) = \int_U \sqrt{g}\, \mathrm{d}(u_1, u_2).$$

You will have wondered why we have chosen to define cumbersomely g as the square of the norm of the cross product, only to write afterwards $\sqrt{g}$ everywhere in the integrands. However, this notation has some advantages. To understand them, it is illuminating to think about the vector-product in different terms: we know that the norm of a vector product $a \times b$ equals the area of the parallelogram spanned by the factors a and b. On the other hand, this (oriented) area equals the determinant $\det(a, b)$ of the two factors. Now we write a and b as column vectors together, obtaining a matrix X. Then we have

$$\begin{aligned}\|a \times b\|^2 &= (\det X)^2 = (\det X) \cdot (\det X) = (\det X^\top) \cdot (\det X) \\ &= \det(X^\top X) = \det \begin{pmatrix} \langle a, a\rangle & \langle a, b\rangle \\ \langle b, a\rangle & \langle b, b\rangle \end{pmatrix}.\end{aligned}$$

In this spirit, we fix $g_{ij} := \left\langle \partial_{u_i}\Phi, \partial_{u_j}\Phi \right\rangle$ (as for the metric tensor) and express g as

$$g = g(u) = \|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\|^2 = \det \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}.$$

Observe that, with the notation from the proof of Lemma 5.33, we have $g = \det B$.

Question: Consider the four-dimensional Euclidean space, and a three-dimensional object in it. How would you parametrise that object ? Guess how the volume formula for that object looks like ! Now you have just found out the key advantage of this function g.

In other literature, you may read the notation of Gauss,

$$E = g_{11}, \qquad F = g_{12} = g_{21}, \qquad G = g_{22}.$$

Then $g = EG - F^2$. The numbers E, F, G are called *coefficients of the first fundamental form*[27].

[26] skalares Oberflächenelement

[27] Koeffizienten der ersten Fundamentalform

Remark 5.37. *Just for completeness, we explain what the first fundamental form is: it is a tool for the easy computation of the length of a tangent vector at a surface patch S. Pick a point u_0 in the parameter domain U; and let $\gamma = \gamma(\tau) = (u_1(\tau), u_2(\tau))$ be a short curve in U, with $\gamma(\tau = 0) = u_0$ and tangent vector $\gamma'(0) = (u_1'(0), u_2'(0))$ at u_0. Then this curve γ gives rise to a tangent vector on the surface S at the point $x_0 = \Phi(u_0)$, namely*

$$\vec{t} = (\partial_{u_1}\Phi) \cdot u_1'(0) + (\partial_{u_2}\Phi) \cdot u_2'(0).$$

The length of this tangent vector can be found via

$$\left\|\vec{t}\right\|^2 = E(u_1'(0))^2 + 2F(u_1'(0))(u_2'(0)) + G(u_2'(0))^2.$$

This quadratic form is called first fundamental form. *Particularly convenient are the parametrisations Φ with $F \equiv 0$: then the tangential vectors $\partial_{u_1}\Phi$ and $\partial_{u_2}\Phi$ are perpendicular to each other everywhere.*

Example: *We compute the area of a sphere with radius R:*

$$S_R = \left\{x \in \mathbb{R}^3 \colon \ \|x\| = R\right\}.$$

We parametrise it with polar coordinates:

$$\Phi(\varphi, \theta) = \begin{pmatrix} R\sin\theta\cos\varphi \\ R\sin\theta\sin\varphi \\ R\cos\theta \end{pmatrix}, \qquad 0 \le \varphi \le 2\pi, \qquad 0 \le \theta \le \pi.$$

Recall that the last condition of Definition 5.23 is violated.

Question: *Can you give a reason why this does not matter ?*

Then we compute

$$\begin{aligned} &\partial_\varphi\Phi = \begin{pmatrix} -R\sin\theta\sin\varphi \\ R\sin\theta\cos\varphi \\ 0 \end{pmatrix}, \qquad \partial_\theta\Phi = \begin{pmatrix} R\cos\theta\cos\varphi \\ R\cos\theta\sin\varphi \\ -R\sin\theta \end{pmatrix}, \\ &E = g_{\varphi\varphi} = R^2\sin^2\theta, \qquad F = g_{\varphi\theta} = 0, \qquad G = g_{\theta\theta} = R^2, \\ &\sqrt{g} = \sqrt{EG - F^2} = R^2\sin\theta. \end{aligned}$$

Then the surface of the sphere is calculated like this:

$$A(S_R) = \int_{\varphi=0}^{\varphi=2\pi}\int_{\theta=0}^{\theta=\pi} R^2\sin\theta \,\mathrm{d}\theta\,\mathrm{d}\varphi = 2\pi R^2\int_{\theta=0}^{\theta=\pi}\sin\theta\,\mathrm{d}\theta = 4\pi R^2.$$

After this preparation, the definition of surface integrals should be plausible:

Definition 5.38 (Surface integral of first kind). *Take a surface patch $S \subset \mathbb{R}^3$ with parametrisation Φ and parameter domain U, and a continuous function $f \colon S \to \mathbb{R}$. Then the expression*

$$\int_S f(x)\,\mathrm{d}\sigma := \int_U f(\Phi(u))\sqrt{g(u)}\,\mathrm{d}(u_1, u_2)$$

is called surface integral of first kind[28] *or* scalar surface integral[29].

We have already proved that the scalar surface element $\mathrm{d}\sigma$ is invariant under re-parametrisations. Therefore, also the values of surface integrals of first kind do not depend on the choice of parametrisation.

[28] Oberflächenintegral erster Art

[29] skalares Oberflächenintegral

5.5.2 Surface Integrals of Second Kind

Consider some thing that can be considered as “flowing”, for instance, the light emanating from the sun. Take a surface patch, for instance, a part of the earth surface. You would like to know how much sunlight arrives at that surface patch in one minute. How to do that ?

You cannot simply multiply “light current per square metres” times “area in square metres”, because then the temperature on earth would be the same everywhere. Instead, you need one factor more, namely the cosine of the angle between the vector of light and the normal vector of the surface.

Hence the **ingredients** to a surface integral of second kind are the following:

a surface patch: it must have a normal vector at every point. In other words, it must be orientable (you cannot define surface integrals of second kind on the Möbius strip).

a vectorial integrand: usually, this is a vector field in $\mathbb{R}^3$, which means that you attach a vector at every point of a three-dimensional domain.

The integral can be roughly **defined** as follows:

- cut the surface patch into many small pieces, each of them looking like a small flat parallelogram,
- for each such piece: take its unit normal vector, compute the scalar product with the vector of the integrand at that point, and multiply with the size of the parallelogram,
- compute the sum over all small parallelograms.

This gives you an approximate value of the integral. Finally, you perform the limit, making the pieces infinitesimally small.

When **computing** a surface integral, you express everything in terms of (u_1, u_2), of course.

Definition 5.39 (Surface integral of second kind). *Take a surface patch $S \subset \mathbb{R}^3$ with positively oriented parametrisation Φ and parameter domain U, and a continuous function $\vec{f}\colon S \to \mathbb{R}^3$. Then the expression*

$$\int_S \vec{f}(x) \cdot \mathrm{d}\vec{\sigma} := \int_S \left\langle \vec{f}(x), \vec{n}(x) \right\rangle \mathrm{d}\sigma = \int_U \left\langle \vec{f} \circ \Phi, (\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi) \right\rangle \mathrm{d}(u_1, u_2)$$

is said to be a surface integral of second kind[30] *or* vectorial surface integral[31]. *The value of the integral is also called* flux of $\vec{f}$ through S[32]. *The expression*

$$\mathrm{d}\vec{\sigma} = (\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\, \mathrm{d}(u_1, u_2)$$

has the name vectorial surface element[33].

Question: Why does the value of the integral not change, when you chose another parametrisation of the surface with the same orientation ?

Example: *Let S be one leaf of the* screw surface[34]*:*

$$S := \left\{(r\cos\varphi, r\sin\varphi, \varphi)^\top : 0 < r < R,\ 0 < \varphi < 2\pi\right\}$$

and let $\vec{f}$ denote the vector field

$$\vec{f}(x, y, z) := \begin{pmatrix} y \\ -x \\ z \end{pmatrix}.$$

[30] Oberflächenintegral zweiter Art
[31] vektorielles Oberflächenintegral
[32] Fluß von $\vec{f}$ durch die Fläche S
[33] vektorielles Oberflächenelement
[34] Schraubenfläche

The integral $\int_S \vec{f}(x)\cdot \mathrm{d}\vec{\sigma}$ will then be calculated as follows:

$$\partial_r\Phi = \begin{pmatrix}\cos\varphi\\ \sin\varphi\\ 0\end{pmatrix}, \qquad \partial_\varphi\Phi = \begin{pmatrix}-r\sin\varphi\\ r\cos\varphi\\ 1\end{pmatrix},$$

$$(\partial_r\Phi)\times(\partial_\varphi\Phi) = \begin{pmatrix}\sin\varphi\\ -\cos\varphi\\ r\end{pmatrix},$$

$$\left\langle \vec{f}\circ\Phi, (\partial_r\Phi)\times(\partial_\varphi\Phi)\right\rangle = \left\langle \begin{pmatrix}r\sin\varphi\\ -r\cos\varphi\\ \varphi\end{pmatrix}, \begin{pmatrix}\sin\varphi\\ -\cos\varphi\\ r\end{pmatrix}\right\rangle = r\sin^2\varphi + r\cos^2\varphi + r\varphi = r(\varphi+1),$$

$$\int_S \vec{f}(x)\cdot\mathrm{d}\vec{\sigma} = \int_{\varphi=0}^{2\pi}\int_{r=0}^{R} r(\varphi+1)\,\mathrm{d}r\,\mathrm{d}\varphi = R^2\pi(\pi+1).$$

5.6 Integral Theorems

We conclude this chapter with some integral theorems. There are several of them, but their main idea is always the same: you transform one type of integral into another one. In the theory of *differential forms*[35]—a theory which we sketch in the outlook at the end—a famous formula is established, namely

$$\int_M \mathrm{d}\omega = \int_{\partial M}\omega,$$

where M is a nice manifold in $\mathbb{R}^n$, ∂M is the nice boundary of M, ω is a differential form on M, and $\mathrm{d}\omega$ is the differential of ω.

All the theorems which we will present below are special cases of the above formula.

Let us list all of them (some are already proved):

the fundamental theorem of calculus:

$$\int_{x=a}^{x=b} f'(x)\,\mathrm{d}x = f(b) - f(a),$$

the path-independence of integrals of gradient fields:

$$\int_A^B \operatorname{grad}\varphi\cdot\mathrm{d}\vec{x} = \varphi(B) - \varphi(A),$$

the Gauss theorem in $\mathbb{R}^2$:

$$\int_\Omega \operatorname{div}\vec{f}\,\mathrm{d}(x,y) = \int_{\partial\Omega}\vec{f}\cdot\vec{\nu}\,\mathrm{d}s, \qquad \vec{\nu} = \text{ outer normal}, \qquad s = \text{ arc-length},$$

the Gauss theorem in $\mathbb{R}^3$:

$$\int_\Omega \operatorname{div}\vec{f}\,\mathrm{d}(x,y,z) = \int_{\partial\Omega}\vec{f}\cdot\vec{\nu}\,\mathrm{d}\sigma,$$

the Stokes theorem in $\mathbb{R}^3$:

$$\int_M \operatorname{rot}\vec{f}\cdot\mathrm{d}\vec{\sigma} = \int_{\partial M}\vec{f}\cdot\mathrm{d}\vec{x}.$$

The latter two formulae are indispensable for the investigation of electric fields and magnetic fields.

The proofs of the above formulas would be quite long if you were asking for full generality. However, we think that it is more important to communicate the main ideas, and for this reason we will go the easy way and assume that all functions and domains are extraordinarily nice.

[35] Differentialformen

Definition 5.40 (Gauss **normal**[36] **domain).** *A set $\Omega \subset \mathbb{R}^2$ is a* Gauss normal domain[37] *if it is open, bounded, connected, and the boundary $\partial\Omega$ has Jordan-measure zero. Additionally, we assume that there are real numbers x_0, x_1, y_0, y_1, and piecewise C^1 functions $\alpha_\pm$, $\beta_\pm$, such that:*

$$\begin{aligned}
&\alpha_-(x) < \alpha_+(x), \qquad && x_0 < x < x_1,\\
&\beta_-(y) < \beta_+(y), \qquad && y_0 < y < y_1,\\
&\Omega = \{(x,y)\colon x_0 < x < x_1,\ \alpha_-(x) < y < \alpha_+(x)\},\\
&\Omega = \{(x,y)\colon y_0 < y < y_1,\ \beta_-(y) < x < \beta_+(y)\}.
\end{aligned}$$

The derivatives of $\alpha_\pm$ and $\beta_\pm$ are supposed to be bounded.

Gauss normal domains in $\mathbb{R}^3$ are defined similarly.

Question: Why is the unit ball not a normal domain ? How can you repair it ?

In the sequel, we will assume that all domains to be considered are normal domains. This makes the proofs easier; however, the theorems are valid also for other domains.

We need some agreements:

The boundary of a domain in the plane is oriented in such a way
that "the domain is to your left hand".

The boundary of a surface patch in $\mathbb{R}^3$ is oriented like this:
if the thumb of your right hand is parallel to the normal of the plane,
then your fingers are parallel to the tangential vector of the boundary of the surface patch.

5.6.1 Integral Theorems in $\mathbb{R}^2$

Proposition 5.41 (Gauss **theorem in** $\mathbb{R}^2$). *Let $\Omega \subset \mathbb{R}^2$ be a Gauss normal domain, and $\vec{f}\colon \overline{\Omega} \to \mathbb{R}^2$ be a C^1 function with bounded first derivative. Let $\vec{\nu}$ denote the outward unit normal on $\partial\Omega$, and* $\mathrm{d}s$ *the arc-length element of $\partial\Omega$. Then the following identity holds:*

$$\int_\Omega \operatorname{div} \vec{f}(x,y)\,\mathrm{d}(x,y) = \int_{\partial\Omega} \vec{f}\cdot\vec{\nu}\,\mathrm{d}s.$$

Proof. Put $\vec{f} = (P,Q)^\top$. Then $\operatorname{div}\vec{f} = P_x + Q_y$, and we can compute easily:

$$\begin{aligned}
\int_\Omega P_x\,\mathrm{d}(x,y) &= \int_{y=y_0}^{y=y_1}\left(\int_{x=\beta_-(y)}^{x=\beta_+(y)} P_x(x,y)\,\mathrm{d}x\right)\mathrm{d}y\\
&= \int_{y=y_0}^{y=y_1} \left(P(\beta_+(y),y) - P(\beta_-(y),y)\right)\mathrm{d}y = \oint_{\partial\Omega} P\,\mathrm{d}y,\\
\int_\Omega Q_y\,\mathrm{d}(x,y) &= \int_{x=x_0}^{x=x_1}\left(\int_{y=\alpha_-(x)}^{y=\alpha_+(x)} Q_y(x,y)\,\mathrm{d}y\right)\mathrm{d}x\\
&= \int_{x=x_0}^{x=x_1} \left(Q(x,\alpha_+(x)) - Q(x,\alpha_-(x))\right)\mathrm{d}x = -\oint_{\partial\Omega} Q\,\mathrm{d}x.
\end{aligned}$$

Summing up, we find

$$\int_\Omega \operatorname{div}\vec{f}\,\mathrm{d}(x,y) = \oint_{\partial\Omega} -Q\,\mathrm{d}x + P\,\mathrm{d}y. \tag{5.7}$$

It remains to re-parametrise $\partial\Omega$ with the arc-length:

$$\partial\Omega = \{(x,y) = (\xi(s),\eta(s))\colon 0 \le s \le L\},$$

[36] "normal" is not related to "normal vector", but to "sane"

[37] Gaußscher Normalbereich

where L is the length of the curve $\partial\Omega$. The unit tangent vector to $\partial\Omega$ is $(\xi'(s), \eta'(s))^\top$, and the outward unit normal vector is $(\eta'(s), -\xi'(s))^\top$. Finally, the differentials are transformed as follows:

$$\mathrm{d}x = \xi'(s)\,\mathrm{d}s, \qquad \mathrm{d}y = \eta'(s)\,\mathrm{d}s.$$

Plugging these equations into each other gives

$$\begin{aligned}\int_\Omega \operatorname{div} \vec{f}\,\mathrm{d}(x,y) &= \int_{s=0}^{s=L} \left(-Q(\xi(s),\eta(s))\xi'(s) + P(\xi(s),\eta(s))\eta'(s)\right)\,\mathrm{d}s \\ &= \int_{s=0}^{s=L} \vec{f}(\xi(s),\eta(s))\cdot\vec{\nu}(s)\,\mathrm{d}s = \int_{\partial\Omega} \vec{f}\cdot\vec{\nu}\,\mathrm{d}s.\end{aligned}$$

□

Question: Can you find formulas for the area of the domain Ω from the Gauss theorem ? As a hint, you could have a look at Proposition 3.68.

Question: Consider a domain with a shape like a horseshoe[38]. This is no Gauss normal domain. Can you prove the Gauss integral theorem for this domain ?

5.6.2 The Gauss Theorem in $\mathbb{R}^3$

Proposition 5.42 (Gauss **theorem in** $\mathbb{R}^3$)**.** *Let $\Omega \subset \mathbb{R}^3$ be a Gauss normal domain, and $\vec{f}\colon \overline{\Omega} \to \mathbb{R}^3$ be a C^1 function with bounded first derivative. Let $\vec{\nu}$ denote the outward unit normal on $\partial\Omega$, and $\mathrm{d}\sigma$ the scalar surface element of $\partial\Omega$. Then the following identity holds:*

$$\int_\Omega \operatorname{div} \vec{f}(x,y,z)\,\mathrm{d}(x,y,z) = \int_{\partial\Omega} \vec{f}\cdot\vec{\nu}\,\mathrm{d}\sigma. \tag{5.8}$$

Proof. It is very similar to the two-dimensional case. Put $\vec{f} = (P,Q,R)^\top$, and write Ω as

$$\Omega = \{(x,y,z)\colon (x,y)\in U,\ \gamma_-(x,y) < z < \gamma_+(x,y)\}, \qquad U \subset \mathbb{R}^2.$$

Obviously, $\operatorname{div}\vec{f} = P_x + Q_y + R_z$. Considering only R_z, we then compute:

$$\begin{aligned}\int_\Omega R_z(x,y,z)\,\mathrm{d}(x,y,z) &= \int_U \left(\int_{z=\gamma_-(x,y)}^{z=\gamma_+(x,y)} R_z(x,y,z)\,\mathrm{d}z\right)\mathrm{d}(x,y) \\ &= \int_U \left(R(x,y,\gamma_+(x,y)) - R(x,y,\gamma_-(x,y))\right)\mathrm{d}(x,y).\end{aligned}$$

Now we rewrite this integral as a surface integral on $\partial\Omega$. The “upper” and “lower” surface of $\partial\Omega$ are parametrised by functions $\Phi_\pm$,

$$\Phi_+(x,y) = \begin{pmatrix} x \\ y \\ \gamma_+(x,y)\end{pmatrix}, \qquad \Phi_-(x,y) = \begin{pmatrix} x \\ y \\ \gamma_-(x,y)\end{pmatrix}.$$

The tangent plane at a point $(x,y,z)^\top \in \partial\Omega$ is spanned by the vectors $\partial_x\Phi_\pm$ and $\partial_y\Phi_\pm$. The cross-product of these two spanning vectors then is

$$\frac{\partial\Phi_\pm}{\partial x} \times \frac{\partial\Phi_\pm}{\partial y} = \begin{pmatrix} 1 \\ 0 \\ \gamma_{\pm,x}\end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ \gamma_{\pm,y}\end{pmatrix} = \begin{pmatrix} -\gamma_{\pm,x} \\ -\gamma_{\pm,y} \\ 1\end{pmatrix},$$

which gives us the outer unit normal vector on $\partial\Omega$:

$$\nu_+(x,y) = \frac{1}{\sqrt{1+\gamma_{+,x}^2+\gamma_{+,y}^2}}\begin{pmatrix} -\gamma_{+,x} \\ -\gamma_{+,y} \\ 1\end{pmatrix}, \qquad \nu_-(x,y) = \frac{1}{\sqrt{1+\gamma_{-,x}^2+\gamma_{-,y}^2}}\begin{pmatrix} \gamma_{-,x} \\ \gamma_{-,y} \\ -1\end{pmatrix}.$$

[38] Hufeisen

Finally, we have the two scalar surface elements

$$\mathrm{d}\sigma = \sqrt{1+\gamma_{\pm,x}^2+\gamma_{\pm,y}^2}\,\mathrm{d}(x,y).$$

Plugging these equations into each other, we then obtain

$$\int_\Omega R_z(x,y,z)\,\mathrm{d}(x,y,z) = \int_{\partial\Omega}\left\langle \begin{pmatrix}0\\0\\R\end{pmatrix}, \vec{\nu}\right\rangle \mathrm{d}\sigma.$$

Repeat with P and Q, and you are done. □

The following formulas are a variation on the partial integration. Green[39] found them 1828 when he was studying the theory of electricity and magnetism.

Proposition 5.43 (Green**'s formulas**). *Let $\Omega \subset \mathbb{R}^3$ be a Gauss normal domain, and $u, v\colon \overline{\Omega} \to \mathbb{R}$ functions from C^2. Then the* Green*'s formulas hold:*

$$\int_\Omega (u \bigtriangleup v + (\operatorname{grad} u)\cdot(\operatorname{grad} v))\,\mathrm{d}(x,y,z) = \int_{\partial\Omega} u\frac{\partial v}{\partial\vec{\nu}}\,\mathrm{d}\sigma, \tag{5.9}$$

$$\int_\Omega (u \bigtriangleup v - v \bigtriangleup u)\,\mathrm{d}(x,y,z) = \int_{\partial\Omega} \left(u\frac{\partial v}{\partial\vec{\nu}} - v\frac{\partial u}{\partial\vec{\nu}}\right)\mathrm{d}\sigma, \tag{5.10}$$

where $\vec{\nu}$ denotes the outward unit normal on $\partial\Omega$.

The expression $\frac{\partial v}{\partial\vec{\nu}}$ is the directional derivative of the function v in the direction of the outward normal.

Proof. Apply the Gauss theorem to $\vec{f} = u \operatorname{grad} v$, and you obtain (5.9). Swap the roles of u and v. Subtraction then gives you (5.10). □

Remark 5.44. *Take U as the vector space of all those functions $f \in C^2(\overline{\Omega} \to \mathbb{R})$ with $f = 0$ on the boundary. Introduce the scalar product $\langle f, g\rangle = \int_\Omega f(x)g(x)\,\mathrm{d}x$ for U (attention: this will* not *make U a Banach space). Then* Green*'s formulas imply $\langle \bigtriangleup f, g\rangle = \langle f, \bigtriangleup g\rangle$ for $f, g \in U$. In this sense, the Laplace operator is symmetric on U. In the context of quantum theory, you will rephrase this as "the Hamilton operator of the free electron is self–adjoint".*

Corollary 5.45. *Choosing $u \equiv 1$ in (5.9) yields the useful identity*

$$\int_\Omega \bigtriangleup v\,\mathrm{d}(x,y,z) = \int_{\partial\Omega} \frac{\partial v}{\partial\vec{\nu}}\,\mathrm{d}\sigma. \tag{5.11}$$

Remark 5.46 (**Fredholm's alternative revisited**). *In contrast to the previous remark, we now take U as the vector space of all those functions $f \in C^2(\overline{\Omega} \to \mathbb{R})$ with vanishing normal derivative on the boundary. Define the usual scalar product and put $\mathcal{A} := \bigtriangleup$. Then $\langle \mathcal{A}f, g\rangle = \langle f, \mathcal{A}g\rangle$ for all $f, g \in U$, and we can think of $\mathcal{A}$ as "self–adjoint". Hence $\mathcal{A} = \mathcal{A}^*$.*

The Fredholm alternative (Corollary 4.21) says that a system $Ax = b$ is solvable if and only if $b \perp \ker A^*$, in case of A being a matrix. Now $\mathcal{A}$ is no matrix, but the Fredholm alternative holds also now: suppose $\mathcal{A}v = b$ and $\frac{\partial}{\partial\vec{\nu}}v = 0$ on $\partial\Omega$. Then (5.11) says $\int_\Omega b\,\mathrm{d}x = 0$, which is equivalent to $\langle 1, b\rangle = 0$. But the function which is everywhere equal to one is exactly a basis of $\ker\mathcal{A}^* = \ker\mathcal{A}$, and then $\langle 1, b\rangle = 0$ simply means $b \perp \ker\mathcal{A}^*$. So we come to the belief that the Fredholm alternative holds also for the Laplace operator (with boundary conditions) on Gauss normal domains. (The details must be skipped.)

[39] George Green, 1793 – 1841

5.6.3 The Stokes Theorem in $\mathbb{R}^3$

Proposition 5.47 (Stokes[40] **theorem).** *Let $S \subset \mathbb{R}^3$ be an orientable surface patch with C^2 parametrisation $\Phi\colon U \to \mathbb{R}^3$ and parameter domain $U \subset \mathbb{R}^2$. We assume that U is a domain for which the Gauss theorem in $\mathbb{R}^2$ is valid. Additionally, suppose that U is simply connected. Moreover, assume that the boundary $\gamma = \partial U$ is a piecewise smooth Jordan curve, and that the parametrisation Φ is injective on a larger open set that contains U.*

Let $\vec{f}$ be a function that is continuously differentiable on a larger domain that contains the surface patch S. Then the following identity holds:

$$\int_S \operatorname{rot} \vec{f} \cdot \mathrm{d}\vec{\sigma} = \int_{\partial S} \vec{f} \cdot \mathrm{d}\vec{x}. \tag{5.12}$$

We can relax the assumptions a bit if we are willing to work harder. However, for simplicity reasons, we stick to this version of the Stokes theorem.

Proof. Formula (5.12) can be written as

$$\int_S \operatorname{rot} \vec{f} \cdot \vec{n} \, \mathrm{d}\sigma = \int_{\partial S} \vec{f} \cdot \mathrm{d}\vec{x},$$

with $\vec{n}(x)$ being the unit normal:

$$\vec{n} = \frac{(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)}{\|(\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\|}.$$

The idea of the proof is to write everything in terms of the parameters (u_1, u_2), and then to apply the Gauss theorem in $\mathbb{R}^2$. We need a good parametrisation of $\gamma = \partial U$, e.g., the arc-length parametrisation:

$$\gamma(s) = \begin{pmatrix} \gamma_1(s) \\ \gamma_2(s) \end{pmatrix}, \qquad 0 \le s \le L.$$

Then a parametrisation of ∂S is given by a function $\phi = \phi(s)\colon [0, L] \to \mathbb{R}^3$,

$$\phi(s) = \Phi(\gamma(s)), \qquad 0 \le s \le L.$$

Rewritten in terms of the parameters (u_1, u_2) and s, equation (5.12) then is

$$\int_U \Big((\operatorname{rot} \vec{f}) \circ \Phi\Big) \cdot \Big((\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)\Big) \, \mathrm{d}(u_1, u_2) = \int_{s=0}^{s=L} \vec{f}(\phi(s)) \cdot \phi'(s) \, \mathrm{d}s. \tag{5.13}$$

The function Φ and its Jacobi matrix are

$$\Phi = \begin{pmatrix} \Phi_1(u_1, u_2) \\ \Phi_2(u_1, u_2) \\ \Phi_3(u_1, u_2) \end{pmatrix}, \qquad \Phi' = \begin{pmatrix} \Phi_{1,1}(u_1, u_2) & \Phi_{1,2}(u_1, u_2) \\ \Phi_{2,1}(u_1, u_2) & \Phi_{2,2}(u_1, u_2) \\ \Phi_{3,1}(u_1, u_2) & \Phi_{3,2}(u_1, u_2) \end{pmatrix},$$

with $\Phi_{i,j}$ being the partial derivative of Φ_i with respect to u_j. Then $\phi' = \phi'(s)$ is given by

$$\phi'(s) = \Phi'(\gamma(s)) \cdot \gamma'(s) = \begin{pmatrix} \Phi_{1,1}(\gamma(s)) & \Phi_{1,2}(\gamma(s)) \\ \Phi_{2,1}(\gamma(s)) & \Phi_{2,2}(\gamma(s)) \\ \Phi_{3,1}(\gamma(s)) & \Phi_{3,2}(\gamma(s)) \end{pmatrix} \cdot \begin{pmatrix} \gamma_1'(s) \\ \gamma_2'(s) \end{pmatrix}.$$

Consider, for instance, the f_3 component in the right–hand side of (5.12). We have to evaluate

$$\int_{\partial S} f_3 \, \mathrm{d}x_3 = \int_{s=0}^{s=L} f_3(\phi(s))(\phi'(s))_3 \, \mathrm{d}s = \int_{s=0}^{s=L} f_3(\phi(s)) \left(\Phi_{3,1}(\gamma(s))\gamma_1'(s) + \Phi_{3,2}(\gamma(s))\gamma_2'(s)\right) \mathrm{d}s.$$

Our goal is to apply the Gauss theorem in $\mathbb{R}^2$ to this equation. The tangent vector to the curve γ is $(\gamma_1'(s), \gamma_2'(s))^\top$, and the outward unit normal vector to that curve is

$$\vec{\nu}(s) = \begin{pmatrix} \cos(-\frac{\pi}{2}) & -\sin(-\frac{\pi}{2}) \\ \sin(-\frac{\pi}{2}) & \cos(-\frac{\pi}{2}) \end{pmatrix} \begin{pmatrix} \gamma_1'(s) \\ \gamma_2'(s) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1'(s) \\ \gamma_2'(s) \end{pmatrix} = \begin{pmatrix} \gamma_2'(s) \\ -\gamma_1'(s) \end{pmatrix}.$$

[40] Sir George Gabriel Stokes, 1819 – 1903

Then we have $\Phi_{3,1}\gamma_1' + \Phi_{3,2}\gamma_2' = \langle (\Phi_{3,2}, -\Phi_{3,1})^\top, \vec{\nu}\rangle$, and, consequently,

$$\begin{aligned}\int_{\partial S} f_3 \,\mathrm{d}x_3 &= \int_{s=0}^{s=L} f_3(\Phi(\gamma(s))) \left\langle \begin{pmatrix} \Phi_{3,2}(\gamma(s)) \\ -\Phi_{3,1}(\gamma(s)) \end{pmatrix}, \vec{\nu}(s) \right\rangle \mathrm{d}s \\ &= \int_{\partial U} f_3(\Phi) \left\langle \begin{pmatrix} \Phi_{3,2} \\ -\Phi_{3,1} \end{pmatrix}, \vec{\nu} \right\rangle \mathrm{d}s \\ &= \int_U \operatorname{div} \begin{pmatrix} f_3(\Phi)\Phi_{3,2} \\ -f_3(\Phi)\Phi_{3,1} \end{pmatrix} \mathrm{d}(u_1, u_2),\end{aligned}$$

the last relation following from the Gauss theorem in $\mathbb{R}^2$.

Now we compute the divergence in the integrand. The vector $\vec{f}$ depends on $\vec{x}$. Let $f_{3,k}$ denote the derivative of f_3 with respect to x_k. We write $\Phi_{3,11}$, $\Phi_{3,12} = \Phi_{3,21}$, $\Phi_{3,22}$ for the second order derivatives of the function Φ_3. Then we infer from the chain rule that

$$\begin{aligned}\operatorname{div} \begin{pmatrix} f_3(\Phi)\Phi_{3,2} \\ -f_3(\Phi)\Phi_{3,1} \end{pmatrix} &= \frac{\partial}{\partial u_1}(f_3(\Phi)\Phi_{3,2}) - \frac{\partial}{\partial u_2}(f_3(\Phi)\Phi_{3,1}) \\ &= \sum_{k=1}^3 (f_{3,k}\Phi_{k,1}\Phi_{3,2} - f_{3,k}\Phi_{k,2}\Phi_{3,1}) + f_3\Phi_{3,21} - f_3\Phi_{3,12} \\ &= f_{3,1}(\Phi_{1,1}\Phi_{3,2} - \Phi_{1,2}\Phi_{3,1}) + f_{3,2}(\Phi_{2,1}\Phi_{3,2} - \Phi_{2,2}\Phi_{3,1}). \end{aligned} \tag{5.14}$$

Next, we compute the integrand of the left–hand side of (5.13). A straight-forward computation reveals

$$\begin{aligned}(\operatorname{rot} \vec{f}) \cdot ((\partial_{u_1}\Phi) \times (\partial_{u_2}\Phi)) &= \begin{pmatrix} f_{3,2} - f_{2,3} \\ f_{1,3} - f_{3,1} \\ f_{2,1} - f_{1,2} \end{pmatrix} \cdot \left(\begin{pmatrix} \Phi_{1,1} \\ \Phi_{2,1} \\ \Phi_{3,1} \end{pmatrix} \times \begin{pmatrix} \Phi_{1,2} \\ \Phi_{2,2} \\ \Phi_{3,2} \end{pmatrix} \right) \\ &= \begin{pmatrix} f_{3,2} - f_{2,3} \\ f_{1,3} - f_{3,1} \\ f_{2,1} - f_{1,2} \end{pmatrix} \cdot \begin{pmatrix} \Phi_{2,1}\Phi_{3,2} - \Phi_{3,1}\Phi_{2,2} \\ \Phi_{3,1}\Phi_{1,2} - \Phi_{1,1}\Phi_{3,2} \\ \Phi_{1,1}\Phi_{2,2} - \Phi_{2,1}\Phi_{1,2} \end{pmatrix}.\end{aligned}$$

We collect only the terms with f_3, they are:

$$f_{3,2}(\Phi_{2,1}\Phi_{3,2} - \Phi_{3,1}\Phi_{2,2}) - f_{3,1}(\Phi_{3,1}\Phi_{1,2} - \Phi_{1,1}\Phi_{3,2}).$$

But this is exactly the same as (5.14).

The other two terms, $\int_{\partial S} f_1 \,\mathrm{d}x_1$ and $\int_{\partial S} f_2 \,\mathrm{d}x_2$, can be considered in the same manner. This completes the proof of the Stokes theorem. □

This proof of Stokes' Theorem in $\mathbb{R}^3$ was by hard computation. We have found out that the formula (5.12) holds, but the deeper reason remains mysterious. The author hopes that the following second proof of the Stokes Theorem might give more insights.

We wish to prove (5.12), which is

$$\int_S \operatorname{rot} \vec{f} \cdot \mathrm{d}\vec{\sigma} = \int_{\partial S} \vec{f} \cdot \mathrm{d}\vec{x}.$$

Our strategy will be:

Step 0: we prepare some tools (Einstein summation convention and the Levi–Civita tensor),

Step 1: we start with the left-hand side and plug in the parametrisation $\Phi\colon U \to \mathbb{R}^3$ of the surface patch S,

Step 2: we obtain an integral $\int_U \boxed{?} \,\mathrm{d}u_1 \,\mathrm{d}u_2$,

Step 3: we apply the Gauss Theorem in $\mathbb{R}^2$,

Step 4: we obtain a curve integral of second kind $\int_{\partial U} \boxed{?} \,\mathrm{d}u_1 + \boxed{?} \,\mathrm{d}u_2$,

Step 5: we plug in the parametrisation $\gamma\colon [0, L] \to \mathbb{R}^2$ for the boundary ∂U,

Step 6: we obtain an integral $\int_{s=0}^{L} \boxed{?}\,\mathrm{d}s$,

Step 7: we see (more or less directly) that this is the desired right-hand side.

Our first tool is the **Einstein summation convention**. This means: whenever in a product one index appears twice, then we sum over this index from one to three (without writing the $\sum$ symbol). For instance, $a_{ij}b_j$ is an abbreviation for $a_{i1}b_1 + a_{i2}b_2 + a_{i3}b_3$.

A first application is the chain rule: when $x = \Phi(u)$, then

$$\frac{\partial f_k(\Phi(u))}{\partial u_l} = \frac{\partial f_k}{\partial x_m}\frac{\partial \Phi_m}{\partial u_l}.$$

Here the relevant summation index is m.

Remark 5.48. *There is a deeper physical reason why in a product a summation index may appear once or twice, but never thrice: the reason is that many such expressions are in fact pairings between an element of one vector space V and an element of the dual vector space V'. Recall the definition of a dual vector space: if V is a vector space over the field K, then the dual space V' contains all the linear mappings from V into (the one-dimensional space) K. We give some examples:*

the dual space V' consists of ...	when the space V contains the ...
wave vectors $k \in \mathbb{R}^3$	position variables $x \in \mathbb{R}^3$
frequencies $\omega \in \mathbb{R}$	time variable $t \in \mathbb{R}$
bra-vectors $\langle\psi\vert$	ket-vectors $\vert\phi\rangle$
differential 1-forms	tangent vectors

Imagine that a typical element of V' just waits for an element of V to come along, then it eats it, and produces a number from K in a linear manner.

Pairing a bra-vector $\langle\psi|$ with a ket-vector $|\phi\rangle$ (both are objects from quantum mechanics) then gives the number $\langle\psi|\phi\rangle$, which is another way of writing the bracket $\langle\psi,\phi\rangle$. And what we call $\langle Ax, y\rangle$ in this lecture, will be written in the quantum mechanics course as $\langle y|A|x\rangle$.

Differential 1-forms will be discussed in the final section of this script.

Our second tool is the **Levi–Civita tensor**. For $i, j, k \in \{1, 2, 3\}$, we define

$$\varepsilon_{ijk} = \begin{cases} +1 & : (i,j,k) \text{ is an even permutation of } (1,2,3), \\ -1 & : (i,j,k) \text{ is an odd permutation of } (1,2,3), \\ 0 & : \text{ else.} \end{cases}$$

In particular, $\varepsilon_{ijk} = 0$ if two of its indices have the same value. The key purpose of this tensor is a simpler expression for the vectorial product in $\mathbb{R}^3$:

$$(\vec{a} \times \vec{b})_i = \varepsilon_{ijk} a_j b_k,$$

with Einstein summation convention with respect to j and k. In particular, we have

$$(\operatorname{rot} \vec{f})_i = \varepsilon_{ijk}\frac{\partial}{\partial x_j} f_k,$$

$$(\,\mathrm{d}\vec{\sigma})_i = n_i \,\mathrm{d}\sigma = \varepsilon_{ipq}\frac{\partial \Phi_p}{\partial u_1}\frac{\partial \Phi_q}{\partial u_2}\,\mathrm{d}u_1\,\mathrm{d}u_2.$$

Lemma 5.49. *Let $j, k, p, q \in \{1, 2, 3\}$ be given. Then it holds*

$$\varepsilon_{ijk}\varepsilon_{ipq} = \delta_{jp}\delta_{kq} - \delta_{jq}\delta_{kp},$$

with $\delta_{..}$ being Kronecker's delta.

Sketch of proof. Consider the case $(j, k) = (2, 3)$ and the case $(j, k) = (2, 2)$. □

Now we prove Stokes' Theorem once again.

Proof. We have, making extensive use of Einstein's summation convention,

$$
\begin{aligned}
\int_S \operatorname{rot} \vec{f} \cdot \mathrm{d}\vec{\sigma} &= \int_S (\operatorname{rot} \vec{f})_i (\,\mathrm{d}\vec{\sigma}_i) \\
&= \int_U \varepsilon_{ijk} \frac{\partial f_k(\Phi(u))}{\partial x_j} \varepsilon_{ipq} \frac{\partial \Phi_p}{\partial u_1} \frac{\partial \Phi_q}{\partial u_2} \,\mathrm{d}u_1 \,\mathrm{d}u_2 \\
&= \int_U (\delta_{jp}\delta_{kq} - \delta_{jq}\delta_{kp}) \frac{\partial f_k(\Phi(u))}{\partial x_j} \frac{\partial \Phi_p}{\partial u_1} \frac{\partial \Phi_q}{\partial u_2} \,\mathrm{d}u_1 \,\mathrm{d}u_2 \qquad \Big| \quad \text{because of Lemma 5.49} \\
&= \int_U \frac{\partial f_k(\Phi(u))}{\partial x_j} \frac{\partial \Phi_j}{\partial u_1} \frac{\partial \Phi_k}{\partial u_2} \,\mathrm{d}u_1 \,\mathrm{d}u_2 - \int_U \frac{\partial f_k(\Phi(u))}{\partial x_j} \frac{\partial \Phi_k}{\partial u_1} \frac{\partial \Phi_j}{\partial u_2} \,\mathrm{d}u_1 \,\mathrm{d}u_2 \\
&= \int_U \frac{\partial f_k(\Phi(u))}{\partial u_1} \frac{\partial \Phi_k}{\partial u_2} \,\mathrm{d}u_1 \,\mathrm{d}u_2 - \int_U \frac{\partial f_k(\Phi(u))}{\partial u_2} \frac{\partial \Phi_k}{\partial u_1} \,\mathrm{d}u_1 \,\mathrm{d}u_2 \qquad \Big| \quad \text{``chain rule backwards''} \\
&= \int_U \frac{\partial}{\partial u_1} \left(f_k(\Phi(u)) \frac{\partial \Phi_k}{\partial u_2} \right) \,\mathrm{d}u_1 \,\mathrm{d}u_2 \\
&\quad - \int_U \frac{\partial}{\partial u_2} \left(f_k(\Phi(u)) \frac{\partial \Phi_k}{\partial u_1} \right) \,\mathrm{d}u_1 \,\mathrm{d}u_2 \qquad \Big| \quad \text{because of the Schwarz theorem on } 2^{\text{nd}} \text{ derivatives} \\
&= \int_{\partial U} \left(f_k(\Phi(u)) \frac{\partial \Phi_k(u)}{\partial u_2} \,\mathrm{d}u_2 + f_k(\Phi(u)) \frac{\partial \Phi_k(u)}{\partial u_1} \,\mathrm{d}u_1 \right) \qquad \Big| \quad \text{because of the Gauss theorem (5.7)} \\
&= \int_{s=0}^{L} \left(f_k(\Phi(\gamma(s))) \frac{\partial \Phi_k(\gamma(s))}{\partial u_2} \frac{\mathrm{d}\gamma_2}{\mathrm{d}s} + f_k(\Phi(\gamma(s))) \frac{\partial \Phi_k(\gamma(s))}{\partial u_1} \frac{\mathrm{d}\gamma_1}{\mathrm{d}s} \right) \mathrm{d}s \qquad \Big| \quad \text{plug in } u = \gamma(s) \\
&= \int_{s=0}^{L} f_k((\Phi \circ \gamma)(s)) \left(\frac{\partial \Phi_k(\gamma(s))}{\partial u_1} \frac{\mathrm{d}\gamma_1(s)}{\mathrm{d}s} + \frac{\partial \Phi_k(\gamma(s))}{\partial u_2} \frac{\mathrm{d}\gamma_2(s)}{\mathrm{d}s} \right) \mathrm{d}s \\
&= \int_{s=0}^{L} f_k(\phi(s)) \frac{\mathrm{d}\phi_k(s)}{\mathrm{d}s} \,\mathrm{d}s \qquad \Big| \quad \text{recall } \phi := \Phi \circ \gamma \text{, and ``chain rule backwards''} \\
&= \int_{s=0}^{L} \vec{f}(\phi(s)) \cdot \vec{\phi}'(s) \,\mathrm{d}s \\
&= \int_{\partial S} \vec{f} \cdot \mathrm{d}\vec{x} \qquad \Big| \quad \text{that is the definition of curve integrals !}
\end{aligned}
$$

The second proof of the Stokes theorem is complete. □

5.7 Outlook: the Stokes Theorem in General Form

(Outlook sections are not relevant for exams.)

The integral theorems proved on the last pages can be brought into a bigger context: the Stokes theorem on *differential forms*. Now we try to explain what this means.

A differential form is an expression like this:

$$\omega := u(x,y,z)\,\mathrm{d}x \wedge \mathrm{d}y + v(x,y,z)\,\mathrm{d}y \wedge \mathrm{d}z + w(x,y,z)\,\mathrm{d}z \wedge \mathrm{d}x. \tag{5.15}$$

This is an example of a 2-form, because of the two differentials "multiplied" using the wedge symbol $\wedge$. Here x, y, z are the usual variables in $\mathbb{R}^3$, and u, v, w are smooth functions. We intentionally do not explain precisely what the d and $\wedge$ symbols *mean*, but only write down how the calculations *work*.

The following presentation unfolds like this:

- first we show the purpose of differential forms and show how to integrate a differential 1-form,
- second we list the properties of the wedge product $\wedge$,
- then we show how to compute the integral of a 2-form in $\mathbb{R}^3$,
- next we present the *exterior derivative* d, and how it relates to grad, rot, div in $\mathbb{R}^3$,

- and finally we show how these ingredients come together, building the celebrated Stokes theorem on differential forms,

$$\int_M \mathrm{d}\omega = \int_{\partial M} \omega. \tag{5.16}$$

The purpose of differential forms

To make a long story short: a differential form is an object which is waiting to be integrated.

Let us think about what we are doing when we integrate a function f (vector-valued or scalar-valued) over a domain M (where M could be a curve or a surface patch or something similar). First we cut M into little parts. Then we need two pieces of information:

- we need to know a typical value of f on such a little piece of M,
- we need to know the size of that little piece of M.

Then we multiply these two pieces of information (in a certain manner, perhaps using a scalar product), and we sum over all the little parts of M.

A differential form like

$$\omega := u(x, y, z)\,\mathrm{d}x + v(x, y, z)\,\mathrm{d}y + w(x, y, z)\,\mathrm{d}z \tag{5.17}$$

unites both mentioned pieces of information in one line. The functions u, v, w can be combined to a vectorial function $\vec{f} = (u, v, w)^\top$, and the expressions $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$ take care of measuring the sizes of the little parts of M, where M will be a curve in $\mathbb{R}^3$. Then $\int_M \omega$ will be a curve integral of second kind in $\mathbb{R}^3$. To compute this integral, we choose one parametrisation $\gamma\colon [0, L] \to \mathbb{R}^3$ — it does not matter which parametrisation, because they all lead to the same final result, which is

$$\int_M \omega = \int_{t=0}^{L} \langle \omega, \dot{\gamma} \rangle \;\mathrm{d}t,$$

with $\langle \omega, \dot{\gamma} \rangle$ being an abbreviation like this:

$$\langle \omega, \dot{\gamma} \rangle := u(\gamma(t))\dot{\gamma}_x(t) + v(\gamma(t))\dot{\gamma}_y(t) + w(\gamma(t))\dot{\gamma}_z(t), \qquad \gamma = (\gamma_x, \gamma_y, \gamma_z)^\top.$$

Pick a point P on M. Whatever parametrisation γ was chosen — the tangential vector $\dot{\gamma}$ must lie on the tangential line at the point P on M. In this sense, all the possible tangential vectors at P form a one-dimensional vector space, and the differential form ω turns such a tangential vector $\dot{\gamma}$ into the real number $\langle \omega, \dot{\gamma} \rangle$. In this sense, the differential 1-forms are members of the dual vector space to the vector space of tangential vectors. This completes the discussion in Remark 5.48.

The wedge product $\wedge$

The product $\wedge$ is called *exterior product* or *outer product* (in contrast to the *inner product* which is the usual scalar product in $\mathbb{R}^n$). An example of a 1-form is (5.17), and an example of a 2-form is (5.15). If you have a k-form ω and an ℓ-form ϱ, then $\omega \wedge \varrho$ will be a $(k + \ell)$-form, which is obtained in the usual manner: you multiply each summand in ω and each summand in ϱ, and then you add up these products. Only one rule is special: The wedge product is anti-commutative in the following sense: if p is one of the variables x, y, z, and if q is also one of the variables x, y, z, then $\mathrm{d}p \wedge \mathrm{d}q = -\,\mathrm{d}q \wedge \mathrm{d}p$. In particular, we get $\mathrm{d}x \wedge \mathrm{d}x = -\,\mathrm{d}x \wedge \mathrm{d}x$, hence $\mathrm{d}x \wedge \mathrm{d}x = 0$.

How to integrate a 2-form in $\mathbb{R}^3$

We wish to integrate

$$\omega = R(x, y, z)\,\mathrm{d}x \wedge \mathrm{d}y + P(x, y, z)\,\mathrm{d}y \wedge \mathrm{d}z + Q(x, y, z)\,\mathrm{d}z \wedge \mathrm{d}x$$

over M. This is a 2-form in $\mathbb{R}^3$, hence M must be a two-dimensional object in $\mathbb{R}^3$, therefore M is a surface patch in $\mathbb{R}^3$, with parametrisation

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \Phi_x(u_1,u_2) \\ \Phi_y(u_1,u_2) \\ \Phi_z(u_1,u_2) \end{pmatrix}, \qquad (u_1,u_2) \in U \subset \mathbb{R}^2.$$

Our calculations are straight-forward:

$$\begin{aligned} \mathrm{d}x &= \mathrm{d}\Phi_x = \frac{\partial \Phi_x}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial \Phi_x}{\partial u_2}\,\mathrm{d}u_2, \\ \mathrm{d}y &= \mathrm{d}\Phi_y = \frac{\partial \Phi_y}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial \Phi_y}{\partial u_2}\,\mathrm{d}u_2, \\ \mathrm{d}z &= \mathrm{d}\Phi_z = \frac{\partial \Phi_z}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial \Phi_z}{\partial u_2}\,\mathrm{d}u_2, \end{aligned} \tag{5.18}$$

hence

$$\begin{aligned} \mathrm{d}x \wedge \mathrm{d}y &= \left(\frac{\partial \Phi_x}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial \Phi_x}{\partial u_2}\,\mathrm{d}u_2\right) \wedge \left(\frac{\partial \Phi_y}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial \Phi_y}{\partial u_2}\,\mathrm{d}u_2\right) \\ &= \frac{\partial \Phi_x}{\partial u_1}\frac{\partial \Phi_y}{\partial u_1}\,\mathrm{d}u_1 \wedge \mathrm{d}u_1 + \frac{\partial \Phi_x}{\partial u_1}\frac{\partial \Phi_y}{\partial u_2}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2 + \frac{\partial \Phi_x}{\partial u_2}\frac{\partial \Phi_y}{\partial u_1}\,\mathrm{d}u_2 \wedge \mathrm{d}u_1 + \frac{\partial \Phi_x}{\partial u_2}\frac{\partial \Phi_y}{\partial u_2}\,\mathrm{d}u_2 \wedge \mathrm{d}u_2 \\ &= 0 + \frac{\partial \Phi_x}{\partial u_1}\frac{\partial \Phi_y}{\partial u_2}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2 + \frac{\partial \Phi_x}{\partial u_2}\frac{\partial \Phi_y}{\partial u_1}\,\mathrm{d}u_2 \wedge \mathrm{d}u_1 + 0 \\ &= \left(\frac{\partial \Phi_x}{\partial u_1}\frac{\partial \Phi_y}{\partial u_2} - \frac{\partial \Phi_x}{\partial u_2}\frac{\partial \Phi_y}{\partial u_1}\right)\mathrm{d}u_1 \wedge \mathrm{d}u_2 \\ &= \left(\frac{\partial \Phi}{\Phi u_1} \times \frac{\partial \Phi}{\Phi u_2}\right)_z \mathrm{d}u_1 \wedge \mathrm{d}u_2. \end{aligned}$$

Similarly we find

$$\mathrm{d}y \wedge \mathrm{d}z = \left(\frac{\partial \Phi}{\partial u_1} \times \frac{\partial \Phi}{\partial u_2}\right)_x \mathrm{d}u_1 \wedge \mathrm{d}u_2, \qquad \mathrm{d}z \wedge \mathrm{d}x = \left(\frac{\partial \Phi}{\partial u_1} \times \frac{\partial \Phi}{\partial u_2}\right)_y \mathrm{d}u_1 \wedge \mathrm{d}u_2,$$

and then $\int_M \omega$ (for our specially written ω) turns into

$$\int_U \left\langle \begin{pmatrix} P \\ Q \\ R \end{pmatrix}, \frac{\partial \Phi}{\partial u_1} \times \frac{\partial \Phi}{\partial u_2} \right\rangle \mathrm{d}u_1 \wedge \mathrm{d}u_2.$$

Since $(u_1,u_2) \in U \subset \mathbb{R}^2$, and U is flat (it has no curvature, but M probably is curved), it is reasonable to make the agreement $\mathrm{d}u_1 \wedge \mathrm{d}u_2 := \mathrm{d}u_1\,\mathrm{d}u_2$. Our result then is $\int_M \omega = \int_M \vec{f} \cdot \mathrm{d}\vec{\sigma}$, as a surface integral of second kind, with $\vec{f} := (P,Q,R)^\top$.

The exterior derivative d

The derivative operator d turns a k-form into a $(k+1)$-form, and we define it as follows: first we make the agreement that a scalar function $\varphi = \varphi(x,y,z)$ on $\mathbb{R}^3$ is a 0-form, to which we can apply the exterior derivative in a natural way:

$$\mathrm{d}\varphi := \frac{\partial \varphi}{\partial x}\,\mathrm{d}x + \frac{\partial \varphi}{\partial y}\,\mathrm{d}y + \frac{\partial \varphi}{\partial z}\,\mathrm{d}z.$$

We have exploited a variant of this rule in (5.18). See also Definition 1.11 for the total differential, which obeys the same formula (but with a slightly different meaning).

And if

$$\omega = \sum_{i_1=1}^{n} \dots \sum_{i_k=1}^{n} f_{i_1 \dots i_k}(x)\,\mathrm{d}x_{i_i} \wedge \ldots \wedge \mathrm{d}x_{i_k}$$

is a k-form in $\mathbb{R}^n$, then we define (with $f_{i_1\dots i_k}$ to be understood as 0-form)

$$\mathrm{d}\omega := \sum_{i_1=1}^{n} \dots \sum_{i_k=1}^{n} (\,\mathrm{d}f_{i_1\dots i_k}) \wedge \mathrm{d}x_{i_i} \wedge \dots \wedge \mathrm{d}x_{i_k}.$$

We discuss examples, and they will tell us how d is related to the operators grad, rot, div in $\mathbb{R}^3$.

$\omega = \varphi$ is a 0-form on $\mathbb{R}^n$: then $\mathrm{d}\omega = \frac{\partial\varphi}{\partial x_1}\,\mathrm{d}x_1 + \dots + \frac{\partial\varphi}{\partial x_n}\,\mathrm{d}x_n$ is a 1-form, and its components look like the components of $\operatorname{grad}\varphi$.

ω is a 1-form on $\mathbb{R}^3$: then $\omega = u(x,y,z)\,\mathrm{d}x + v(x,y,z)\,\mathrm{d}y + w(x,y,z)\,\mathrm{d}z$, hence

$$\begin{aligned}
\mathrm{d}\omega &= (\,\mathrm{d}u) \wedge \mathrm{d}x + (\,\mathrm{d}v) \wedge \mathrm{d}y + (\,\mathrm{d}w) \wedge \mathrm{d}z \\
&= \left(\frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y + \frac{\partial u}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}x \\
&\quad + \left(\frac{\partial v}{\partial x}\,\mathrm{d}x + \frac{\partial v}{\partial y}\,\mathrm{d}y + \frac{\partial v}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}y \\
&\quad + \left(\frac{\partial w}{\partial x}\,\mathrm{d}x + \frac{\partial w}{\partial y}\,\mathrm{d}y + \frac{\partial w}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}z \\
&= \frac{\partial u}{\partial x}\,\mathrm{d}x \wedge \mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y \wedge \mathrm{d}x + \frac{\partial u}{\partial z}\,\mathrm{d}z \wedge \mathrm{d}x \\
&\quad + \frac{\partial v}{\partial x}\,\mathrm{d}x \wedge \mathrm{d}y + \frac{\partial v}{\partial y}\,\mathrm{d}y \wedge \mathrm{d}y + \frac{\partial v}{\partial z}\,\mathrm{d}z \wedge \mathrm{d}y \\
&\quad + \frac{\partial w}{\partial x}\,\mathrm{d}x \wedge \mathrm{d}z + \frac{\partial w}{\partial y}\,\mathrm{d}y \wedge \mathrm{d}z + \frac{\partial w}{\partial z}\,\mathrm{d}z \wedge \mathrm{d}z \\
&= 0 - \frac{\partial u}{\partial y}\,\mathrm{d}x \wedge \mathrm{d}y + \frac{\partial u}{\partial z}\,\mathrm{d}z \wedge \mathrm{d}x \\
&\quad + \frac{\partial v}{\partial x}\,\mathrm{d}x \wedge \mathrm{d}y + 0 - \frac{\partial v}{\partial z}\,\mathrm{d}y \wedge \mathrm{d}z \\
&\quad - \frac{\partial w}{\partial x}\,\mathrm{d}z \wedge \mathrm{d}x + \frac{\partial w}{\partial y}\,\mathrm{d}y \wedge \mathrm{d}z + 0 \\
&= \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\,\mathrm{d}x \wedge \mathrm{d}y + \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right)\,\mathrm{d}y \wedge \mathrm{d}z + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right)\,\mathrm{d}z \wedge \mathrm{d}x,
\end{aligned}$$

and now the three components of this 2-form look like the components of $\operatorname{rot}\vec{f}$, with $\vec{f} = (u,v,w)^\top$. In that sense: d applied to a 1-form in $\mathbb{R}^3$ gives a 2-form, and your calculation has similarities to the calculation of the rotation of a vector field.

ω is a 2-form on $\mathbb{R}^3$: then $\omega = R(x,y,z)\,\mathrm{d}x \wedge \mathrm{d}y + P(x,y,z)\,\mathrm{d}y \wedge \mathrm{d}z + Q(x,y,z)\,\mathrm{d}z \wedge \mathrm{d}x$, and we put $\vec{f} = (P,Q,R)^\top$. Then our computing rules give

$$\begin{aligned}
\mathrm{d}\omega &= \left(\frac{\partial R}{\partial x}\,\mathrm{d}x + \frac{\partial R}{\partial y}\,\mathrm{d}y + \frac{\partial R}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}x \wedge \mathrm{d}y \\
&\quad + \left(\frac{\partial P}{\partial x}\,\mathrm{d}x + \frac{\partial P}{\partial y}\,\mathrm{d}y + \frac{\partial P}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}y \wedge \mathrm{d}z \\
&\quad + \left(\frac{\partial Q}{\partial x}\,\mathrm{d}x + \frac{\partial Q}{\partial y}\,\mathrm{d}y + \frac{\partial Q}{\partial z}\,\mathrm{d}z\right) \wedge \mathrm{d}z \wedge \mathrm{d}x \\
&= \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right)\,\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = (\operatorname{div}\vec{f})\,\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z.
\end{aligned}$$

A key property of the exterior derivative d is $\mathrm{d} \circ \mathrm{d} = 0$. Applying this key property to scalar functions in $\mathbb{R}^3$, we get the well-known rule $\operatorname{rot}\operatorname{grad} = 0$. And applied to 1-forms in $\mathbb{R}^3$, it corresponds to the rule $\operatorname{div}\operatorname{rot} = 0$.

The Stokes Theorem on differential forms

Now we discuss examples of the general Stokes theorem (5.16).

$M = \Gamma$ is a curve from A to B, and $\omega = \varphi$ is a function: then $\partial M = \{A, B\}$ is the set of the two end-points of the curve, and $\mathrm{d}\omega = \mathrm{d}\varphi = \varphi_x \,\mathrm{d}x + \varphi_y \,\mathrm{d}y + \varphi_z \,\mathrm{d}z$ (with φ_x, φ_y, φ_z being the partial derivatives of φ), and the general Stokes theorem (5.16) turns into

$$\int_\Gamma \varphi_x \,\mathrm{d}x + \varphi_y \,\mathrm{d}y + \varphi_z \,\mathrm{d}z = \varphi(B) - \varphi(A),$$

which is the path–independence of the integral over a gradient field, proved in Proposition 3.77.

$M = \Omega$ is a domain in $\mathbb{R}^2$, and ω is a 1-form: then ∂M is the boundary of Ω, which is a curve in the plane, and ω can be written as

$$\omega = -Q(x, y) \,\mathrm{d}x + P(x, y) \,\mathrm{d}y,$$

with certain functions P and Q. Put $\vec{f} = (P, Q)^\top$. Then our computing rules give

$$\begin{aligned} \mathrm{d}\omega &= (-Q_x \,\mathrm{d}x - Q_y \,\mathrm{d}y) \wedge \mathrm{d}x + (P_x \,\mathrm{d}x + P_y \,\mathrm{d}y) \wedge \mathrm{d}y \\ &= (P_x + Q_y) \,\mathrm{d}x \wedge \mathrm{d}y = (\operatorname{div} \vec{f}) \,\mathrm{d}x \wedge \mathrm{d}y, \end{aligned}$$

and the general Stokes theorem (5.16) turns into (5.7).

$M = \Omega$ is a domain in $\mathbb{R}^3$, and ω is a 2-form: then ∂M is the boundary of Ω, which is a surface in the space, and ω can be written as

$$\omega = R(x, y, z) \,\mathrm{d}x \wedge \mathrm{d}y + P(x, y, z) \,\mathrm{d}y \wedge \mathrm{d}z + Q(x, y, z) \,\mathrm{d}z \wedge \mathrm{d}x,$$

with certain functions P, Q, R. Put $\vec{f} = (P, Q, R)^\top$. Then our computing rules give $\mathrm{d}\omega = (\operatorname{div} \vec{f}) \,\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, and the general Stokes theorem (5.16) turns into (5.8).

$M = S$ is a surface patch in $\mathbb{R}^3$, and ω is a 1-form: then ∂M is the boundary of S, which is a curve in $\mathbb{R}^3$, and ω can be written as

$$\omega = f_1 \,\mathrm{d}x_1 + f_2 \,\mathrm{d}x_2 + f_3 \,\mathrm{d}x_3,$$

with scalar functions f_1, f_2, f_3 and $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Writing $f_{j,k}$ for the derivative of f_j with respect to x_k, from our computing rules we then get

$$\begin{aligned} \mathrm{d}\omega = {} & (f_{1,1} \,\mathrm{d}x_1 + f_{1,2} \,\mathrm{d}x_2 + f_{1,3} \,\mathrm{d}x_3) \wedge \mathrm{d}x_1 \\ & + (f_{2,1} \,\mathrm{d}x_1 + f_{2,2} \,\mathrm{d}x_2 + f_{2,3} \,\mathrm{d}x_3) \wedge \mathrm{d}x_2 \\ & + (f_{3,1} \,\mathrm{d}x_1 + f_{3,2} \,\mathrm{d}x_2 + f_{3,3} \,\mathrm{d}x_3) \wedge \mathrm{d}x_3 \\ = {} & (f_{2,1} - f_{1,2}) \,\mathrm{d}x_1 \wedge \mathrm{d}x_2 + (f_{3,2} - f_{2,3}) \,\mathrm{d}x_2 \wedge \mathrm{d}x_3 + (f_{1,3} - f_{3,2}) \,\mathrm{d}x_3 \wedge \mathrm{d}x_1, \end{aligned}$$

and the general Stokes theorem (5.16) turns into (5.12).

Of course you can generalize this to higher dimensions. Perhaps it is a good idea to read Section 1.5 once again. Several important physical theories become beautiful (to the trained eye, admittedly) if you write their equations in terms of differential forms, for instance the electro-magnetism (see the *Teubner-Taschenbuch der Mathematik, Teil II*, 7. Auflage 1995, 10.2.9) and also the theory of special relativity (see: Hubert Goenner, *Spezielle Relativitätstheorie und die klassische Feldtheorie*, 1. Auflage 2004).

5.8 Keywords

- Fubini's theorem,
- substitution rule,
- surfaces and how to parametrise them,
- two kinds of surface integrals,
- the three integral theorems.

Bibliography

[1] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions.* Harri Deutsch, Frankfurt, 1984.

[2] I.N. Bronstein, K.A. Semendjajew, G. Musiol, and H. Mühlig. *Handbook of mathematics. Transl. from the German. 4th ed.* Julius Springer Verlag, 2004.

[3] Klemens Burg, Herbert Haf, and Friedrich Wille. *Höhere Mathematik für Ingenieure 1-5.* B.G. Teubner Stuttgart, 1989.

[4] Harro Heuser. *Lehrbuch der Analysis.* B.G.Teubner Stuttgart Leipzig, 1998.

[5] Klaus Jänich. *Analysis für Physiker und Ingenieure.* Julius Springer Verlag, 2001.

[6] Klaus Jänich. *Mathematik 1,2. Geschrieben für Physiker.* Julius Springer Verlag, 2001.

[7] Ivan Kuscer and Alojz Kodre. *Mathematik in Physik und Technik.* Julius Springer Verlag, 1993.

Index