Analysis II
for
computer scientists

Script of the lecture

held at the Cooperative State University Stuttgart,
Baden-Wuerttemberg

Course: STG-TINF16A
Lecturer: Dipl.-Ing. Tim Lindemann

September 4, 2017

Contents

Chapter I

Functions in multiple variables

1 Introduction

Last semester, we invested a significant amount of work in studying the behavior of functions

$$f : \mathbb{R} \to \mathbb{R}, \;\; x \mapsto f(x),$$

i.e., of functions depending solely on one independent variable $x \in \mathbb{R}$. Unfortunately, it is rather seldom the case that real world systems depend solely on one variable. In general, you will find more complex inter-variable relationships for the outcome f you are interested in. We list some examples:

(i) The trajectory and associated throwing range of a thrown body depends on the initial speed v_0 and initial angle α. Therefore it can be modeled by a two-variable function

$$f(v_0, \alpha) = \frac{v_0^2 \sin(2\alpha)}{g},$$

where g is the acceleration of gravity. We can associate a unique throwing range for every pair (v_0, α). A reasonable domain for this function would be $\mathbb{R}_+ \times (0, 90°]$.

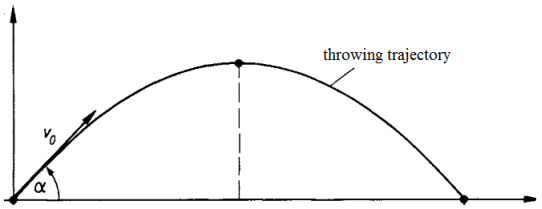
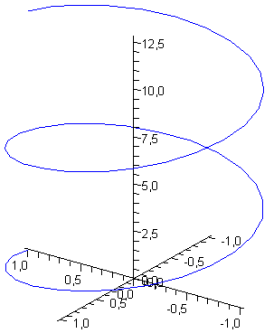


Figure I.1: Trajectory of thrown body.

(ii) An easy task already encountered in school is the determination of length characteristics for simple objects, e.g., the perimeter of squares, triangles, polygons or circles. However, there are many more complicated objects in our environment: what about the length of a spiral or the length of a rollercoaster (as shown in Figure I.2)?



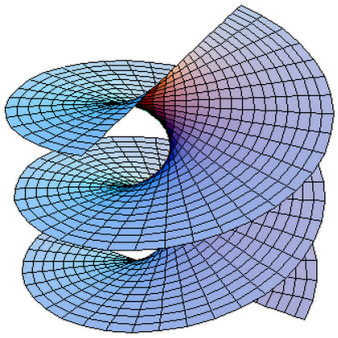
(a) Spiral

(b) Rollercoaster

Figure I.2: More complex Perimeters?

The calculation of lengths and curvatures of more generally shaped objects is performed using the so-called *contour integral*. We will introduce this in the context of multi-dimensional integration.

(iii) A similar problem appears when it comes to the calculation of surface and volume. Of course we all know about the surface and volume of cubes, pyramides or balls. But what about complex objects like the ones shown in Figure I.3)?



(a) Helicoid

(b) Torus

Figure I.3: Complex Surface & Volume?

The calculation of surfaces and volumes for such objects is performed using multi-dimensional integration.

(iv) A continuous quadratic grey-value picture can be represented by a mapping

$$f : [0,1] \times [0,1] \to \mathbb{R},$$

where the distribution of brightness is assigned. Its analysis and manipulation can be performed by examining analytic properties of this function, e.g., edge detection using the derivative.

Figure I.4: An example of a grey-value picture.

(v) In computer graphics objects in the plane are moving in time. The position of an object at time t can be represented by a function

$$f : \mathbb{R} \to \mathbb{R}^6,$$

where the (x, y)-coordinates of three characteristic points are assigned for a given time t. We have already seen that such movements are often given in terms of differential equations. In this lecture, the multi-dimensional case will be covered by systems of differential equations.

(vi) In economic applications, optimization theory is an important mathematical tool to determine optimal configurations minimizing costs. For example, if one were to minimize the operational costs, more than one factor must be considered, e.g., the costs for advertising, wages for employees, etc. In particular, a several variable function, e.g.,

$$f : \mathbb{R}^5 \to \mathbb{R}, \;\; (x_1, \ldots, x_5) \mapsto 3x_1^2 + 4x_2 + \sqrt{x_3} + \frac{1}{x_4} + 6x_5$$

should be used for the modeling, where the variables x_i underly certain constraints. In this lecture, we will discuss the *method of Lagrange multipliers* used to optimize a function where the variables underly equality constraints and the *multi-dimensional Newton method* to solve systems of non-linear equations and unconstrained optimization problems numerically.

In the last semester we have already introduced important concepts, such as continuity, differentiability and integrability. Moreover, we provided strong theorems for function examinations like Taylor's theorem or curve sketching procedures. In this lecture we will generalize these topics to functions of several variables thereby encountering new phenomena and analogous notions in $\mathbb{R}^n$.

2 Representation of multi-variable functions

2.1 Definition

*Let $n \in \mathbb{N}_{\geq 1}$. A **real-valued multi-variable function** is a mapping*

$$f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}, \ (x_1, \ x_2, \ldots, \ x_n) \mapsto f(x_1, \ x_2, \ldots, \ x_n).$$

*The **domain** $\mathbb{D}$ is a subset of n-tuples of real numbers for which the function rule is evaluated. The **image** $\mathrm{Im}(f)$ is a set of real numbers defining all possible outcomes.*

Let us consider some examples.

2.2 Remarks/Examples

(i) *The domain of a multi-variable function can have a multitude of formats like sectors, rectangles or some domains with holes as depicted in* Figure I.5.

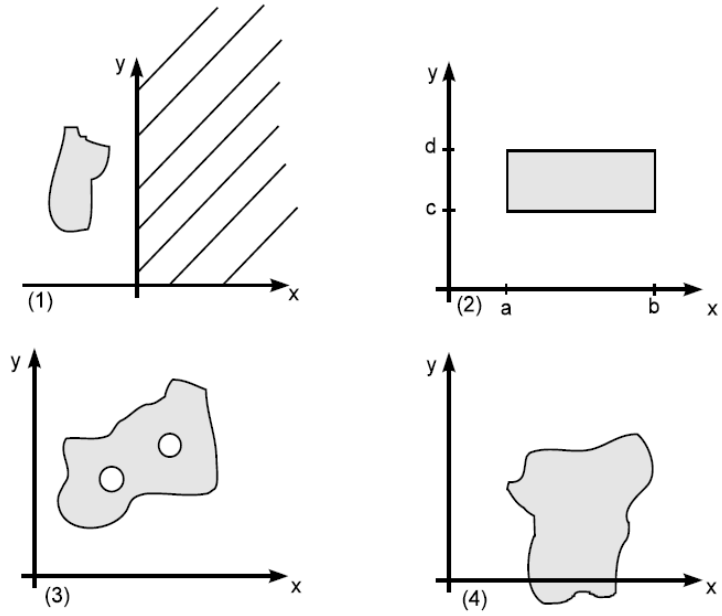


Figure I.5: Some example domains in $\mathbb{R}^2$.

(ii) *Consider the functions*

$$f : \mathbb{R}^2 \to \mathbb{R}, \ (x, y) \mapsto 2 - \frac{1}{2}x - \frac{2}{3}y,$$
$$g : \mathbb{R}^2 \to \mathbb{R}, \ (x, y) \mapsto x^2 + y^2.$$

Then the maximum domain of f is $\mathbb{D} = \mathbb{R}^2$ thereby being surjective. Recall (from school) that it represents a plane in the x-y-z-coordinate system which can be illustrated by sketching the traces like in Figure I.6.
The function g has the same maximum domain as f but the range is precisely $\mathbb{R}_{\geq 0}$.

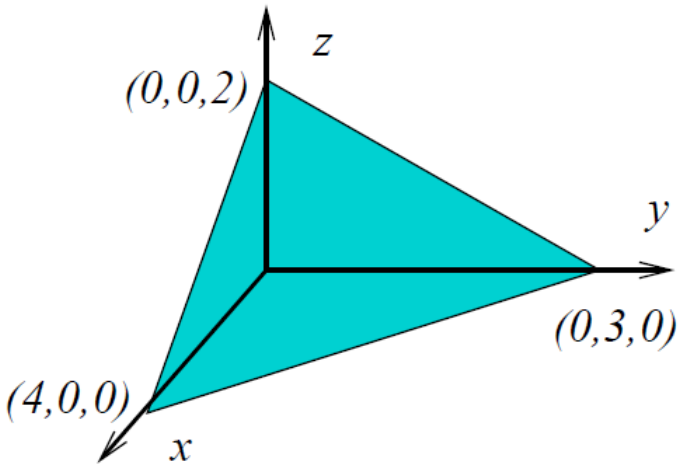


Figure I.6: The function f in the first octant.

(iii) *The function* $f(x,y) = \frac{\sin(x^2+y^2)}{x^2+y^2}$ *has a formal domain of* $\mathbb{R}^2\backslash\{\mathbb{0}\}$.

(iv) *The function* $f(x,y) := \sqrt{25-x^2-y^2}$ *is a function of two variables with maximum domain*

$$\begin{aligned} & 25 - x^2 - y^2 \geq 0 \\ \iff & x^2 + y^2 \leq 5^2, \end{aligned}$$

i.e., all points in the x-y-plane in the ball of radius $r = 5$ *form the maximum domain of the function. The range of values is here simply* $\mathrm{Im}(f) = [0,5]$.

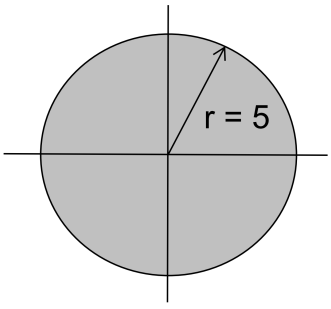


Figure I.7: The domain of f in the x-y-plane.

Since many important phenomena can be observed in two dimensions already, we restrict the discussion to the case of two-variable functions. We will generalize some of the derived material later.
For real-valued single-variable functions we had an easy way of visualizing it by means of the function graph in an x-y-diagram. For the two-variable case we will now derive similar techniques for visualization.

2.3 Definition
Let f be a real-valued two-variable function.

(i) *The **graph** of a function f is the set*

$$G_f = \{(x, y) \in \mathbb{D} \times \mathbb{R} \ni z \mid z = f(x, y)\} \subset \mathbb{R}^3.$$

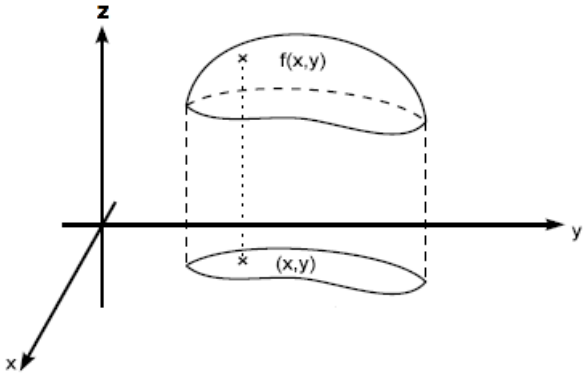


Figure I.8: The graph of a two-variable function in $\mathbb{R}^3$.

(ii) *A **level set** (niveau line) of f is a set of the form*

$$L_c(f) = \{(x, y) \in \mathbb{D} \mid f(x, y) = c\} \subseteq \mathbb{R}^2,$$

where $c \in \mathbb{R}$.

(iii) *For fixed $(a, b) \in \mathbb{D}$, a **coordinate curve** of f associated with $x = a$ or $y = b$ is the function*

$$f_{x=a}(y) := f(a, y)$$
$$f_{y=b}(x) := f(x, b)$$

where y and x run through the corresponding feasible range of values.

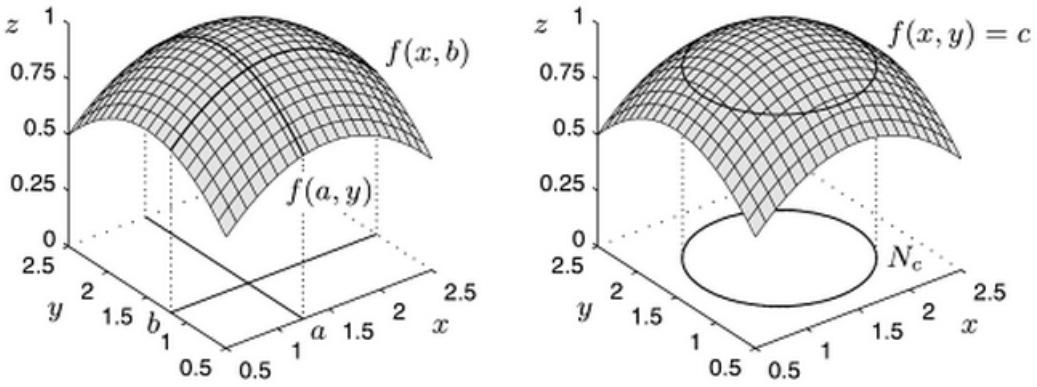


Figure I.9: Coordinate curves (left) and level set (right) of a two-variable function.

We will now demonstrate how these three objects can be used to graphically visualize a function.

2.4 Examples

(i) *Consider the function* $f : \mathbb{R}^2 \to \mathbb{R},\ (x, y) \mapsto x^2 + y^2$. *An excerpt of the graph of this function is depicted in* Figure I.10.

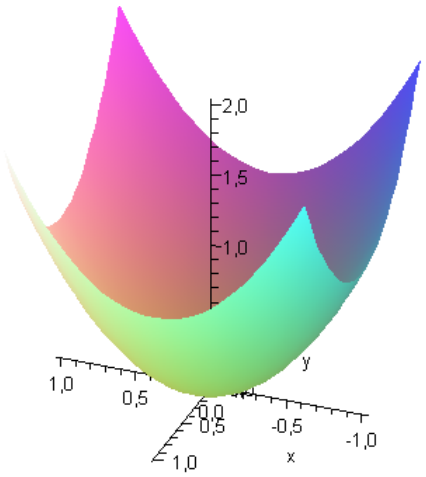
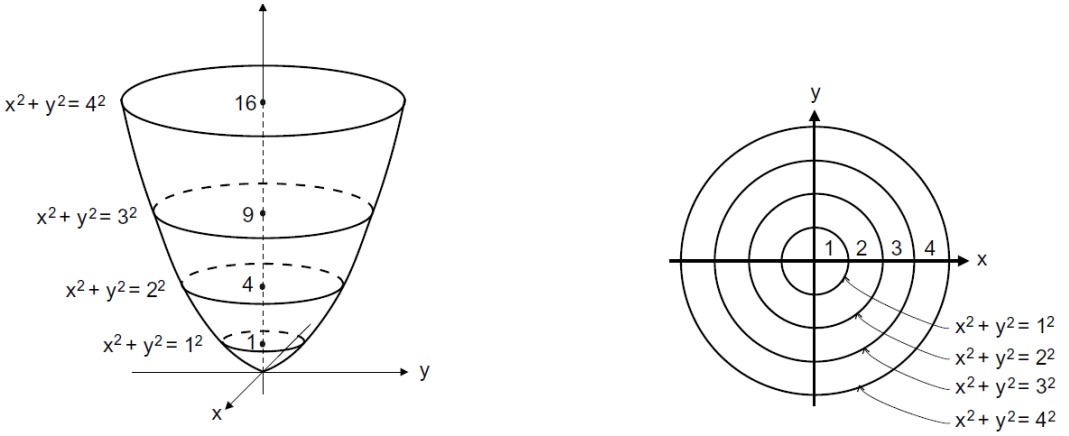


Figure I.10: Graph of f.

We now include cutting planes for this function (Figure I.11(a)). *So for the four cutting curves* $f(x, y) = k^2$ $(k = 1, 2, 3, 4)$ *we obtain as a result the level sets*

$$L_{k^2}(f) = \left\{(x, y) \in \mathbb{D} \mid x^2 + y^2 = k^2\right\},$$

which are simple circles. With this we can illustrate the tuples that are mapped to fixed values precisely as the level sets in the x-y-*plane* (Figure I.11(b)).



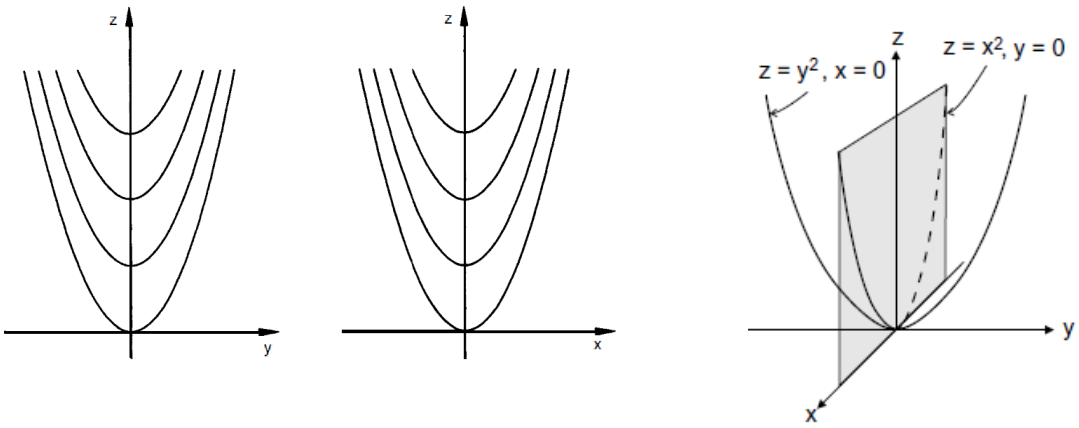
(a) graph of f with cutting curves.

(b) Level sets in $x - y$-plane

Figure I.11: Visualization of level sets

Moreover, we can use a coordinate curve for $x = 0$ *thereby letting* y *vary and we can see that the resulting projection to this plane (the* y-z-*plane) is a parable, whereas the coordinate curve for* $y = 0$ *by varying* x *yields a parable in the* x-z-*plane. We*

have depicted some coordinate curves for different values $x = a$ in Figure I.12 *(left) and for different values $y = b$ in* Figure I.12 *(right). We always obtain different parabolas shifted on the z-axis.*



(a) In $2D$-planes. (b) In 3D-plot.

Figure I.12: Coordinate curves of f.

(ii) *Consider the function $f(x, y) = \frac{x^2}{4} - \frac{y^2}{5}$. Setting $f(x, y) = 1$, we obtain the level set*

$$L_1(f) = \left\{ (x, y) \in \mathbb{D} \mid y = \pm\sqrt{\frac{5}{4}x^2 - 5} \right\},$$

i.e., a hyperbola. The same happens for different cutting planes and we obtain a picture as shown in Figure I.13.

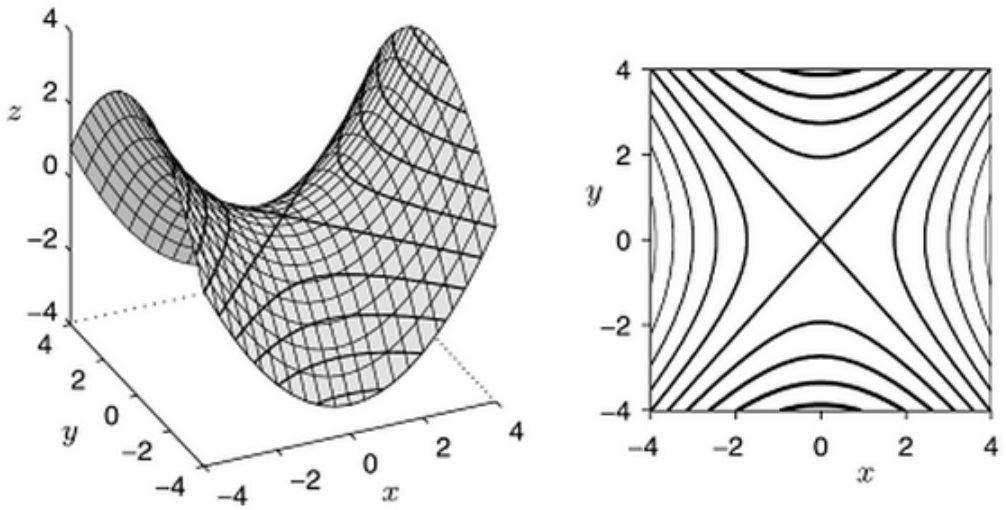


Figure I.13: 3D-plot and level sets of f.

3 Convergence in $\mathbb{R}^n$

Infinite sequences formed the base to define the important concepts of single-variable analysis (such as continuity, differentiability, approximation, etc.). We will need to extend our notions to the multi-dimensional case.

3.1 Definition

(i) *Consider the tuples $(x_k, y_k) \in \mathbb{R}^2$ for all $k \in \mathbb{N}$. We denote an* ***infinite sequence of $\mathbb{R}$-tuples*** *(or vectors) by $((x_k, y_k))_{k\in\mathbb{N}}$.*

(ii) *An infinite sequence of tuples $((x_k, y_k))_{k\in\mathbb{N}}$ is called* ***convergent*** *to $(x, y) \in \mathbb{R}^2$, if*

$$\lim_{k\to\infty} x_k = x, \quad i.e., \ |x - x_k| \to 0 \quad \textbf{\textit{and}} \quad \lim_{k\to\infty} y_k = y, \quad i.e., \ |y - y_k| \to 0.$$

Then, we write $\lim_{k\to\infty}(x_k, y_k) = (x, y)$. Otherwise, the sequence is called ***divergent****.*

To illustrate the multidimensional case, we consider some examples.

3.2 Examples

(i) *The sequence $\left(\left(\frac{1}{k}, \frac{k}{k+1}\right)\right)_{k\in\mathbb{N}}$ converges to $(0, 1)$, since*

$$\lim_{k\to\infty} \frac{1}{k} = 0 \quad \textit{and} \quad \lim_{k\to\infty} \frac{k}{k+1} = 1.$$

(ii) *On the other hand, if we consider $\left(\left(\frac{1}{k}, \sum_{j=1}^{k} \frac{1}{j}\right)\right)_{k\in\mathbb{N}}$ we have*

$$\lim_{k\to\infty} \frac{1}{k} = 0 \quad \textit{and} \quad \lim_{k\to\infty} \sum_{j=1}^{k} \frac{1}{j} = \infty$$

and so the sequence is divergent, since the second coordinate fails to meet the one-dimensional convergence criterion.

The notion *converges in every coordinate* is a little cumbersome to handle analytically. It is in deed beneficial to introduce a distance measure for two dimensions as well, just like we have the absolute value in one dimension (for $\mathbb{R}$ and $\mathbb{C}$). In this case, we have several possibilities.

3.3 Definition

(i) *The* ***maximum norm*** *(or ∞-norm) of a vector $(x, y) \in \mathbb{R}^2$ is defined by*

$$\|(x, y)\|_\infty := \max(|x|, |y|).$$

(ii) *The* ***euclidian norm*** *(or 2-norm) of a vector $(x, y) \in \mathbb{R}^2$ is defined by*

$$\|(x, y)\|_2 := \sqrt{|x|^2 + |y|^2} = \sqrt{\langle (x, y), (x, y)\rangle}.$$

(iii) *The* ***taxicab norm*** *(or 1-norm) of a vector $(x, y) \in \mathbb{R}^2$ is defined by*

$$\|(x, y)\|_1 := |x| + |y|.$$

3.4 Examples

(i) *Please note that the norms can be interpreted as a function*

$$\|\cdot\|_p : \mathbb{R}^2 \to \mathbb{R}, \quad (x,y) \mapsto \|(x,y)\|_p \quad \text{for } p = 1, 2, \infty.$$

In particular, the quantity $\|(x,y)\|_p$ *is always a real positive number.*

(ii) *We explicitly visualize the unit spheres*

$$S^1_{(p)} := \left\{(x,y) \in \mathbb{R}^2 \mid \|(x,y)\|_p = 1\right\} \text{ for } p = 1, 2, \infty$$

in $\mathbb{R}^2$ *in* Figure I.14.

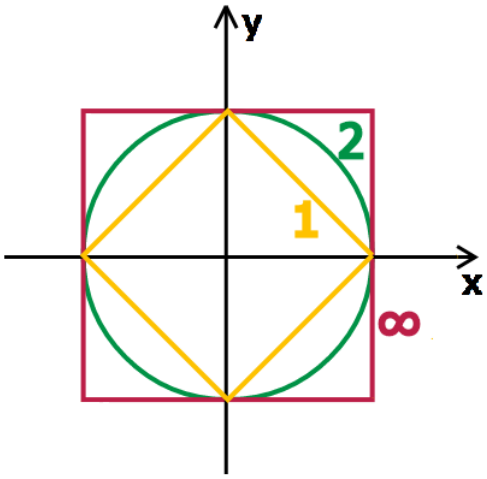


Figure I.14: The unit balls in $\mathbb{R}^2$.

(iii) *The (closed) ball of radius* r *around* (a,b) *is defined by*

$$B^r_{(p)}((a,b)) := \left\{(x,y) \in \mathbb{R}^2 \mid \|(x,y) - (a,b)\|_p \leq r\right\} \text{ for } p = 1, 2, \infty.$$

The open balls are defined by a strict "<".

The important point is now to note that all three norms define the same convergence notion. This is the content of the next theorem.

3.5 Theorem

(i) *The following statements are equivalent:*

(1) *The sequence* $((x_k, y_k))_{k\in\mathbb{N}}$ *converges to* $(x,y) \in \mathbb{R}^2$.

(2) $\|(x_k,y_k) - (x,y)\|_2 = \sqrt{|x_k - x|^2 + |y_k - y|^2} \to 0$ *as* $k \to \infty$.

(3) $\|(x_k,y_k) - (x,y)\|_1 = (|x_k - x| + |y_k - y|) \to 0$ *as* $k \to \infty$.

(4) $\|(x_k,y_k) - (x,y)\|_\infty = \max(|x_k - x|, |y_k - y|) \to 0$ *as* $k \to \infty$.

(ii) *For* $(x,y) \in \mathbb{R}^2$, *we have the relationship*

$$\|(x,y)\|_\infty \leq \|(x,y)\|_2 \leq \|(x,y)\|_1.$$

3.6 Remarks/Examples

(i) *The last theorem classifies convergence completely by means of norm convergence, i.e.,*

$$\lim_{k\to\infty}(x_k, y_k) = (x, y) \Longleftrightarrow \|(x_k, y_k) - (x, y)\|_p \to 0.$$

In fact, it also states that not even the choice of p-norm is important, convergence for one is equivalent to convergence for the other one.

(ii) *The only difference that arises is the concrete measure of distance. Let us consider the following example: Say an iterative algorithm produces tuples of data and we know that iterating often enough will lead us to a desired result. For example the algorithm could produce the sequence* $\left(\left(1/k^2, 1/k^2\right)\right)_{k\in\mathbb{N}}$*, the desired result would be here* $(0,0)$ *and we need an accuracy of* 0.25*. Calculating the norms of the iterates yields*

$$\left(\left\| \left(\frac{1}{k^2}, \frac{1}{k^2}\right) - (0,0) \right\|_2 \right)_{k\in\mathbb{N}} = \left(\frac{\sqrt{2}}{k^2}\right)_{k\in\mathbb{N}} = \left(\frac{\sqrt{2}}{1}, \ \frac{\sqrt{2}}{4}, \ \frac{\sqrt{2}}{9}, \ \frac{\sqrt{2}}{16}, \dots\right),$$

$$\left(\left\| \left(\frac{1}{k^2}, \frac{1}{k^2}\right) - (0,0) \right\|_\infty \right)_{k\in\mathbb{N}} = \left(\frac{1}{k^2}\right)_{k\in\mathbb{N}} = \left(\frac{1}{4}, \ \frac{1}{9}, \ \frac{1}{16}, \dots\right).$$

Therefore, we would have stopped the algorithm at the second iteration, if we would measure with the ∞*-norm, but we would need three iterations (for the same result) until the algorithm stops with the euclidian norm.*

(iii) *The transfer from the two-dimensional to general n-dimensional sequences poses only a notational difference:*

Let $x^{(k)} \in \mathbb{R}^n$ *for* $k \in \mathbb{N}$*. Then we can define an infinite sequence of k-tuples by* $(x^{(k)})_{k\in\mathbb{N}}$ *and this sequence converges to the k-tuple* $(a_1, \dots, a_k)$*, if it converges in all components, or, if*

$$\begin{aligned} & \|x^{(k)} - (a_1, \dots, a_k)\|_\infty = \max_{m=1,\dots,k} \left|x_m^{(k)} - a_m\right| \to 0, \\ \Longleftrightarrow \ & \|x^{(k)} - (a_1, \dots, a_k)\|_2 = \sqrt{\sum_{m=1}^{k} |x_m^{(k)} - a_m|^2} \to 0, \\ \Longleftrightarrow \ & \|x^{(k)} - (a_1, \dots, a_k)\|_1 = \sum_{m=1}^{k} |x_m^{(k)} - a_m| \to 0. \end{aligned}$$

(iv) *Finally, it is noteworthy to mention that all the known theorems from the one-dimensional case also hold in analog. For instance:*

- *(1)* *The limit is linear and constants can be extracted.*
- *(2)* *Convergent sequences are bounded (where bounded means* $\|(x_k, y_k)\|_p \leq M$ *for all k and ONE fixed M).*

4 Continuous multi-variable functions

One of the fundamental notions of analysis is continuity. The definition will be completely analogous to the one-dimensional case: convergence of a sequence in the domain must force convergence of the sequence of function values in the range. The only difference is to replace the measure $|\cdot|$ by the norm $\|\cdot\|$.

4.1 Definition

(i) *Let $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. The function f is called* ***continuous in*** *a, if we have*

$$\lim_{n\to\infty} f\left(x^{(n)}\right) = f(a), \quad \textit{i.e.,} \quad |f(x^{(n)}) - f(a)| \to 0$$

for all *sequences of n-tuples $(x^{(n)})_{n\in\mathbb{N}} \subset \mathbb{D}$, where*

$$\lim_{n\to\infty} x^{(n)} = a, \quad \textit{i.e., where} \quad \|x^{(n)} - a\|_p \to 0$$

holds.

(ii) *If f is continuous for all $a \in \mathbb{D}$, then the function f is called* ***continuous****.*

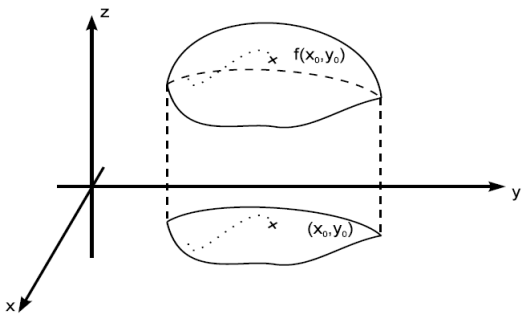


Figure I.15: Illustration of continuity for a real-valued two-variable function.

First of all, we note that everything is like in the one-dimensional case.

4.2 Proposition

(i) *The linear combination $\alpha \cdot f + g$ of continuous functions f, $g : \mathbb{D} \to \mathbb{R}$ is continuous ($\alpha \in \mathbb{R}$).*

(ii) *The product and quotient $f \cdot g$ and f/g of continuous functions is continuous (where defined).*

(iii) *The composition $g \circ f$ of continuous functions is continuous (note that f must map into the domain of g).*

We immediately consider some examples.

4.3 Examples

(i) *The projection function*

$$\pi_j : \mathbb{R}^n \to \mathbb{R},\ (x_1, \ldots, x_n) \mapsto x_j \quad j \in \{1, \ldots, n\}$$

is continuous, since $\|(x^{(k)}) - (a_1, \ldots, a_n)\|_\infty$ *implies* $|\pi_j((x^{(k)})) - \pi_j(a_1, \ldots, a_n)| \to 0$.

(ii) *All polynomials in multiple variables are continuous (as sums and products of continuous functions and compositions with projections), e.g.,*

$$f(x, y) = 2 - x \cdot y + 5 \cdot y \cdot x^3.$$

Moreover, as quotients of polynomials, all rational functions are continuous (where defined).

(iii) *The following functions are also continuous (as compositions of continuous functions):*

$$e^{2x+y},\ \ln(1x^2 + y^4),\ \sin(x \cdot y),\ \sqrt{x^2 + y^2}.$$

(iv) *The "mexican hat" function visualized in* Figure I.16 *with*

$$f(x, y) = \begin{cases} \frac{\sin(x^2+y^2)}{x^2+y^2}, & (x, y) \neq \mathbb{0}, \\ 1, & \textit{else}, \end{cases}$$

from Example I.2.2(iv) *is certainly continuous on* $\mathbb{R}^2 \backslash \{\mathbb{0}\}$.

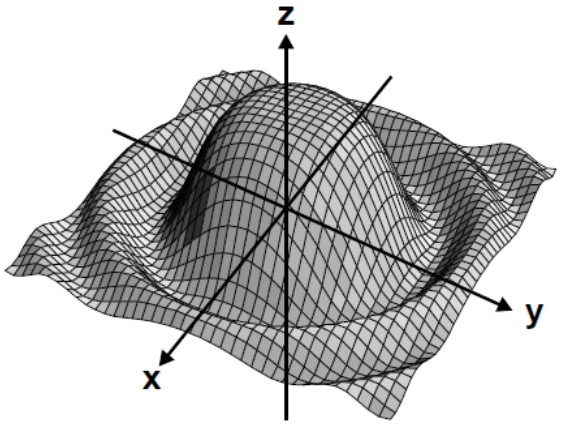


Figure I.16: "Mexican hat" function

We will now prove what the picture already suggests, namely the continuity in $\mathbb{0}$:

Suppose that $(x_k, y_k) \to \mathbb{0}$. *We have to prove that* $f(x_k, y_k) \to 1$. *At first it is clear that*

$$v_k := (\|(x_k, y_k)\|_2)^2 = x_k^2 + y_k^2 \to 0$$

as well if $(x_k, y_k) \to \mathbb{0}$. *Then we can substitute*

$$f(x_k, y_k) = \frac{\sin(x_k^2 + y_k^2)}{x_k^2 + y_k^2} = \frac{\sin(v_k)}{v_k} \to 1,$$

where we used de l'Hospital's rule in the last step (remember $(v_k)_{k\in\mathbb{N}}$ *is a sequence of real numbers!), which proves the continuity of* f *in* $\mathbb{0}$.

In contrast to the above examples, the notion of continuity for functions of several variables yields many counter-intuitive results not demonstrated by single-variable functions. This shall be motivated by the following counterexamples.

4.4 Counterexamples

(i) The function

$$f(x,y) = \begin{cases} \frac{x^2}{x^2+y^2}, & (x,y) \neq \mathbb{0}, \\ 0, & \text{else}, \end{cases}$$

is certainly continuous on $\mathbb{R}^2\backslash\{\mathbb{0}\}$. *It is visualized in* Figure I.17.

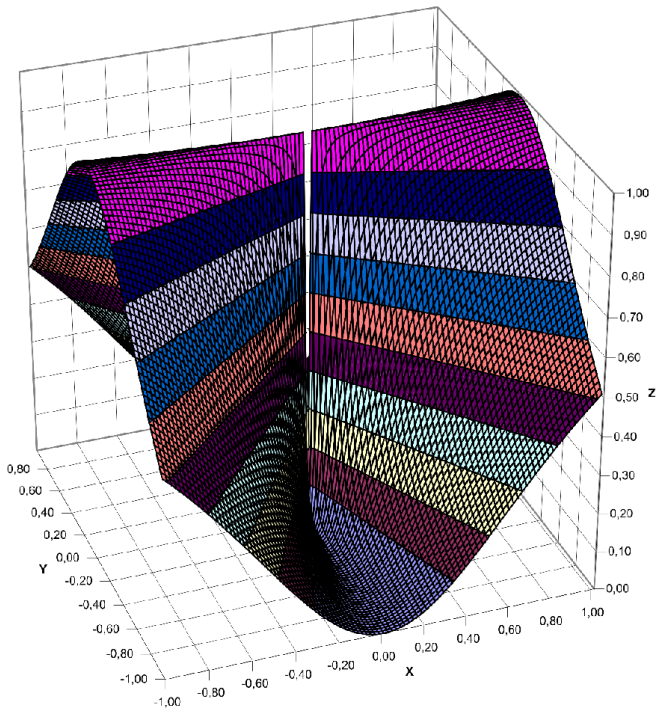


Figure I.17: Illustration of the non-continuous function f.

This function is not continuous in $\mathbb{0}$: *If we use a sequence* $(0, y_k) \to \mathbb{0}$ *(i.e., we approach* $\mathbb{0}$ *along the y-axis), we see that* $f(0, y_k) = 0 \to 0$. *However, using the sequence* $(x_k, 0) \to \mathbb{0}$ *(i.e., approaching* $\mathbb{0}$ *along the x-axis), we observe that* $f(x_k, 0) = 1 \to 1$ *and so* f *has no unique limit in* $\mathbb{0}$ *and can therefore not be continuous.*

(ii) *Now we consider the function*

$$g(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2}, & (x,y) \neq \mathbb{0}, \\ 0, & else, \end{cases}$$

depicted in Figure I.18, *which is certainly continuous on* $\mathbb{R}^2 \backslash \{\mathbb{0}\}$.

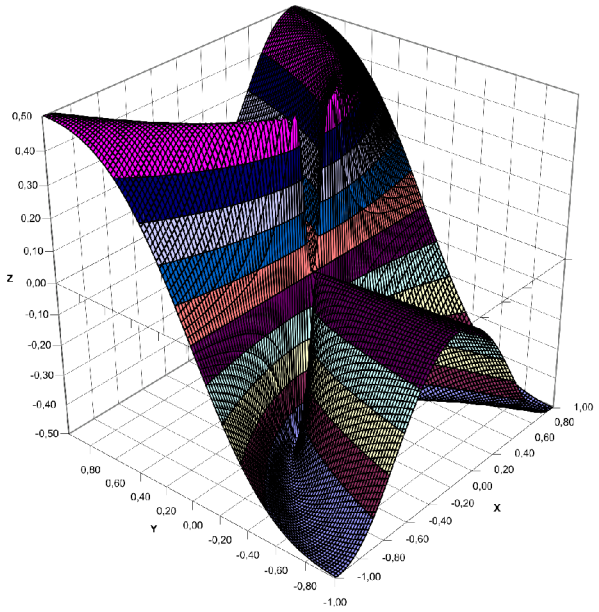


Figure I.18: Illustration of the non-continuous function g.

We will show that g *has the same limit if we approach* $\mathbb{0}$ *along any arbitrary line* $y = ax$ *in the* x*-*y*-plane, but that it is still* ***not*** *continuous:*

Consider a sequence $(x_k, ax_k) \to \mathbb{0}$. *Then we see that*

$$g(x_k, ax_k) = \frac{x_k^2 \cdot ax_k}{x_k^4 + (ax_k)^2} = \frac{ax_k}{x_k^2 + a^2} \to 0$$

for every $a \in \mathbb{R}$. *However, if we take the sequence* $(x_k, x_k^2) \to \mathbb{0}$ *(i.e., we approach* $\mathbb{0}$ *along a parabola), we obtain*

$$g(x_k, x_k^2) = \frac{x_k^2 x_k^2}{x_k^4 + (x_k^2)^2} = \frac{1}{2} \to \frac{1}{2},$$

and so g *has no unique limit in* $\mathbb{0}$ *and can therefore not be continuous. Remember that the function values must converge to the same value for* ***all*** *sequences converging to* $\mathbb{0}$.

Chapter II

Multivariable differential calculus

1 Introduction

The derivative of a function $f\colon \mathbb{R} \to \mathbb{R}$ at position x is the real number the difference quotients

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

tend towards, as h goes to zero.

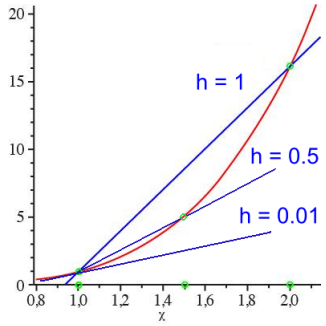


Figure II.1: Secant converging to tangent.

We have already seen that it is a key ingredient to successfully examine functions in one variable analytically. This is, e.g., reflected in the following applications:

(i) Curve sketching: The mathematical treatment of visual properties of a function such as extreme points, curvature and slope was only possible using derivatives.

(ii) Newton's method: The determination of zeros of non-trivial functions by means of a numerically stable, fast and easy to implement algorithm incorporating the slope of a function.

(iii) Approximation: Taylor's theorem served to approximate function values of complex functions by means of easy to calculate polynomial expressions.

The situation is similar for multivariable functions. We will see that similar geometric interpretations of the derivative (tangent line in one dimension) are possible and a variety of applications will follow.

2 Partial derivatives

To define the derivative of a multi-variable function we reduce the situation to the one-dimensional case. More precisely, we define it by taking derivatives of the coordinate curves.

2.1 Definition
Let $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ be a multivariable function, e_i denote the i-th unit vector in $\mathbb{R}^n$ ($i \in \{1, \dots, n\}$) and $x = (x_1, \dots, x_n) \in \mathbb{D}$ be arbitrary.

(i) *The function f is called **partially differentiable w.r.t.** x_i ($i \in \{1, \dots, n\}$), if the limit*

$$\lim_{h \to 0} \frac{f(x + h \cdot e_i) - f(x)}{h}$$

*exists in $\mathbb{R}$. In this case, we call this number the **partial derivative of f in x w.r.t.** x_i and denote it by $\frac{\partial f}{\partial x_i}(x)$ or $f_{x_i}(x)$.*

(ii) *If f is partially differentiable w.r.t. x_i on $\mathbb{D}_i \subseteq \mathbb{D}$, we define the function*

$$\frac{\partial f}{\partial x_i} : \mathbb{D}_i \subseteq \mathbb{R}^n \to \mathbb{R}, \;\; x \mapsto \frac{\partial f}{\partial x_i}(x)$$

*called the **partial derivative of f w.r.t.** x_i.*

(iii) *If f is partially differentiable w.r.t. all x_i ($i = 1, \dots, n$) in x, then we call f **partially differentiable**.*

We illustrate the situation first for a two-variable function and give some practical hints for the calculation.

2.2 Remarks
(i) In practice, partial differentiation of $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ w.r.t. x_i is executed as follows:

(1) In the function rule, all independent variables x_k, aside from x_i are considered as constants. So we interpret f as a single-variable function of the independent variable x_i.

(2) Differentiate f as if it was a function of one variable only.

*Since we are reducing the differentiation process to the one-dimensional case, **all known differentiation techniques from one-dimensional calculus can be used** (product rule, chain rule, inverse function rule, etc.).*

(ii) It is rather practical to use differential operators

$$\frac{\partial}{\partial x_i}\left(f(x, y)\right) = \frac{\partial f}{\partial x_i}(x, y)$$
$$\frac{\partial}{\partial x_i} f(\cdot) = \frac{\partial f}{\partial x_i}(\cdot).$$

(iii) *Consider a two-variable function* $f\colon \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto f(x,y)$. *In this situation, we can define partial derivatives w.r.t. to* x *and* y *in a point* $(x_0, y_0) \in \mathbb{R}^2$

$$\frac{\partial f}{\partial x}(x_0,y_0) = \lim_{h\to 0} \frac{f(x_0+h,y_0)-f(x_0,y_0)}{h}, \quad \text{where limit exists,} \tag{2.1}$$

$$\frac{\partial f}{\partial y}(x_0,y_0) = \lim_{h\to 0} \frac{f(x_0,y_0+h)-f(x_0,y_0)}{h}, \quad \text{where limit exists.} \tag{2.2}$$

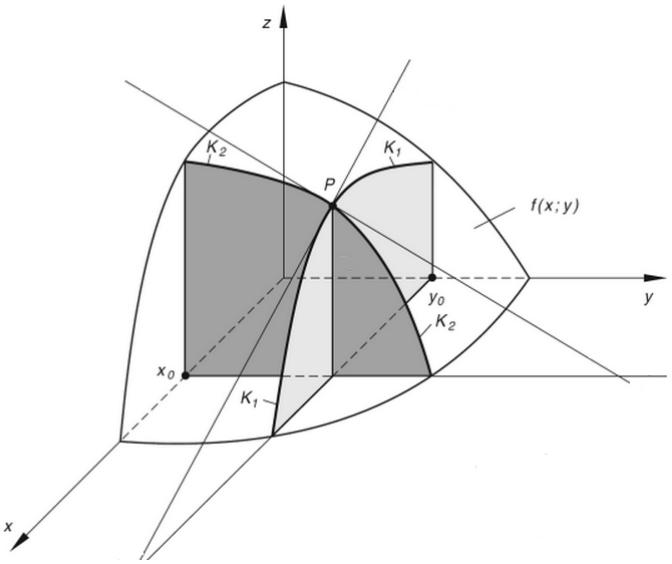


Figure II.2: Partial derivatives for a two-variable function

Geometrically, we can interpret the partial derivative given in (2.1) *as follows:*

1. *Consider the* 3*D-graph of* f *as depicted in* Figure II.2.
2. *At fixed* y_0, *we consider the coordinate curve* $f_{y=y_0}(x) = f(x,y_0)$. *This is depicted by* K_1 *in* Figure II.2.
3. *Calculating the derivative of* $f_{y=y_0}$ *in* x_0 *yields*

$$\begin{aligned} f'_{y=y_0}(x_0) &= \lim_{h\to 0} \frac{f_{y=y_0}(x_0+h) - f_{y=y_0}(x_0)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0+h,y_0)-f(x_0,y_0)}{h} \\ &= \frac{\partial f}{\partial x}(x_0,y_0), \end{aligned}$$

i.e., the partial derivative of f *w.r.t.* x *in* (x_0,y_0) *is simply the ordinary derivative of the coordinate curve corresponding to* $y = y_0$ *at position* x_0.

Analogously, we obtain that the partial derivative of f *w.r.t.* y *in* (x_0,y_0) *is simply the ordinary derivative of the coordinate curve corresponding to* $x = x_0$ *(depicted by* K_2 *in* Figure II.2*) at position* y_0.
Consequently, the value of the partial derivative $f_x(x_0,y_0)$ *can be interpreted as the slope of the tangent of* f *with* y*-values* $y = y_0$ *at the point* P.

Now, we consider some concrete examples.

2.3 Examples

(i) *We test the procedure for the example* $f(x,y) = -4x^3y^2 + 3xy^4 - 3x + 2y + 5$. *Then we derive*

$$\begin{aligned} f_x(x,y) &= \frac{\partial}{\partial x}\left(-4x^3y^2 + 3xy^4 - 3x + 2y + 5\right) \\ &= -4 \cdot (3x^2)y^2 + 3y^4 - 3 + 0 + 0 = -12x^2y^2 + 3y^4 - 3. \\ f_y(x,y) &= \frac{\partial}{\partial y}\left(-4x^3y^2 + 3xy^4 - 3x + 2y + 5\right) \\ &= -4x^3(2y) + 3x(4y^3) - 0 + 2 + 0 = -8x^3y + 12xy^3 + 2. \end{aligned}$$

In the point $(1,2)$ *we get the partial derivatives* $f_x(1,2) = -3$ *and* $f_y(1,2) = 82$, *whereas* $f(1,2) = 38$. *Therefore, in terms of* Figure II.2 *we have* $P(1|2|38)$, *the slope of the tangent of* f *in this point in* x*-direction is* -3 *and the slope of tangent of* f *in this point in* y *direction is* 82.

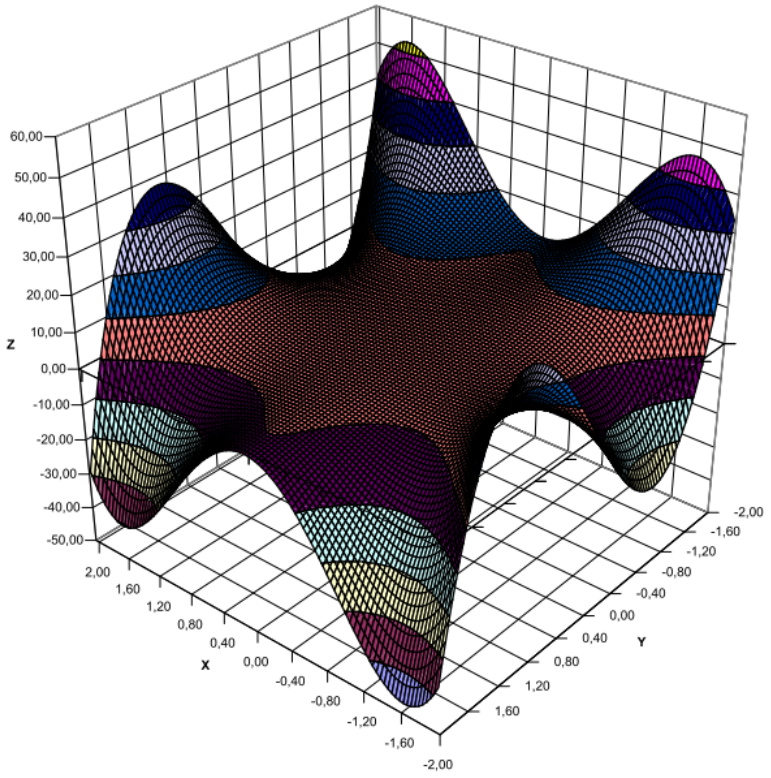


Figure II.3: The graph of f

(ii) *Consider the function* $f : \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto \|(x,y)\|_2 = \sqrt{x^2+y^2}$. *This function is partially differentiable for all* $(x,y) \neq \mathbb{0}$. *We derive the partial derivatives*

$$\begin{aligned} f_x(x,y) &= \frac{\partial}{\partial x}\left(\sqrt{x^2+y^2}\right) = \frac{1}{2} \cdot \frac{2x}{\sqrt{x^2+y^2}} = \frac{x}{\|(x,y)\|_2}, \\ f_y(x,y) &= \frac{\partial}{\partial y}\left(\sqrt{x^2+y^2}\right) = \frac{1}{2} \cdot \frac{2y}{\sqrt{x^2+y^2}} = \frac{y}{\|(x,y)\|_2}. \end{aligned}$$

(iii) Consider the function f as depicted in Figure II.4 *with*

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y^2}, & (x,y) \neq \mathbb{0}, \\ 0, & \textit{else.} \end{cases}$$

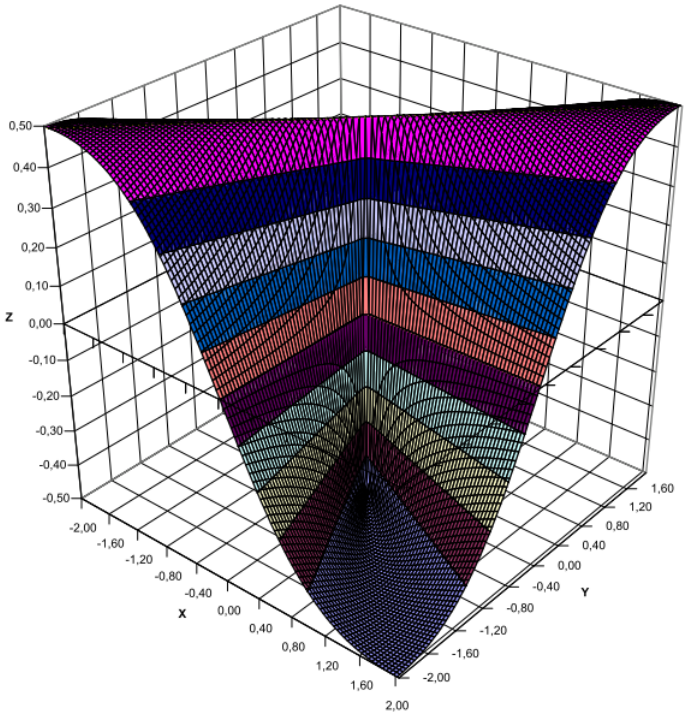


Figure II.4: The graph of f

Clearly, this function is partially differentiable for all $(x,y) \neq \mathbb{0}$ and there we have

$$f_x(x,y) = \frac{\partial}{\partial x}\left(\frac{xy}{x^2+y^2}\right) = \frac{y(x^2+y^2)-(xy)(2x)}{(x^2+y^2)^2} = \frac{-y(x^2-y^2)}{(x^2+y^2)^2},$$

$$f_y(x,y) = \frac{\partial}{\partial y}\left(\frac{xy}{x^2+y^2}\right) = \frac{x(x^2+y^2)-(xy)(2y)}{(x^2+y^2)^2} = \frac{x(x^2-y^2)}{(x^2+y^2)^2},$$

where we used the quotient rule. For $(x,y) = \mathbb{0}$ we derive

$$f_x(0,0) = \lim_{h\to 0} \frac{f(0+h,0)-f(0,0)}{h} = \lim_{h\to 0} \frac{0-0}{h} = 0$$

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

So f is partially differentiable w.r.t. x and y everywhere. However, it is now interesting to observe that f is not continuous (in $\mathbb{0}$), since

$$\lim_{n\to\infty} f\left(\frac{1}{n},\frac{1}{n}\right) = \lim_{n\to\infty} \frac{1/n^2}{2/n^2} = \frac{1}{2},$$

$$\lim_{n\to\infty} f\left(\frac{1}{n}, -\frac{1}{n}\right) = \lim_{n\to\infty} \frac{-1/n^2}{2/n^2} = -\frac{1}{2}$$

does not lead to the same limit. Therefore, in contrast to the single variable case:

f partially differentiable $\not\Rightarrow$ f continuous (in general)

The partial derivatives obtained from a function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ define themselves real-valued multivariable functions of the same variables

$$\frac{\partial f}{\partial x_i} : \mathbb{D}_i \subseteq \mathbb{R}^n \to \mathbb{R}, \;\; x \mapsto \frac{\partial f}{\partial x_i}(x)$$

on $\mathbb{D}_i$, the set of points where the partial derivatives w.r.t. x_i exist. Therefore, we can take partial derivatives again and define higher order derivatives.

2.4 Definition

(i) *Assume that a function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is (partially) differentiable w.r.t. x_i and let $\frac{\partial f}{\partial x_i} : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ be (partially) differentiable w.r.t. x_j. Then, the function*

$$f_{x_i,x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}, \;\; x \mapsto \frac{\partial^2 f}{\partial x_j \partial x_i}(x) := \left(\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)\right)(x)$$

is called a partial derivative of f of order 2. In total we have n^2 partial derivatives of order 2. Higher order derivatives are defined inductively.

(ii) *A function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called k-times partially differentiable, if all partial derivatives of order k exist. If these are also continuous, then the function is called k-times continuously partially differentiable, denoted $f \in \mathcal{C}^k(D)$.*

(iii) *If $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ has arbitrary many partial derivatives and they are all continuous, then f is called infinitely times partially continuously differentiable and we write $f \in \mathcal{C}^\infty(D)$.*

2.5 Remarks/Examples

(i) *Note that, e.g., for a function $f : \mathbb{D} \subseteq \mathbb{R}^5 \to \mathbb{R}$, the expression f_{x_1,x_2,x_4} means to*

1. *first derive f w.r.t. x_1 partially and obtain f_{x_1},*
2. *then derive f_{x_1} w.r.t. x_2 partially and obtain f_{x_1,x_2} and*
3. *at last derive f_{x_1,x_2} w.r.t. x_4 partially and obtain f_{x_1,x_2,x_4}.*

However, using the representation with the differential operator, we have the opposite order

$$f_{x_1,x_2,x_4}(\cdot) = \left(\frac{\partial^3 f}{\partial x_4 \partial x_2 \partial x_1}\right)(\cdot) = \left(\frac{\partial}{\partial x_4}\left(\frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right)\right)\right)(\cdot).$$

(ii) *For a two-variable function* $f\colon \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto f(x,y)$*, all possible partial derivatives up to order* 3 *are depicted in* Figure II.5*. So, we have (at most):* 2 *first order partial derivatives,* 2^2 *second order derivatives and* 2^3 *third order derivatives.*

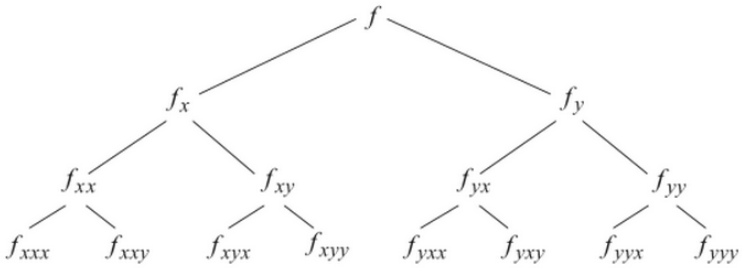


Figure II.5: Higher order partial derivatives

(iii) *We consider the function* $f(x,y) = e^{xy^2}$*. Taking partial derivatives of arbitrary order is possible since the function in one independent variable has arbitrary many derivatives. We derive, e.g.,*

$$f_x(x,y) = \frac{\partial f}{\partial x}(x,y) = e^{xy^2}y^2 \qquad f_y(x,y) = \frac{\partial f}{\partial y}(x,y) = e^{xy^2}2xy$$

$$f_{xx}(x,y) = \frac{\partial^2 f}{\partial x \partial x}(x,y) = e^{xy^2}y^4 \qquad f_{yy}(x,y) = \frac{\partial^2 f}{\partial y \partial y}(x,y) = e^{xy^2}\left(4x^2y^2 + 2x\right)$$

$$f_{xy}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = e^{xy^2}\left(2xy^3 + 2y\right) \quad f_{yx}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) = e^{xy^2}\left(2xy^3 + 2y\right)$$

Please note that all derived functions are continuous on $\mathbb{R}^2$*. Therefore, we even have* $f \in \mathcal{C}^\infty(\mathbb{R}^2)$*. Moreover, please note that the identity*

$$f_{xy}(x,y) = f_{yx}(x,y) \tag{2.3}$$

holds true in this example. From a computational aspect this would be very beneficial, since we can save the calculation of many (possibly expensive) derivatives.

(iv) *This example shall demonstrate that* Equation (2.3) *is not always true. Consider the function*

$$f(x,y) = \begin{cases} xy \cdot \dfrac{x^2 - y^2}{x^2 + y^2}, & (x,y) \neq \mathbb{0}, \\ 0, & \text{else.} \end{cases}$$

Then we have

$$f_x(x,y) = \frac{\partial f}{\partial x}(x,y) = \frac{y(x^4 + 4x^2y^2 - y^4)}{(x^2+y^2)^2}$$

$$f_y(x,y) = \frac{\partial f}{\partial y}(x,y) = \frac{x(x^4 - 4x^2y^2 - y^4)}{(x^2+y^2)^2}$$

$$f_x(0,0) = 0 = f_y(0,0).$$

Both functions are continuous everywhere. Taking partial derivatives in $(0,0)$ *yields*

$$f_{xy}(0,0) = \lim_{h\to 0} \frac{f_x(0,h) - f_x(0,0)}{h} = \lim_{h\to 0} \frac{h\cdot(-h^4) - 0}{h\cdot h^4} = -1$$
$$f_{yx}(0,0) = \lim_{h\to 0} \frac{f_y(h,0) - f_y(0,0)}{h} = \lim_{h\to 0} \frac{h\cdot h^4 - 0}{h\cdot h^4} = \ 1$$

and so $f_{xy} \neq f_{yx}$ *in this case. It will turn out that the problem lies in the second derivatives: In deed, we have for* $(x,y) \neq \mathbb{0}$

$$f_{xy}(x,y) = \frac{x^6 + 9x^4y^2 - 9y^4x^2 - y^6}{(x^2+y^2)^3}$$

and approaching the origin on the x*-axis yields*

$$\lim_{h\to 0} f_{xy}(h,0) = \frac{h^6}{h^6} = 1.$$

So f_{xy} *is not continuous in* $(0,0)$*,* $f \notin \mathcal{C}^2(\mathbb{R}^2)$ *and this is what causes the trouble.*

Even though the last example is quite discouraging, the commutativity of partial differentiation actions in the calculation of partial derivatives is a computationally important aspect. In deed, in practice examples as above where $f_{xy} \neq f_{yx}$ are not common. This is assured by Schwarz' theorem (Hermann Schwarz, 1843-1921, German mathematician).

2.6 Theorem (Schwarz' theorem)
Consider the multi-variable function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$*.*

(i) *If* $f \in \mathcal{C}^2(\mathbb{D})$*. Then we have*

$$f_{x_i,x_j}(x) = f_{x_j,x_i}(x)$$

for all $x \in \mathbb{D}$*.*

(ii) *More generally, if* $f \in \mathcal{C}^k(\mathbb{D})$*, then the sequence of differentiation of higher order derivatives can be permuted arbitrarily without changing the result.*

2.7 Remark

(i) The theorem uncovers the differences between Example II.2.5(iii) *and* (iv)*: In the first case the second derivatives are continuous, in the second case this property is not given.*

(ii) From a computational point of view, we can reduce the amount of work for higher order partial derivative calculations. For a two-variable function

$$f : \mathbb{R}^2 \to \mathbb{R},\ \ (x,y) \mapsto f(x,y),$$

we observe from Figure II.5 *that we need three calculations instead of four for second order derivatives and four instead of eight for third order derivatives.*

3 The total derivative

Example II.2.3(iii) showed that partial differentiability is not strong enough to enforce continuity of a function, whereas differentiability implied continuity in the one-dimensional case. Moreover, in Example II.2.5(iv) the second order derivative calculations were not commutative, which we resolved by additional regularity requirements on f. However, another possibility to tackle the situation is to reconsider the differentiability concept.

Currently, we simplify the situation completely by assuming that differentiation in the direction of coordinate axes is the superior concept for the multidimensional case, even though for continuity (in multiple dimensions) the limit process was not restricted to the coordinate axes. In Counterexample I.4.4(ii), we saw that this would not have been sufficient either, because f was *partially continuous* but still had a jump in the origin.
In this section we aim to generalize the differentiability concept by allowing a variance of all coordinates simultaneously. This will help to better understand the multidimensional situation.

3.1 Discussion (One-dimensional differentiation)
In order to introduce differentiability in multiple dimensions, we recall the one-dimensional case. First note that the tangent of a differentiable function $f\colon (a,b) \to \mathbb{R}$ at the point $(x_0, f(x_0))$ is given by

$$y(x) = f'(x_0)(x - x_0) + f(x_0) \qquad \text{or} \qquad f'(x_0) = \frac{y(x) - f(x_0)}{x - x_0}.$$

We know that if f is differentiable. Then

$$\begin{aligned} & \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) \\ \iff & \lim_{x \to x_0} \left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| = 0 \\ \iff & \lim_{x \to x_0} \frac{|f(x) - \overbrace{[f(x_0) + f'(x_0)(x - x_0)]}^{tangent}|}{|x - x_0|} = 0 \end{aligned} \tag{3.1}$$

This reformulation offers two crucial insights:

(i) *Differentiability at x_0 means that the function value close to this point can be well approximated by the tangent function value. This can be deduced since the numerator goes to zero if $x \to x_0$.*

(ii) *But the first point is not enough, if we are precise differentiability means even more: Since the complete expression in* Equation (3.1) *goes to zero, the error given by approximating the function value in x by the tangent value goes faster to zero than x goes to x_0. Thus, we can infer that a function that is differentiable in x_0 has to be continuous there.*

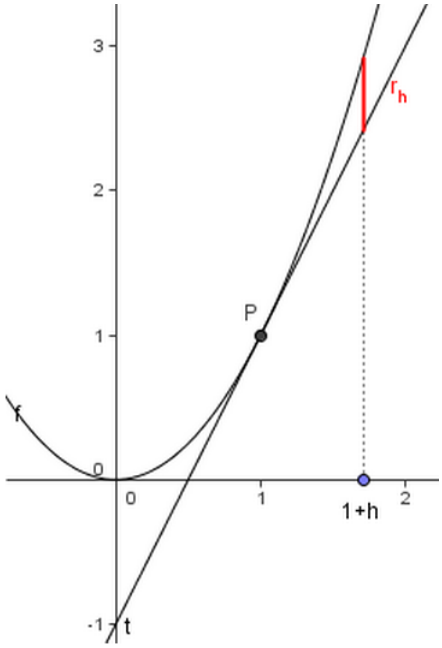


Figure II.6: The tangent at a function.

We state our observations about the one-dimensional case in the following proposition.

3.2 Proposition
Let $f\colon (a,b) \to \mathbb{R}$ be a function and $x_0 \in (a,b)$.

if f is differentiable in $x_0 \Rightarrow f$ is continuous in x_0, i.e,

$$\text{if } \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} \text{ exists} \Rightarrow \lim_{x\to x_0} \frac{|f(x)-[f(x_0)+f'(x_0)(x-x_0)]|}{|x-x_0|} = 0$$

We will now discuss the multi-dimensional case in complete analogy. Our aim is to find a suitable definition for differentiability of a multi-variable function.

3.3 Discussion (Two-dimensional differentiation)
In order to approximate a two-variable function

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

in a small ball around a point (x_0, y_0) in the x-y-plane, we cannot use a straight line like the tangent anymore, because we have more than one direction where we want to approximate it. The analogous concept is here a ***tangent plane****!*

We will set up an equation for it in the point (x_0, y_0) using the standard coordinate form (compare Example I.2.2(ii)*)*

$$t : \mathbb{R}^2 \to \mathbb{R}, \quad (x,y) \mapsto t(x,y) := a \cdot x + b \cdot y + c, \quad a,b,c \in \mathbb{R}.$$

Therefore we have to determine the coefficients a, b and c such that:

(i) *The point $f(x_0, y_0)$ is in the tangent plane, i.e.,*

$$f(x_0, y_0) = t(x_0, y_0) = ax_0 + by_0 + c \iff c = f(x_0, y_0) - ax_0 - by_0.$$

(ii) *The slope of the tangent plane in x-direction and in y-direction must be equal to the slope of the function in these directions:*

$$t_x(x_0, y_0) = a = f_x(x_0, y_0) \tag{3.2}$$
$$t_y(x_0, y_0) = b = f_y(x_0, y_0) \tag{3.3}$$

Altogether, the tangent plane of f at (x_0, y_0) if it exists is given by

$$t : \mathbb{R}^2 \to \mathbb{R}, \ (x, y) \mapsto t(x, y) = f(x_0, y_0) + f_x(x_0, y_0) \cdot (x - x_0) + f_y(x_0, y_0) \cdot (y - y_0)$$

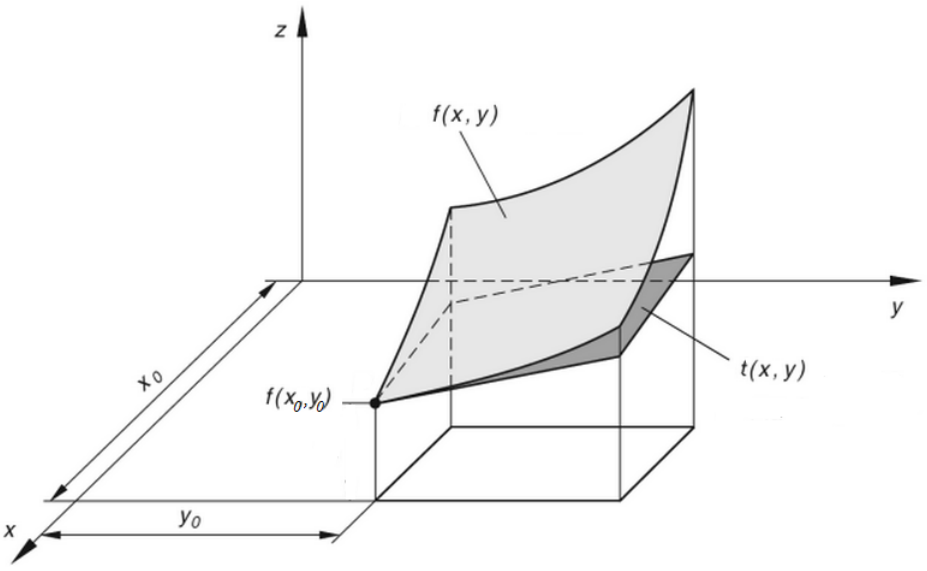


Figure II.7: Tangent plane of f at (x_0, y_0).

We can now use these observations to define differentiability in analogy to the one-dimensional case.
Here, the definition will offer the same two conclusions like above:

(i) Differentiability at (x_0, y_0) means that the function value close to this point can be well approximated by the tangent plane function value.

(ii) Furthermore, we will infer that differentiability in a point (x_0, y_0) guarantees continuity there.

3.4 Definition
Let $f\colon \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ be partially differentiable in $(x_0, y_0) \in \mathbb{D}$. Then f is called ***totally differentiable*** *in (x_0, y_0), if*

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x,y) - t(x,y)}{\|(x,y) - (x_0,y_0)\|_2} = 0,$$

where $t(x,y)$ is the tangent plane defined by the function rule

$$\begin{aligned} t : \mathbb{R}^2 \to \mathbb{R},\ \ (x,y) \mapsto t(x,y) &= f(x_0,y_0) + f_x(x_0,y_0)\cdot(x-x_0) + f_y(x_0,y_0)\cdot(y-y_0) \\ &= f(x_0,y_0) + \underbrace{(f_x(x_0,y_0),\ \ f_y(x_0,y_0))}_{:=f'(x_0,y_0)} \cdot \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} \end{aligned}$$

In this case, the 1×2-matrix $f'(x_0,y_0)$ (sometimes also denoted $J_f(x_0,y_0)$) is called the ***Jacobian matrix****, the associated linear map is called the* ***(total) differential*** *or simply the derivative.*

Even though this definition is rather intuitive, a more practical criterion to identify totally differentiable functions is provided by the following proposition.

3.5 Proposition
Let $f\colon \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ be a given function.

(i) *f is totally differentiable in $(x_0,y_0) \in \mathbb{D}$, if the following criteria are fulfilled:*

(1) There is a ball $B_\varepsilon(x_0,y_0)$ where f is partially differentiable.

(2) f_x and f_y are continuous in (x_0,y_0).

(ii) *If $f \in \mathcal{C}^1(\mathbb{D})$ then f is totally differentiable on $\mathbb{D}$.*

We consider some examples.

3.6 Remarks/Examples

(i) The generalization of total differentiability to functions $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is simple: For $x_0 \in \mathbb{D}$ f is called totally differentiable, if

$$\lim_{x\to x_0} \frac{f(x) - \left[f(x_0) + J_f(x_0)\cdot(x-x_0)^\top\right]}{\|x-x_0\|_2} = 0,$$

where $J_f(x_0) = (f_{x_1}(x_0), f_{x_2}(x_0), \ldots, f_{x_n}(x_0))$ is the Jacobian again.

(ii) Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R},\ f(x,y) = x\cdot y$. We observe that f is partially differentiable in $\mathbb{R}^2$ with partial derivatives

$$\begin{aligned} f_x(x,y) &= y & \qquad f_x(1,-1) &= -1 \\ f_y(x,y) &= x & \qquad f_y(1,-1) &= 1. \end{aligned}$$

Clearly, these functions are continuous as well, so f is totally differentiable according to Proposition II.3.5. *Now we set up the tangent plane in $(1,-1)$. In addition to the partial derivatives, we calculate $f(1,-1)=-1$ and obtain*

$$t:\mathbb{R}^2\to\mathbb{R},\ \ (x,y)\mapsto t(x,y)=-1+(-1,1)\cdot\begin{pmatrix}x-1\\y+1\end{pmatrix}=1-x+y.$$

The Jacobian (or derivative) of f is the matrix

$$J_f(x,y)=(y,x)\ \text{ and we have } J_f(1,-1)=(-1,1).$$

(iii) Consider the function $f\colon \mathbb{D}=\mathbb{R}^2\setminus(\mathbb{R}\times\{0\})\to\mathbb{R},\ \ f(x,y)=\frac{x}{y}$. We observe that f is partially differentiable on $\mathbb{D}$ with partial derivatives

$$\begin{aligned} f_x(x,y)&=\frac{1}{y} & f_x(1,-1)&=-1\\ f_y(x,y)&=-\frac{x}{y^2} & f_y(1,-1)&=-1. \end{aligned}$$

Clearly, these functions are continuous on $\mathbb{D}$ as well, so f is totally differentiable according to Proposition II.3.5. *Now we set up the tangent plane in $(1,-1)$. In addition to the partial derivatives, we calculate $f(1,-1)=-1$ and obtain*

$$t:\mathbb{R}^2\to\mathbb{R},\ \ (x,y)\mapsto t(x,y)=-1+(-1,-1)\cdot\begin{pmatrix}x-1\\y+1\end{pmatrix}=-1-x-y.$$

The Jacobian (or derivative) of f is the matrix

$$J_f(x,y)=\left(\frac{1}{y},-\frac{x}{y^2}\right)\ \text{ and we have}J_f(1,-1)=(-1,-1).$$

*(iv) A general representation of a **linear-affine function** $f\colon \mathbb{R}^2\to\mathbb{R}$ is*

$$f:\mathbb{R}^2\to\mathbb{R},\ \ (x,y)\mapsto a\cdot x+b\cdot y+c=(a,b)\begin{pmatrix}x\\y\end{pmatrix}+c.$$

These functions are totally differentiable with derivative

$$f'(x,y)=J_f(x,y)=(a,b).$$

*(v) A general representation of a **quadratic function** $f\colon \mathbb{R}^2\to\mathbb{R}$ is*

$$\begin{aligned} f(x_1,x_2)&=\alpha x_1^2+2\beta x_1x_2+\gamma x_2^2+\delta x_1+\varepsilon x_2+c\\ &=\underbrace{(x_1,x_1)}_{=:x}\underbrace{\begin{pmatrix}\alpha&\beta\\\beta&\gamma\end{pmatrix}}_{=:A}\begin{pmatrix}x_1\\x_2\end{pmatrix}+\underbrace{(\delta,\varepsilon)}_{:=b}\begin{pmatrix}x_1\\x_2\end{pmatrix}+c\\ &=x^\top Ax+b^\top x+c. \end{aligned}$$

These functions are totally differentiable with derivative

$$\begin{aligned} f'(x_1,x_2) = J_f(x_1,x_2) &= (2\alpha x_1 + 2\beta x_2 + \delta,\ \ 2\beta x_2 + 2\gamma x_2 + \varepsilon) \\ &= 2\cdot(x_1,x_2)\begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} + (\delta,\varepsilon) \\ &= 2x^\top A + b^\top. \end{aligned}$$

An example of such a function is

$$f(x,y) = x^2 - y^2 = (x,y)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$g(x,y) = x^2 + 4xy - 2y^2 + x = (x,y)\begin{pmatrix} 1 & 2 \\ 2 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + (1,0)\begin{pmatrix} x \\ y \end{pmatrix}.$$

(vi) Here we want to emphasize the approximation character exhibited by the total derivative. In deed, the tangent plane can be seen as the linearisation of a function around a point and it can be used to approximate function values around that point. Consider, for example, two ohmic resistances R_1 and R_2 (measured in Ω) in parallel connection. Then, the overall resistance is given by the (non-linear) function

$$R : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R},\ \ (R_1,R_2) \mapsto \frac{R_1 R_2}{R_1 + R_2}$$

Say, we select resistances $R_1 = 100$ and $R_2 = 400$. Then the overall resistance of the parallel circuit is

$$R(100,400) = \frac{100\cdot 400}{100+400} = 80\,\Omega$$

If we set up the tangential plane in this point, we derive

$$t(R_1,R_2) = R(100,400) + \left(\frac{\partial R}{\partial R_1}(100,400), \frac{\partial R}{\partial R_2}(100,400)\right)\begin{pmatrix} R_1 - 100 \\ R_2 - 400 \end{pmatrix}$$

$$= 80 + (0.64, 0.04)\begin{pmatrix} R_1 - 100 \\ R_2 - 400 \end{pmatrix}.$$

The total resistance for a similar parallel circuit with $R_1 = 110$ and $R_2 = 380$ can then be approximated by the tangent plane and we derive

$$R(110,380) \approx 80 + (0.64, 0.04)(110-100, 380-400)^\top = 85.6\,\Omega$$

$$R(110,380) = \frac{110\cdot 380}{110+380} = 85.3\,\Omega$$

where the second row contains the exact value.

In applications the function rule may not even be known such as it is the case in this example but one may be able to calculate it for a few fixed values around a given point. Then the tangent plane can still be set up using finite differences for the partial derivatives! In this case the linear approximation is then the only choice to estimate surrounding values.

As announced in the introduction of this section, the following important property holds, just like for single-variable functions.

3.7 Theorem
If $f\colon \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ is totally differentiable in (x_0, y_0), then it is also continuous in (x_0, y_0).

Proof.
Let (x, y) converge to (x_0, y_0). Since f is totally differentiable, we have

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x,y) - t(x,y)}{\|(x,y) - (x_0,y_0)\|_2} = 0$$

and the difference between function values and the tangent plane in (x_0, y_0) goes faster to zero than the distance between the points (x_0, y_0) and (x, y):

$$\lim_{(x,y)\to(x_0,y_0)} \big[f(x,y) - (f(x_0,y_0) + f'(x_0,y_0)(x - x_0, y - y_0)^\top)\big] = 0.$$

Now we can reformulate this term:

$$\begin{aligned}
\lim_{(x,y)\to(x_0,y_0)} \big[f(x,y) - f'(x_0,y_0)(x - x_0, y - y_0)^\top)\big] - \lim_{(x,y)\to(x_0,y_0)} f(x_0,y_0) &= 0 \\
\lim_{(x,y)\to(x_0,y_0)} f(x,y) \;\; - \lim_{(x,y)\to(x_0,y_0)} \big(f'(x_0,y_0)(x - x_0, y - y_0)^\top\big) &= \lim_{(x,y)\to(x_0,y_0)} f(x_0,y_0). \\
\lim_{(x,y)\to(x_0,y_0)} f(x,y) &= f(x_0,y_0),
\end{aligned}$$

which means that f is continuous. □

At the end of this section, we reconsider Example II.2.3(iii).

3.8 Example
We proved that the function

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y^2}, & (x,y) \neq \mathbb{0}, \\ 0, & \textit{else}, \end{cases}$$

is partially differentiable everywhere, but not continuous. Therefore, according to Theorem II.3.7*, this function cannot be totally differentiable (in $\mathbb{0}$). In deed, the tangent plane in $\mathbb{0}$ would be*

$$t : \mathbb{R}^2 \to \mathbb{R},\;\; t(x,y) = f(0,0) + (f_x(0,0), f_y(0,0)) \cdot \begin{pmatrix} x \\ y \end{pmatrix} = 0 + (0,0) \begin{pmatrix} x \\ y \end{pmatrix} = 0$$

which cannot be a tangential plane for f (compare Figure II.4*!).*

4 The chain rule - Differentiation w.r.t. a parameter

Recall that for differentiable $f\colon (a,b) \to \mathbb{R}$ and $g\colon \mathbb{R} \to (a,b)$, we can differentiate the composition

$$f \circ g : \mathbb{R} \to \mathbb{R},\ \ t \mapsto (f \circ g)(t) := f(g(t))$$

by the formula

$$(f \circ g)'(t) = f'(g(t)) \cdot g'(t).$$

This rule is extremely useful: Recall that we determined the derivative for basic functions like polynomials, $\sqrt{(\cdot)}$, $\exp(\cdot)$ or $\sin(\cdot)$, etc. explicitly and, more or less, it was important to learn these derivatives by heart. Just imagine, we would have to do the same thing for $\cos(t^2)$, $e^{-\sin(t)}$ or $\frac{x^2}{\sin(x)}$. This would mean a significant overflow for our mind, but in particular, a significant overhead for derivatives built with a computer (look up the topic Automatic differentiation!). The chain rule simplifies things a lot, since the knowledge of the basic derivatives and the possibility to write the functions as compositions of simpler ones enables an effective algorithmic treatment.
In this section we want to generalize the chain rule to the multi-variable case. We immediately start with the main theorem.

4.1 Theorem (Chain rule)
Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be totally differentiable and assume that $g_i\colon (a,b) \to \mathbb{R}$ $(i = 1,2)$ are differentiable as well. Then the function

$$F : (a,b) \to \mathbb{R},\ \ t \mapsto F(t) := f\left(g_1(t), g_2(t)\right)$$

is also differentiable and we have

$$F'(t) = J_f\left(g_1(t), g_2(t)\right) \cdot \begin{pmatrix} g_1'(t) \\ g_2'(t) \end{pmatrix}.$$

Proof.
Since we are dealing with a single variable function, we simply consider the differential quotient of F at a fixed point $t \in \mathbb{R}$

$$F'(t) = \lim_{h \to 0} \frac{F(t+h) - F(t)}{h} = \lim_{h \to 0} \frac{f(g_1(t+h), g_2(t+h)) - f(g_1(t), g_2(t))}{h}$$

and use a little trick by subtracting and adding the term

$$\frac{1}{h} \cdot J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix}$$

to the diffence quotient. This leads to the equation

$$\begin{aligned} \frac{F(t+h) - F(t)}{h} = {} & \frac{f(g_1(t+h), g_2(t+h)) - f(g_1(t), g_2(t))}{h} \\ & - \frac{1}{h} \cdot J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix} \\ & + \frac{1}{h} \cdot J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix} \end{aligned}$$

$$= \frac{f(g_1(t+h), g_2(t+h)) - f(g_1(t), g_2(t)) - J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix}}{h}$$
$$+ \frac{1}{h} \cdot J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix}$$

$$= \frac{f(g_1(t+h), g_2(t+h)) - \left[f(g_1(t), g_2(t)) + J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix} \right]}{h}$$
$$+ J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} \frac{g_1(t+h) - g_1(t)}{h} \\ \frac{g_2(t+h) - g_2(t)}{h} \end{pmatrix}.$$

Finally, we expand the first term with

$$\|(g_1(t+h), g_2(t+h)) - (g_1(t), g_2(t))\|_2$$

and change the position of the denominators:

$$= \underbrace{\frac{f(g_1(t+h), g_2(t+h)) - \left[f(g_1(t), g_2(t)) + J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1(t+h) - g_1(t) \\ g_2(t+h) - g_2(t) \end{pmatrix} \right]}{\|(g_1(t+h), g_2(t+h)) - (g_1(t), g_2(t))\|_2}}_{(I)}$$
$$\cdot \underbrace{\frac{\|(g_1(t+h), g_2(t+h)) - (g_1(t), g_2(t))\|_2}{h}}_{(II)} + \underbrace{J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} \frac{g_1(t+h) - g_1(t)}{h} \\ \frac{g_2(t+h) - g_2(t)}{h} \end{pmatrix}}_{(III)}.$$

Now, let us consider the three terms (I), (II) and (III) separately:

(I) For the first term we note that the continuity of g_1 and g_2 implies that

$$(g_1(t+h), g_2(t+h)) \to (g_1(t), g_2(t)).$$

Therefore, term (I) goes to zero since f is totally differentiable (just set $x = g_1(t+h)$, $y = g_2(t+h)$, $x_0 = g_1(t)$ and $y_0 = g_2(t)$ in the definition of total differentiability).

(II) This term can be reformulated to

$$\sqrt{\left(\frac{g_1(t+h) - g_1(t)}{h} \right)^2 + \left(\frac{g_2(t+h) - g_2(t)}{h} \right)^2} \to \sqrt{(g_1'(t))^2 + (g_2'(t))^2}$$

where we used the differentiability of g_i and the continuity of $\sqrt{(\cdot)}$ and $(\cdot)^2$. So this term is bounded and the complete product $(I) \cdot (II)$ still goes to zero.

(III) Because of the differentiability of g_i we have for the this term as h goes to zero

$$J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1'(t) \\ g_2'(t) \end{pmatrix}$$

This completes the proof:

$$\begin{aligned} F'(t) &= \lim_{h\to 0} \frac{F(t+h) - F(t)}{h} \\ &= \lim_{h\to 0} [(I) \cdot (II) + (III)] \\ &= \lim_{h\to 0}(I) \cdot \lim_{h\to 0}(II) + \lim_{h\to 0}(III) \\ &= 0 \cdot \sqrt{(g_1'(t))^2 + (g_2'(t))^2} + J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1'(t) \\ g_2'(t) \end{pmatrix} \\ &= J_f(g_1(t), g_2(t)) \cdot \begin{pmatrix} g_1'(t) \\ g_2'(t) \end{pmatrix}. \end{aligned}$$

□

4.2 Remark
For the multi-dimensional case, the situation is similar. Let f: $\mathbb{R}^n \to \mathbb{R}$ be totally differentiable and assume that g_i: $(a,b) \to \mathbb{R}$ $(i = 1, \ldots, n)$ are differentiable as well. Then the function

$$F : (a,b) \to \mathbb{R}, \quad t \mapsto F(t) := f\left(g_1(t), g_2(t), \ldots, g_n(t)\right)$$

is also differentiable and we have

$$F'(t) = J_f\left(g_1(t), g_2(t), \ldots, g_n(t)\right) \cdot \begin{pmatrix} g_1'(t) \\ g_2'(t) \\ \vdots \\ g_n'(t) \end{pmatrix}.$$

4.3 Example (Rock-climbing)
A very nice illustration of parameter-dependent functions and the chain rule is the following. Assume that f: $\mathbb{R}^2 \to \mathbb{R}$ is the height profile of a mountain, so $f(x,y)$ is the height above sea-level of the mountain at position (x,y) in the plane.

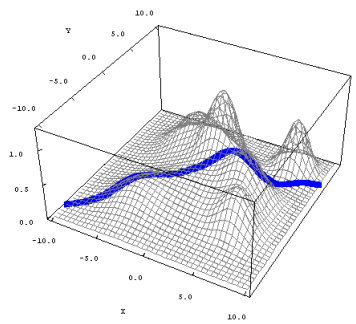


Figure II.8: A mountain and your path

As a young mountain climber you decided to climb the mountain and so you chose a special path to get to the top. The functions

$$x : \mathbb{R}_+ \to \mathbb{R}, \;\; t \mapsto x(t)$$
$$y : \mathbb{R}_+ \to \mathbb{R}, \;\; t \mapsto y(t)$$

shall determine your x- and y-position at time t, respectively. Then, you are of course interested in the height you have already climbed at time t. This is precisely expressed by the composed function

$$F : \mathbb{R}_+ \mapsto \mathbb{R}, \;\; t \mapsto f(x(t), y(t)),$$

because the coordinates of your walk at time t are $(x(t), y(t))$ and inserting these into f gives you the associated height.
But what does the derivative of F tell you? As always, the derivative is interpreted as a rate of change, so in this case it tells you how fast your height is changing along the path you have chosen.
As a concrete example consider a mountain given by the function

$$f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) = \frac{100}{1 + x^2 + y^2}.$$

visualized in Figure II.9(a).

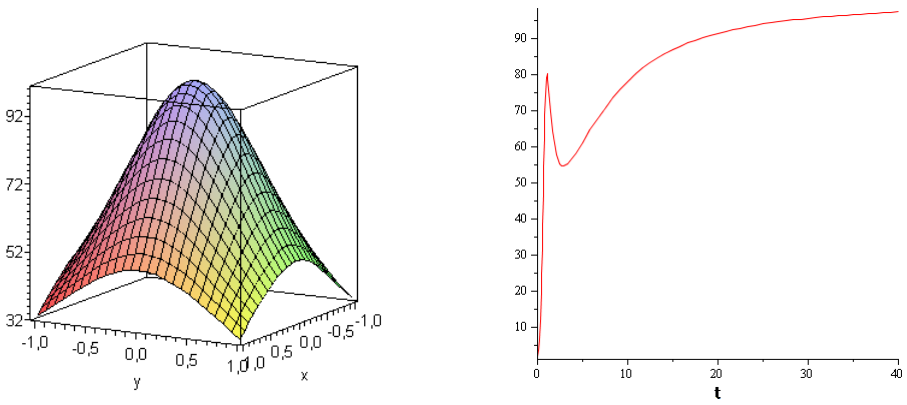


Figure II.9: The mountain to climb and the height profile

The mountain top is in the origin at height $f(0, 0) = 100\,\mathrm{m}$. We have chosen to walk the path where we reach x- and y-coordinates at time t as follows:

$$x : \mathbb{R}_+ \to \mathbb{R}, \;\; t \mapsto x(t) = \frac{1}{t+1}$$
$$y : \mathbb{R}_+ \to \mathbb{R}, \;\; t \mapsto y(t) = -\frac{7(t-1)}{(t+1)^2}$$

So, our height F (in m*) at time t (in* min*) is*

$$F(t) = \frac{100}{1 + \frac{1}{(t+1)^2} + \frac{49(t-1)^2}{(t+1)^4}}.$$

To calculate how fast our height is changing, we need to take the derivative of F, which is relatively ugly here. So we decompose this function and calculate

$$x'(t) = -\frac{1}{(t+1)^2} \qquad y'(t) = \frac{7(t-3)}{(t+1)^3}$$
$$f_x(x,y) = -\frac{200x}{(1+x^2+y^2)^2} \qquad f_y(x,y) = -\frac{200y}{(1+x^2+y^2)^2}.$$

Note that x and y are differentiated with respect to the time t, so g' is a climbing velocity measured in m/min*, whereas f is differentiated with respect to a length, so the partial derivatives f_x and f_y can be interpreted as a percentage slope with unit* $\mathrm{m}/\mathrm{m} = 1$ *relating an altitude difference to the way along the x- or y-axis.*

Using the chain rule, we can calculate the derivative as

$$F'(t) = (f_x(x(t),y(t)), f_y(x(t),y(t))) \cdot \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix}.$$

The calculation of explicit values is then easy for a computer if, e.g., these parts are supplied by individual functions in a code. For example, if we were to calculate the change in height at time $t = 5$ min*, we can proceed as follows:*

(i) *Calculate*

$$x(5) = \frac{1}{6} \qquad y(5) = -\frac{7}{9}$$
$$x'(5) = -\frac{1}{36} \qquad y'(5) = \frac{7}{108}$$

(ii) *Additionally determine*

$$f_x\left(\frac{1}{6}, -\frac{7}{9}\right) \approx -12.50 \qquad f_y\left(\frac{1}{6}, -\frac{7}{9}\right) \approx 58.35$$

(iii) *Finalize the calculation by*

$$F'(5) = (-12.504, 58.353) \cdot \left(-\frac{1}{36}, \frac{7}{108}\right)^\top = 4.13 \text{ m/min}.$$

4.4 Examples

(i) Consider the function $f(x,y) = x^2 + y^2$. We are interested in the derivatives along the circle of radius $r = 1$. Therefore, we choose

$$x : \mathbb{R}_+ \to \mathbb{R}, t \mapsto x(t) = \cos(t)$$
$$y : \mathbb{R}_+ \to \mathbb{R}, t \mapsto y(t) = \sin(t)$$

and calculate

$$x'(t) = -\sin(t) \qquad y'(t) = \cos(t)$$

$$f_x(x,y) = 2x \qquad f_y(x,y) = 2y$$

and derive

$$F'(t) = (2\cos(t), 2\sin(t)) \cdot (-\sin(t), \cos(t))^\top = 0.$$

This is precisely the change in height that we would expect along the circle, since we are actually moving along the level set $L_1(f)$,

(ii) Consider the function $f(x,y) = x^2 y$. We are interested in the derivatives along the curve given by the coordinates

$$\begin{aligned} x &: \mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto x(t) = t^3 \\ y &: \mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto y(t) = t^4. \end{aligned}$$

So, we calculate

$$\begin{aligned} x'(t) &= 3t^2 & y'(t) &= 4t^3 \\ f_x(x,y) &= 2xy & f_y(x,y) &= x^2 \end{aligned}$$

and derive

$$F'(t) = (2t^3 t^4, (t^3)^2) \cdot \left(3t^2, 4t^3\right)^\top = 10t^9.$$

As an application of the multi-dimensional chain rule, we now provide a method to differentiate single variable functions that are given in an implicit form, i.e.,

$$F(x, f(x)) = 0.$$

In such cases it might not be easy to solve for $f(x)$ explicitly or, using the techniques provided here, it might be easier to differentiate the function in the given form. We formulate the method first and provide some examples afterwards.

4.5 Corollary (Implicit differentiation)
Let $f\colon \mathbb{R} \to \mathbb{R}$ be a differentiable function satisfying

$$F(x, f(x)) = 0, \;\; x \in \mathbb{R},$$

for some totally differentiable function $F\colon \mathbb{R}^2 \to \mathbb{R}$. For given $(x_0, f(x_0))$ assume further that $F_y(x_0, f(x_0)) \neq 0$. Then we have

$$f'(x_0) = -\frac{F_x(x_0, f(x_0))}{F_y(x_0, f(x_0))}.$$

Proof.
Defining the functions

$$\begin{aligned} x &: \mathbb{R} \;\to \mathbb{R}, \quad t \mapsto x(t) = t \\ y &: \mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto y(t) = f(t) \end{aligned}$$

we can apply the chain rule for

$$\tilde{F}(t) = F(x(t), y(t)) = F(t, f(t))$$

and derive

$$\begin{aligned}\tilde{F}'(t) &= (F_x(x(t), y(t)), F_y(x(t), y(t))) \cdot (x'(t), y'(t))^\top \\ &= F_x(t, f(t)) + F_y(t, f(t)) f'(t).\end{aligned}$$

By assumption we know that $\tilde{F}(t) = F(t, f(t)) = 0$ for all t, so it is constant. Therefore, its derivative must also equal zero, i.e.,

$$0 = \tilde{F}'(t) = F_x(t, f(t)) + F_y(t, f(t)) f'(t).$$

Evaluating this in x_0 and solving for f' yields the result

$$f'(x_0) = -\frac{F_x(x_0, f(x_0))}{F_y(x_0, f(x_0))}.$$

□

4.6 Examples

(i) *Consider the ellipse with the characteristic equation*

$$\frac{x^2}{36} + \frac{y^2}{16} = 1$$

depicted in an x-y-diagram in Figure II.10.

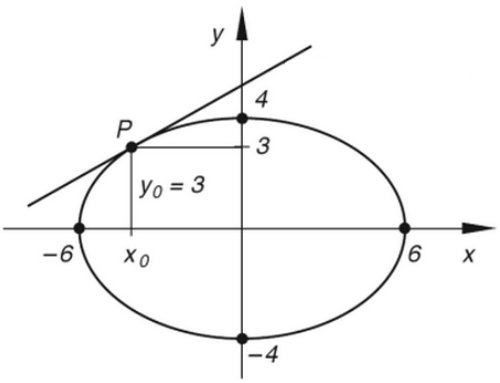


Figure II.10: Ellipse and tangent

We want to determine the slope of the ellipse in the point $(-\frac{3}{2}\sqrt{7}, 3)$*. In the region where* $x \leq 0$ *and* $y \geq 0$*, the ellipse can be described by a function* $f\colon \mathbb{R}_{\leq 0} \to \mathbb{R}$ *satisfying*

$$\frac{x^2}{36} + \frac{f(x)^2}{16} - 1 = 0.$$

So in our terminology we have

$$F(x, y) = \frac{x^2}{36} + \frac{y^2}{16} - 1,$$

which is of course totally differentiable. Then we may calculate

$$f'(x) = -\frac{F_x(x, f(x))}{F_y(x, f(x))} = -\frac{x/18}{f(x)/8} = -\frac{4}{9} \cdot \frac{x}{f(x)}.$$

Since $(-\frac{3}{2}\sqrt{7}, 3) = (x, f(x))$ *is on the ellipse (as it satisfies its equation) we can calculate the slope of the tangent to be*

$$-\frac{4 \cdot (-\frac{3}{2}\sqrt{7})}{9 \cdot 3} = \frac{2}{9}\sqrt{7}.$$

(ii) Consider the so-called Cartesian leaf curve with the characteristic equation

$$x^3 + y^3 = \frac{9}{2} x \cdot y$$

defining a third-order curve named after the French mathematician René Descartes. Its trace is depicted in an x-y-diagram in Figure II.11.

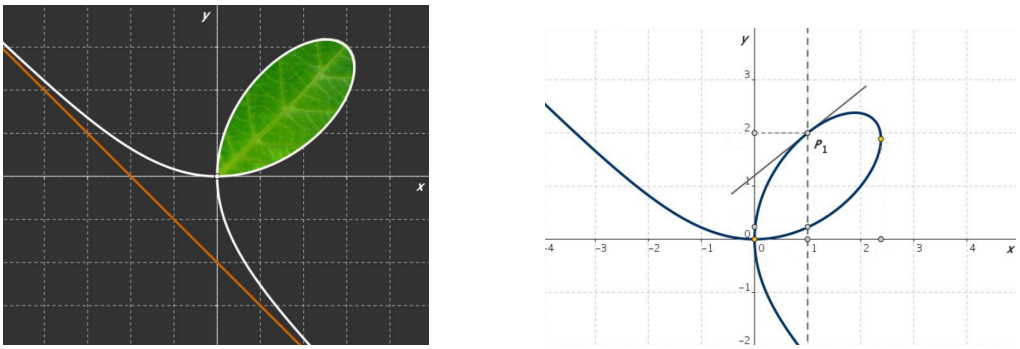


Figure II.11: Cartesian leaf

We want to determine the slope of the curve in the point $(1, 2)$*. In this region, the curve can be described by a function* $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisfying*

$$x^3 + f(x)^3 - \frac{9}{2} x \cdot f(x) = 0.$$

So in our terminology we have $F(x, y) = x^3 + y^3 - \frac{9}{2} x \cdot y$*, which is of course totally differentiable. Then we may calculate*

$$f'(x) = -\frac{F_x(x, f(x))}{F_y(x, f(x))} = -\frac{3x^2 - (9/2)f(x)}{3f(x)^2 - (9/2)x}.$$

Since $(1, 2) = (x, f(x))$ *is on the curve (as it satisfies its equation) we can calculate the slope of the tangent to be*

$$-\frac{3 \cdot 1^2 - (9/2)2}{3 \cdot 2^2 - (9/2)} = \frac{4}{5}.$$

(iii) Please note that the calculation of derivatives as demonstrated in the last two examples only works because we consider points that satisfy the characteristic equation of the curve, i.e., we consider derivatives at points where we actually have the function defining the curve there.

5 Directional derivatives & gradient

In Section II.2 we have learnt how to differentiate a multi-variable function. More precisely, we introduced partial derivatives, which is simply an ordinary single-variable derivative along one of the coordinate axes. But since there are more dimensions available, this is not the only possibility to define a (partial) derivative. We can also go into another direction.

5.1 Definition
Given $f\colon \mathbb{D} \subset \mathbb{R}^2 \to \mathbb{R}$, $x = (x_0, y_0) \in \mathbb{D}$ *and* $v^\top \in \mathbb{R}^2$ *with* $\|v\|_p = 1$. *If the limit*

$$\lim_{h \to 0} \frac{f(x + h \cdot v) - f(x)}{h}$$

exists, this real number is called the ***directional derivative of*** f ***w.r.t.*** v. *In this case we denote it by* $\partial_v f(x_0, y_0)$ *or* $\dfrac{\partial f}{\partial v}(x_0, y_0)$.

5.2 Remarks

(i) *The directional derivative describes how the function values of* f *are changing when we move along a straight line in the x-y-plane in direction* v *(compare* Figure II.12 *for a visualization).*

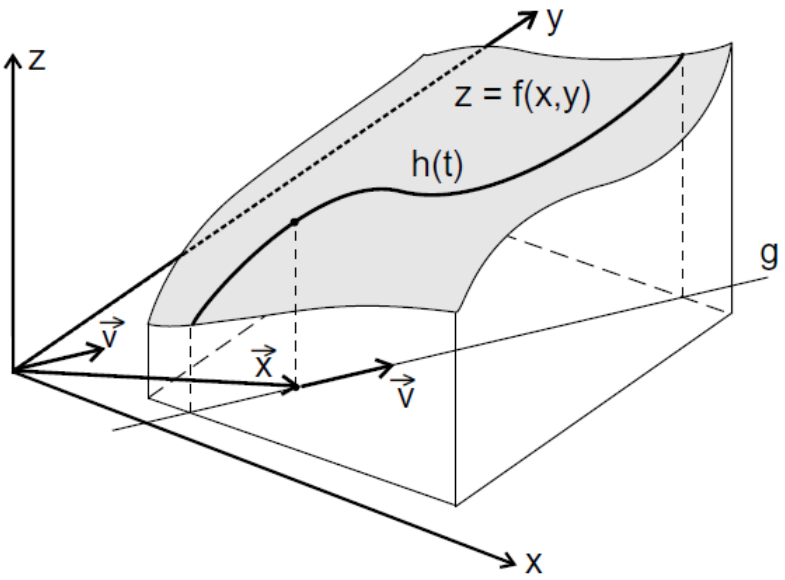


Figure II.12: Directional derivative

To see this we consider the straight line

$$g := \{x + t \cdot v \mid t \in \mathbb{R}\}$$

in the x-y-plane and define the function obtained from f *by inserting a cutting plane along this line, orthogonal to the x-y-plane, i.e.,*

$$h : \mathbb{R} \to \mathbb{R}, \quad t \mapsto h(t) := f(x + t \cdot v).$$

This function is simply the trace in this cutting plane (just like a coordinate curve) and describes the function values in direction v. Its derivative in $t = 0$ is then given by

$$h'(0) = \lim_{t\to 0} \frac{h(t) - h(0)}{t} = \lim_{t\to 0} \frac{f(x + t \cdot v) - f(x)}{t} = \partial_v f(x),$$

which means that the change in direction v is described by the directional derivative in that direction.

(ii) Special cases for the directional derivative are given by the partial derivatives:

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h\to 0} \frac{f((x_0, y_0) + h \cdot e_1) - f(x_0, y_0)}{h} = \partial_{e_1} f(x_0, y_0)$$
$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h\to 0} \frac{f((x_0, y_0) + h \cdot e_2) - f(x_0, y_0)}{h} = \partial_{e_2} f(x_0, y_0)$$

(iii) A mountain climber is at a certain point with coordinates (x_0, y_0) in the mountains. Then the directional derivative tells him how steep the path along a given direction will be.

So far, we don't have a reasonable method to calculate directional derivatives, since the method via difference quotients cannot be the viable way. We will see that the following construct will be very helpful.

5.3 Definition
Let $f\colon \mathbb{D} \subset \mathbb{R}^2 \to \mathbb{R}$ be partially differentiable and $(x_0, y_0) \in \mathbb{D}$. The vector

$$\nabla f(x_0, y_0) = \operatorname{grad}(f) := \begin{pmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \end{pmatrix} = J_f(x_0, y_0)^\top$$

*is called the **gradient of** f at (x_0, y_0).*

This means that applying the so-called "Nabla"-operator ∇ to a function f at position (x_0, y_0) yields a vector containing the partial derivatives at this point as an output. This helps to reformulate our differentiation rules.

5.4 Corollary
Let f, $g\colon \mathbb{D} \subset \mathbb{R}^2 \to \mathbb{R}$ be partially differentiable, $(x, y) \in \mathbb{D}$ and $\alpha \in \mathbb{R}$. Then

(i) $\nabla\left[(\alpha f + g)(x, y)\right] = \alpha \nabla f(x, y) + \nabla g(x, y)$ **(linearity of differentiation)**,

(ii) $\nabla\left[(f \cdot g)(x, y)\right] = \nabla f(x, y) \cdot g(x, y) + f(x, y) \cdot \nabla g(x, y)$ **(product rule)**,

(iii) $\nabla \dfrac{f(x, y)}{g(x, y)} = \dfrac{\nabla f(x, y) \cdot g(x, y) - f(x, y) \cdot \nabla g(x, y)}{g(x, y)^2}$ **(quotient rule)**.

The next proposition tells us how to calculate partial derivatives simply by using the gradient.

5.5 Proposition
Let $f\colon \mathbb{D} \subset \mathbb{R}^2 \to \mathbb{R}$ be totally differentiable, $x \in \mathbb{D}$ and $v^\top \in \mathbb{R}^2$ with $\|v\|_p = 1$. Then

$$\partial_v f(x) = \langle \nabla f(x), v \rangle = J_f(x) \cdot v.$$

Proof.
Since f is totally differentiable, we have

$$\begin{aligned}
\partial_v f(x) &= \lim_{h\to 0} \frac{f(x+h\cdot v) - f(x)}{h} \\
&= \lim_{h\to 0} \left(\frac{f(x+h\cdot v) - f(x)}{h} - \frac{1}{h}\left(J_f(x)(x+hv-x)^\top\right) + \frac{1}{h}\left(J_f(x)(x+hv-x)^\top\right)\right) \\
&= \lim_{h\to 0} \left(\frac{f(x+h\cdot v) - \left[f(x) + h\cdot J_f(x)v^\top\right]}{h}\right) + \lim_{h\to 0}\left(J_f(x)v^\top\right) \\
&= J_f(x)v^\top.
\end{aligned}$$

Here $f(x)+h\cdot J_f(x)v^\top$ is the tangent plane in x at position $x+hv$ and therefore the first limit is zero as we approach along the tangent plane towards the point $f(x)$ faster than h goes to zero. □

We calculate some examples.

5.6 Examples

(i) Consider the function $f(x,y) = x^2y^2 + y + 1$. We are interested in the directional derivative in $(0,0)$ along $v^\top = \left(1/\sqrt{2}, 1/\sqrt{2}\right)^\top$. Then we can calculate

$$\begin{aligned}
f_x(x,y) &= 2xy^2 & f_x(0,0) &= 0 \\
f_y(x,y) &= 2x^2y + 1 & f_y(0,0) &= 1 \\
J_f(x,y) &= (2xy^2, 2x^2y+1) & J_f(0,0) &= (0,1) \\
\nabla f(x,y) &= J_f^\top & \nabla f(0,0) &= (0,1)^\top
\end{aligned}$$

and so we apply the last proposition to calculate the directional derivative

$$\partial f_v(x,y) = \left\langle \nabla f(x,y), v^\top \right\rangle = \left\langle \begin{pmatrix} 2xy^2 \\ 2x^2y+1 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \frac{2xy^2}{\sqrt{2}} + \frac{2x^2y+1}{\sqrt{2}},$$

$$\partial f_v(0,0) = \left\langle \nabla f(0,0), v^\top \right\rangle = \left\langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = 0\cdot\frac{1}{\sqrt{2}} + 1\cdot\frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}.$$

(ii) Consider the function $f(x,y) = \|(x,y)\|_2$. We are interested in the directional derivative along $v^\top = \left(1/\sqrt{2}, 1/\sqrt{2}\right)^\top$. Then we can calculate

$$f_x(x,y) = \frac{x}{\|(x,y)\|_2}$$

$$\begin{aligned} f_y(x,y) &= \frac{y}{\|(x,y)\|_2} \\ J_f(x,y) &= \left(\frac{x}{\|(x,y)\|_2}, \frac{y}{\|(x,y)\|_2} \right) \\ \nabla f(x,y) &= J_f^\top(x,y) \end{aligned}$$

and so we apply the last proposition to calculate the directional derivative

$$\partial f_v(x,y) = \left\langle \nabla f(x,y), v^\top \right\rangle = \left\langle \begin{pmatrix} \frac{x}{\|(x,y)\|_2} \\ \frac{y}{\|(x,y)\|_2} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \frac{x+y}{\sqrt{2} \cdot \|(x,y)\|_2}.$$

The gradient has more very appealing properties relevant to applications. The most important ones are summarized in the following theorem.

5.7 Theorem
Let f be totally differentiable in $x = (x_0, y_0)$ and $\nabla f(x_0, y_0) \neq \mathbb{0}$.

(i) *The gradient points in the direction of steepest ascent, i.e., for $g := \dfrac{\nabla f(x)}{\|\nabla f(x)\|_2}$, we have*

$$\partial_g f(x) \geq \partial_w f(x) \qquad \text{for all } w \in \mathbb{R}^2 \text{ with } \|w\|_2 = 1.$$

The value $\|\nabla f(x)\|_2$ is a measure for the steepness of f in $x = (x_0, y_0)$.

(ii) *Analogously, the direction of steepest descent is $-\nabla f(x)$, i.e., for $g := -\dfrac{\nabla f(x)}{\|\nabla f(x)\|_2}$, we have*

$$\partial_g f(x) \leq \partial_w f(x) \qquad \text{for all } w \in \mathbb{R}^2 \text{ with } \|w\|_2 = 1.$$

(iii) *If v is a tangential vector at the level set*

$$L_{f(x_0,y_0)}(f) = \{(x,y) \in \mathbb{R}^2 \mid f(x,y) = f(x_0,y_0)\},$$

then v is orthogonal to $\nabla f(x_0, y_0)$.

Proof.
The proof for (i) and (ii) is really simple. Consider an arbitrary vector w with $\|w\|_2 = 1$. Then we know from linear algebra that the angle α between this vector and the vector $\nabla f(x_0, y_0)$ is given by

$$\cos\alpha = \frac{\langle \nabla f(x_0,y_0), w \rangle}{\|\nabla f(x_0,y_0)\|_2 \cdot \|w\|_2}$$

and so, using Proposition II.5.5 we can write the directional derivative as

$$\begin{aligned} \partial_w f(x_0,y_0) &= \langle \nabla f(x_0,y_0), w \rangle \\ &= \|\nabla f(x_0,y_0)\|_2 \cdot \|w\|_2 \cdot \cos\alpha \\ &= \|\nabla f(x_0,y_0)\|_2 \cdot \cos\alpha \end{aligned}$$

(i) The steepness in direction w is given by this directional derivative and it is obviously maximal, if $\cos\alpha = 1$, which occurs only when the angle between $\nabla f(x_0, y_0)$ and w is 0, i.e., when the direction w we choose is the gradient direction.

(ii) Moreover, the steepness is minimal, if the directional derivative becomes minimal, which happens for $\cos\alpha = -1$, i.e., if w points in the opposite direction of $\nabla f(x_0, y_0)$ (at $\alpha = 180°$).

□

5.8 Remarks/Examples

(i) Please note that the gradient is situated in the x-y-plane. So the direction it points to must be depicted in this plane. The directional derivative, which is a real number, then indicates the change in function values along the gradient direction.

(ii) The situation is illustrated in Figure II.13. *The first picture shows the x-y-plane and a level set corresponding to the function value $f(x)$. It shall be illustrated that the gradient is orthogonal to a tangent vector v on the level set at point x. The second picture shows the graph of the function together with the level set and one may observe that $-\nabla f(x)$ points in a direction where the function value is decreased the most: the steepest descent direction.*

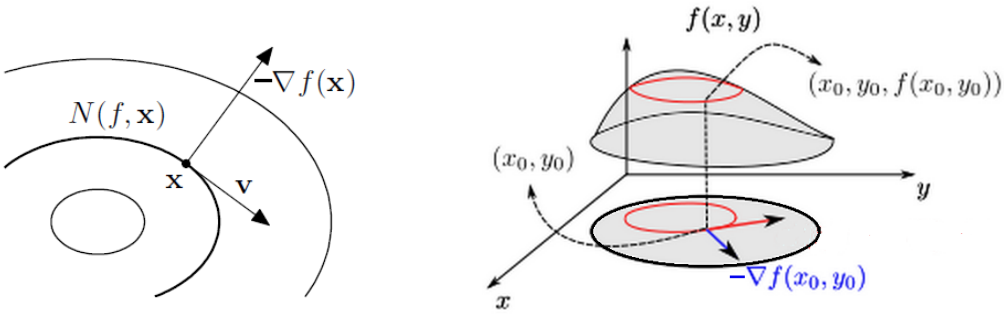


Figure II.13: Illustration of Theorem 5.7

(iii) A nice illustration for the gradient is the fastest path down a mountain. Assume you are almost on top of a mountain. The fastest way to get down is to walk orthogonal to the height-lines on the map. This is done by always taking the path along the direction $-\nabla f(x, y)$. Such a path is illustrated in Figure II.14.

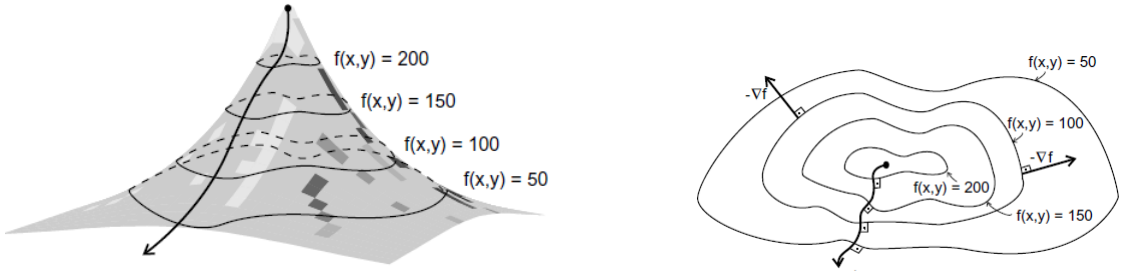


Figure II.14: Climbing down a mountain - steepest descent

(iv) *A continuous grey-value picture can be interpreted as a function $f(x, y)$ of its coordinates. If we cover the picture by a raster and calculate the gradients at all these points, then the size of the values $\|\nabla f(x, y)\|_2$ tells us how strong the change in colors is at a particular point. Edges in the picture are located along points where this value differs very strongly. The edge is then located orthogonal to the gradients we calculated. All in all, the gradient can be seen as a simple edge detector.*

(v) *All theorems and notations generalize naturally to functions $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$. Just note that the gradient is then defined as*

$$\nabla f(x) = \begin{pmatrix} f_{x_1}(x) \\ f_{x_2}(x) \\ \vdots \\ f_{x_n}(x) \end{pmatrix}.$$

Chapter III

Optimization of multivariable functions

The global objective of this chapter is to find the points where a given function f: $\mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ attains its maximum and minimum function values. These points are of major interest for many applications, since they may represent a combination of minimum cost for economical situations or minimum errors when it comes to model parameter estimation.
In the first section we will introduce some basic knowledge on tools that we need to generalize the optimization statements from the single variable situation, e.g., to obtain results of type:

$$f : (a, b) \to \mathbb{R} \text{ has a local maximum in } x_0, \text{ if } f'(x_0) = 0 \text{ and } f''(x_0) < 0.$$

To do so, we will generalize Taylor's formula to the multi-dimensional case and recall some linear algebra.

1 Definiteness of matrices

Since we will need it later, we introduce the so-called Hessian matrix first.

1.1 Definition
For $f : \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ with $f \in \mathcal{C}^2(\mathbb{D})$ and $(x, y) \in \mathbb{D}$, the (symmetric!) matrix

$$H_f(x, y) := \begin{pmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{pmatrix}$$

*is called the **Hessian matrix** of f in (x, y).*

Please note that $H_f(x, y)$ is necessarily symmetric ($H_f^\top(x, y) = H_f(x, y)$) by Schwarz Theorem (compare II.2.6).
For single variable functions, Taylor's theorem was an important tool to approximate function values of complex functions around a given point using a polynomial whose coefficients are composed of the derivatives in that point. We will apply this now to two-variable functions.

1.2 Discussion (Taylor formula)
Consider a function $f : \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ with $f \in \mathcal{C}^3(\mathbb{D})$. Recall that for sufficiently smooth single variable functions we have

$$g(x+h) = g(x) + \frac{g'(x)}{1!}h + \frac{g''(x)}{2!}h^2 + O(h^3).$$

Now, we want to apply this representation to determine the function value $f(x+h, y+k)$ close to $f(x,y)$ using only evaluations of f and its derivatives in (x,y). Therefore, we fix the second coordinate first and develop f only w.r.t. x into a Taylor series. We derive

$$f(x+h, y+k) = f(x, y+k) + \frac{f_x(x,y+k)}{1!}h + \frac{f_{xx}(x,y+k)}{2!}h^2 + O(h^3). \tag{1.1}$$

On the right hand side we still have terms of f and its derivatives that involve the displacement of y, so we fix x now and develop these terms with respect to y. We derive

$$\begin{aligned}
f(x, y+k) &= f(x,y) + \frac{f_y(x,y)}{1!}k + \frac{f_{yy}(x,y)}{2!}k^2 + O(k^3) && \text{(order 2)}\\
f_x(x, y+k) &= f_x(x,y) + \frac{f_{xy}(x,y)}{1!}k + O(k^2) && \text{(order 1)}\\
f_{xx}(x, y+k) &= f_{xx}(x,y) + O(k) && \text{(order 0)}
\end{aligned}$$

Including all this in Equation (1.1)*, we obtain*

$$\begin{aligned}
f(x+h, y+k) = f(x,y) &+ f_x(x,y)h + f_y(x,y)k\\
&+ \frac{f_{xx}(x,y)}{2!}h^2 + f_{xy}(x,y)h\cdot k + \frac{f_{yy}(x,y)}{2!}k^2\\
&+ O(h^3) + O(h^2k) + O(hk^2) + O(k^3),
\end{aligned}$$

or in matrix-vector notation

$$f(x+h, y+k) = f(x,y) + J_f(x,y)\begin{pmatrix} h \\ k \end{pmatrix} + \frac{1}{2}(h,k)\underbrace{\begin{pmatrix} f_{xx}(x,y) & f_{xy}(x,y) \\ f_{yx}(x,y) & f_{yy}(x,y) \end{pmatrix}}_{H_f(x,y)}\begin{pmatrix} h \\ k \end{pmatrix} + \dots$$

We summarize our observations in the following theorem.

1.3 Theorem
For $f : \mathbb{D} \subseteq \mathbb{R}^2 \to \mathbb{R}$ with $f \in \mathcal{C}^2(\mathbb{D})$, $\{(x_0,y_0) + t\cdot[(x,y)-(x_0,y_0)] \mid t \in [0,1]\} \subset \mathbb{D}$, we have

$$f(x,y) = f(x_0,y_0) + J_f(x_0,y_0)\begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} + \frac{1}{2}(x-x_0,\ y-y_0)H_f(x_0,y_0)\begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} + \dots$$

1.4 Remarks/Examples

(i) *Consider the function given by* $f(x,y) = x^2\sin(y)$.

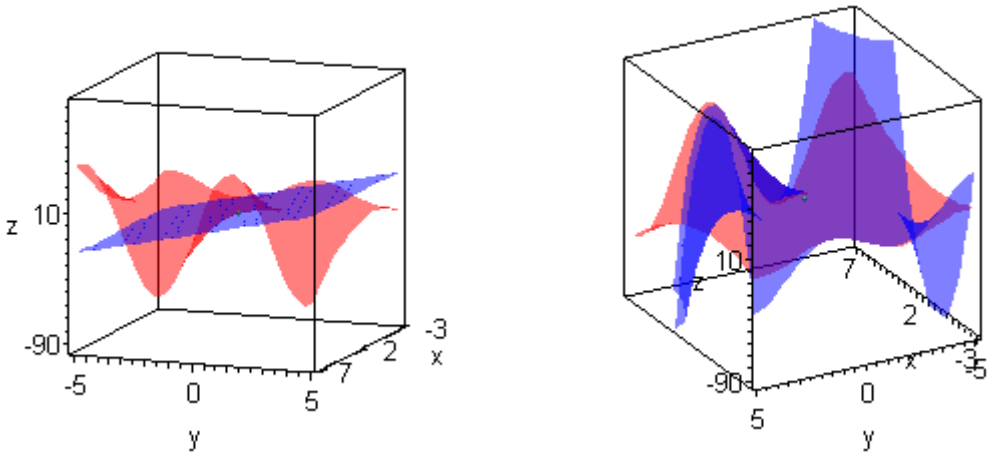


Figure III.1: Illustration of f and 1st and 2nd order approximation

In order to find a 2nd order approximation around the point $(x_0,y_0) = (2,0)$ *by a Taylor polynomial, we need to calculate*

	f	f_x	f_y	f_{xx}	f_{xy}	f_{yy}
general	$x^2\sin(y)$	$2x\sin(y)$	$x^2\cos(y)$	$2\sin(y)$	$2x\cos(y)$	$-x^2\sin(y)$
in $(2,0)$	0	0	4	0	4	0

Then, we use Theorem III.1.3 *and obtain*

$$\begin{aligned} f(x,y) &\approx f(2,0) + J_f(2,0)\cdot\begin{pmatrix} x-2 \\ y-0 \end{pmatrix} + \frac{1}{2}\cdot(x-2,\ \ y-0)\cdot H_f(2,0)\begin{pmatrix} x-2 \\ y-0 \end{pmatrix} \\ &= 0 + (0,\ \ 4)\begin{pmatrix} x-2 \\ y \end{pmatrix} + \frac{1}{2}\cdot(x-2,\ \ y)\cdot\begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix}\begin{pmatrix} x-2 \\ y \end{pmatrix} \\ &= 4y + 4y(x-2) = 4y(x-1). \end{aligned}$$

For example, we could calculate $f(2.2,0.2) \approx 4\cdot 0.2(2.2-1) = 0.96$ *which is exact for the first two digits after the comma already.*

(ii) *Consider the function given by* $f(x,y) = \sin(x^2+2y)$. *In order to find a 2nd order approximation around the point* $(x_0,y_0) = \left(0,\frac{\pi}{4}\right)$ *by a Taylor polynomial, we need to calculate*

	f	f_x	f_y
general	$\sin(x^2+2y)$	$2x\cos(x^2+2y)$	$2\cos(x^2+2y)$
in $\left(0,\frac{\pi}{4}\right)$	1	0	0
	f_{xx}	f_{xy}	f_{yy}
general	$-4x^2\sin(x^2+2y)+2\cos(x^2+2y)$	$-4\sin(x^2+2y)$	$-4x\sin(x^2+2y)$
in $\left(0,\frac{\pi}{4}\right)$	0	0	-4

Then, we use Theorem III.1.3 *and obtain*

$$\begin{aligned} f(x,y) &\approx f\left(0,\tfrac{\pi}{4}\right) + J_f\left(0,\tfrac{\pi}{4}\right)\cdot\begin{pmatrix} x-0 \\ y-\frac{\pi}{4}\end{pmatrix} + \frac{1}{2}\cdot(x-0,\ y-\frac{\pi}{4})\cdot H_f\left(0,\frac{\pi}{4}\right)\begin{pmatrix} x-0 \\ y-\frac{\pi}{4}\end{pmatrix} \\ &= 1 + (0,\ 0)\cdot\begin{pmatrix} x \\ y-\frac{\pi}{4}\end{pmatrix} + \frac{1}{2}\left(x,\ y-\frac{\pi}{4}\right)\cdot\begin{pmatrix} 0 & 0 \\ 0 & -4\end{pmatrix}\begin{pmatrix} x \\ y-\frac{\pi}{4}\end{pmatrix} \\ &= 1 - 2\cdot\left(y-\frac{\pi}{4}\right)^2 \end{aligned}$$

(iii) *The Taylor approximation easily generalizes to functions $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$. For $\{x_0 + t\cdot[x-x_0] \mid t\in[0,1]\} \subset \mathbb{D}$, we have*

$$f(x) = f(x_0) + J_f(x_0)\cdot(x-x_0) + \frac{1}{2}\cdot(x-x_0)^\top\cdot H_f(x_0)\,(x-x_0) + \dots$$

where the (symmetric!) Hessian is then given by

$$H_f(x_0) = \begin{pmatrix} f_{x_1x_1}(x_0) & \dots & f_{x_nx_1}(x_0) \\ \vdots & \ddots & \vdots \\ f_{x_1x_n}(x_0) & \dots & f_{x_nx_n}(x_0)\end{pmatrix}.$$

In order to examine functions on extreme points, it is necessary to enhance our linear algebra knowledge about matrices. We give a short summary of the topics that we will need on matrices.

1.5 Definition
For a matrix $A \in M_{n\times n}(\mathbb{R})$, we say that

(i) *A is **positive definite**, if $x^\top Ax = \langle x, Ax\rangle > 0$ for all $\mathbb{0} \neq x \in \mathbb{R}^n$,*

(ii) *A is **negative definite**, if $x^\top Ax = \langle x, Ax\rangle < 0$ for all $\mathbb{0} \neq x \in \mathbb{R}^n$,*

(iii) *A is **positive semi-definite**, if $\langle x, Ax\rangle \geq 0$ for all $\mathbb{0} \neq x \in \mathbb{R}^n$.*

(iv) *A is **negative semi-definite**, if $\langle x, Ax\rangle \leq 0$ for all $\mathbb{0} \neq x \in \mathbb{R}^n$.*

1.6 Examples
(i) *The identity matrix $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1\end{pmatrix}$ is positive definite, since*

$$x^\top Id_2 x = (x_1,x_2)\cdot\begin{pmatrix} x_1 \\ x_2\end{pmatrix} = \|x\|^2 > 0 \quad \text{for } \mathbb{0} \neq x.$$

(ii) *The matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & -1\end{pmatrix}$ is neither positive nor negative (semi)definite, since*

$$(1,0)A(1,0)^\top = (1,0)\cdot\begin{pmatrix} 1 \\ 0\end{pmatrix} = 1$$

$$(0,1)A(0,1)^\top = (0,-1)\cdot\begin{pmatrix} 0 \\ 1\end{pmatrix} = -1.$$

We say that the matrix is indefinite.

(iii) *For every real matrix A, the associated matrix $A^\top A$ is positive semi-definite, since*

$$x^\top (A^\top A)x = (Ax)^\top (Ax) = \|Ax\|_2^2 \geq 0 \quad \text{for } \mathbb{0} \neq x.$$

Moreover, if A is invertible, we can even ensure that $A^\top A$ is positive definite! In multiple dimensions the Newton method $x_{k+1} = x_k - H_f(x_k)^{-1}\nabla f(x_k)$ is often applied to find maxima of a function $f : \mathbb{R}^n \to \mathbb{R}$. The positive definiteness of the Hessian is a necessary requirement for the convergence of this method. It can be enforced by replacing $H_f(x_k)$ by $H_f(x_k)^\top H_f(x_k)$.

Positive definiteness seems to be very hard to determine for a given matrix. But for symmetric matrices ($A = A^\top$) the following theorem gives a very simple characterization. It is an immediate consequence of the spectral theorem.

1.7 Theorem
A symmetric matrix $A \in M_{n\times n}(\mathbb{R})$ is

(i) *positive definite $\iff$ all eigenvalues are > 0*

(ii) *negative definite $\iff$ all eigenvalues are < 0*

(iii) *positive (negative) semi-definite $\iff$ all eigenvalues are ≥ 0 (≤ 0).*

(iv) *indefinite $\iff$ A has negative and positive eigenvalues.*

For symmetric 2×2-matrices even more can be said.

1.8 Proposition
Let $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$ be a symmetric real 2×2-matrix.

(i) *A is positive definite $\iff \det(A) > 0$ and $a_{11} > 0$.*

(ii) *A is negative definite $\iff \det(A) > 0$ and $a_{11} < 0$.*

(iii) *A is semidefinite $\iff \det(A) \geq 0$.*

(iv) *A is indefinite $\iff \det(A) < 0$.*

Proof.
From linear algebra we know that the determinant is the product of all eigenvalues (denoted λ_1, λ_2), say

$$\det(A) = a_{11} \cdot a_{22} - a_{12}^2 = \lambda_1 \cdot \lambda_2$$

(i) So, if $\det(A) > 0$, we know that $\lambda_1 \cdot \lambda_2 > 0$ and both eigenvalues are either positive or negative (therefore the matrix is in any case positive or negative definite). Moreover, we see from

$$a_{11} \cdot a_{22} - a_{12}^2 > 0 \iff a_{11} \cdot a_{22} > a_{12}^2 \geq 0$$

that in case of $a_{11} > 0$, we must also have $a_{22} > 0$ or if $a_{11} < 0$ we must also have $a_{22} < 0$. Since we also know, that the sum of the eigenvalues is the trace of A, we have

$$a_{11} + a_{22} = \lambda_1 + \lambda_2.$$

We know that in case of $a_{11} > 0$ (and then also $a_{22} > 0$) we must have $\lambda_1 + \lambda_2 > 0$ which is only possible when both are positive, since we already noted that they have the same sign. Therefore A is positive definite in this case.

(ii) In case of $a_{11} < 0$ (and then also $a_{22} < 0$) we must have $\lambda_1 + \lambda_2 < 0$ which is only possible when both are negative, since we already noted that they have the same sign. Therefore A is negative definite in this case.

(iii) We already noted that $\det(A) \geq 0$ means that the eigenvalues are either both positive or both negative (or zero).

(iv) If $\lambda_1 \cdot \lambda_2 = \det(A) < 0$, we must have one positive and one negative eigenvalue and A is consequently indefinite.

□

Let us consider some examples.

1.9 Examples

(i) *The Hessian matrix* $H_f(2,0) = \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix}$ *from* Examples III.1.4(i) *is symmetric (as it is a Hessian!) and of dimension* 2. *Therefore, we can apply* Proposition III.1.8. *Calculating the determinant yields*

$$\det(H_f(2,0)) = 0 - 4^2 = -16$$

yields that the matrix is indefinite, i.e., we have positive and negative eigenvalues. In deed, calculating the characteristic polynomial gives

$$p_{char}(\lambda) = \lambda^2 - 16 = (\lambda - 4)(\lambda + 4),$$

and we see that the eigenvalues have opposite sign.

(ii) *The matrix* $A = \begin{pmatrix} 4 & -4 \\ -4 & 5 \end{pmatrix}$ *is symmetric and of dimension* 2. *Therefore, we can apply* Proposition III.1.8. *Calculating the determinant*

$$\det(A) = 20 - (-4)^2 = 4$$

and observing that $a_{11} = 4 > 0$ *yields that the matrix is positive definite.*

(iii) *The matrix* $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ *is positive definite, since it is a symmetric diagonal matrix containing only positive entries. Its characteristic polynomial is*

$$p_{char}(\lambda) = (\lambda - 1)(\lambda - 2)(\lambda - 3).$$

2 Unconstrained optimization - Local extreme points

Last semester we have learnt about necessary and sufficient criteria to determine whether a function $f : (a, b) \to \mathbb{R}$ has a local maximum or minimum in a point $x_0 \in (a, b)$. The key for all the theorems that we provided was the derivative which was used to translate geometric curve characteristics into analytic formulas (curve sketching). We will now do the same for a function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$.
We begin by defining local extreme points.

2.1 Definition
*A function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ has a **local maximum (minimum)** in $x_0 \in \mathbb{D}$, if*

$$f(x_0) \geq f(x) \qquad (f(x_0) \leq f(x)) \qquad \forall\, x \in B_\varepsilon(x_0).$$

*The local extreme point is called **isolated**, if the inequalities are strict.*

In the one-dimensional case, a function had a horizontal tangent at a local extreme point or, differently stated $f'(x) = 0$. In two dimensions this would mean to have a horizontal tangent plane at such a point, i.e., in its equation

$$t(x, y) = f(x_0, y_0) + \underbrace{(f_x(x_0, y_0), f_y(x_0, y_0))}_{=f'(x_0,y_0)} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}$$

we would expect $f'(x_0, y_0) = J_f(x_0, y_0) = \mathbb{0}$. This is also what we see in Figure III.3 and in deed the content of the next proposition.

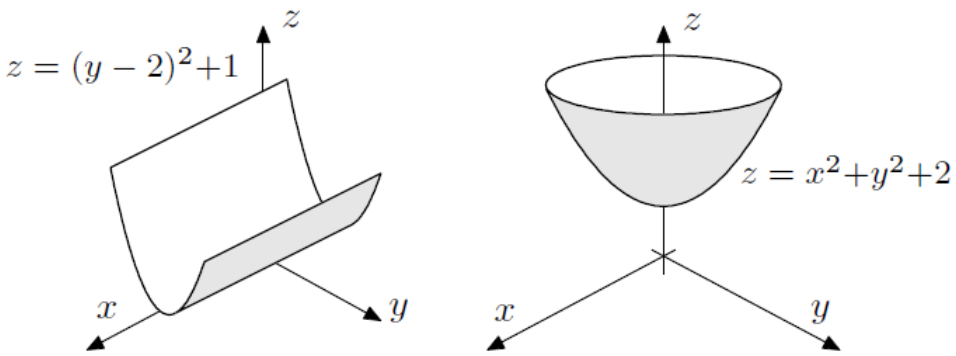


Figure III.2: Some examples of local extreme points

2.2 Proposition
If $f : \mathbb{R}^n \to \mathbb{R}$ has a local extreme point in x_0, then we have

$$f'(x_0) = J_f(x_0) = \nabla f(x_0)^\top = \mathbb{0}.$$

Proof.
The basic idea, say in two dimensions, is the following: If the surface f has a minimum

in x_0 then any coordinate curve in this point must also have a minimum there.
More precisely, we consider the coordinate curves

$$F_j : \mathbb{R} \to \mathbb{R}, \;\; t \mapsto F(t) := f(x_0 + t \cdot e_j) \quad (j = 1, \ldots, n).$$

Its derivatives are simply $F_j'(t) = f_{x_j}(x_0 + t \cdot e_j)$ for $j = 1, \ldots, n$, e.g., by the chain rule. Since f has a local extreme point in x_0, F_j must also have a local extreme point in $t = 0$. Therefore

$$f_{x_j}(x_0) = F_j'(0) = 0 \quad \text{(one-dimensional result!)}$$

and we see that all partial derivatives are zero in x_0, i.e., $J_f(x_0) = \mathbb{0}$. □

Let us consider some examples.

2.3 Remarks/Examples

(i) *Consider the function given by* $f(x, y) = x(y - 1) + x^3$. *Then we can determine*

$$\nabla f(x, y) = \begin{pmatrix} y - 1 + 3x^2 \\ x \end{pmatrix}$$

and all points where this quantity vanishes are potential candidates for local extreme points. In this case, we can solve the (non-linear) system $\nabla f(x, y) = \mathbb{0}$*:*

$$\begin{aligned} x &= 0 && \text{(from equation 2)} \\ y &= 1 && \text{(from equation 1)} \end{aligned}$$

and obtain that a local extreme point could be in $x_0 = (0, 1)$, *since it is a zero of the gradient.*

(ii) *Consider the function given by* $f(x, y) = x^2 \sin(y)$ *shown in* Figure III.3.

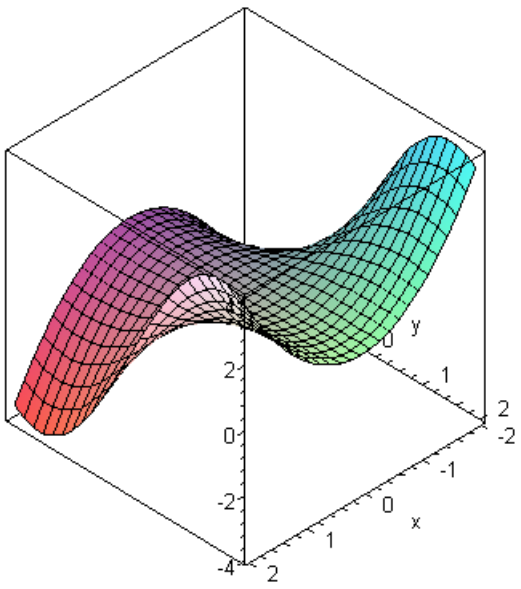


Figure III.3: gradient vanishes in $\mathbb{0}$, but no extreme point!

Then we can determine

$$\nabla f(x,y) = \begin{pmatrix} 2x\sin(y) \\ x^2\cos(y) \end{pmatrix}$$

and we clearly have a zero of the gradient in $(0,0)$*. However, looking at the graph reveals that there is no extreme point!*
In the picture we see that we can place a vertical cutting plane through $\mathbb{0}$ *thereby obtaining a maximum in the projected curve of* f *in this plane, while we could also place a cutting plane such that the function has a minimum in the projected plane. Such a point is called a* ***saddle point****.*

(iii) *As we have seen the vanishing of the first derivative does not necessarily imply the presence of a local extreme point. But this was also not the case in the one-dimensional situation (think of* $f(t) = t^3$*). We recall the that a point* $x_0 \in \mathbb{R}^n$*, where*

$$\nabla f(x_0) = \mathbb{0}$$

is called a ***stationary point****.*

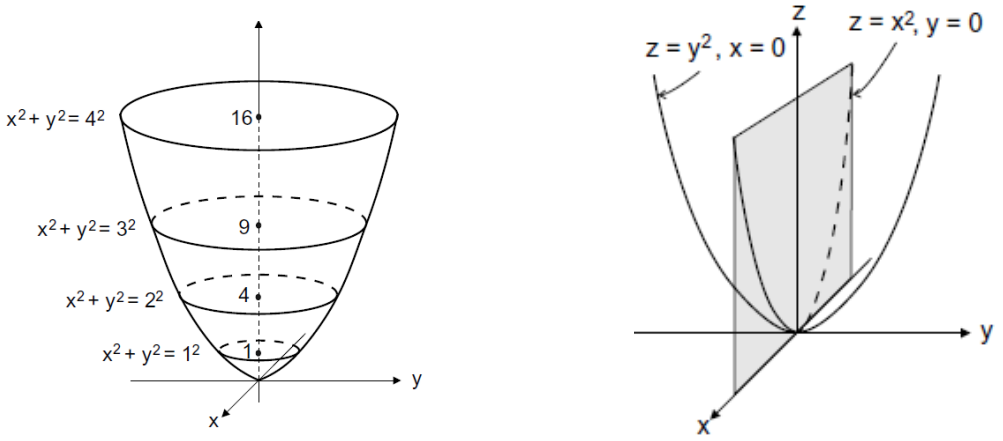
The stationary points are, beside the boundary points and points of discontinuity, our candidates for local extreme points. In order to answer whether this is the case, we have to look at higher order derivatives, like in the one-dimensional situation. We prove the following theorem.

2.4 Theorem
Given $f \in \mathcal{C}^2(\mathbb{R}^n)$ *and a stationary point of* f *in* $x_0 \in \mathbb{R}^n$*. Then*

(i) *f has an isolated local maximum in x_0, if $H_f(x_0)$ is negative definite.*

(ii) *f has an isolated local minimum in x_0, if $H_f(x_0)$ is positive definite.*

Proof.
We will show that the function projected to any vertical plane has a local minimum or maximum in the one-dimensional sense in x_0 (as indicated in Figure III.4)



(a) f has minimum in $\mathbb{0}$

(b) projected function has minimum in $\mathbb{0}$

Figure III.4: 3-d function and projected function

Therefore, we consider an arbitrary unit vector $v \in \mathbb{R}^n$ indicating a direction and the projected function

$$F : \mathbb{R} \to \mathbb{R}, \;\; t \mapsto F(t) := f(x_0 + t \cdot v).$$

This function F has a stationary point at $t = 0$, since

$$F'(0) = \partial_v f(x_0) = \langle \underbrace{\nabla f(x_0)}_{=\mathbb{0}}, \;\; v \rangle = 0,$$

as f has a stationary point in x_0 by assumption.
From one-dimensional theory, we know that F has a local maximum in $t = 0$, if $F''(0) < 0$ and a minimum if $F''(0) > 0$. We will see that this is ensured by our assumptions on the definiteness of the Hessian. Therefore, we develop F into a Taylor series around $t = 0$,

$$F(t) = F(0) + \frac{1}{1!}F'(0) \cdot t + \frac{1}{2!}F''(0) \cdot t^2 + O(t^3), \tag{2.1}$$

where we used the one-dimensional Taylor expansion! On the other side, we can apply the multi-dimensional Taylor expansion from Theorem III.1.3 to calculate $f(x_0 + t \cdot v)$, using a series around x_0, i.e.,

$$F(t) = f(x_0 + tv) = f(x_0) + J_f(x_0) \cdot v \cdot t + \frac{1}{2} \cdot v^\top H_f(x_0) \cdot v \cdot t^2 + O(t^3). \tag{2.2}$$

Comparing coefficients in Equations (2.1) and (2.2) gives

$$\begin{aligned} F(0) &= f(x_0) \\ F'(0) &= J_f(x_0) \cdot v = 0 \\ F''(0) &= v^\top \cdot H_f(x_0) \cdot v = \langle v, \;\; H_f(x_0) \cdot v \rangle. \end{aligned}$$

The first two equations were already verified, but the last one links $F''(0)$ to a definiteness condition on $H_f(x_0)$:

(i) If $H_f(x_0)$ is negative definite, then $F''(0) = \langle v, \;\; H_f(x_0) \cdot v \rangle < 0$ for all directions $v \in \mathbb{R}^n$. Therefore F has a local maximum in $t = 0$ in whichever projected plane we consider it, i.e., independent of v. Then f must have a local maximum.

(ii) If $H_f(x_0)$ is positive definite, then $F''(0) = \langle v, H_f(x_0)v \rangle > 0$ for all directions $v \in \mathbb{R}^n$. Therefore F has a local minimum in $t = 0$ in whichever projected plane we consider it (i.e., independent of v). Then f must have a local minimum.

□

So, to check whether a function has a local extreme point in a stationary point, the corresponding Hessian matrix must be examined for positive or negative definiteness. In the last section we already provided useful methods to check this for a matrix. More precisely, since the Hessian is always symmetric, we may use Theorem III.1.7 to formulate an explicit calculation method for local extreme points of a multidimensional function f.

2.5 Corollary
Given $f \in \mathcal{C}^2(\mathbb{R}^n)$ and a stationary point of f in $x_0 \in \mathbb{R}^n$. Then

(i) *f has an isolated local maximum in x_0, if all eigenvalues of $H_f(x_0)$ are < 0.*

(ii) *f has an isolated local minimum in x_0, if all eigenvalues of $H_f(x_0)$ are > 0.*

For two-variable functions, Proposition III.1.8 makes it even simpler:

Given $f \in \mathcal{C}^2(\mathbb{R}^2)$ and a stationary point of f in $x_0 \in \mathbb{R}^2$. Then

(i) *f has an isolated local extreme point, if $\det(H_f(x_0)) > 0$.*

(ii) *f has an isolated local maximum in x_0, if $\det(H_f(x_0)) > 0$ and $f_{xx}(x_0) < 0$*

(iii) *f has an isolated local minimum in x_0, if $\det(H_f(x_0)) > 0$ and $f_{xx}(x_0) > 0$*

After this exhausting procedure it is time for some examples.

2.6 Remarks/Examples
(i) We want to find the local extreme points of the function given by

$$f(x,y) = e^{-\left(x^2+y^2\right)}.$$

Then all stationary points are potential candidates, i.e., we calculate

$$\nabla f(x,y) = \begin{pmatrix} -2xe^{-\left(x^2+y^2\right)} \\ -2ye^{-\left(x^2+y^2\right)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} x = 0 \\ y = 0 \end{cases}$$

and obtain the only stationary point $x_0 = (0,0)$. We need the Hessian to check whether we really have an extreme point, i.e., we calculate

$$H_f(x,y) = \begin{pmatrix} f_{xx}(x,y) & f_{yx}(x,y) \\ f_{xy}(x,y) & f_{yy}(x,y) \end{pmatrix} = \begin{pmatrix} (4x^2-2)e^{-\left(x^2+y^2\right)} & 4xye^{-\left(x^2+y^2\right)} \\ 4xye^{-\left(x^2+y^2\right)} & (4y^2-2)e^{-\left(x^2+y^2\right)} \end{pmatrix}$$

and since we want to check explicitly the point $x_0 = (0,0)$, we calculate

$$H_f(0,0) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$$

By Corollary III.2.5*, it suffices to determine the determinant $\det(H_f(0,0))$ and the entry $f_{xx}(0,0)$ of this matrix. We see that*

$$\det(H_f(0,0)) = (-2)^2 - 0 = 4 > 0,$$

$$f_{xx}(0,0) = -1 < 0,$$

i.e., f has a local maximum in $x_0 = (0,0)$ with corresponding function value $f(0,0) = 1$.

By the way, this is even the global maximum. A quick look at the function rule reveals: for all points $(x, y) \neq (0, 0)$ *the exponent is negative and so all function values* $f(x, y)$ *are lower than the one in the origin:*

$$f(x, y) \leq f(0, 0) \quad \forall (x, y) \in \mathbb{R}^2.$$

This is verified by the graph of f *illustrated in* III.5.

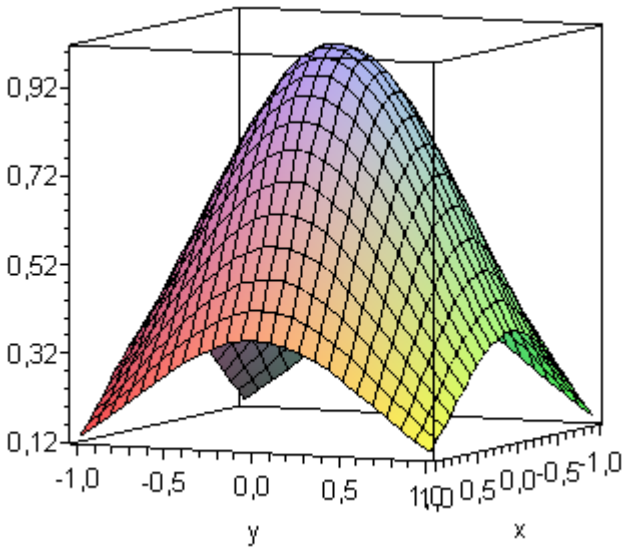


Figure III.5: The graph of the function f

(ii) *It is also relatively easy to locate a saddle point* x_0 *for a two-variable function* f*. The criterion is that* $H_f(x_0)$ *is indefinite or, even easier,* $\det(H_f(x_0)) < 0$*.*

(iii) *We want to find the stationary points of the function given by*

$$f(x, y) = x^6 + y^6 - 3x^2 - 3y^2$$

and obtain the non-linear system

$$\nabla f(x, y) = \begin{pmatrix} 6x^5 - 6x \\ 6y^5 - 6y \end{pmatrix} = \begin{pmatrix} 6x(x^2 - 1)(x^2 + 1) \\ 6y(y^2 - 1)(y^2 + 1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We see that $x = 0, -1, 1$ *and* $y = 0, -1, 1$ *are possible solutions of this equation. Combining these values we get the stationary points*

$$\begin{array}{lll} x_0 = (0, 0), & x_1 = (0, 1), & x_2 = (0, -1), \\ x_3 = (1, 0), & x_4 = (1, 1), & x_5 = (1, -1), \\ x_6 = (-1, 0), & x_7 = (-1, 1), & x_8 = (-1, -1). \end{array}$$

To determine where we have local extreme points, we calculate the Hessian

$$H_f(x,y) = \begin{pmatrix} f_{xx}(x,y) & f_{yx}(x,y) \\ f_{xy}(x,y) & f_{yy}(x,y) \end{pmatrix} = \begin{pmatrix} 30x^4 - 6 & 0 \\ 0 & 30y^4 - 6 \end{pmatrix}$$

and evaluate it together with its determinant for the stationary points x_i. Where necessary, we consider the top left entry $f_{xx}(x_i)$ to distinguish between a local maximum or minimum.

x_0: $H_f(x_0) = \begin{pmatrix} -6 & 0 \\ 0 & -6 \end{pmatrix}$, $\det(H_f(x_0)) = 36$ *and* $f_{xx}(0,0) = -6 < 0$
$\Longrightarrow$ *local maximum in* x_0.

x_1: $H_f(x_1) = \begin{pmatrix} -6 & 0 \\ 0 & 24 \end{pmatrix}$, $\det(H_f(x_1)) = -144 < 0$
$\Longrightarrow$ *saddle point in* x_1. *The same holds for* x_2, x_3 *and* x_6.

x_4: $H_f(x_4) = \begin{pmatrix} 24 & 0 \\ 0 & 24 \end{pmatrix}$, $\det(H_f(x_4)) > 0$ *and* $f_{xx}(1,1) = 24 > 0$
$\Longrightarrow$ *local minimum* x_4. *The same holds for* x_5, x_7 *and* x_8.

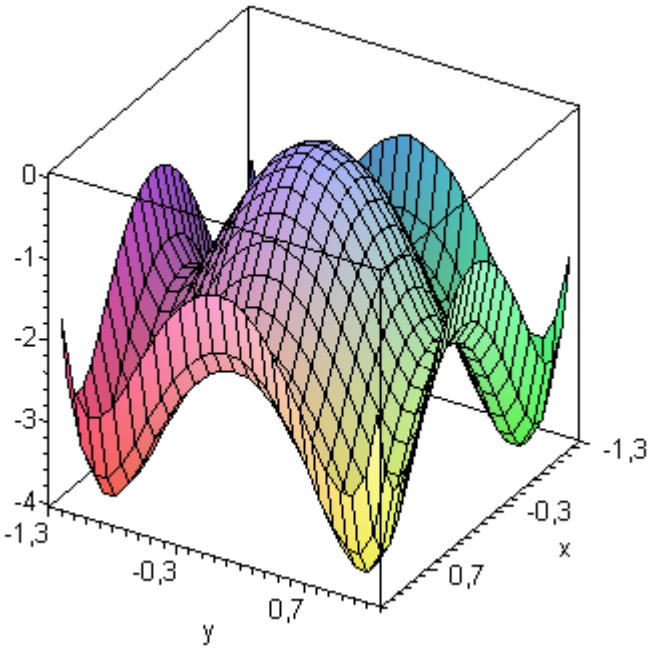


Figure III.6: The graph of the function f

So far, we have only tried to find a combination $(x_1, \ldots, x_n) \in \mathbb{D}$ that maximizes or minimizes a given function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, without taking care of dependencies that typically exist between the independent variables. In the next section, these interdependencies (also called constraints) will be covered.

3 Constrained optimization

Consider the example of maximizing the profit of a manufacturer, who has a given amount of money m available to fulfill this task. Say he can invest this money m either on advertising (variable x_1), on the employees wages (variable x_2) or on new machines (variable x_3). Then every combination he chooses to maximize the profit should naturally link these two quantities by the equation

$$x_1 + x_2 + x_3 = m.$$

In particular, money that is spent on advertising can't be used for wages anymore! We are confronted with a **constrained optimization problem**.
In this case we have to maximize the profit, e.g., given by the function

$$p(x, y) = \frac{1}{3}x_1^2 + \ln(x_2) + \sqrt{x_3}.$$

under the side constraint

$$x_1 + x_2 + x_3 = m. \tag{3.1}$$

One method for solving such problems is to simply solve Equation (3.1) for x_3 and inserted the result in $p(\cdot)$, thereby eliminating x_3. Then minimizing the new function

$$\tilde{p}(x_1, x_2) = \frac{1}{3}x_1^2 + \ln(x_2) + \sqrt{m - x_1 - x_2},$$

that is now depending only on the variables x_1 and x_2 is possible by unconstrained optimization techniques presented in Section III.2. The constraint (3.1) was implicitly included.

The problems with this approach are the following:

(i) We must be able to uniquely solve the constraint for one variable. This is in general not possible if the variables appear in a non-linear relation like $x^2+y\sin(x\ln(y)) = 1$.

(ii) How shall we incorporate multiple constraints?

The **Lagrange multiplier method** provides an elegant and powerful approach to tackle such problems overcoming the named difficulties and will be subject to this section.

3.1 Definition
For $f : \mathbb{R}^n \to \mathbb{R}$, an integer $k < n$ and $c_i \in \mathbb{R}$ $(i = 1, \dots k)$, the problem defined by

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{or} \quad \max_{x\in\mathbb{R}^n} f(x) \qquad \text{such that} \qquad g_i(x) = c_i, \quad \text{for} \quad i = 1, \dots, k$$

*is referred to as an **equality-constrained optimization problem** or a minimization/maximization problem under equality constraints.*

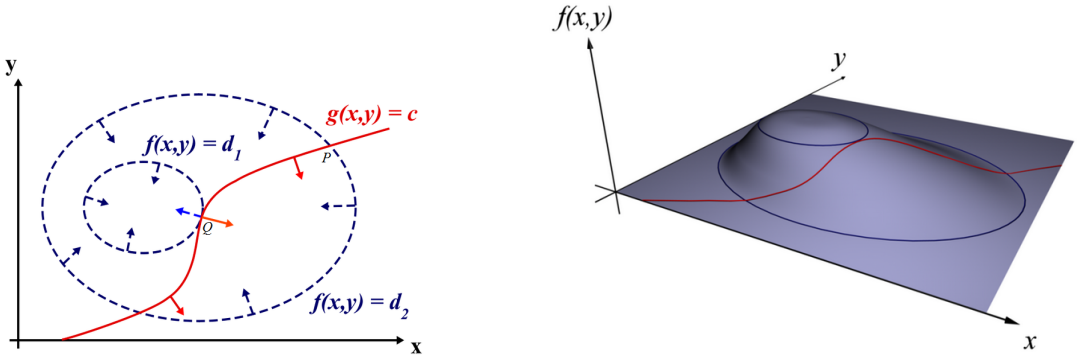
We motivate the Lagrange multiplier method, the main method of this section, in the following discussion.

3.2 Discussion

(i) *Consider first, the following equality-constrained optimization problem*

$$\text{(3.2)} \qquad \begin{cases} \max\limits_{(x,y)\in\mathbb{R}^2} f(x,y) \\ \text{such that} \quad g(x,y) = c \end{cases}$$

for $f, g \in \mathcal{C}^1(\mathbb{R}^2)$. *In* Figure III.7(a) *we have visualized two level sets of* f *for* d_1 *and* d_2. *The blue gradient arrows shall indicate that the function values increase towards the interior. We also included the level set of* g *corresponding to the value* c.



(a) Contour lines of f and g (b) Contour lines in 3-d

Figure III.7: Max. of f under equality g

Since a solution to Problem (3.2) *only allows pairs* (x, y) *that satisfy* $g(x, y) = c$, *only points* (x, y) *on this level set of* g *are feasible points for our maximization. What are the characteristics of a maximum point* (x_0, y_0)*?*

(a) *We must be in a feasible point, i.e.,* $g(x_0, y_0) = c$.

(b) *When we change our position locally in* $L_c(g)$*(!), then we can't increase the function value of* f *anymore.*

The second condition can be reformulated: For any vector $v \in \mathbb{R}^2$ *towards which we might vary our position, we must stay in* $L_c(g)$, *i.e., the directional derivative in that direction must vanish:*

$$\partial_v g(x_0, y_0) = \langle \nabla g(x_0, y_0),\ v\rangle = 0.$$

In a local maximum, a change in such a direction does not increase the function values of f, *i.e., the directional derivative of* f *also vanishes. This means that*

$$\langle \nabla g(x_0, y_0),\ v\rangle = 0 \quad \Longrightarrow \quad \partial_v f(x_0, y_0) = \langle \nabla f(x_0, y_0),\ v\rangle = 0,$$

which can also be stated as

$$\text{(3.3)} \qquad \nabla f(x_0, y_0) = -\lambda \nabla g(x_0, y_0) \iff \nabla f(x_0, y_0) + \lambda \nabla g(x_0, y_0) = \mathbb{0}.$$

The multiplier λ is required, since the length and orientation of ∇f and ∇g may be different (they are only in parallel).

Now assume we are walking along the contour line $L_c(g)$ starting from the top right. If we arrive at P we observe that continuing to walk on $L_c(g)$ increases the function values of f, since we are moving on towards the interior. In deed, this is confirmed by our Criterion (3.3)*, since ∇g and ∇f are by far not in parallel here. The situation does not change until we arrive at Q, which we only "touch".*
If we continue to walk along the path from here on, we will reach a lower level set again. This is also what our Criterion (3.3) *is telling us: it is satisfied here!*

(ii) Lagrange's approach tells us now how to combine items (a) and (b) from above. Form the Lagrange function

$$\mathcal{L}(x, y, \lambda) := f(x, y) + \lambda \left(g(x, y) - c \right).$$

Its stationary points are precisely the ones satisfying item (a) and (b) from above, since such a stationary point is characterized by

$$\left.\begin{array}{l} \frac{\partial \mathcal{L}}{\partial x}(x, y, \lambda) = f_x(x, y) + \lambda g_x(x, y) \\ \frac{\partial \mathcal{L}}{\partial y}(x, y, \lambda) = f_y(x, y) + \lambda g_y(x, y) \end{array}\right\} = \nabla f(x, y) + \lambda \nabla g(x, y) = \mathbb{0}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda}(x, y, \lambda) = g(x, y) - c = 0.$$

The generalization of the above discussion to multiple equality constraints can be done in complete analogy, we only need one multiplier per equality constraint. This leads to the following algorithm.

3.3 Algorithm (Lagrange multiplier method)
To solve an equality-constrained optimization problem for $\mathcal{C}^1(\mathbb{R}^n)$-functions $f, g_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 1, \ldots, k)$, proceed as follows

(i) *Form the* ***Lagrangian (function)*** *by*

$$\mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k) := f(x_1, \ldots, x_n) + \sum_{j=1}^{k} \lambda_j \left(g_j(x_1, \ldots, x_n) - c_j \right).$$

The new auxiliary variables $\lambda_j \in \mathbb{R}$ are called ***Lagrange multipliers****.*

(ii) *Setup the gradient of $\mathcal{L}$ by*

$$\nabla \mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k) = \begin{pmatrix} \nabla f(x_1, \ldots, x_n) + \sum_{j=1}^{k} \lambda_j \nabla g_j(x_1, \ldots, x_n) \\ (g_1(x_1, \ldots, x_n) - c_1) \\ \vdots \\ (g_k(x_1, \ldots, x_n) - c_k) \end{pmatrix}$$

(iii) *Determine the stationary points of* $\mathcal{L}$ *by solving the non-linear system*

$$\begin{aligned} \nabla f(x_1,\dots,x_n) + \sum_{j=1}^{k} \lambda_j \nabla g_j(x_1,\dots,x_n) &= \mathbb{0} && (n \text{ equations}) \\ g_j(x_1,\dots,x_n) - c_j &= 0 && (j = 1,\dots,k.) \end{aligned}$$

containing $n+k$ *equations and* $n+k$ *variables.*

(iv) *The vector* $x \in \mathbb{R}^n$ *formed by the first coordinates of a stationary point is a potential candidate for the solution of the original equality constrained optimization problem.*

3.4 Remarks/Examples

(i) We want to solve the constrained maximization problem

$$\max_{(x,y)\in\mathbb{R}^2} (x+y) \quad \textit{such that} \quad x^2+y^2 = 1$$

depicted in Figure III.8 *and apply the Lagrangian method:*

1. The Lagrangian function is

$$\mathcal{L}(x,y,\lambda) = x + y + \lambda\left(x^2+y^2-1\right).$$

2. We calculate the gradient of $\mathcal{L}$ *and set it to zero:*

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x}(x,y,\lambda) &= 1 + 2\lambda x &&= 0 \\ \frac{\partial \mathcal{L}}{\partial y}(x,y,\lambda) &= 1 + 2\lambda y &&= 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda}(x,y,\lambda) &= x^2 + y^2 - 1 &&= 0 \end{aligned}$$

3. To determine the stationary points we solve the above nonlinear system. Note that we cannot apply the Gauss- algorithm, since the system in the unknowns (x,y,λ) *is non-linear.*

The first equation yields $x = -\frac{1}{2\lambda}$ *(*$\lambda \neq 0$*, since equation* (I) *would then read as* $1+0=0$*). In the same way, we arrive at* $y = -\frac{1}{2\lambda}$ *from equation* (II)*. Inserting this in equation* (III) *yields the Lagrange multipliers*

$$\begin{aligned} & x^2+y^2-1 = \left(-\frac{1}{2\lambda}\right)^2 + \left(-\frac{1}{2\lambda}\right)^2 - 1 = \frac{1}{2\lambda^2} - 1 = 0 \\ \iff \quad & \lambda_{1/2} = \pm\frac{1}{\sqrt{2}}. \end{aligned}$$

Inserting λ_1 *in equations* (I) *and* (II) *yields the stationary point* $\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$*, while inserting* λ_2 *yields* $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$.

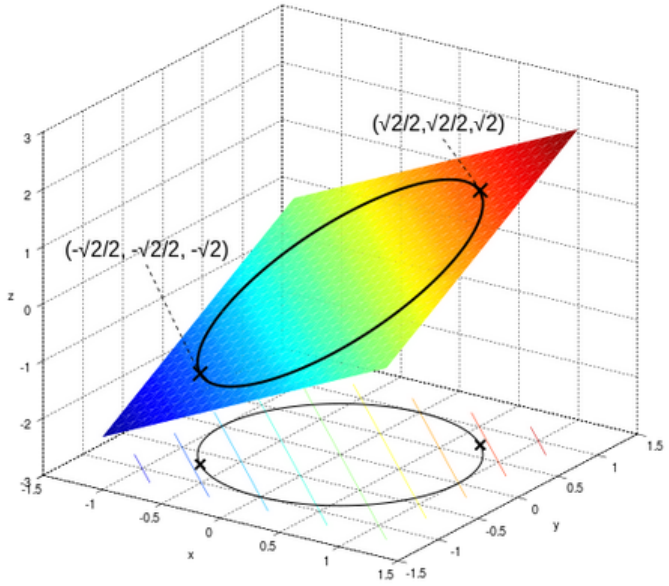


Figure III.8: The feasible set and f

Inserting the point $\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ *now into* f *yields a function value of* $-\sqrt{2}$*, while the point* $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ *obtains a function value of* $\sqrt{2}$*. We can observe in the plot that the second point solves our maximization problem.*

(ii) *Assume that we want to illuminate a famous opera singer on stage. For the solo part she will stand at point A of the stage and we are* 10 m *in front of the stage at height h creating the spotlight from a light source L (compare* Figure III.9*).*

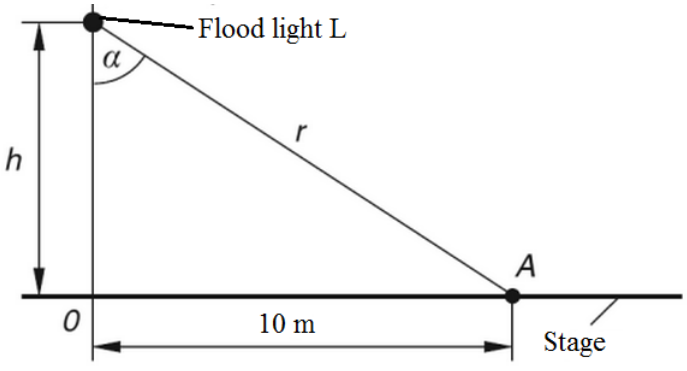


Figure III.9: Illumination of a stage

The illumination power created by L at A depends on the angle α *and the distance to the point A. It is given by Lambert's law*

$$I(\alpha, r) = \frac{I_0 \cos(\alpha)}{r^2},$$

where I_0 is the constant luminosity of L. The task is now to find the maximum illumination power in order to optimally illuminate the opera singer, when we have the possibility to vary our light source along a vertical line at 10 m distance.
First note that changing the height leads to a change of the angle α. However, the distance r is also affected by this change in height. In fact, we must always have the relation

$$\sin(\alpha) = \frac{10}{r} \quad \text{or} \quad \underbrace{r\sin(\alpha)}_{=:g(\alpha,r)} = \underbrace{10}_{=:c}.$$

So we actually have to solve the constrained optimization problem

$$\max_{(\alpha,r)\in\mathbb{R}^2} \frac{I_0\cos(\alpha)}{r^2} \qquad \textit{such that} \qquad r\sin(\alpha) = 10$$

We then apply the Lagrangian method:

1. *The Lagrangian function is*

$$\mathcal{L}(\alpha, r, \lambda) = \frac{I_0\cos(\alpha)}{r^2} + \lambda\left(r\sin(\alpha) - 10\right).$$

2. *We calculate the gradient of $\mathcal{L}$ and set it to zero:*

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\alpha}(\alpha, r, \lambda) &= -\frac{I_0\sin(\alpha)}{r^2} + \lambda r\cos(\alpha) &&= 0\\
\frac{\partial\mathcal{L}}{\partial r}(\alpha, r, \lambda) &= -\frac{2I_0\cos(\alpha)}{r^3} + \lambda\sin(\alpha) &&= 0\\
\frac{\partial\mathcal{L}}{\partial\lambda}(\alpha, r, \lambda) &= r\sin(\alpha) - 10 &&= 0
\end{aligned}$$

3. *To determine the stationary points we solve the above nonlinear system. The first and second equations yield*

$$\lambda = \frac{I_0\sin(\alpha)}{r^3\cos(\alpha)} \quad \textit{and} \quad \lambda = \frac{2I_0\cos(\alpha)}{r^3\sin(\alpha)}.$$

Eliminating λ gives

$$\frac{I_0}{r^3}\tan(\alpha) = \frac{2I_0}{r^3}\cdot\frac{1}{\tan(\alpha)} \quad \Longleftrightarrow \quad \tan(\alpha) = \pm\sqrt{2}$$

Since $\alpha \geq 0$, we only have to solve

$$\alpha = \arctan(\sqrt{2}) \approx 0.955 = 54.74^\circ,$$

For this angle, we obtain

$$r = \frac{10}{\sin(0.955)} \approx 12.25 \quad \textit{and} \quad h = r\cos(\alpha) \approx 7.07.$$

In conclusion the optimum illumination is obtained at approximately 7 m *above the ground.*

(iii) Please note that the derivative of the Lagrangian with respect to a multiplier λ_j *always results in* $g_j(x_1,\dots,x_n) - c_j$*, so you don't really have to calculate these derivatives. Moreover, since this is set to zero to find the stationary points of* $\mathcal{L}$ *we obtain the original constraint* $g_j(x_1,\dots,x_n) = c_j$*.*

(iv) Please note that a solution of a constrained optimization problem is a stationary point of the Lagrangian, but it need not be an extreme point of the Lagrangian.

The Lagrange method only yields potential candidates for the solution of the constrained optimization problem. More precisely, any local extreme point of the constrained optimization problem must necessarily be a stationary point of $\mathcal{L}$. However, the criterion is not sufficient. Sufficient criteria also exist, a very easy one is the following.

3.5 Proposition
Assume that $(\bar{x}_1,\dots,\bar{x}_n,\bar{\lambda}_1,\dots,\bar{\lambda}_k)$ *is a stationary point of the Lagrangian function* $\mathcal{L}$ *of to the constrained optimization problem given by* $f, g_j : \mathbb{R}^n \to \mathbb{R}$ *(*$j = 1,\dots,k$*).*
Also define the ***modified Lagrangian*** *by*

$$\bar{\mathcal{L}}(x_1,\dots,x_n) := \mathcal{L}(x_1,\dots,x_n,\bar{\lambda}_1,\dots,\bar{\lambda}_k).$$

Then, the point $(\bar{x}_1,\dots,\bar{x}_n)$ *is a*

(i) *minimum of the original constrained optimization problem, if the Hessian of the modified Lagrangian at* $(\bar{x}_1,\dots,\bar{x}_n)$ *is positive definite, i.e.,* $H_{\bar{\mathcal{L}}}(\bar{x}_1,\dots,\bar{x}_n)$ *is positive definite.*

(ii) *maximum of the original constrained optimization problem, if* $H_{\bar{\mathcal{L}}}(\bar{x}_1,\dots,\bar{x}_n)$ *is negative definite.*

3.6 Example
We check whether the criterion of Proposition III3.5 *is conclusive for* Example III3.4(i)*. Therefore, we calculate the modified Lagrangian*

$$\bar{\mathcal{L}}(x,y) = x + y + \bar{\lambda}(x^2 + y^2 - 1)$$

where $\bar{\lambda}_{1/2} = \pm\frac{1}{\sqrt{2}}$ *together with its Hessian*

$$H_{\bar{\mathcal{L}}}(x,y) = \begin{pmatrix} 2\bar{\lambda} & 0 \\ 0 & 2\bar{\lambda} \end{pmatrix}.$$

If we evaluate the stationary point $\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)$*, we derive the Hessian*

$$H_{\bar{\mathcal{L}}}(x,y) = \begin{pmatrix} -2\frac{1}{\sqrt{2}} & 0 \\ 0 & -2\frac{1}{\sqrt{2}} \end{pmatrix},$$

which is of course negative definite (both eigenvalues are negative and can be read from the diagonal). Therefore, the original system has a local maximum here.
If we evaluate the stationary point $\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$*, we derive the Hessian*

$$H_{\bar{\mathcal{L}}}(x,y) = \begin{pmatrix} 2\frac{1}{\sqrt{2}} & 0 \\ 0 & 2\frac{1}{\sqrt{2}} \end{pmatrix},$$

which is positive definite, i.e., the original system has a local minimum here.

Chapter IV

Vector-valued functions

1 Introduction

So far, we have only considered functions with one-dimensional range, i.e.,

$$f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^1$$

In this chapter we generalize the main concepts continuity and differentiability for **vector-valued functions** (also called **vector fields** if $n = m$)

$$f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^m, \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}.$$

We start with some examples:

(i) If a manufacturer produces two products, his profit shall be given by $f = (f_1, f_2)$, where the coordinate f_1 in the vector f represents the profit due to product 1 and f_2 represents the profit due to product 2. Using the notation

p_1 prize for product 1,
s_1 sold amount of product 1,
p_2 prize for product 2,
s_2 sold amount of product 2,

we could have the profit functions

$$f_1(p_1, s_1, p_2, s_2) = 5s_1 - 4p_1 - 6 - s_2$$
$$f_2(p_1, s_1, p_2, s_2) = 7s_2 - 2p_2 - 4 - s_1$$

which can be summarized in a vector-valued function

$$f : \mathbb{R}^4 \to \mathbb{R}^2, \quad \begin{pmatrix} p_1 \\ s_1 \\ p_2 \\ s_2 \end{pmatrix} \mapsto \begin{pmatrix} 5s_1 - 4p_1 - 6 - s_2 \\ 7s_2 - 2p_2 - 4 - s_1 \end{pmatrix}$$

(ii) We already saw a very important example of a vector-valued function. For a partially differentiable function

$$f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^1$$

we defined the gradient at a position x, $\nabla f(x)$. If we evaluate it at x it becomes a vector in $\mathbb{R}^n$ and since we can vary $x \in \mathbb{D}$, it is indeed a vector-valued function

$$\nabla f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n, \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \nabla f(x_1, \ldots, x_n) = \begin{pmatrix} f_{x_1}(x_1, \ldots, x_n) \\ \vdots \\ f_{x_n}(x_1, \ldots, x_n) \end{pmatrix}.$$

(iii) Parameterized curves in $\mathbb{R}^2$ or $\mathbb{R}^3$ are given by vector-valued functions with a one-dimensional domain. Some examples are the **Euler spiral** in $\mathbb{R}^2$ (compare Figure IV.1 (left))

$$f : \mathbb{R}_+ \to \mathbb{R}^2, \quad t \mapsto f(t) = \begin{pmatrix} \int_0^t \cos(s^2)\, ds \\ \int_0^t \sin(s^2)\, ds \end{pmatrix},$$

a spiral in $\mathbb{R}^3$ (compare Figure IV.1 (middle))

$$g : \mathbb{R}_+ \to \mathbb{R}^3, \quad t \mapsto g(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \\ t \end{pmatrix}$$

or the figure eight shaped **Viviani's curve**. This curve illustrated in Figure IV.1 (right) is the intersection of a sphere of radius $r = 2a$, i.e,

$$x^2 + y^2 + z^2 = 4a^2$$

with a cylinder of radius a centered at $P(a, 0, 0)$ given by

$$(x - a)^2 + y^2 = a^2$$

that is tangent to the sphere and that passes through the center of the sphere. The resulting curve of intersection V can be parameterized by t to give the parametric equation of Viviani's curve:

$$V : [0, 2\pi] \to \mathbb{R}^3, \quad t \mapsto V(t) = \begin{pmatrix} a(1 + \cos(t)) \\ a \sin^2(t) \\ 2a \sin(\frac{t}{2}) \end{pmatrix}.$$

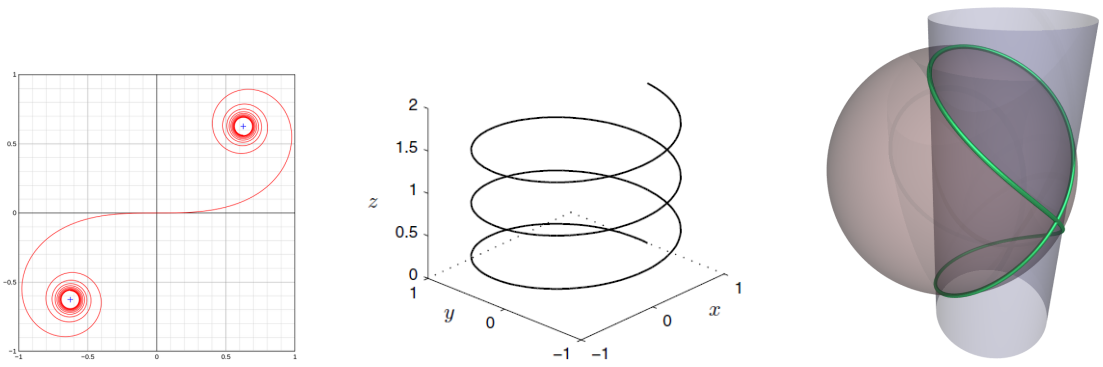


Figure IV.1: Euler spiral in $\mathbb{R}^2$ (left), spiral in $\mathbb{R}^3$ (middle) and Viviani's curve (right)

2 Continuity & differentiation of vector-valued functions

In our generalization of the main notions to vector-valued functions, we start with continuity.

2.1 Definition
A function

$$f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^m, \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}.$$

is called ***continuous*** *in $x_0 \in \mathbb{D}$, if every coordinate f_j is continuous in x_0.*

2.2 Remarks/Examples

(i) *By* Theorem I.3.5*, we can express the definition of continuity for a vector-valued function as follows. For all sequences $(x_n)_{n\in\mathbb{N}}$ with*

$$\|x_n - x\|_{\mathbb{R}^n} \to 0$$

we have

$$\|f(x_n) - f(x)\|_{\mathbb{R}^m} \to 0.$$

If we recall the definition of continuity of a single variable function $f : \mathbb{R} \to \mathbb{R}$, i.e., for all sequences $(x_n)_{n\in\mathbb{N}}$ with

$$|x_n - x| \to 0$$

we have

$$|f(x_n) - f(x)| \to 0.$$

we see that in the multidimensional case the absolute values have simply been replaced by norms.

(ii) *The vector-valued function*

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto f(x, y) = \begin{pmatrix} f_1(x, y) = x^2 + y^2 - 4 \\ f_2(x, y) = xy - 1 \end{pmatrix}$$

is continuous since both coordinates are polynomials and therefore continuous.

Next we treat partial differentiability.

2.3 Definition
Given a vector-valued function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $x_0 \in \mathbb{D}$.

(i) *f is called* ***partially differentiable*** *at x, if all partial derivatives of the coordinates f_j $(j = 1, \ldots, m)$ of f exist at x.*

(ii) *If f is partially differentiable at x, the $m \times n$ matrix*

$$J_f(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}$$

is called the ***Jacobian matrix*** *of f.*

(iii) *f is called k* ***times continuously differentiable****, denoted $f \in \mathcal{C}^k(\mathbb{D}, \mathbb{R}^m)$, if f and all partial derivatives of order k exist and are continuous.*

(iv) *f is called* ***totally differentiable****, if all coordinate functions f_j are totally differentiable.*

This generalizes the notion of the one-dimensional range case naturally. We consider some examples.

2.4 Remarks/Examples

(i) Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto f(x,y) = \begin{pmatrix} f_1(x,y) = x^2 + y^2 - 4 \\ f_2(x,y) = xy - 1 \end{pmatrix}$$

again. Since both coordinates are polynomials, they are partially differentiable and so f is partially differentiable everywhere. We calculate the Jacobian as

$$J_f(x,y) = \begin{pmatrix} 2x & 2y \\ y & x \end{pmatrix}$$

and see that $f \in \mathcal{C}^1(\mathbb{R}^2, \mathbb{R}^2)$, since all coordinates of the Jacobian are continuous functions as well.

(ii) The transformation of polar coordinates (r, φ) (describing a point by the distance r to the origin and the angle it builds with the x-axis) to cartesian coordinates (x, y) is given by

$$f : \mathbb{R}_{\geq 0} \times [0, 2\pi) \to \mathbb{R}^2, \ \begin{pmatrix} r \\ \varphi \end{pmatrix} \mapsto f(r,\varphi) = \begin{pmatrix} x(r,\varphi) = r\cos(\varphi) \\ y(r,\varphi) = r\sin(\varphi) \end{pmatrix}.$$

Since this transformation is partially differentiable everywhere again, we calculate the Jacobian as

$$J_f(r,\varphi) = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{pmatrix}$$

and see that $f \in \mathcal{C}^1(\mathbb{R}^2, \mathbb{R}^2)$, since all coordinates of the Jacobian are continuous functions as well.

We want to reformulate total differentiability using the Jacobian introduced above. This is motivated in the following discussion.

2.5 Discussion (Total differentiability)
A two-variable function

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto f(x,y) = \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix}$$

is totally differentiable, if f_1 and f_2 are totally differentiable in (x_0, y_0), i.e., if we have

$$\lim_{(x,y)\to(x_0,y_0)} \frac{1}{\|(x,y)-(x_0,y_0)\|_2} \cdot \begin{pmatrix} f_1(x,y) - \left[f_1(x_0,y_0) + \underbrace{\left(f_{1x}(x_0,y_0), f_{1y}(x_0,y_0)\right)}_{J_{f_1}(x_0,y_0)} \cdot (x-x_0, y-y_0)^\top \right] \\ f_2(x,y) - \left[f_2(x_0,y_0) + \underbrace{\left(f_{2x}(x_0,y_0), f_{2y}(x_0,y_0)\right)}_{J_{f_2}(x_0,y_0)} \cdot (x-x_0, y-y_0)^\top \right] \end{pmatrix} = \mathbb{0}.$$

Using vector-valued notation we can rewrite this using the Jacobian matrix of f by

$$\lim_{(x,y)\to(x_0,y_0)} \frac{1}{\|(x,y)-(x_0,y_0)\|_2} \cdot \left\| \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix} - \left[\begin{pmatrix} f_1(x_0,y_0) \\ f_2(x_0,y_0) \end{pmatrix} + \underbrace{\begin{pmatrix} f_{1x}(x_0,y_0) & f_{1y}(x_0,y_0) \\ f_{2x}(x_0,y_0) & f_{2y}(x_0,y_0) \end{pmatrix}}_{=J_f} \cdot \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} \right] \right\|_2 = 0$$

In general terms for $f : \mathbb{R}^n \to \mathbb{R}^m$ we can say, that f is totally differentiable at x_0 if

$$\lim_{x\to x_0} \frac{\|f(x) - [f(x_0) - J_f(x_0)(x-x_0)]\|_2}{\|(x,y)-(x_0,y_0)\|_2} = 0$$

and call the Jacobian matrix the (total) derivative (or total differential) of f like in the one-dimensional range situation:

$$f'(x,y) = J_f(x,y) = \begin{pmatrix} f_{1x}(x,y) & f_{1y}(x,y) \\ f_{2x}(x,y) & f_{2y}(x,y) \end{pmatrix}.$$

In the last section of this chapter differential operators like the already introduced gradient will be presented.

3 Differential operations

Vector calculus studies various differential operators defined on scalar or vector fields, which are typically expressed in terms of the already introduced "nabla"-operator utilized for the definition of the gradient. Further important differential operations in vector calculus are the *Divergence*, the *Curl* and the *Laplacian*.

3.1 Definition
Let x, y, z be a system of Cartesian coordinates in 3-dimensional Euclidean space, and let i, j, k be the corresponding basis of unit vectors. The ***divergence*** *of a continuously differentiable vector field $F = Ui + Vj + Wk$ is defined as the scalar-valued function:*

$$div\, F = \nabla \cdot F = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) \cdot (U, V, W) = \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} + \frac{\partial W}{\partial z}.$$

3.2 Proposition
The following properties can all be derived from the ordinary differentiation rules of calculus.

(i) *Most importantly, the divergence is a* ***linear operator****, i.e.,*

$$div(a \cdot F + b \cdot G) = div(a \cdot F) + div(b \cdot G)$$

for all vector fields F and G and all $a, b \in \mathbb{R}$.

(ii) *There is a* ***product rule*** *of the following type: if λ is a scalar valued function and F is a vector field, then*

$$div(\lambda F) = grad(\lambda) \cdot F + \lambda \cdot div(F)$$

or in more suggestive notation

$$\nabla(\lambda F) = (\nabla\lambda) \cdot F + \lambda \cdot (\nabla \cdot F).$$

3.3 Remarks/Examples

(i) In physical terms, the divergence of a three-dimensional vector field is the extent to which the vector field flow behaves like a source at a given point. If the divergence is positive at some point then there must be a source, if it is negative there must be a sink at that position.
Note that we are imagining the vector field to be like the velocity vector field of a fluid in motion when we use the terms flow, source and so on.

(ii) The intuition that the sum of all sources minus the sum of all sinks should give the net flow outwards of a region is made by the divergence theorem we will discuss later on.

(iii) *A generalization to the more dimensional case is simple. If $F = (F_1, F_2, \ldots, F_n)$ is an n-dimensional vector field in the Euclidian coordinate system, then*

$$div(F) = \nabla \cdot F = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + \ldots \frac{\partial F_n}{\partial x_n}.$$

(iv) *For $F(x, y, z) = (xy,\ \ xz,\ \ x^2yz^2)$ we calculate $div(F(x, y, z))$ and $div(F(1, 2, 3))$:*

$$div(F(x, y, z)) = \frac{\partial xy}{\partial x} + \frac{\partial xz}{\partial y} + \frac{\partial x^2yz^2}{\partial z} = y + 0 + 2x^2yz = y \cdot (1 + 2x^2z)$$

$$div(F(1, 2, 3)) = 2 \cdot (1 + 2 \cdot 1^2 \cdot 3) = 14.$$

Since at $P(1, 2, 3)$ the divergence is positive, there is a source at this point.

(v) *If $div(F) = 0$ for all vectors $F \in G$, then the environment G is called* ***solenoidal****.*

3.4 Definition
Let F: $\mathbb{R}^3 \to \mathbb{R}^3$ be a vector field, i.e., $F = (F_1(x, y, z),\ \ F_2(x, y, z),\ \ F_3(x, y, z))$, then

$$curl(F) = \nabla \times F = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z},\ \ \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x},\ \ \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right)$$

is called the ***curl*** *of the vector field.*

3.5 Proposition
The following properties can all be derived from the ordinary differentiation rules of calculus.

(i) *Most importantly, the curl is a linear operator, i.e.*

$$curl(a \cdot F + b \cdot G) = curl(a \cdot F) + curl(b \cdot G)$$

for all vector fields F and G and all $a, b \in \mathbb{R}$.

(ii) *There is a product rule of the following type: if λ is a scalar valued function and F is a vector field, then*

$$curl(\lambda F) = grad(\lambda) \times F + \lambda curl(F)$$

3.6 Remarks/Examples

(i) *In vector calculus, the curl is a vector operator that describes the infinitesimal rotation of a 3-dimensional vector field. At every point in the field, the curl of that point is represented by a vector. The attributes of this vector like the length and the direction characterize the rotation at that point.*

(ii) *The direction of the curl is the axis of rotation, as determined by the right-hand rule, and the magnitude of the curl is the magnitude of rotation.*

(iii) *A vector field whose curl is zero is called irrotational.*

(iv) *The curl is a form of differentiation for vector fields. The corresponding form of the fundamental theorem of calculus is* ***Stokes' theorem****, we will discuss later on. It relates the surface integral of the curl of a vector field to the line integral of the vector field around the boundary curve.*

(v) *An alternative terminology for the curl is* ***rotor*** *or* ***rotational*** *denoted by* $rot(F)$.

(vi) *Unlike the gradient and divergence, curl does not generalize as simply to other dimensions.*

(vii) *Considering a vector field that describes the velocity field of a fluid flow, the curl can by intuitively interpreted. Suppose a small ball is located within the fluid and the centre of the ball is fixed at a certain point. If the ball has a rough surface, the fluid flowing past it will make it rotate.*
The rotation axis oriented according to the right hand rule points in the direction of the curl of the field at the centre of the ball. The angular speed of the rotation is half the magnitude of the curl at this point.

(viii) *For* $F(x,y,z) = (xy,\ xz,\ x^2yz^2)$ *we calculate* $curl(F(x,y,z))$ *and* $curl(F(1,2,3))$*:*

$$\begin{aligned} curl(F(x,y,z)) &= \left(\frac{\partial}{\partial y}(x^2yz^2) - \frac{\partial}{\partial z}(xz),\ \frac{\partial}{\partial z}(xy) - \frac{\partial}{\partial x}(x^2yz^2),\ \frac{\partial}{\partial x}(xz) - \frac{\partial}{\partial y}(xy)\right) \\ &= (x^2z^2 - x,\ -2xyz^2,\ z - x) \end{aligned}$$

$$curl(F(1,2,3)) = (1^2 \cdot 3^2 - 1,\ -2 \cdot 1 \cdot 2 \cdot 3^2,\ 3 - 1) = (8,\ -36,\ 2)$$

We additionally calculate the magnitude of $curl(F)$ *at* $(1,2,3)$ *which is twice the magnitude of the angular speed at this point:*

$$|curl(F(1,2,3))| = |(8,\ -36,\ 2)| = \sqrt{8^2 + (-36)^2 + 2^2} \approx 36.93$$

and thus the angular speed is about 18.47.

(ix) *Consider a rigid body rotating with constant angular velocity* $\omega = (\omega_1,\ \omega_2,\ \omega_3)$ *about a fixed rotation axis. Then its track velocity* v *in a point* (x,y,z) *is defined by*

$$v = \omega \times x = (\omega_2 z - \omega_3 y,\ \omega_3 x - \omega_1 z,\ \omega_1 y - \omega_2 x)$$

and the curl of v *can be calculated:*

$$\begin{aligned} curl(v) &= \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z},\ \frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x},\ \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\right) \\ &= (\omega_1 - (-\omega_1),\ \omega_2 - (-\omega_2),\ \omega_3 - (-\omega_3)) \\ &= (2\omega_1,\ 2\omega_2,\ 2\omega_3) = 2\omega \\ \Rightarrow \omega &= \frac{1}{2} curl(v). \end{aligned}$$

3.7 Definition
If f is a twice-differentiable real-valued function, then the ***Laplacian*** *is defined by*

$$\Delta f := div(grad(f)) = \nabla^2 f = \nabla \cdot \nabla f,$$

where the latter notations derive from formally writing $\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$.

3.8 Remarks/Examples

(i) *The Laplace operator or Laplacian is a second order differential operator given by the divergence of the gradient of a function on Euclidean space.*

(ii) *The Laplacian $\Delta f(x, y, z)$ of a function f at a point $P(x_P, y_P, z_P)$ is the rate at which the average value of f over spheres centered at P deviates from $f(x_P, y_P, z_P)$ as the radius of the sphere grows.*

(iii) *In a Cartesian coordinate system, the Laplacian is given by the sum of second partial derivatives of the function with respect to each independent variable. In the* 3*-dimensional case we have*

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

In other coordinate systems such as cylindrical and spherical coordinates, the Laplacian also has a useful form.

(iv) *Note: besides the Laplacian there exists the* vector Laplacian *which will not be discussed further here.*

3.9 Survey of differential operators
In the following table the differential operators are summarized. Note that the curl and divergence differ because the former uses a cross product and the latter a dot product. Furthermore, f denotes a scalar field and F denotes a vector field.

Operation	Notation	Description	Domain/Range
Gradient	$grad(f) = \nabla f$	Measures the rate and direction of change in a scalar field.	Maps scalar fields to vector fields.
Divergence	$div(F) = \nabla \cdot F$	Measures the scalar of a source or sink at a given point in a vector field.	Maps vector fields to scalar fields.
Curl	$curl(F) = \nabla \times F$	Measures the tendency to rotate about a point in a vector field.	Maps vector fields to vector fields.
Laplacian	$\Delta f = \nabla^2 f$	Measures the difference between the value of the scalar field with its average on infinitesimal balls.	Maps between scalar fields.

Chapter V

Numerical optimization techniques

1 Introduction

In Section III.2 we discussed unconstrained optimization problems, where the objective was to find a local (isolated) extreme point $x \in \mathbb{R}^n$ of a certain function $f\colon \mathbb{R}^n \to \mathbb{R}$. These optimal points are the zeros of the vector-valued function ∇f, i.e.,

$$\nabla f(x) = \mathbb{0}$$

and the stationary points of f are the potential candidates for extreme points. Unfortunately, this leads to a non-linear system of equations, which is in general hard to solve. Remember, even in one space-dimension finding the zeros of $f(x) = e^x - \sin(x)$ is non-trivial!
In this chapter we introduce two important numerical iteration methods to minimize or maximize a multi-variable function.

Using a simple trick these techniques can then also be used to solve a non-linear system of n equations

$$h(x_1, \ldots, x_n) = \begin{pmatrix} h_1(x_1, \ldots, x_n) \\ \vdots \\ h_n(x_1, \ldots, x_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We only need to observe that a solution of the non-linear system is necessarily a minimizer of the function

$$\tilde{h} : \mathbb{R}^n \to \mathbb{R}, \;\; (x_1, \ldots, x_n) \mapsto \frac{1}{2} \left\| \begin{pmatrix} h_1(x_1, \ldots, x_n) \\ \vdots \\ h_n(x_1, \ldots, x_n) \end{pmatrix} \right\|_2^2$$

and consequently we can apply the techniques to minimize $\tilde{h}$ in order to find the zero of the non-linear system. This has two important applications:

(i) As an alternative to applying the techniques directly to find the maximum or minimum of $f\colon \mathbb{R}^n \to \mathbb{R}$, we can use the above trick to solve the non-linear system

$$\nabla f(x) = \begin{pmatrix} f_{x_1}(x_1, \ldots, x_n) \\ \vdots \\ f_{x_n}(x_1, \ldots, x_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

by minimizing the function

$$\tilde{f}(x_1, \ldots, x_n) = \frac{1}{2}\|\nabla f(x_1, \ldots, x_n)\|_2^2.$$

(ii) In constrained optimization problems, where the objective is to find a point $x \in \mathbb{R}^n$ that minimizes (or maximizes) $f\colon \mathbb{R}^n \to \mathbb{R}$ subject to certain side-constraints $g_j(x) = c_j$ $(j = 1, \ldots, k)$, we are lead to finding the stationary points of the Lagrangian, i.e., we need to solve the non-linear system

$$\nabla \mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k) = \mathbb{0}$$

in order to find candidates for our local extreme points. This non-linear system can then again be solved by minimizing the function

$$\tilde{\mathcal{L}}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k) = \frac{1}{2}\|\nabla \mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k)\|_2^2.$$

2 Steepest descent method

2.1 Remarks

(i) The problem we want to solve in this section is

$$\min_{x \in \mathbb{D}} f(x)$$

for a (totally) differentiable $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$.

(ii) Please note that a point x^* that maximizes the function f, i.e., $\max\limits_{x \in \mathbb{D}} f(x) = f(x^*)$ necessarily minimizes the function $-f$, i.e.,

$$f(x^*) = \max_{x \in \mathbb{D}} f(x) = \min_{x \in \mathbb{D}} \left(-f(x)\right).$$

Therefore, any maximization problem can be reformulated as a minimization problem.

We motivate the main method of this section in the following discussion.

2.2 Discussion

The idea of the steepest descent method is the following:

To get to a minimum point x^* from the current point x_0, always take the path of steepest descent.

Recall that for a (totally) differentiable multi-variable function $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ the gradient expression $-\nabla f(x_0)$ is orthogonal to the current level set and points in the direction of steepest descent (compare Figure V.1*)*

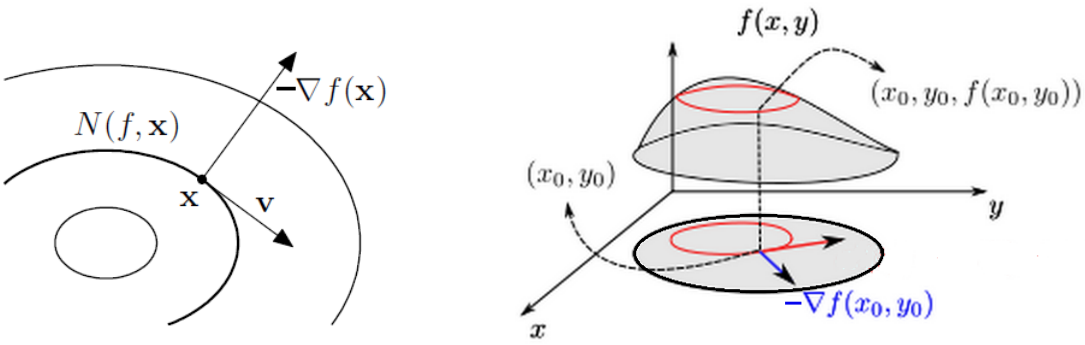


Figure V.1: The steepest descent direction

Having the mountain climber example in mind, we can argue that the fastest path down a mountain is to walk orthogonal to the height-lines on the map after each step we have taken. Such a path is illustrated in Figure V.2.

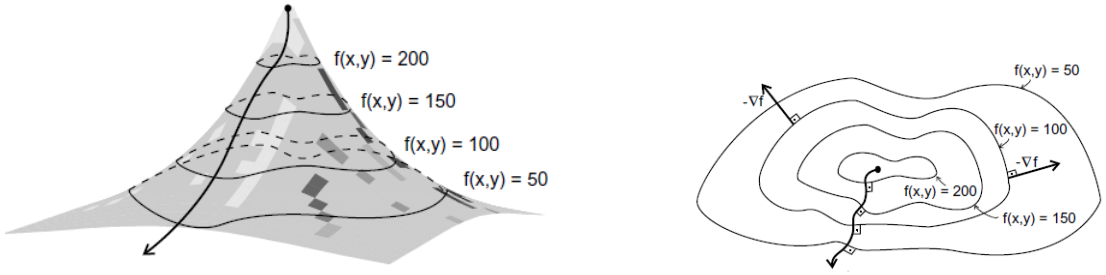


Figure V.2: Climbing down a mountain - steepest descent

The iteration procedure

$$x_{k+1} = x_k - \nabla f(x_k) \qquad k = 0, 1, \ldots \tag{2.1}$$

is the outcome of this motivation. However, note that we always take a step of a certain length from x_k to x_{k+1}. But the steepest descent occurs only locally at x_k, which means that by taking a step as long as given by Equation (2.1) *might be too long and we may have climbed upwards again. Therefore, we introduce a step-size at every step and modify iteration* (2.1) *to*

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) \qquad k = 0, 1 \ldots .$$

This leads to the following algorithm.

2.3 Algorithm (Steepest descent method)
To solve the problem $\min_{x\in\mathbb{D}} f(x)$ *for (totally) differentiable* $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, *the* ***steepest descent iteration*** *is applied as follows:*

(i) *Choose an* ***initial iterate*** $x_0 \in \mathbb{D}$ *and a* ***tolerance*** $\tau > 0$.

(ii) *Calculate the iterates* $(x_k)_{k\in\mathbb{N}}$ *by*

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) \qquad\qquad k = 0, 1, \ldots,$$

where the ***step-sizes*** *are chosen such that*

(a) $\lambda_k \in [0, 1]$ *for all* $k = 0, 1, \ldots,$

(b) $f(x_{k+1}) < (1 - \alpha) f(x_k)$ *for some small* $\alpha > 0$, *e.g.,* $\alpha = 10^{-4}$.

(iii) ***Terminate*** *the iteration, if* $\|\nabla f(x_k)\|_2 \leq \tau$.

We remark on this method and calculate an example.

2.4 Remarks/Examples
(i) We apply the steepest descent method to find the minimizer of

$$f(x, y) = 4x^2 - 4xy + 2y^2$$

using the initial guess $x_0 = (2, 3)$.

(a) So, we calculate

$$\nabla f(x, y) = \begin{pmatrix} 8x - 4y \\ 4y - 4x \end{pmatrix}$$

and obtain the current steepest descent direction $-\nabla f(2, 3) = (-4, -4)$. *The algorithm tells us to determine the next iterate by*

$$x_1 = (2,\ 3) - \lambda_0 (4,\ 4).$$

(b) We need to calculate λ_0 *such that we are really decreasing the overall function value by moving to* x_1, *i.e.,* $f(x_1) < f(x_0)$. *This is ensured by choosing* λ_0 *such that*

$$\varphi(\lambda) = f\left((2,\ 3) - \lambda(4,\ 4)\right) = f(2 - 4\lambda,\ 3 - 4\lambda)$$

is minimized, i.e., we choose the step size in the steepest descent direction such that we minimize the function value w.r.t. the step in that direction.

(b1) To find the minimum points of the single-variable function φ *we calculate the derivative by*

$$\begin{aligned}
\varphi'(\lambda) &= J_f(2 - 4\lambda,\ 3 - 4\lambda) \cdot \begin{pmatrix} (2 - 4\lambda)' \\ (3 - 4\lambda)' \end{pmatrix} \\
&= (8(2 - 4\lambda) - 4(3 - 4\lambda),\ 4(3 - 4\lambda) - 4(2 - 4\lambda)) \begin{pmatrix} -4 \\ -4 \end{pmatrix} \\
&= 64\lambda - 32.
\end{aligned}$$

using the chain rule.

(b2) By setting $\varphi'(\lambda) = 0$ and noting that $\varphi''(\lambda) = 64 > 0$) we see that the minimizer of φ is consequently $\lambda_0 = \frac{1}{2}$.

(c) The new iterate is calculated by $x_1 = (2,\ 3) - \dfrac{1}{2}(4,\ 4) = (0,\ 1)$.

(d) Further iterating yields

$$\begin{aligned} x_2 &= \left(\frac{2}{5},\ \frac{3}{5}\right) && \text{with } \lambda_1 = \frac{1}{10} \\ x_3 &= \left(0,\ \frac{2}{10}\right) \\ &\downarrow \quad k \to \infty \\ x^* &= (0,0). \end{aligned}$$

(ii) The iteration

$$x_{k+1} = x_k + \lambda_k \nabla f(x_k) \qquad\qquad k = 0, 1, \ldots,$$

with step-sizes such that

(a) $\lambda_k \in [0,1]$ *for all* $k = 0, \ldots,$

(b) $f(x_{k+1}) > (1+\alpha) f(x_k)$ *for some (small)* $\alpha > 0$, *e.g.,* $\alpha = 10^{-4}$,

is called the ***steepest ascent method*** *and can be directly applied to find maximizers of a function* f.

(iii) The method of steepest descent has one major advantage over other numerical optimization methods. It is guaranteed to converge to the global minimizer, if f *is*

(a) totally differentiable,

(b) strictly convex, i.e.,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y) \ \text{ for } \ t \in (0,1),$$

so a straight connection line between x *and* y *must always be above the function values,*

(c) coercive, i.e., $\lim\limits_{k\to\infty} \|x_k\| = \infty \Longrightarrow \lim\limits_{k\to\infty} f(x_k) = \infty$.

For example, quadratic functions

$$f(x) = a + b^\top x + \frac{1}{2} x^\top A x$$

with positive definite A *are satisfying these criteria.*

(iv) The steepest descent method is inferior in terms of convergence speed compared to, e.g., the Newton method (discussed in the next section). It also shows erratic zigzagging steps for some hard problems like the so-called Rosenbrock function

$$f(x,y) = (1-x)^2 + 100(y - x^2)^2,$$

which is designed to test the quality of numerical algorithms. The iteration steps starting at $x_0 = (-0.5, 0.5)$ *towards the global minimum* $x^* = (1, 1)$ *are depicted in* Figure V.3.

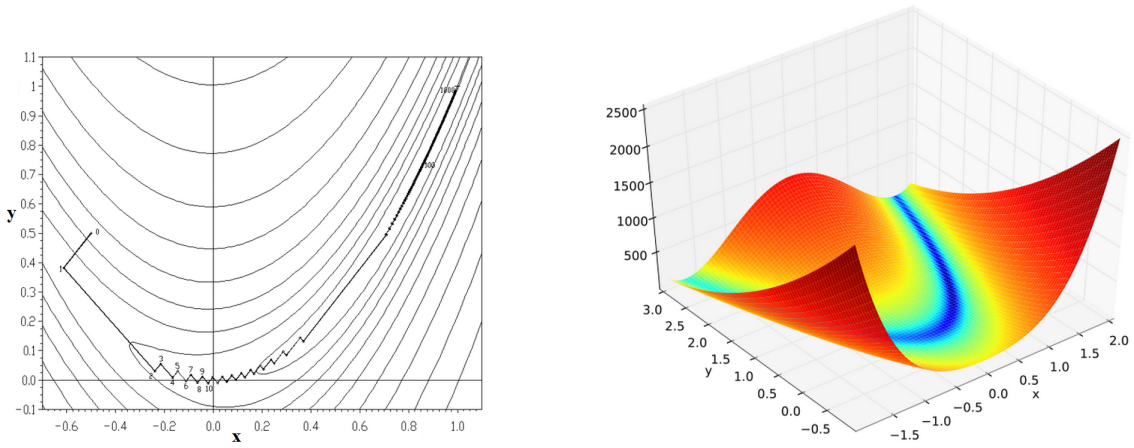


Figure V.3: Steepest descent for the Rosenbrock function

3 Newton's method

Newton's method is based on finding the minimizer of a quadratic function. For its motivation we recall that for a single variable function

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto f(x) = \frac{1}{2}a(x - x_0)^2 + b(x - x_0) + c$$

we have

$$\begin{aligned} f'(x) &= a(x - x_0) + b \\ f''(x) &= a \end{aligned}$$

and so, if $a > 0$, the (unique global) minimizer of f is at

$$f'(x^*) = 0 \iff x^* = x_0 - \frac{1}{a}b.$$

The same reasoning leads to the following Lemma in multiple dimensions.

3.1 Lemma

The unique global minimum of a quadratic function

$$f : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \frac{1}{2}(x - x_0)^\top A(x - x_0) + b^\top(x - x_0) + c$$

with positive definite, symmetric $A \in M_{n\times n}(\mathbb{R})$*, given column-vectors* $b, x_0 \in \mathbb{R}^n$ *and a constant* $c \in \mathbb{R}$ *is*

$$x^* = x_0 - A^{-1}b$$

This insight suffices to motivate Newton's method.

3.2 Discussion

The main task is to solve the problem

$$\min_{x\in\mathbb{D}} f(x)$$

for $f\colon \mathbb{D}\subseteq\mathbb{R}^n\to\mathbb{R}$, *where we assume that* $f\in\mathcal{C}^3(\mathbb{D})$. *Newton's method is an iterative scheme, where the next iterate is simply determined under the assumption that the function is a quadratic function. More precisely, for a given iterate* x_0, *we use Taylor's formula to provide a quadratic local approximation around* x_0 *(compare* III.1.4(iii)*), i.e.,*

$$f(x)\approx f(x_0)+(\nabla f(x_0))^\top\cdot(x-x_0)+\frac{1}{2}\cdot(x-x_0)^\top\cdot H_f(x_0)\cdot(x-x_0)\,.$$

Then, by Lemma 3.1, *the minimizer of this quadratic is*

$$x^*=x_0-(H_f(x_0))^{-1}\cdot\nabla f(x_0).$$

Newton's approach is now that this exact minimizer x^* *of the local quadratic approximation is a good estimate for the actual minimizer of* f, *i.e., we set* $x_1=x^*$.

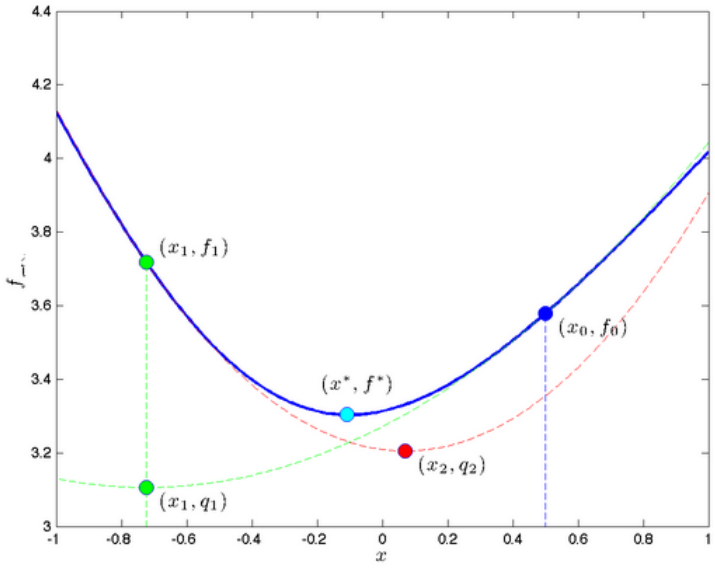


Figure V.4: Quadratic approximations for Newton steps

As an example, in Figure V.4, *we depicted two steps of the Newton iteration for*

$$f:\mathbb{R}\to\mathbb{R},\quad x\mapsto f(x)=\ln\left(e^{x-3}+e^{-2x+2}\right)$$

starting at $x_0=0.5$. *The quadratic Taylor approximation of* f *at* x_0 *is shown by the green graph and its minimizer provides the first iterate of the Newton step. Building the quadratic approximation in* x_1 *leads to the red curve and its minimizer* x_2 *gives the next Newton iterate. Further iteration will approach the point* x^* *(in light blue), which is the minimizer of* f.

The previous discussion leads to the Newton(-Raphson) method.

3.3 Algorithm (Newton's method)
To solve the problem $\min_{x\in\mathbb{D}} f(x)$ *for* $f : \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ *where* $f \in \mathcal{C}^3(\mathbb{D})$ ***Newton's method*** *is as follows:*

(i) *Choose an* ***initial iterate*** $x_0 \in \mathbb{D}$ *and a* ***tolerance*** $\tau > 0$.

(ii) *Calculate the iterates* $(x_k)_{k\in\mathbb{N}}$ *by*

$$x_{k+1} = x_k \underbrace{- (H_f(x_k))^{-1} \cdot \nabla f(x_k)}_{=:\Delta x_k = \text{Newton correction}} \qquad k = 0, 1, \ldots.$$

To do so,

(a) solve $H_f(x_k) \cdot \Delta x_k = -\nabla f(x_k)$ *by Gaussian elimination, i.e., determine the* LU*- factorization of* $H_f(x_k)$*:*

$$H_f(x_k) \cdot \Delta x_k = L \cdot \underbrace{U \cdot \Delta x_k}_{=:y} = -\nabla f(x_k)$$

(b) Solve the triangular system $L \cdot y = -\nabla f(x_k)$ *by forward substitution.*

(c) Solve the triangular system $U \cdot \Delta x_k = y$ *by backward substitution.*

(iii) *Terminate the iteration, if* $\|\nabla f(x_k)\|_2 \leq \tau$.

3.4 Remarks/Examples

(i) We calculate a simple example for illustration. Consider the function

$$f(x, y) = x^4 + 2x^2y^2 + y^4.$$

Then, we have the following gradient and (always symmetrical) Hessian

$$\nabla f(x) = \begin{pmatrix} 4x^3 + 4xy^2 \\ 4x^2y + 4y^3 \end{pmatrix},$$
$$H_f(x) = \begin{pmatrix} 12x^2 + 4y^2 & 8xy \\ 8xy & 4x^2 + 12y^2 \end{pmatrix}$$

and, using the initial iterate $(1, 1)$*,*

$$\nabla f(1, 1) = \begin{pmatrix} 8 \\ 8 \end{pmatrix}, \qquad H_f(1, 1) = \begin{pmatrix} 16 & 8 \\ 8 & 16 \end{pmatrix}.$$

The generated Newton iterate is then

$$x_1 = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{=x_0} - \underbrace{\underbrace{\frac{1}{16^2 - 8^2}}_{\frac{1}{\det H_f(1,1)}} \begin{pmatrix} 16 & -8 \\ -8 & 16 \end{pmatrix}}_{=H_f(1,1)^{-1}} \cdot \underbrace{\begin{pmatrix} 8 \\ 8 \end{pmatrix}}_{\nabla f(1,1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{2}{3} \end{pmatrix}.$$

In deed, it is easy to see that the Newton iterates are $x_k = \left(\left(\frac{2}{3}\right)^k, \left(\frac{2}{3}\right)^k\right)$*, converging to the unique global minimizer* $x^* = (0, 0)$.

(ii) *Newton's method assumes that the Hessian $H_f(x_k)$ $(k = 0, 1, \ldots)$ is always positive definite and then automatically invertible, since all eigenvalues are > 0. If the matrix is not invertible, the algorithm breaks down. If the matrix $H_f(x_k)$ $(k = 0, 1, \ldots)$ is always negative definite, we will actually maximize the objective function.*

(iii) *The convergence of the method is typically quadratic, i.e., the number of correct digits doubles in each iteration and therefore faster than the steepest descent method, where the convergence is only linear. However, this is only the case when we are reasonably close to the optimal point. In practice, the choice of an initial iterate in the "quadratic convergence area" can be achieved because people know typical values for the results. However, a proper choice of initial values is an important problem in applications.*

(iv) *If the initial iterate is far from the solution, the steps taken by the Newton method may be inaccurately large and, in particular, they may not reduce the objective function value. In order to prevent from this phenomena and to* ***globalize*** *the method (convergence from arbitrary starting points!) a damping parameter λ_k is included in the iteration, i.e.,*

$$x_{k+1} = x_0 - \lambda_k \left(H_f(x_0)\right)^{-1} \cdot \nabla f(x_0) \qquad\qquad k = 0, 1 \ldots,$$

where we choose λ_k such that

(i) $\lambda_k \in [0, 1]$ *for all* $k = 0, \ldots,$

(ii) $f(x_{k+1}) < (1 - \alpha) f(x_k)$ *for some (small)* $\alpha > 0$ *(e.g.* $\alpha = 10^{-4}$*).*

Typically, a so-called line search is executed to find a reasonable value for λ_k (keyword: cubic backtracking). In a well-designed algorithm, the damping factors will initially be smaller than 1, but if we reach the quadratic convergence area, full Newton steps will be taken.

(v) *The partial derivatives used for the Newton iteration are often calculated by (central) difference quotients. Please note that this may slow down the convergence speed of the method or even lead the method to fail, if calculated inaccurately. Furthermore, since this calculation may also be too expensive, a so-called* ***simplified Newton method*** *is applied. The idea is to keep the calculated Hessian matrix of one step for several iterations and update either never or when the convergence behavior becomes bad.*

(vi) *By construction, the Newton method finds the minimum (or the maximum) of a quadratic function*

$$f : \mathbb{R}^n \to \mathbb{R}, \;\; x \mapsto x^\top A x + b^\top x + c$$

in one iteration, starting from every initial point!

We close this section with a short discussion of non-linear equations.

3.5 Discussion (Newton method for non-linear equations)
Consider the problem of solving a system of n non-linear equations

$$F(x_1,\dots,x_n) = \begin{pmatrix} f_1(x_1,\dots,x_n) \\ \vdots \\ f_n(x_1,\dots,x_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

As motivated in the introduction, this system can be solved by finding the (global) minimizer of the function

$$\begin{aligned} f:\mathbb{R}^n \to \mathbb{R},\ (x_1,\dots,x_n) \mapsto f(x) &:= \frac{1}{2}\left\| \begin{pmatrix} f_1(x_1,\dots,x_n) \\ \vdots \\ f_n(x_1,\dots,x_n) \end{pmatrix} \right\|_2^2 \\ &= \frac{1}{2} F(x_1,\dots,x_n)^\top F(x_1,\dots,x_n). \end{aligned}$$

The obvious way to do this is to apply Newton's method to the function f directly now. However, there is a simpler more commonly used way: Observe that, at a given iterate x_k,

$$\nabla f(x_k) = J_F(x_k)^\top F(x_k)$$

and calculating the directional derivative w.r.t. the direction $v_N = -\left(J_F(x_k)\right)^{-1} F(x_k)$ gives

$$\begin{aligned} \partial_{v_N} f(x_k) &= \nabla f(x_k)^\top v_N \\ &= \left(J_F(x_k)^\top F(x_k)\right)^\top \left(-\left(J_F(x_k)\right)^{-1} F(x_k)\right) \\ &= -\underbrace{F(x_k)^\top F(x_k)}_{\|F(x_k)\|_2^2} < 0, \end{aligned}$$

i.e., we have a descent in this direction. Consequently, iterating by

$$x_{k+1} = x_k - \left(J_F(x_k)\right)^{-1} F(x_k)$$

produces a sequence that reduces function values, which is known as the ***Newton iteration for non-linear systems****.*
Please note that this choice also results, when the function F is linearized around x_k by a Taylor series approximation

$$F(x) \approx F(x_k) + J_F(x_k)(x - x_k)$$

and the zero of this linear approximation

$$x^* = x_k - \left(J_F(x_k)\right)^{-1} F(x_k)$$

is used for the next iterate.
As an application, the method is effectively used to calculate zeros of $\nabla\mathcal{L}(x,\lambda)$ in order to solve constrained optimization problems.

We calculate an example for demonstration.

3.6 Example (Intersection of curves)
Consider the non-linear system

$$F(x,y) = \begin{pmatrix} f_1(x,y) = x^3 + y^2 \\ f_2(x,y) = x^2 - y^3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

which we want to solve using the Newton method starting at $x_0 = (-1, 0.4)$ *up to a tolerance of* $\tau = 10^{-11}$. *The curves are visualized in* Figure V.5, *together with the two solutions* $(0,0)$ *and* $(-1,1)$.

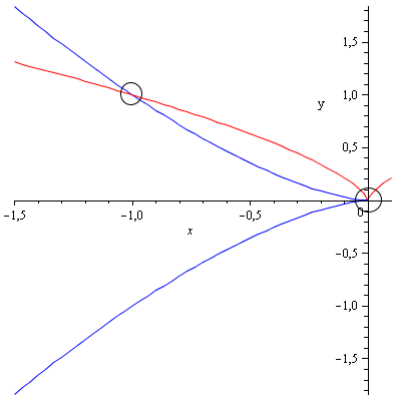


Figure V.5: Curves in $x-y$-plane

In order to do so, we need to proceed as follows.

(i) *Calculate the Jacobian matrix of* $F : \mathbb{R}^2 \to \mathbb{R}^2$*:*

$$J_F(x,y) = \begin{pmatrix} \frac{\partial f_1}{\partial x}(x,y) & \frac{\partial f_1}{\partial y}(x,y) \\ \frac{\partial f_2}{\partial x}(x,y) & \frac{\partial f_2}{\partial y}(x,y) \end{pmatrix} = \begin{pmatrix} 3x^2 & 2y \\ 2x & -3y^2 \end{pmatrix}$$

(ii) *Invert the Jacobian:*

$$(J_F(x,y))^{-1} = \frac{1}{-xy(9xy+4)} \begin{pmatrix} -3y^2 & -2y \\ -2x & 3x^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3x^2 + \frac{4x}{3y}} & \frac{1}{\frac{9x^2y}{2} + 2x} \\ \frac{1}{\frac{9xy^2}{2} + 2y} & \frac{1}{-3y^2 - \frac{4y}{3x}} \end{pmatrix}$$

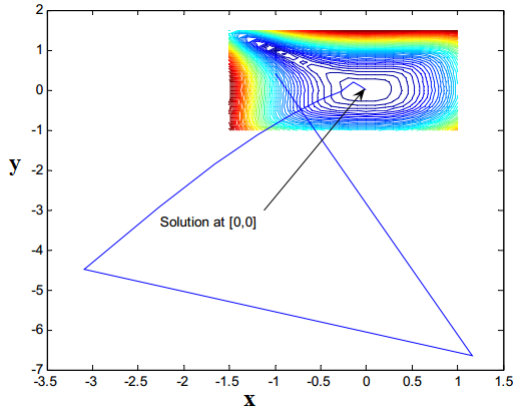
(iii) *Compute the Newton step*

$$\begin{aligned} x_1 &= x_0 - (J_F(-1,0.4))^{-1} F(-1,0.4) \\ &= \begin{pmatrix} -1 \\ 0.4 \end{pmatrix} - \begin{pmatrix} -3 & -5 \\ -12.5 & 18.75 \end{pmatrix} \begin{pmatrix} -0.84 \\ 0.936 \end{pmatrix} \\ &= \begin{pmatrix} 1.16 \\ -6.65 \end{pmatrix} \end{aligned}$$

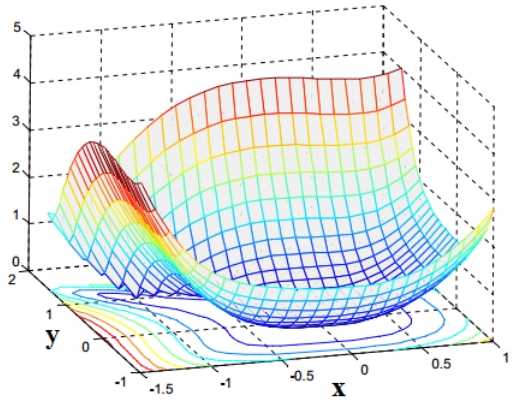
(iv) *Further iterating yields*

$$\begin{aligned} x_2 &= x_1 - \left(J_F(1.16,-6.65)\right)^{-1} F(1.16,-6.65) && \approx \begin{pmatrix} -3.0897 \\ -4.4975 \end{pmatrix} \\ x_3 &= x_2 - \left(J_F(x_1)\right)^{-1} F(x_1) && \approx \begin{pmatrix} -3.0897 \\ -4.4975 \end{pmatrix} \\ &\vdots \quad \dots \\ x_{16} &\approx \begin{pmatrix} -0.0006758 \\ 0.00135670 \end{pmatrix}. \end{aligned}$$

We have depicted the Newton path in Figure V.6(a).



(a) Newton path in the $x-y$-plane



(b) Objective function f

(v) *Now note that we actually minimize the objective function*

$$f(x,y) = \frac{1}{2}\left\| \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix} \right\|_2^2 = 0.5\left(\left(x^3+y^2\right)^2 + \left(x^2-y^3\right)^2 \right)$$

and since $f(x_{16}) = 1.79 \cdot 10^{-12}$*, we may terminate our algorithm at this point.*

Let us examine the path we took a little closer by looking at the objective function (shown in Figure V.6(b)*). We know that the Newton direction is always a descending one for* f*, but are the step sizes OK? We calculate*

$$\begin{aligned} f(-1,0.4) &\approx 0.8908480 & f(1.16,-6.65) &\approx 44686.09146 \\ f(x_2) &\approx 5095.07528 & f(x_3) &\approx 460.1279 \\ f(x_4) &\approx 42.63729 & f(x_5) &\approx 4.1196 \\ f(x_6) &\approx 0.417762 & f(x_7) &\approx 0.04191 \end{aligned}$$

and observe that we are initially drifting away from the correct solution and arrive in the area of quick convergence only afterwards. This slows down the convergence of the algorithm significantly and can even lead to divergence.

To deal with this problem we introduce a simple line search: We accept the step x_{k+1} from x_k only if $f(x_{k+1}) < f(x_k)$. If this is not the case we trim the Newton step by a factor 0.5 *and try whether this Newton correction reduces the objective function. This trimming is executed until we found a reducing step. This means, that we apply the following line search at step k:*

Line Search algorithm

Set $\lambda = 1.0$
while $(f(x_k + \lambda \Delta x_k) \geq f(x_k))$: $\quad \lambda = \frac{1}{2}\lambda$
Set $x_{k+1} = x_k + \lambda x_k$

We depicted the Newton path with line search in Figure V.6. *Observe that we are always reducing the objective functions (look at the contour lines!) and that we approach the zero of F on a "shortcut".*

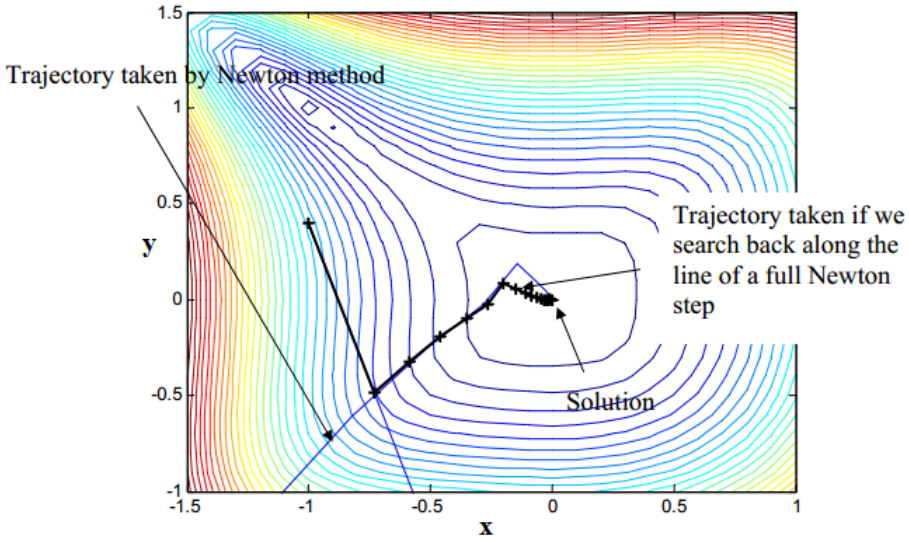


Figure V.6: Newton path with line search

Chapter VI

Multivariable integral calculus

1 Introduction

Last semester we introduced the integral of a function motivated by the task of calculating areas enclosed by the function's curve. The main idea was to approximate the area under a function by Darboux sums formed by easy to calculate rectangles of a certain width Δx_i (compare Figure VI.1). Then, the integral was the value obtained by increasing the number of rectangles and letting their corresponding width go to zero.

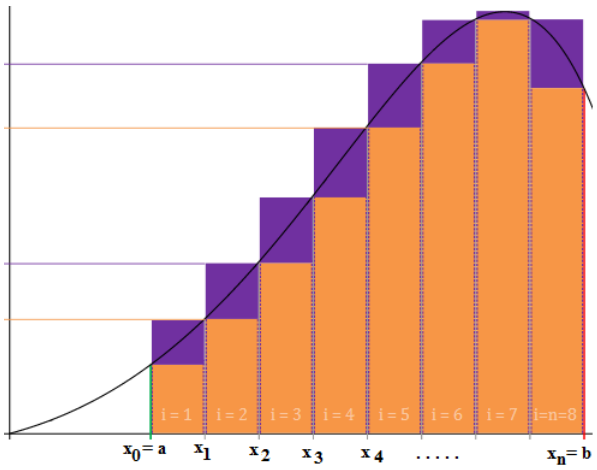


Figure VI.1: Upper & lower Darboux sums

Now, we take a similar path: In order to calculate complex surfaces and volumes, we generalize the integral notion to multi-variable functions $f\colon \mathbb{R}^n \to \mathbb{R}$. The outcome will be so-called multiple integrals, where integrals of a function of two variables over a region in R^2 are called double integrals, and integrals of a function of three variables over a region of R^3 are called triple integrals. They have extensive applications in physics (calculation of center of mass, moments of inertia) and probability theory (calculation of probabilities of combined product-experiments, characteristics of standard normal distribution).
The general guideline will be to reduce the calculation of the multi-variable integral to several single-variable integrals, where we already know how to determine their values.

2 Double integrals

In this section we introduce the integral for functions $f\colon \mathbb{R}^2 \to \mathbb{R}$, starting with the following motivational discussion.

2.1 Discussion (Double integrals)
The main task for this motivation is to calculate the volume V of the cylindric body enclosed by the surface A in the x-y- plane and the function values surface

$$\{(x, y, f(x, y)) \mid (x, y) \in A\}$$

given by a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ with $f(x, y) \geq 0$ for $(x, y) \in A$ (compare Figure VI.2*).*

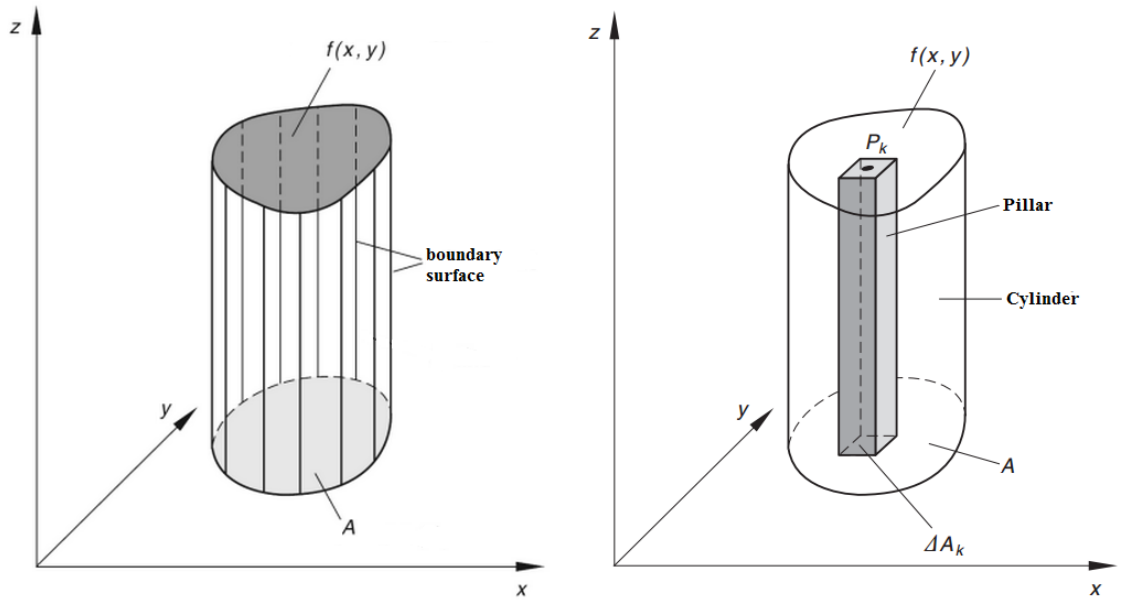


Figure VI.2: Approximation of volume by infinitesimal pillars

The idea for the volume determination is now the same as for the single-variable integral: Approximate the volume of the complex body by easy to calculate ones. Therefore, we choose the following approach:

(i) *We raster the surface A in the x-y-plane by including horizontal and vertical lines thereby decomposing it into n smaller surfaces A_k (disjointly adding up to A, i.e., a partition P_0 of A). The ones in the interior become rectangles (the ones at the boundary may be little different, but this is not important to us). The areas of the surfaces A_k are easy to calculate (they are basically rectangles) and we denote them by ΔA_k ($k = 0, \ldots, n-1$). The* ***norm of the partition*** *P_0 is defined by the maximum area*

$$N(P_0) := \max_{i=0\ldots n-1} \Delta A_i.$$

(ii) *For a surface A_k, we choose a point (x_k, y_k) inside and build a pillar of height $f(x_k, y_k)$. The volume of these pillars is given by*

$$\Delta V_k = \Delta A_k \cdot f(x_k, y_k), \quad k = 0, \ldots n-1.$$

We can sum all of these volumes to get an approximation to the volume V of the cylindrical body we are interested in

$$V \approx R(P_0) := \sum_{k=0}^{n-1} \Delta V_k = \Delta A_k \cdot f(x_k, y_k).$$

The obtained approximating sum is called a ***Riemann-sum****. Note that this is in general only an approximation, since we have chosen a fixed height $f(x_k, y_k)$ for a pillar in contrast to the correct pillar where the height varies with the values of f over A_k.*

(iii) *Now we refine the partition P_0 in* (i) *to a new partition P_1 by adding more raster lines. If we consider the new Riemann sums, we now get a better approximation of the volume, which is reflected by a new norm $N(P_2)$, i.e., a new maximum area of the subareas A_k will become smaller.*

(iv) *Repeating steps* (ii)-(iii) *thereby letting $N(P_k) \to 0$ in the refinement will tend to approximate the volume V better and better.*

The approach just presented is due to the German mathematician Bernhard Riemann (1826-1866) and it leads to the definition of the double integral.

2.2 Definition (Double integrals)

(i) *A bounded function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is called* ***(Riemann-)integrable*** *over a domain $A \subset \mathbb{R}^2$, if the Riemann sums converge to the same real number for every sequence of partitions P_k with $N(P_k) \to 0$.*

(ii) *If $f : \mathbb{R}^2 \to \mathbb{R}$ is Riemann-integrable over A, we call the limit of the Riemann sums the* ***double integral of*** *f over the surface A and denote it by*

$$\int_A f(x,y)\, dA = \int_A f(x,y)\, d(x,y) = \iint_A f(x,y)\, dA := \lim_{k\to\infty} R(P_k),$$

where $(P_k)_{k\in\mathbb{N}}$ is any sequence of partitions satisfying $N(P_k) \to 0$.

2.3 Remarks

(i) *Continuous functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ are (Riemann-)integrable, like in the one-dimensional case.*

(ii) *Using the definition, it is easy to see that the following calculation rules also apply to the double integral (for integrable functions):*

(a) $$\int_A \alpha f(x,y)\, d(x,y) = \alpha \int_A f(x,y)\, d(x,y) \text{ for } \alpha \in \mathbb{R}$$

(b) $$\int_A \left(f(x,y) + g(x,y)\right)\, d(x,y) = \int_A f(x,y)\, d(x,y) + \int_A g(x,y)\, d(x,y).$$

(c) *If $A = A_1 \cup A_2$ and A_1, A_2 have only boundary points in common, then*

$$\int_A f(x,y)\,d(x,y) = \int_{A_1} f(x,y)\,d(x,y) + \int_{A_2} f(x,y)\,d(x,y).$$

(iii) *The value of the double integral of f over A only represents the volume of the cylindric body enclosed by the surface A and the function values surface when $f(x,y) \geq 0$ on A. This is similar as for the one-dimensional integral, where the integral only represents the area between f and the x-axis, if $f(x) \geq 0$.*

(iv) *The double integral*

$$\int_A 1\,d(x,y)$$

represents the volume of a cylinder with base A and height 1. Therefore, this numerical value is precisely the area of A.

(v) *All aspects of double integrals can be extended to triple or multiple integrals. An important application of triple integrals is the calculation of the volume V, the mass m, the momenents of inertia T and the center of mass $(x_m, y_m, z_m))$ of an inhomogenuous body K with mass density $\rho = \rho(x,y,z)$:*

$$V = \iiint_K 1\ dV \qquad x_m = \frac{1}{m}\iiint_K x\cdot\rho\ dV$$
$$m = \iiint_K \rho\ dV \qquad y_m = \frac{1}{m}\iiint_K y\cdot\rho\ dV$$
$$T = \iiint_K a^2\rho\ dV \qquad z_m = \frac{1}{m}\iiint_K z\cdot\rho\ dV,$$

where a is the distance from the rotational axis.

The definition of the double integral is not very practical to calculate given integrals explicitly. We will develop calculation methods for domains of the following form.

2.4 Definition (Cartesian normal domain)
A set $A \subseteq \mathbb{R}^2$ is called a ***(cartesian) normal domain***

(i) *of* ***type I****, if $\exists\ a,b \in \mathbb{R}$ and $\mathcal{C}^1$-functions $o,u : [a,b] \to \mathbb{R}$ such that*

$$A = \left\{(x,y) \in \mathbb{R}^2 \mid a \leq x \leq b,\ u(x) \leq y \leq o(x)\right\},$$

(ii) *of* ***type II****, if $\exists\ c,d \in \mathbb{R}$ and $\mathcal{C}^1$-functions $l,r : [c,d] \to \mathbb{R}$ such that*

$$A = \left\{(x,y) \in \mathbb{R}^2 \mid c \leq y \leq d,\ l(y) \leq x \leq r(y)\right\}.$$

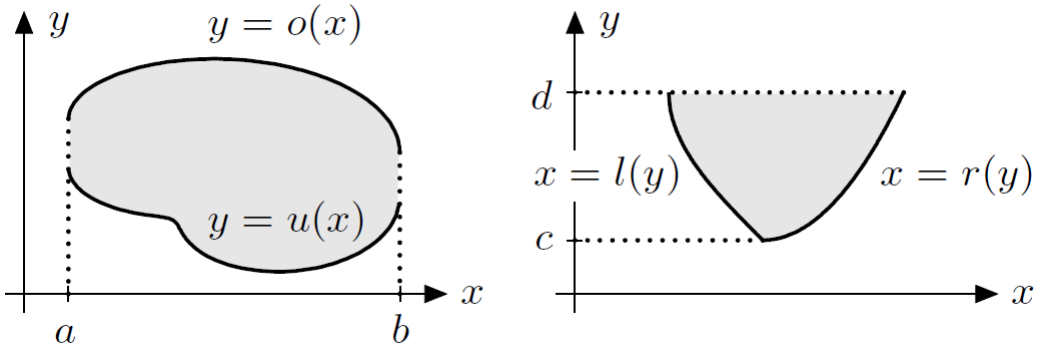


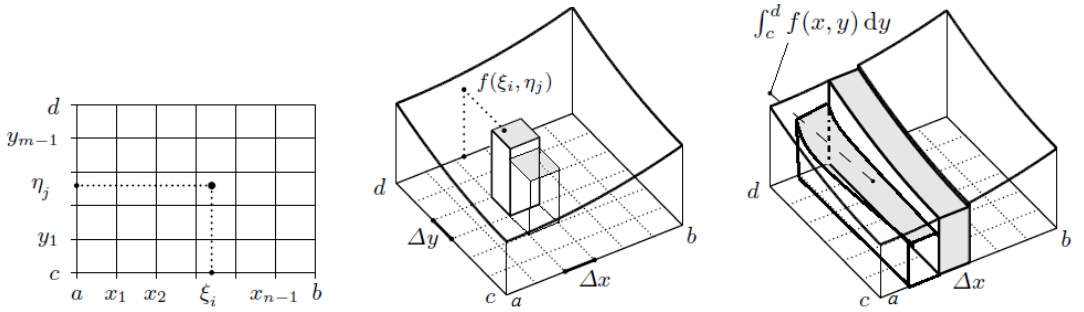
Figure VI.3: Cartesian normal domains of type I and II

The following discussion shall motivate the concrete calculation of double integrals.

2.5 Discussion (Calculating double integrals)
Assume we want to calculate the value of

$$V = \int_A f(x,y)\, d(x,y)$$

for integrable and positive f over a normal domain A of type I. To determine its value we choose the following approach corresponding to Figure VI.4.



(a) Rastered rectangle (b) A Pilar in the slice (c) Area & volume of the slice

Figure VI.4: Approximating the volume slice-wise

(i) *At first, we assume that the domain A is a rectangle, say $A = [a,b] \times [c,d]$. This rectangle is rastered by horizontal and vertical lines which yields a partition into smaller rectangles $A_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$ ($i = 0, \ldots, n-1$, $j = 0, \ldots, m-1$), where*

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$
$$c = y_0 < y_1 < \ldots < y_{n-1} < y_m = d.$$

This is illustrated in Figure VI.4(a).

(ii) *In order to approximate V, we use Riemann sums again. Choosing the heights of the pillars corresponding to the A_{ij} by $f(\xi_i, \eta_j)$, where $(\xi_i, \eta_j) \in A_{ij}$, we get*

$$V = \int_A f(x,y)\, d(x,y) \approx \sum_{i=0}^{n-1} \underbrace{\left(\sum_{j=0}^{m-1} f(\xi_i, \eta_j)\, (y_{j+1} - y_j) \right) (x_{i+1} - x_i)}_{=:L_i}.$$

Note that the quantity L_i is the sum of the volumes of all pillars in one layer of width $(x_{i+1} - x_i)$ (compare Figure VI.4(b)*) and so it is an approximation to the volume V in this layer.*

(iii) *If we add more horizontal raster lines (increase m!) we may approximate the volume V included in the layer more accurately. In this way, the rectangular pillars we have added up (depicted in* Figure VI.4(b)*) become a "smooth" slice as depicted in* Figure VI.4(c)*. This is precisely the one-dimensional crossover from the sum to the integral, i.e., by adding more and more lines, we get*

$$(x_{i+1} - x_i) \sum_{j=0}^{m-1} f(\xi_i, \eta_j)\, (y_{j+1} - y_j) \to (x_{i+1} - x_i) \int_c^d f(\xi_i, y)\, dy.$$

(iv) *Since we actually wanted to integrate over a normal domain with boundaries $o(x)$ and $u(x)$, depending on the current x-value, rather than fixed c and d everywhere, we get the approximation*

$$\int_A f(x,y)\, d(x,y) \approx \sum_{i=0}^{n-1} \left(\int_{u(\xi_i)}^{o(\xi_i)} f(x_i, y)\, dy \right) (x_{i+1} - x_i).$$

Please note that we simply add up the volumes of the layers of width $(x_{i+1} - x_i)$ to approximate V (compare Figure VI.4(c)*)*

(v) *In the last step, we include more vertical lines (increase n!) thereby making the slices to approximate V smaller and smaller. This is again the one-dimensional crossover from the sum over x-intervals to the integral over x, i.e., by adding more and more lines, we get the approximation*

$$\sum_{i=0}^{n-1} \left(\int_{u(\xi_i)}^{o(\xi_i)} f(x_i, y)\, dy \right) (x_{i+1} - x_i) \to \int_a^b \left(\int_{u(x)}^{o(x)} f(\xi_i, y)\, dy \right) dx \approx V.$$

The mathematically correct proof is a little more involved, but it simply recycles the above approach.

As a result of this discussion, we get the following theorem.

2.6 Theorem
Let $f : \mathbb{R}^2 \to \mathbb{R}$ be integrable over the normal domain A.

(i) *If A is a normal domain of type I, enclosed by $a, b \in \mathbb{R}$ and the functions $o, u \in \mathcal{C}^1([a,b])$, then*

$$\int\limits_A f(x,y)\, d(x,y) = \int\limits_{x=a}^{b} \left(\int\limits_{y=u(x)}^{o(x)} f(x,y)\, dy \right) dx. \tag{2.1}$$

(ii) *If A is a normal domain of type II, enclosed by $c, d \in \mathbb{R}$ and the functions $l, r \in \mathcal{C}^1([c,d])$, then*

$$\int\limits_A f(x,y)\, d(x,y) = \int\limits_{y=c}^{d} \left(\int\limits_{x=l(y)}^{r(y)} f(x,y)\, dx \right) dy. \tag{2.2}$$

2.7 Remarks/Examples

(i) *The formula given by* Equation (2.1) *tells us to execute the integration over the domain A successively. At first the inner integral w.r.t. y is determined, then the resulting expression is integrated w.r.t. x. Note that the variable x may appear in the inner integral. It should be treated like a constant parameter for the integration process. The limits of the outer integral are constants independent of x or y.*

(ii) *The order of differentiation in* Equations (2.1) *and* (2.2) *is clearly determined by the integration variables, i.e., from the inner to the outer part. Changing the order without changing the limits of the integrals is in general* ***not*** *correct. One particular situation, where the integration order may be exchanged is when integrating a continuous function over a rectangle $A = [a,b] \times [c,d]$. Then the following formula holds true:*

$$\int\limits_A f(x,y)\, d(x,y) = \int\limits_{x=a}^{b} \left(\int\limits_{y=c}^{d} f(x,y)\, dy \right) dx = \int\limits_{y=c}^{d} \left(\int\limits_{x=a}^{b} f(x,y)\, dx \right) dy. \tag{2.3}$$

This is the content of ***Fubini's theorem****.*

(iii) *We calculate the volume formed by the square $A = [0,3] \times [0,\pi]$ and the function $f(x,y) = x^2 \sin(y)$ above it, i.e., using* Theorem 2.6(i),

$$\int\limits_A x^2 \sin(y)\, d(x,y) = \int\limits_0^3 \int\limits_0^\pi x^2 \sin(y)\, dy\, dx.$$

The inner integral must be determined first

$$\int\limits_0^\pi x^2 \sin(y)\, dy = x^2 \int\limits_0^\pi \sin(y)\, dy = x^2 \left[-\cos(y) \right]_0^\pi = 2x^2$$

followed by the outer integral

$$\int_A x^2 \sin(y)\, d(x,y) = \int_0^3 2x^2\, dx = \left[\frac{2}{3}x^3\right]_0^3 = 18.$$

The volume is depicted in Figure VI.5.

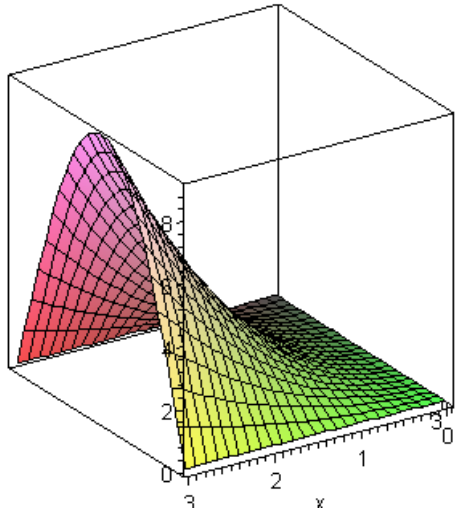


Figure VI.5: Volume formed by f and A

(iv) We calculate the volume beneath the paraboloid surface $f(x,y) = x^2 + y^2$, given the triangular area $A = \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 1,\ 0 \le y \le 1 - x\}$. Here, we have a cartesian normal domain with boundaries given by

$$a = 0 \qquad b = 1,$$
$$u(x) = 0 \qquad o(x) = 1 - x.$$

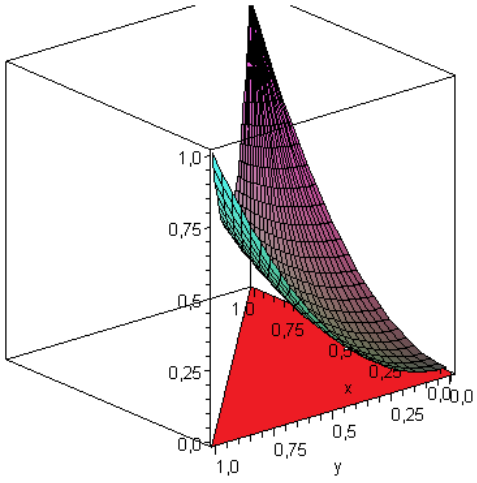


Figure VI.6: Volume formed by f and A

Therefore, we apply Theorem 2.6(i) *and get*

$$\int_A (x^2+y^2)\,d(x,y) = \int_0^1 \left(\int_0^{1-x} (x^2+y^2)\,dy \right) dx.$$

Calculating the inner integral first gives

$$\int_0^{1-x} (x^2+y^2)\,dy = + \left[x^2 y + \frac{1}{3}y^3 \right]_0^{1-x} = x^2(1-x) + \frac{(1-x)^3}{3}.$$

Note that the resulting expression still depends on x! Calculating the outer integral gives the final value

$$\int_A (x^2+y^2)\,d(x,y) = \int_0^1 \left(x^2(1-x) + \frac{(1-x)^3}{3} \right) dx = \frac{1}{6}.$$

Note that this value does not depend on x or y anymore.

So far, we have only covered cartesian normal domains. It is also very common to describe domains in the x-y-plane by polar coordinates. This leads to polar normal domains.

2.8 Definition (Polar normal domain)
*A set $A \subseteq \mathbb{R}^2$ is called a **(polar) normal domain**, if there exist angles $\varphi_1, \varphi_2 \in [0, 2\pi)$ and angle- dependent radii $r_i(\varphi), r_a(\varphi) : [\varphi_1, \varphi_2] \to \mathbb{R}_+$ such that*

$$A = \{(r,\varphi) \in \mathbb{R}_+ \times [0,2\pi) \mid \varphi_1 \le \varphi \le \varphi_2 \text{ and } r_i(\varphi) \le r \le r_a(\varphi)\}$$

So, polar normal domains are always enclosed between fixed angles φ_1 and φ_2, whereas they may have a non-circular radius line (compare Figure VI.7(right)).

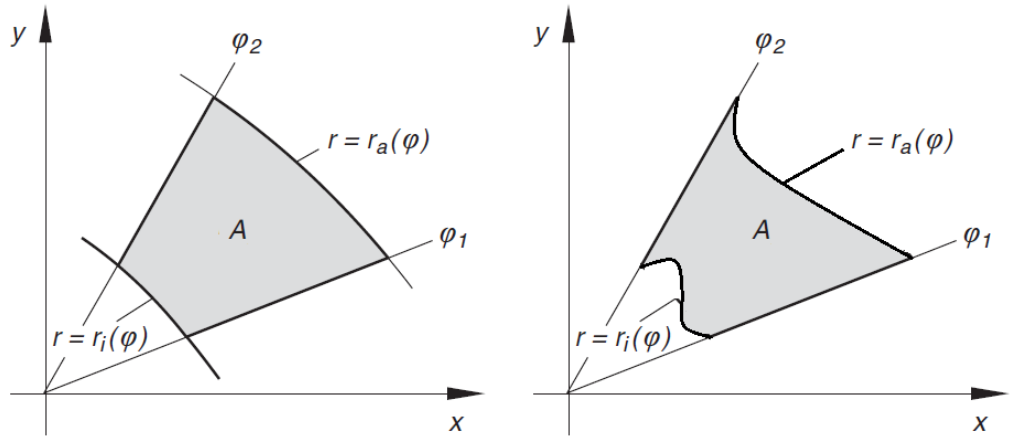


Figure VI.7: Polar normal domains

How to integrate w.r.t. polar coordinates is answered by the following Proposition, which is based on the famous transformation formula using the determinant of the Jacobian:

$$dA = dxdy = \det(J_f)\,drd\varphi = r\,drd\varphi.$$

2.9 Proposition
Assume that A is a polar normal domain and $f : \mathbb{R}^2 \to \mathbb{R},\ (x, y) \mapsto f(x, y)$ is integrable. Using the transformations

$$x = r \cdot \cos(\varphi) \quad y = r \cdot \sin(\varphi)$$

to polar coordinates, we have

$$\int_A f(x,y)\, d(x,y) = \int_{\varphi=\varphi_1}^{\varphi_2} \left(\int_{r=r_i(\varphi)}^{r_a(\varphi)} f(r\cos(\varphi), r\sin(\varphi)) \cdot r\, dr \right) d\varphi.$$

We immediately consider some interesting examples.

2.10 Remarks/Examples

(i) *Since $\int_A 1\, d(x,y)$ is used to calculate the area of the domain A (compare* Remarks 2.3(iv)*), we may use the formula*

$$\int_A 1\, d(x,y) = \int_{\varphi=\varphi_1}^{\varphi_2} \left(\int_{r=r_u(\varphi)}^{r_o(\varphi)} r\, dr \right) d\varphi$$

to determine the area of a polar normal domain.

(ii) *By rotating the parabola $z = 4 - x^2$ around the z-axis, we derive the rotational parabola*

$$f : \mathbb{R}^2 \to \mathbb{R},\ (x, y) \to f(x, y) = 4 - (x^2 + y^2)$$

depicted in Figure VI.8(a)*. We are interested in its volume enclosed with the $x-y$-plane. Therefore, we can easily express the domain A in polar coordinates by*

$$A = \{(r,\varphi) \in \mathbb{R}_+ \times [0, 2\pi) \mid \underbrace{0}_{\varphi_1} \le \varphi < \underbrace{2\pi}_{\varphi_2} \text{ and } \underbrace{0}_{r_i(\varphi)} \le r \le \underbrace{2}_{r_a(\varphi)}\},$$

which is a polar normal domain. Using Proposition 2.9*, we can calculate*

$$\begin{aligned}
\int_A (4 - x^2 - y^2)\, d(x,y) &= \int_{\varphi=0}^{2\pi} \left(\int_{r=0}^{2} r \cdot (4 - (r\cos(\varphi))^2 - r\sin(\varphi)^2)\, dr \right) d\varphi \\
&= \int_{\varphi=0}^{2\pi} \left(\int_{r=0}^{2} r \cdot \left(4 - r^2 \left(\cos^2(\varphi) + \sin^2(\varphi)\right)\right)\, dr \right) d\varphi \\
&= \int_{\varphi=0}^{2\pi} \left(\int_{r=0}^{2} r(4 - r^2)\, dr \right) d\varphi \\
&= \int_{\varphi=0}^{2\pi} \left[2r^2 - \frac{1}{4} r^4 \right]_{r=0}^{2} d\varphi = \int_{\varphi=0}^{2\pi} 4\, d\varphi = [4\varphi]_{\varphi=0}^{2\pi} = 8\pi.
\end{aligned}$$

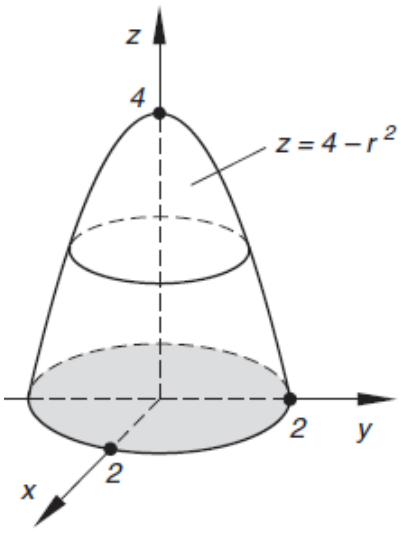
(iii) *We want to calculate the area A of the cardioid as depicted in* Figure VI.8(b). *The characteristic equation of the outer boundary is known to be*

$$r_a(\varphi) = 1 + \cos(\varphi).$$

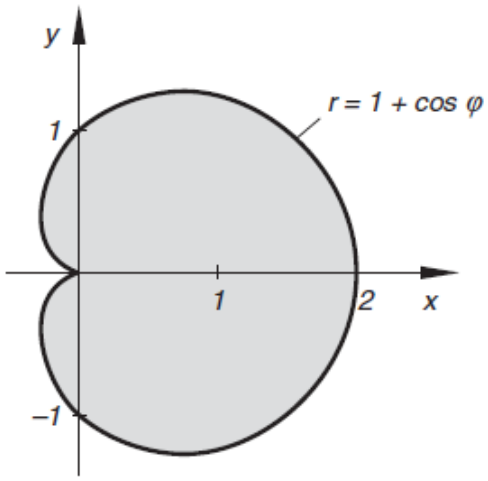
Together with $r_i(\varphi) = 0$ and $\varphi_1 = 0$, $\varphi_2 = 2\pi$, we see that it is a polar normal domain. To calculate the area, we use the previous remark and obtain

$$\begin{aligned}
\int_A 1 \, d(x,y) &= \int_{\varphi=0}^{2\pi} \left(\int_{r=0}^{1+\cos(\varphi)} r \, dr \right) d\varphi \\
&= \int_{\varphi=0}^{2\pi} \left[\frac{1}{2} r^2 \right]_{r=0}^{1+\cos(\varphi)} d\varphi \\
&= \int_{\varphi=0}^{2\pi} \frac{1}{2} \left(1 + \cos(\varphi)\right)^2 d\varphi \\
&= \frac{1}{2} \int_{\varphi=0}^{2\pi} \left(1 + 2\cos(\varphi) + \cos^2(\varphi)\right) d\varphi \\
&= \frac{1}{2} \left[\varphi + 2\sin(\varphi) + \frac{1}{2}\varphi + \frac{1}{4}\sin(2\varphi) \right]_{\varphi=0}^{2\pi} = \frac{3}{2}\pi,
\end{aligned}$$

where we used partial integration to calculate a primitive of $\varphi \mapsto \cos^2(\varphi)$.



(a) Rotational parabola



(b) Cardioid

(iv) *The gaussian bell curve*

$$g : \mathbb{R} \to \mathbb{R}, \quad x \mapsto g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

is used extensively in probability theory, since it is the density function of the famous standard normal distribution. Similar processes which are executed many times are normally distributed, irrespectively of the original process.

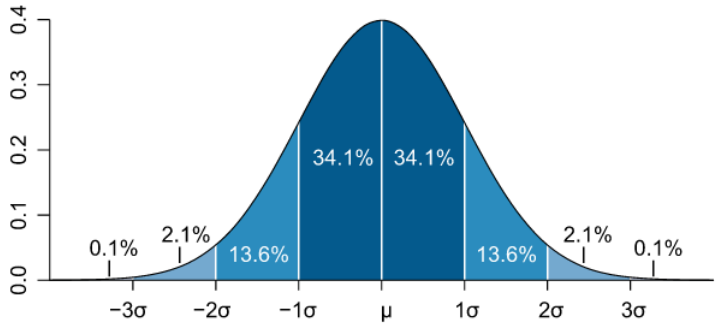


Figure VI.8: Probabilities from Gaussian bell function

As a consequence, probabilities for events $E \subseteq \mathbb{R}$ can be calculated by

$$P(E) = \int_E \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \, dx.$$

It is clear that $P(\emptyset) = 0$. Now, we want to show that the definite event $P(\mathbb{R})$ has probability 1, which (partially) justifies to calculate probabilities with this formula at all.
Then, with a little trick

$$\begin{aligned}
P(\mathbb{R})^2 &= \left(\int_{-\infty}^{\infty} g(x)\, dx \right)^2 \\
&= \left(\int_{-\infty}^{\infty} g(x)\, dx \right) \cdot \left(\int_{-\infty}^{\infty} g(y)\, dy \right) \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x)\, dx \right) \cdot g(y)\, dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \underbrace{g(x) \cdot g(y)}_{=f(x,y)}\, dx \right) dy \\
&= \int_{\mathbb{R}^2} f(x,y)\, d(x,y)
\end{aligned}$$

we obtain an ordinary double integral for the two-variable function f. Instead of calculating the integral over $\mathbb{R}^2$, we first choose the circle area of radius $n > 0$, denoted A, as polar normal domain. Later we let $n \to \infty$ to obtain the result. Then,

$$\begin{aligned}
\int_A f(x,y)\,d(x,y) &= \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,d(x,y) \\
&= \frac{1}{2\pi} \int_A e^{-\frac{1}{2}(x^2+y^2)}\,d(x,y) \\
&= \frac{1}{2\pi} \int_{\varphi=0}^{2\pi} \left(\int_{r=0}^{n} r e^{-\frac{1}{2}r^2}\,dr \right) d\varphi \\
&= \frac{1}{2\pi} \int_{\varphi=0}^{2\pi} \left[-e^{-\frac{1}{2}r^2} \right]_{r=0}^{n} d\varphi \\
&= \frac{1}{2\pi} \left(-e^{-\frac{1}{2}n^2} + 1 \right) \int_{\varphi=0}^{2\pi} 1\,d\varphi \\
&= \frac{1}{2\pi} \left(-e^{-\frac{1}{2}n^2} + 1 \right) 2\pi = \underbrace{-e^{-\frac{1}{2}n^2} + 1}_{\to 1 \quad (n\to\infty)},
\end{aligned}$$

where we used that $x^2 + y^2$ becomes $r^2\cos^2(\varphi) + r^2\sin^2(\varphi) = r^2$, when transformed into polar coordinates.

3 Line Integrals

A line integral, also known as curve or path integral, is an integral where the function to be integrated is evaluated along a planar or spatial curve.
The function to be integrated may be a scalar field or a vector field, where the value of the line integral is the sum of values of the field at all points on the curve, weighted by some scalar function on the curve.
This weighting distinguishes the line integral from simpler integrals defined on intervals as discussed in the last semester.

In physics, e.g., the work W needed to move an object through a gravitational or electrical field can be calculated by evaluating the force F along the path C with infinitesimal arc length ds:

$$W = \int_C F\,ds.$$

3.1 Definition (Line integral of a scalar field)
Let $f\colon U \subseteq \mathbb{R}^n \to \mathbb{R}$ be a scalar field. Then the ***line integral of the scalar field*** *f along the piecewise smooth curve $C \subset U$ is defined as*

$$\int_C f\, ds = \int_a^b f(r(t))\, \|r'(t)\|_2\, dt,$$

where $r\colon [a,b] \to C$ is an arbitrary bijective parametrization of the curve C such that $r(a)$ and $r(b)$ give the endpoints of C with $a < b$ and $ds = \|r'(t)\|_2\, dt$ is a scalar arc element.

Now, we try to interpret the features of the line integral for a scalar field and consider some examples.

3.2 Remarks/Examples

(i) *Geometrically, when the scalar field f is defined over a plane ($\mathbb{R}^2$), its graph is a surface $z = f(x,y)$ in space, and the line integral gives the (signed) cross-sectional area bounded by the curve C and the graph of f (compare* Figure VI.9*).*

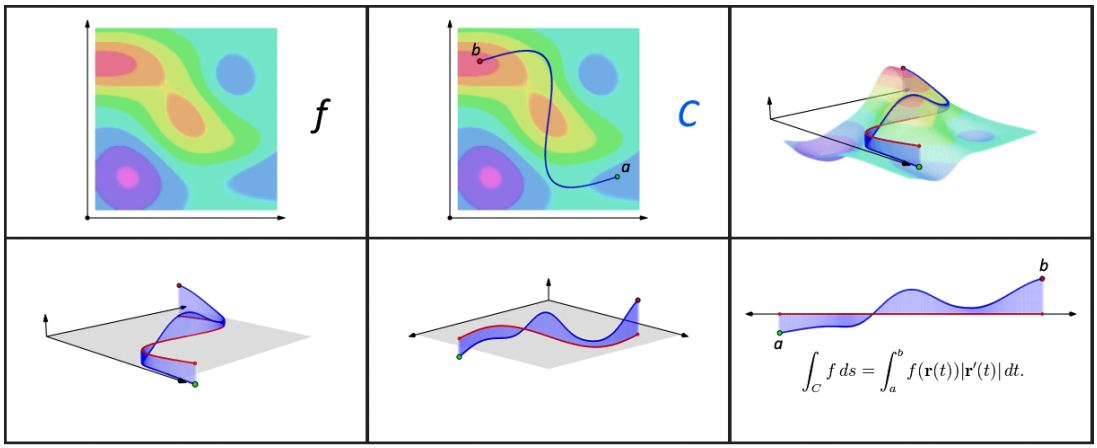


Figure VI.9: The line integral over a scalar field f

(ii) *Line integrals of scalar fields over a curve C do not depend on the chosen parametrization r of C.*

(iii) *If we interpret $r(t)$ as the position of an object at time t, then $\|r'(t)\|_2$ is the absolute value of the speed of the object along the curve C and the line integral over the function $f = 1$ is the distance the object covered from $t_1 = a$ to $t_2 = b$.*

(iv) *Example: calculating the length of a curve C.*
Let f be a function $f\colon [a,b] \to \mathbb{R}$, then its graph C is parameterized by

$$r:\ [a,b] \to \mathbb{R}^2, \quad t \mapsto (t,\ f(t)).$$

Calculating the norm of the derivative of the parameter function $r(t)$ yields

$$\|r'(t)\|_2 = \|(1,\ f'(t))\|_2 = \sqrt{1 + (f'(t))^2}$$

and therefore the length of the graph from a to b is

$$\int_C ds = \int_a^b \sqrt{1 + (f'(t))^2}\, dt.$$

(v) For the function $f(x) = \frac{2}{3}\sqrt{(x-1)^3}$ we want to calculate the length of the graph between $x_1 = a = 2$ and $x_2 = b = 4$. Using $x = t$ and $f = f(t)$ we have a simple parameterization $t \mapsto (t,\ f(t))$. Calculating the derivative of f yields

$$f'(t) = \frac{2}{3} \cdot \frac{3}{2}\sqrt{t-1} = \sqrt{t-1}$$

and therefore the length of the graph is

$$\begin{aligned}
\int_C ds &= \int_2^4 \sqrt{1 + f'(t)^2}\, dt = \int_2^4 \sqrt{1 + t - 1}\, dt = \int_2^4 \sqrt{t}\, dt \\
&= \left[\frac{2}{3}\sqrt{t^3}\right]_2^4 = \frac{2}{3}\left(\sqrt{4^3} - \sqrt{2^3}\right) \\
&= \frac{2}{3}\left(8 - \sqrt{8}\right) = \frac{2}{3}\left(8 - 2\sqrt{2}\right) \\
&= \frac{4}{3}\left(4 - \sqrt{2}\right).
\end{aligned}$$

3.3 Definition (Line integral of a vector field)
Let $F\colon U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a vector field. Then the ***line integral of the vector field*** *F along the piecewise smooth curve $C \subset U$ in the direction of the vector r is defined as*

$$\int_C F(r)\, dr = \int_a^b F(r(t)) \cdot r'(t)\, dt,$$

where $r\colon [a,b] \to C$ is an arbitrary bijective parametrization of the curve C such that $r(a)$ and $r(b)$ give the endpoints of C with $a < b$ and $dr = r'(t)\, dt$ is a vector arc element.

Now, we remark on the line integral for a vector field, try to relate it to a scalar field line integral and we conclude with some examples.

3.4 Remarks/Examples

(i) A line integral of a scalar field is a line integral of a vector field where the vectors are always tangential to the line. For example the friction force is always tangential to the moving direction, whereas the force caused by wind may differ from the intended sailing course. Therefore, the former corresponds to a scalar computation when we are interested in the amount of work we had to supply to push an object along a certain path. The latter leads to a vector field calculation for the thrust.

(ii) In Figure VI.10 *the trajectory of a particle (in red) along a curve C inside a vector field F is illustrated.*

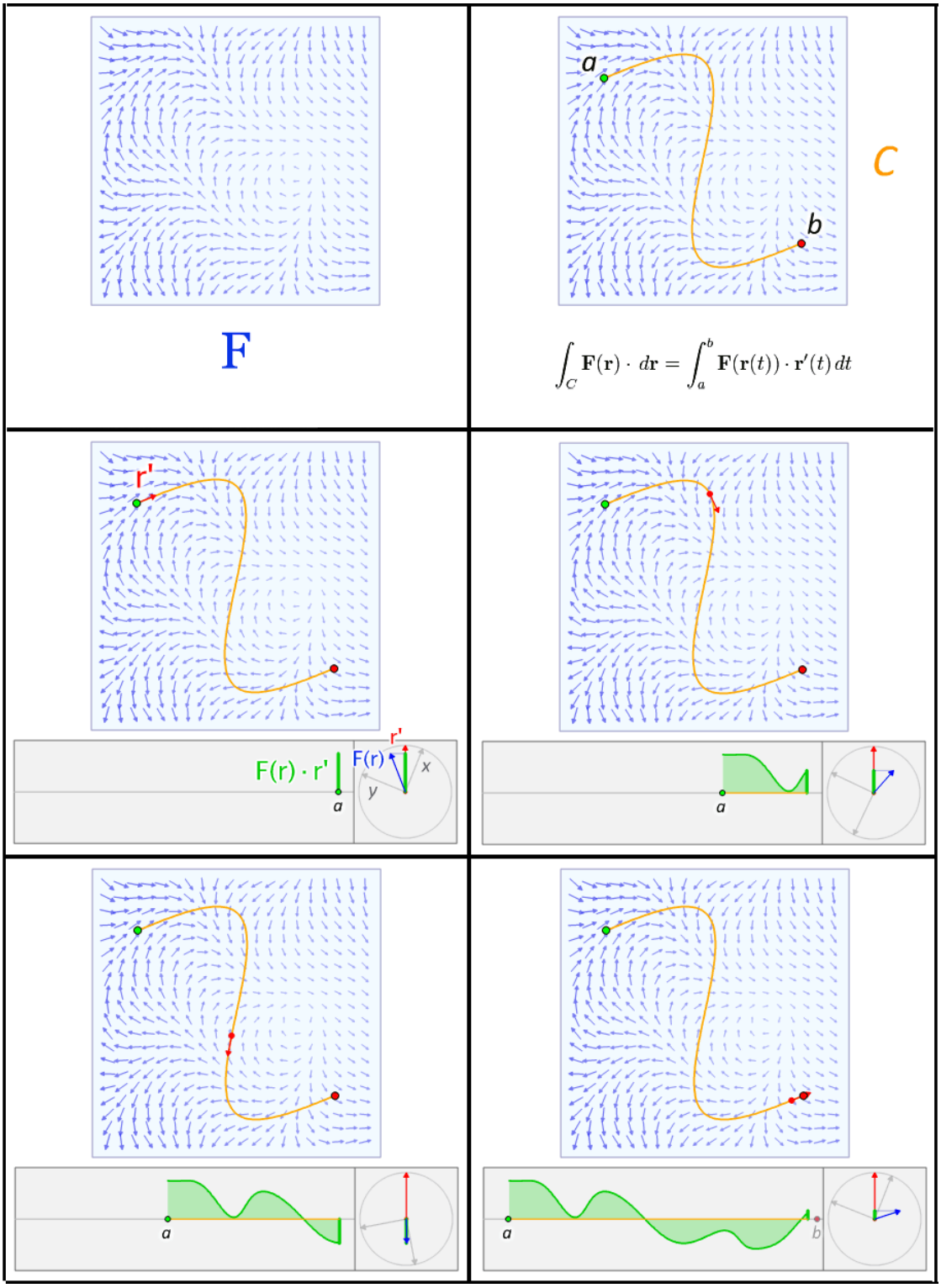


Figure VI.10: The line integral over a scalar field f

Starting from a, the particle traces the path C along the vector field F. The dot product (green line) of its displacement vector (red arrow) and the field vector (blue arrow) defines an area under a curve, which is equivalent to the path's line integral.

(iii) *Note that $F(r(t)) \cdot r'(t)$ is a dot product of two vectors which is defined as*

$$F(r(t)) \cdot r'(t) = \|F(r(t))\|_2 \cdot \|r'(t)\|_2 \cdot \cos(\alpha)$$

involving the angle α between the scalar field and the vector $r'(t)$. In case of the sailing comparison $F(r(t))$ is the wind force vector at a point $r(t)$ and $r'(t)$ is the sailing velocity.

(iv) *Line integrals of vector fields are independent of the parametrization r in absolute value, but they do depend on its orientation. Specifically, a reversal in the orientation of the parametrization changes the sign of the line integral (responsible: $\cos(\alpha)$!).*

(v) *When C is a closed curve, i.e., its initial and final points coincide, the notation*

$$\int_C f(r(t))\, dr = \oint_C f(r(t))\, dr$$

is often used for the line integral of f along C. A closed curve line integral is sometimes referred to as a ***cyclic integral*** *or circulation in engineering applications, especially in aeronautics when it comes to the lift calculation of a wing by integrating the pressure alongside of the wing surface.*

(vi) *We calculate the line integral $\int_C F\, dr$ along the curve*

$$C = \{r(t) = (\cos(t), \sin(t), 2t) \mid t \in [0, 2\pi]\}$$

for the vector field

$$F(x, y, z) = (xy,\ yz,\ xz).$$

For that purpose, we first calculate $r'(t) = (-\sin(t), \cos(t), 2)$, then the scalar product:

$$F(r(t)) \cdot r'(t) = \begin{pmatrix} \cos(t)\sin(t) \\ 2t \cdot \sin(t) \\ 2t \cdot \cos(t) \end{pmatrix} \cdot \begin{pmatrix} -\sin(t) \\ \cos(t) \\ 2 \end{pmatrix}$$

$$= -\sin^2(t)\cos(t) + 2t\sin(t)\cos(t) + 4t\cos(t)$$

$$\int_C F\, dr = \int_0^{2\pi} \left(-\sin^2(t)\cos(t) + 2t\sin(t)\cos(t) + 4t\cos(t)\right)\, dt = -\pi.$$

4 Surface Integrals

In section VI.2 we discussed the definition of double integrals and how we can calculate them. We saw that the double integral of a positive function of two variables represents the volume of the region between the surface defined by a function $f(x, y)$ and the domain in the x-y-plane.
In the last section we learned about line integrals, where the scalar- or vector-valued function that has to be integrated is evaluated along a curve. The value of the line integral is the sum of the field at all points on the curve, weighted by a factor resulting from the scalar function value or the scalar product of the vector field F and a differential vector dr. This weighting factor distinguishes the line integral from the simpler integral of a function defined on intervals.

The *surface integral* can be seen as the double integral analog of the line integral where the function to be integrated can also be a scalar field or a vector field and the value of the surface integral is the sum of the field at all points on the surface. This can be achieved by splitting the surface into small surface elements, which provide the partitioning for Riemann sums.

Given a vector field $\vec{v}(\vec{x})$ describing the velocity of a fluid flowing through the surface S. The so-called *flux* is defined as the quantity of fluid flowing through S per unit time. To find the flux, we need to take the scalar product of $\vec{v}$ with the normal vector $\vec{n}_S$ of S at each point, which will give us a scalar field, which we integrate over the surface:

$$\int_S \vec{v}(\vec{x})\, d\vec{S} = \int_S \vec{v}(\vec{x}) \cdot \vec{n}_S\, dS.$$

The fluid flux in this example may be from a physical fluid such as water or air, or from electrical or magnetic flux. Thus, surface integrals have applications in physics, particularly with the classical theory of electromagnetism.

Before we explicitly define the surface integrals, we need some preliminary work concerning parametrization.

4.1 Discussion (Parametrization)
From school you might remember the paramatric representation of a straight line g defined by two points A and B or a plane E spanned by points A, B and C in the x_1-x_2-x_3-coordinate system:

$$g:\ \vec{X} = \vec{A} + r \cdot \vec{AB},$$

$$E:\ \vec{X} = \vec{A} + s \cdot \vec{AB} + t \cdot \vec{AC},$$

where $r, s, t \in \mathbb{R}$ are the parameters, $\vec{A}$ is the position vector and $\vec{AB}, \vec{AC}$ are direction vectors spanning the line and plane respectively. By changing the parameter values we can reach every point on the line and plane.

Beyond that, even more complicated curves or surfaces can be specified by parameters like the already mentioned Euler spiral *or* Vivani's curve *illustrated in* Figure VI.11.

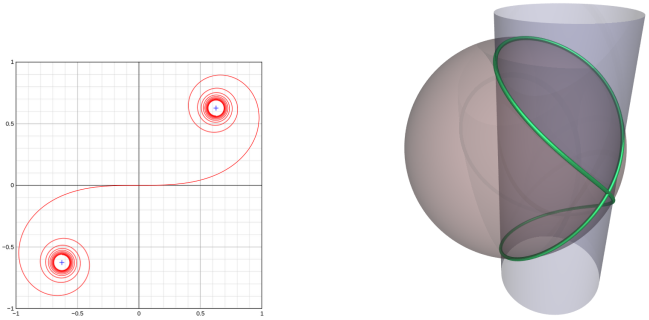


Figure VI.11: Euler spiral in $\mathbb{R}^2$ (left) and Viviani's curve (right)

The parametrization of the Euler spiral is given by

$$E : \mathbb{R}_+ \to \mathbb{R}^2, \;\; t \mapsto E(t) = \begin{pmatrix} \int_0^t \cos(s^2)\, ds \\ \int_0^t \sin(s^2)\, ds \end{pmatrix},$$

whereas Vivian's curve is parameterized in the following way:

$$V : [0, 2\pi] \to \mathbb{R}^3, \;\; t \mapsto V(t) = \begin{pmatrix} a(1+\cos(t)) \\ a\sin^2(t) \\ 2a\sin(\frac{t}{2}) \end{pmatrix}.$$

In general, curves and surfaces can be parameterized in different ways. A frequently chosen parameter for a curve is its arc length measured from a certain starting point. Parameters for surfaces are often chosen such that their parametric lines are orthogonal. For instance, think about the isogonal Mercator projection of a map, where lines of latitude and lines of longitude are perpendicular to each other.
To find an explicit formula for the surface integral, we need to parameterize the surface of interest S by considering a system of curvilinear coordinates on S.

4.2 Definition (Surface integral of a scalar field)
Let $f(\vec{x})$ be a scalar field to be integrated on a surface $S = \{\vec{x}(u,v) \mid (u,v) \in B\}$, where B is the parameter domain of the parameters u and v. Then the ***surface integral of the scalar field*** *f is defined as*

$$\int_S f \, dS = \int_B f(\vec{x}(u,v)) \cdot \|\vec{x}_u(u,v) \times \vec{x}_v(u,v)\|_2 \; dB,$$

where

$\vec{n}$	$=$	$\|\vec{x}_u(u,v) \times \vec{x}_v(u,v)\|_2$	is the two-norm of the normal vector on the surface S,
dS	$=$	$\|\vec{x}_u(u,v) \times \vec{x}_v(u,v)\|_2 \, dB$	is an infinitesimal scalar surface element and
dB	$=$	$d(u,v) = du\, dv$	is an infinitesimal surface element in the parameter domain.

4.3 Remarks/Examples

(i) *If we integrate over the scalar field $f(\vec{x}) = 1$, then the surface integral gives the size of the surface S.*

(ii) *The vectors $\vec{x}_u$ and $\vec{x}_v$ are the partial derivatives of the parametrization $\vec{x}(u, v)$.*

(iii) *Remember that the cross product $\vec{x}_u \times \vec{x}_v$ is the normal vector of the plane spanned by $\vec{x}_u$ and $\vec{x}_v$ and its two-norm $\|\vec{x}_u(u, v) \times \vec{x}_v(u, v)\|_2$ is the length of the normal vector, which is a real number. This value is a scaling factor between dS and dB depending on the position of (u, v) on B.*

(iv) *For the scalar field $f = x^2 z$ and the cylinder barrel*

$$S = \{(x, y, z) \mid x^2 + y^2 = 1,\ 0 \le z \le 1\}$$

we calculate the surface integral. For this purpose we need a parametrization of S:

$$\vec{x} = \vec{x}(u, v) = (\cos(u),\ \sin(u),\ v) \quad \textit{with} \quad (u, v) \in B := [0, 2\pi] \times [0, 1]\,.$$

Then we determine the partial derivatives

$$\vec{x}_u(u, v) = (-\sin(u),\ \cos(u),\ 0) \quad \textit{and} \quad \vec{x}_v(v) = (0,\ 0,\ 1),$$

the cross product and its two-norm

$$\begin{pmatrix} -\sin(u) \\ \cos(u) \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \cos(u) \\ \sin(u) \\ 0 \end{pmatrix}$$

$$\Rightarrow \left\| \begin{pmatrix} \cos(u) \\ \sin(v) \\ 0 \end{pmatrix} \right\|_2 = \sqrt{\cos^2(u) + \sin^2(u) + 0^2} = 1.$$

Now we can calculate the surface integral

$$\begin{aligned}
\int_S x^2 z \, dS &= \int_B \cos^2(u) \cdot v \cdot 1 \; dB \\
&= \int_0^{2\pi} \left(\int_0^1 \cos^2(u) \cdot v \, dv \right) du \\
&= \int_0^{2\pi} \cos^2(u) \cdot \left(\int_0^1 v \, dv \right) du \\
&= \int_0^{2\pi} \cos^2(u) \cdot \left[\tfrac{1}{2} v^2 \right]_0^1 du \\
&= \int_0^{2\pi} \cos^2(u) \cdot \tfrac{1}{2} \, du \\
&= \frac{1}{2} \left[\frac{u}{2} + \frac{1}{4} \sin(2u) \right]_0^{2\pi} \\
&= \frac{1}{2} \pi.
\end{aligned}$$

4.4 Definition (Surface integral of a vector field)
Let $\vec{F}(\vec{x})$ be a vector field to be integrated on an orientated surface $S = \{\vec{x}(u,v) \mid (u,v) \in B\}$, where B is the parameter domain of the parameters u and v. Then the ***surface integral of the vector field*** *$\vec{F}$ is defined as*

$$\int_S \vec{F}\, d\vec{S} = \int_B \vec{F}(\vec{x}(u,v)) \cdot (\vec{x}_u(u,v) \times \vec{x}_v(u,v))\ dB,$$

where the integrand is the scalar product of the vector field $\vec{F}$ and the normal vector $\vec{n} = \vec{x}_u \times \vec{x}_v$ on the surface S and

$\vec{n}$	$=$	$\vec{x}_u(u,v) \times \vec{x}_v(u,v)$	is the normal vector on the surface S,
$d\vec{S}$	$=$	$(\vec{x}_u(u,v) \times \vec{x}_v(u,v))\ dB$	is an infinitesimal vectorial surface element and
dB	$=$	$d(u,v) = du\, dv$	is an infinitesimal surface element in the parameter domain.

4.5 Remarks/Examples

(i) *If $\vec{F}$ is a vector field describing the velocity of a flowing fluid, then the surface integral on the surface S is the quantity of fluid flowing through the surface per unit time interval, which is the definition of the so-called* ***flux*** *$\Phi_S(\vec{F})$. That is why the surface integral is sometimes referred to as the* ***flux integral****.*
In the integral the component of the velocity perpendicular to the surface is integrated. If the velocity vector is tangential to the surface, then the flux is zero, because the fluid just flows parallel to S. Consequently, the flux reaches a maximum if the velocity vector is orthogonal to the surface.

(ii) *We calculate the flux $\Phi_C(\vec{F})$ of the vector field $\vec{F} = (z,\ y,\ z{+}1)$ through the surface (barrel and bottom) of the cone $C = \{(x,y,z) \mid 0 \le z \le 2 - \sqrt{x^2+y^2}\ \}$, whereas the normal vector on the surface points outside the cone.*

1.) *In a first step we parameterize the cone barrel S (compare* Figure VI.12*):*

$$S = \{(x,y,z) \mid z = 2 - \sqrt{x^2+y^2}\ ,\ z > 0\ \}$$
$$\vec{x} = (r\cos(\varphi),\ r\sin(\varphi),\ 2-r) \quad \textit{with} \quad (r,\varphi) \in [0,2] \times [0,2\pi] =: B.$$

Then we calculate the partial derivatives

$$\vec{x}_r = (\cos(\varphi),\ \sin(\varphi),\ -1)$$
$$\vec{x}_\varphi = (-r\sin(\varphi),\ r\cos(\varphi),\ 0)$$

and the cross product, i.e., the normal vector to the cone surface

$$\begin{pmatrix} \cos(\varphi) \\ \sin(\varphi) \\ -1 \end{pmatrix} \times \begin{pmatrix} -r\sin(\varphi) \\ r\cos(\varphi) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & - & (-r\cos(\varphi)) \\ r\sin(\varphi) & - & 0 \\ r\cos^2(\varphi) & - & (-r\sin^2(\varphi)) \end{pmatrix} = \begin{pmatrix} r\cos(\varphi) \\ r\sin(\varphi) \\ r \end{pmatrix}.$$

Note: the normal vector points out of the cone!

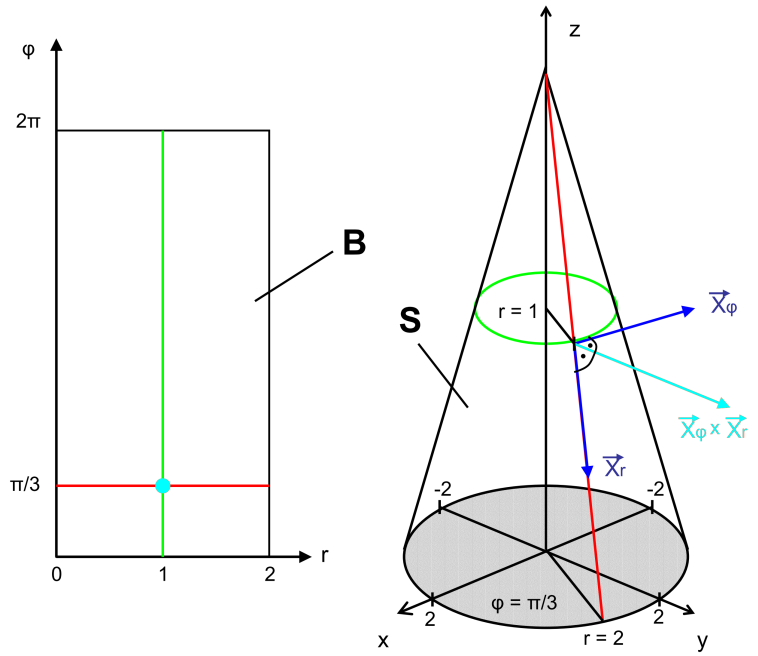


Figure VI.12: Parametrization B of the cone barrel S

Now, we determine the integrand by scalar multiplication of the vector field

$$\vec{F} = (z,\ y,\ z+1) = (2-r,\ r\sin(\varphi),\ 3-r)$$

with the normal vector to get the orthogonal flow velocity component through the surface element dB.

$$\begin{aligned}\vec{F}(\vec{x}(r,\phi)) \cdot (\vec{x}_r \times \vec{x}_\varphi) &= \begin{pmatrix} 2-r \\ r\sin(\varphi) \\ 3-r \end{pmatrix} \cdot \begin{pmatrix} r\cos(\varphi) \\ r\sin(\varphi) \\ r \end{pmatrix} \\ &= (2r-r^2)\cdot\cos(\varphi) + r^2\sin^2(\varphi) + (3r-r^2)\end{aligned}$$

To calculate the flux through the cone barrel we integrate

$$\begin{aligned}\int_S \vec{F}\,d\vec{S} &= \int_0^2 \left(\int_0^{2\pi} \left((2r-r^2)\cos(\varphi) + r^2\sin^2(\varphi) + 3r - r^2 \right)\,d\varphi \right)\,dr \\ &= \int_0^2 \left(\left[(2r-r^2)\sin(\varphi) + \tfrac{1}{2}(r^2 + \sin(\varphi)\cos(\varphi)) + (3r-r^2)\varphi \right]_0^{2\pi} \right)\,dr \\ &= \int_0^2 \left(\tfrac{1}{2}r^2 + (3r-r^2)\cdot 2\pi \right)\,dr \\ &= 2\pi\ \left[\tfrac{3}{2}r^2 - \tfrac{1}{6}r^3 \right]_0^2 \\ &= 2\pi(6 - \tfrac{8}{6} - 0) \\ &= \frac{28}{3}\pi.\end{aligned}$$

Note that we can also utilize the parametrization $\vec{x} = (x,\ y,\ 2-\sqrt{x^2+y^2})$ with $(x,y) \in B = \{(x,y) \mid x^2+y^2 \le 4\}$.

2.) *In a second step we parameterize the bottom surface G*

$$G = \{(x, y, z) \mid x^2 + y^2 \leq 4 \ , z = 0 \ \}$$

with polar coordinates $x = r\cos(\varphi)$ and $y = r\sin(\varphi)$, i.e.,

$$\vec{x} = (r\cos(\varphi),\ r\sin(\varphi), 0) \quad \textit{with} \quad (r, \varphi) \in [0, 2] \times [0, 2\pi] =: B.$$

Then we calculate the partial derivatives

$$\vec{x}_r = (\cos(\varphi),\ \sin(\varphi),\ 0)$$
$$\vec{x}_\varphi = (-r\sin(\varphi),\ r\cos(\varphi),\ 0)$$

and the cross product, i.e., the normal vector to the bottom surface

$$\begin{pmatrix} \cos(\varphi) \\ \sin(\varphi) \\ 0 \end{pmatrix} \times \begin{pmatrix} -r\sin(\varphi) \\ r\cos(\varphi) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix}.$$

Note: the normal vector points into the cone!

Now, we determine the integrand by scalar multiplication of the vector field

$$\vec{F} = (z,\ y,\ z + 1) = (0,\ r\sin(\varphi),\ 1)$$

with the normal vector to get the orthogonal flow velocity component through the surface element dB.

$$\vec{F}(\vec{x}(r, \phi)) \cdot (\vec{x}_r \times \vec{x}_\varphi) = \begin{pmatrix} 0 \\ r\sin(\varphi) \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix} = r$$

We integrate to calculate the flux through the cone bottom and since the normal vector points into the cone we must change the sign in the flux calculation:

$$\begin{aligned} \int_G \vec{F}\, d\vec{G} &= -\int_0^2 \left(\int_0^{2\pi} r\, d\varphi \right) dr \\ &= -\int_0^2 \left([r\varphi]_0^{2\pi} \right) dr = -2\pi \left[\tfrac{1}{2} r^2 \right]_0^2 dr = -4\pi. \end{aligned}$$

3.) *As a final result we get the total flux by adding the fluxes through barrel and bottom:*

$$\Phi_C(\vec{F}) = \Phi_B(\vec{F}) + \Phi_G(\vec{F}) = \frac{28}{3}\pi + (-4\pi) = \frac{16}{3}\pi.$$

Note:
For the bottom flux we can take a shortcut, because we see that the normal vector on G pointing outside of the cone is $\vec{n}_G = (0,\ 0,\ -1)$. Thus, we have

$$\int_G \vec{F}\, d\vec{G} = \int_G \vec{F} \cdot \vec{n}_G\, dG = \int_G \begin{pmatrix} 0 \\ y \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} dG = \int_G -1\, dG = -4\pi.$$

The last example shows, that calculating surface integrals may be expensive. Therefore, we introduce two vector calculus theorems, known as divergence theorem and curl theorem, in the next section.

5 Vector calculus theorems

Vector calculus, or vector analysis, is a branch of mathematics concerned with differentiation and integration of vector fields, primarily in 3-dimensional Euclidean space $\mathbb{R}^3$. The term vector calculus is sometimes used as a synonym for the broader subject of multivariable calculus, which includes vector calculus as well as partial differentiation and multiple integration.

In this section we introduce the *divergence theorem*, also known as Gauss's theorem, and the *curl theorem*, also known Kelvin-Stokes theorem.

5.1 Discussion (Divergence theorem)
In section IV.3.1 *we discussed the divergence of a continuously differentiable vector field and we could relate this differential operator to the physical terms of sources and sinks. According to this, the divergence describes the quantity of sources and sink within the vector field.*

If we now combine our knowledge of line and surface integrals with the divergence, we can connect the flux through a surface with the behavior of vector field inside the surface. This relation is defined by the divergence theorem.

The divergence theorem is an important result for the mathematics of engineering, in particular in electrostatics and fluid dynamics. In physics and engineering, the divergence theorem is usually applied in three dimensions, where it states that the outward flux of a vector field through the regions's surfaces is equal to the volume integral of the divergence over the region inside the surface.

However, it generalizes to any number of dimensions. In one dimension, it is equivalent to the fundamental theorem of calculus we know from last semester. In two dimensions, it is equivalent to Green's theorem, stating that the flux through the boundary curve of a surface is equivalent to the sum of all sources minus the sum of all sinks within the region.

The following proposition contains the divergence theorem in two and three dimensions.

5.2 Proposition (Divergence theorem)

(i) *Let the area A be a subset of $\mathbb{R}^2$, which is compact and has a piecewise smooth and positively orientated (counterclockwise) boundary curve C, also indicated with ∂A. If F is a continuously differentiable* ***two-dimensional*** *vector field defined on a neighborhood of V, then the* ***divergence theorem*** *states*

$$\iint_A (\nabla \cdot \vec{F})\, dA = \iint_A div\, \vec{F}\, dA = \oint_C (\vec{F} \cdot \vec{n}_0)\, ds,$$

where $\nabla \cdot \vec{F}$ is the divergence on the 2-dimensional vector field $\vec{F}$ and $\vec{n}_0$ is the outward-pointing unit normal vector on the boundary.

(ii) *Let the volume V be a subset of $\mathbb{R}^3$, which is compact and has a piecewise smooth boundary S, also indicated with ∂V. If F is a continuously differentiable* ***three-dimensional*** *vector field defined on a neighborhood of V, then the* ***divergence theorem*** *states*

$$\iiint_V (\nabla \cdot \vec{F})\, dV = \iiint_V div\vec{F}\, dV = \oiint_S (\vec{F} \cdot \vec{n})\, dS,$$

where $\nabla \cdot \vec{F}$ is the divergence on the 3-dimensional vector field $\vec{F}$ and $\vec{n}_0$ is the outward-pointing unit normal vector on the boundary surface.

5.3 Remarks/Examples

(i) *In case of the 2-dimensional theorem, the left side is a surface integral over the area A, the right side is the line integral over the boundary of the area A.*

(ii) *In case of the 3-dimensional theorem, the left side is a volume integral over the volume V, the right side is the surface integral over the boundary of the volume V.*

(iii) *In both cases the left hand side represents the total of the sources and sinks in the region, whereas the right hand side represents the total flow across the boundary of the region.*

(iv) *As an example we consider the vector field $\vec{F}(x, y, z) = (2x,\ y^2,\ z^2)$ and we try to calculate the flux through the surface of the unit sphere*

$$S = \left\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\right\},$$

i.e., we wish to evaluate

$$\oiint_S \vec{F}\, d\vec{S} = \oiint_S \vec{F}\vec{n}\, dS.$$

The direct computation of this integral is quite difficult, but we can simplify the derivation of the result using the unit ball W

$$W = \left\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 \leq 1\right\},$$

and the divergence theorem, because the divergence theorem says that the integral is equal to:

$$\begin{aligned}\oiint_S \vec{F}\, d\vec{S} &= \iiint_W (\nabla \cdot \vec{F})\, dV \\ &= 2 \iiint_W (1 + y + z)\, dV \\ &= 2 \iiint_W 1\, dV + 2 \iiint_W y\, dV + 2 \iiint_W z\, dV\end{aligned}$$

Since the function y is positive in one hemisphere of W and negative in the other, in an equal and opposite way, its total integral over W is zero. The same is true for z:

$$\iiint_W y\, dV = \iiint_W z\, dV = 0.$$

Therefore,

$$\oiint_S \vec{F} \cdot \vec{n}\, dS = 2 \iiint_W dV = \frac{8\pi}{3}$$

because the unit ball W has volume $\frac{4\pi}{3}$.

(v) We reconsider example VI.4.5(ii) *and calculate the flux $\Phi_C(\vec{F})$ through the barrel of the cone by means of the divergence theorem. For that purpose, we calculate the divergence of the vector field $\vec{F} = (z,\ y,\ z+1)$:*

$$div\, \vec{F} = \nabla \cdot \vec{F} = \frac{\partial z}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial (z+1)}{\partial z} = 0 + 1 + 1 = 2.$$

Then the calculation of the flux reduces to

$$\int_C div\vec{F}\, dV = 2 \cdot \int_C dV = 2 \cdot \left(\tfrac{1}{3}\pi 2^2 \cdot 2\right) =$$

where we used the formula for the volume of a cone

$$V_C = \frac{1}{3}\pi r^2 h$$

with $r = 2$ and $h = 2$. Remember, a volume integral with the integrand 1 is simply the volume of the body (compare VI.2.3(v)*).*

5.4 Discussion (Curl theorem)
The curl theorem, also known as Kelvin-Stokes theorem (named for Lord Kelvin and George Stokes), is a theorem in vector calculus on $\mathbb{R}^3$. Given a vector field, the theorem relates the integral of the curl of the vector field over some surface, to the line integral of the vector field tangential to the closed boundary of the surface.
It is a helpful tool, since the curve integral is in general easier to solve than the surface integral.

5.5 Definition (Curl theorem)
Let the Volume V be an open subset of the $\mathbb{R}^3$, i.e. $V \subset \mathbb{R}^3$, and $\vec{F}$: $V \to \mathbb{R}^3$ be a continuously differentiable vector field defined on V. Besides, let $\Sigma \subset V$ be a two-dimensional piecewise smooth surface contained in V and orientated by a unit normal vector field $\vec{n}$. Then the ***curl theorem*** *states*

$$\iint_\Sigma curl\vec{F} \cdot\, d\vec{S} = \iint_\Sigma (\nabla \times \vec{F}) \cdot\, d\vec{S} = \oint_{\partial\Sigma} \vec{F} \cdot\, d\vec{r} \quad or$$

$$\iint_\Sigma curl\vec{F} \cdot \vec{n}\, dS = \iint_\Sigma (\nabla \times \vec{F}) \cdot \vec{n}\, dS = \oint_{\partial\Sigma} \vec{F} \cdot \vec{\tau}\, dr,$$

where $\nabla \times \vec{F}$ is the curl of the 3-dimensional vector field $\vec{F}$, $\vec{n}$ is the outward-pointing unit normal vector on the boundary surface and τ is a unit tangent vector of the boundary curve $\partial\Sigma$ of the surface Σ.

5.6 Remarks/Examples

(i) *Like the divergence theorem the curl theorem can be interpreted physically as well. The flux of a curl of a vector field through a surface equals the circulation of the field along the boundary curve of the surface.*

(ii) *We verify the curl theorem by calculating both the surface integral and the line integral for the vector field* $\vec{F} = (xy,\ yz,\ xz)$ *and a quarter of the unit sphere (compare* Figure VI.13*)*

$$\Sigma = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1 \ , y, z \geq 0\}.$$

In analogy to polar coordinates we utilize spherical coordinates to solve the problem.

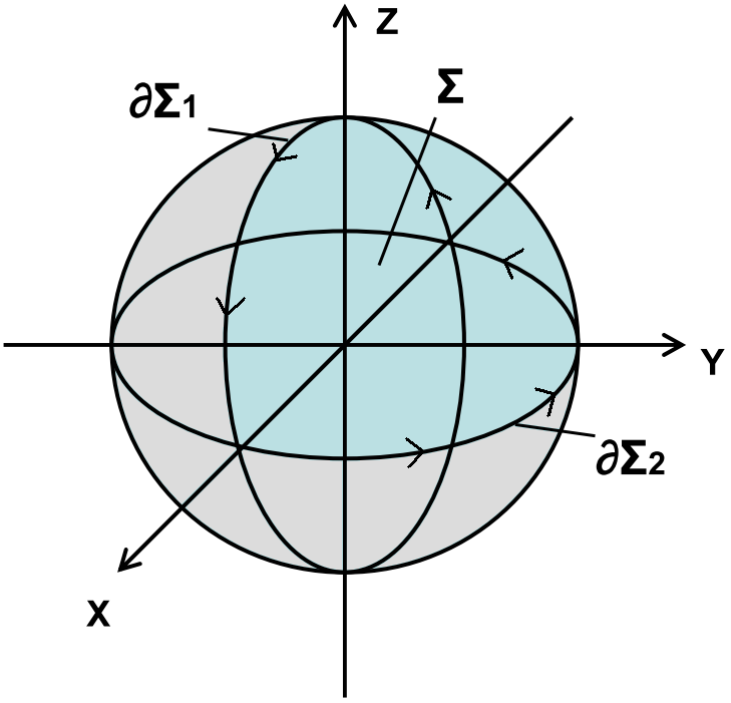


Figure VI.13: Quarter sphere Σ with boundary curves $\partial\Sigma_1$, $\partial\Sigma_2$

In this case a point P *can be described by the radius* r *and the angles* φ *and* θ *(compare* Figure VI.14*) and the relation to the cartesian coordinates* x*,* y *and* z *is*

$$\begin{aligned} x &= r \cdot \cos(\theta) \cdot \cos(\varphi) \\ y &= r \cdot \cos(\theta) \cdot \sin(\varphi) \\ z &= r \cdot \sin(\theta). \end{aligned}$$

To verify the curl theorem we calculate both the surface integral and the line integral for the vector field and compare the results.

1.) *We calculate the flux integral of the curl using spherical coordinates with radius* $r = 1$ *and* $\Sigma := [0, \frac{\pi}{2}] \times [0, \pi]$.

$$\vec{x} = \begin{pmatrix} \cos(\theta) \cdot \cos(\varphi) \\ \cos(\theta) \cdot \sin(\varphi) \\ \sin(\theta) \end{pmatrix}$$

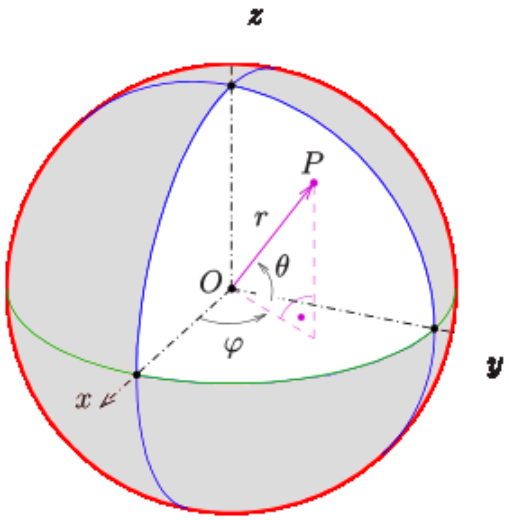


Figure VI.14: Illustration of spherical coordinates

$$\vec{x}_\theta = \begin{pmatrix} -\sin(\theta)\cdot\cos(\varphi) \\ -\sin(\theta)\cdot\sin(\varphi) \\ \cos(\theta) \end{pmatrix}$$

$$\vec{x}_\varphi = \begin{pmatrix} -\cos(\theta)\cdot\sin(\varphi) \\ \cos(\theta)\cdot\cos(\varphi) \\ 0 \end{pmatrix}$$

$$\vec{x}_\theta \times \vec{x}_\varphi = \begin{pmatrix} -\cos^2(\theta)\cdot\cos(\varphi) \\ -\cos^2(\theta)\cdot\sin(\varphi) \\ -\sin(\theta)\cos(\theta) \end{pmatrix} = -\cos(\theta)\cdot\vec{x}$$

$$curl\vec{F} = \nabla\times\vec{F} = \begin{pmatrix} -y \\ -z \\ -x \end{pmatrix} = \begin{pmatrix} -\cos(\theta)\cdot\sin(\varphi) \\ -\sin(\theta) \\ -\sin(\theta)\cdot\cos(\varphi) \end{pmatrix}$$

$$\iint_\Sigma curl\vec{F}\,d\vec{S} = \iint_\Sigma (\nabla\times\vec{F})\cdot(\vec{x}_\theta\times\vec{x}_\varphi)\,d(\theta,\varphi)$$

$$= \iint_\Sigma \begin{pmatrix} -\cos(\theta)\cdot\sin(\varphi) \\ -\sin(\theta) \\ -\sin(\theta)\cdot\cos(\varphi) \end{pmatrix} \cdot \begin{pmatrix} \cos^2(\theta)\cdot\cos(\varphi) \\ \cos^2(\theta)\cdot\sin(\varphi) \\ \cos(\theta)\cos(\theta) \end{pmatrix} d(\theta,\varphi)$$

$$= \int_0^\pi \int_0^{\frac{\pi}{2}} \big(\cos^2(\theta)\cos(\varphi)\sin(\varphi) + \cos(\theta)\sin(\theta)(\cos(\varphi)+\sin(\varphi))\big)\cdot\cos(\theta)\,d\theta\,d\varphi$$

$$= \int_0^{\frac{\pi}{2}} \cos^3(\theta)\,d\theta \cdot \int_0^\pi \cos(\varphi)\sin(\varphi)\,d\varphi$$

$$+ \int_0^{\frac{\pi}{2}} \cos^2(\theta)\sin(\theta)\,d\theta \cdot \int_0^\pi (\cos(\varphi)+\sin(\varphi))\,d\varphi$$

$$= \frac{2}{3}\cdot 0 + \left[-\tfrac{1}{3}\cos^3(\varphi)\right]_0^{\frac{\pi}{2}} \cdot [\sin(\varphi)-\cos(\varphi)]_0^\pi$$

$$= \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 2 = \frac{2}{3}$$

2.) We calculate the circulation along the boundary curves $\partial\Sigma_1$ and $\partial\Sigma_2$. The parametrization of the curves is:

$$\begin{aligned} \partial\Sigma_1 : \quad & \vec{x} = (-\cos(t), \sin(t), 0) \quad \textit{with} \quad 0 \le t \le \pi \\ \partial\Sigma_2 : \quad & \vec{x} = (\cos(t), 0, \sin(t)) \quad \textit{with} \quad 0 \le t \le \pi \end{aligned}$$

$$\oint_{\partial\Sigma} \vec{F}\, d\vec{S} = \int_{\partial\Sigma_1} \vec{F}\, d\vec{S} + \int_{\partial\Sigma_2} \vec{F}\, d\vec{S}$$

$$\begin{aligned} \int_{\partial\Sigma_1} \vec{F}\, d\vec{S} &= \int_{\partial\Sigma_1} \vec{F}(\vec{x}(t)) \cdot \vec{x}'(t)\, dt \\ &= \int_0^{\pi} (-\cos(t)\sin(t),\ \ 0,\ \ 0) \cdot (\sin(t),\ \ \cos(t),\ \ 0)\, dt \\ &= -\int \sin^2(t)\cos(t)\, dt \\ &= \left[\tfrac{1}{3}\sin^3(t)\right]_0^{\pi} = 0 \end{aligned}$$

$$\begin{aligned} \int_{\partial\Sigma_2} \vec{F}\, d\vec{S} &= \int_{\partial\Sigma_2} \vec{F}(\vec{x}(t))\vec{x}'(t)\, dt \\ &= \int_0^{\pi} (0,\ \ 0,\ \ \cos(t)\sin(t)) \cdot (-\sin(t),\ \ 0,\ \ \cos(t))\, dt \\ &= \int_0^{\pi} \cos^2(t)\sin(t)\, dt \\ &= \left[\tfrac{1}{3}\cos^3(t)\right]_0^{\pi} = \frac{2}{3} \end{aligned}$$

$$\Rightarrow \quad \oint_{\partial\Sigma} \vec{F}\, d\vec{S} = \int_{\partial\Sigma_1} \vec{F}\, d\vec{S} + \int_{\partial\Sigma_2} \vec{F}\, d\vec{S} = \frac{2}{3}$$

Chapter VII

Systems of ordinary differential equations

1 Introduction

Last semester we already discussed ordinary differential equations and their importance in applications. Then, the main task was to find a function

$$y : \mathbb{R} \to \mathbb{R}$$

satisfying

$$y^{(n)}(t) = F\left(t, y(t), y'(t), \ldots, y^{(n-1)}(t)\right)$$

for some given multi-variable function $F : \mathbb{R}^{n+1} \to \mathbb{R}$. In industrial applications, there is typically more than one unknown function to determine, that is, one is actually interested in a vector function

$$y : \mathbb{R} \to \mathbb{R}^n, \;\; t \mapsto \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix}. \tag{1.1}$$

Moreover, the behavior of the unknown functions is not separated, but they influence, disturb and govern each other. The outcome is a *system of ordinary differential equations* that needs to be solved. If only first order derivatives are involved, it has the form

$$\left.\begin{array}{rcl} y_1'(t) &=& f_1(t, y_1(t), \ldots, y_n(t)) \\ \vdots & \vdots & \vdots \\ y_n'(t) &=& f_n(t, y_1(t), \ldots, y_n(t)) \end{array}\right\} \quad y(t) = F(t, y(t)), \tag{1.2}$$

where now $F : \mathbb{R}^{n+1} \to \mathbb{R}^n$ and y is as in Equation (1.1).
We want to motivate such systems by the following example from medicine.

We consider the following model describing how drugs are metabolized by the human body. A certain part of the drug is transferred from the bloodstream to the tissue, while

another part is taken to the outside world. On the other side, there is a part that is passed back to the bloodstream again from the tissue. The situation is illustrated in Figure VII.1,

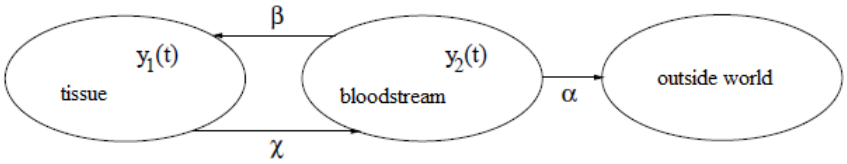


Figure VII.1: Metabolism of a drug

where $y_1(t)$, $y_2(t)$ are the amounts of the drug in the tissue and bloodstream at time t, whereas α, β and $\chi > 0$ are the transfer constants, i.e., at time t, the change in the bloodstream is that $\beta y_2(t)$ go to the tissue and $\alpha y_2(t)$ are flooded out, while $\chi y_1(t)$ are coming in from the tissue. The change in the tissue at time t is that $\beta y_2(t)$ is added, while $\chi y_1(t)$ is going to the bloodstream. This leads to the following system of ODEs:

$$\begin{aligned} y_1'(t) &= \beta y_2(t) - \chi y_1(t), && t \geq 0, \\ y_2'(t) &= \chi y_1(t) - \alpha y_2(t) - \beta y_2(t), && t \geq 0. \end{aligned}$$

Please note that this system can be rewritten in matrix form

$$\underbrace{\begin{pmatrix} y_1'(t) \\ y_2'(t) \end{pmatrix}}_{=:y'(t)} = \underbrace{\begin{pmatrix} -\chi & \beta \\ \chi & -(\alpha+\beta) \end{pmatrix}}_{=:A} \underbrace{\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}}_{=:y(t)}.$$

The interesting questions we should answer now are the following

(i) What are the amounts of the drug contained in the bloodstream and tissue after some time t? This means to solve the system by determining $y_1(t)$ and $y_2(t)$.

(ii) Say we have given initial concentrations of the drug in the bloodstream and the tissue. Is the drug leaving the body eventually or do we stay polluted?

This is only a very tiny toy example. The following industrial applications are dealing with real-world systems where systems of differential equations appear in XXL-dimensions:

(i) To carry out simulations on a computer such as crash test simulations, fluid motion calculations or nuclear weapon tests.

(ii) In game physics, the motion of an object (i.e., its position as it moves in time) is calculated by solving differential equations.

(iii) The behavior of components of an electrical network and its interaction among each other is modeled by differential equation systems. In order to simulate the evolution of the electrical network over time a large system of differential equations has to be solved. This is used to examine the system on stability.

In this chapter we want to lay some theoretical groundwork for such systems. In particular, we discuss the important application case of systems of linear ODEs. For its solution, we will introduce the matrix exponential function and discuss its characteristic properties. The chapter will be closed with an overview of numerical methods to solve ODE-systems.

2 Systems of linear ODEs

In this section we will discuss systems of ODEs that are linear in constant coefficients. Even more, we will focus only on first order ODEs, which, at first sight, form only a small subset of linear ODE systems. The reason for this restriction is not that higher order linear ODE-systems are not interesting, it is rather the intriguing result, that all higher order (linear) ODE-systems can be transformed into a system of first order ODEs!
We start with the definition of a linear ODE system. For notational reasons only, we further restrict ourselves to linear ODEs with constant coefficients. Further, for the rest of the chapter we denote by $\mathbb{I}$ a real interval, $\mathbb{R}$ or $\mathbb{R}_+$.

2.1 Definition
Let $A \in M_{n\times n}(\mathbb{R})$ and y be differentiable such that

$$y : \mathbb{I} \to \mathbb{R}^n,\ t \mapsto \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix}, \quad y' : \mathbb{I} \to \mathbb{R}^n,\ t \mapsto \begin{pmatrix} y_1'(t) \\ \vdots \\ y_n'(t) \end{pmatrix}.$$

*Further, let $B : \mathbb{I} \to \mathbb{R}^n$ be another function, called the **inhomogeneity**. Then the equation system*

$$y'(t) = Ay(t) + B(t),\ t \in \mathbb{I}, \tag{2.1}$$

*is called a **first order, linear system of ODEs with constant coefficients** (given by A). If $B \equiv \mathbb{0}$, the system is called **homogeneous**, otherwise **inhomogeneous**.*

The first interesting observation is now that any linear ODE of order n (as discussed last semester) can be reduced to a first order ODE system.

2.2 Theorem
Let $a_k \in \mathbb{R}$ $(k = 0, \dots, n-1)$, $b : \mathbb{I} \to \mathbb{R}$ and consider the ODE

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + a_{n-2}y^{(n-2)}(t) + \ldots + a_0 y(t) = b(t),\ t \in \mathbb{I}. \tag{2.2}$$

Further, consider the (first order) system of n ODEs

$$Y'(t) = AY(t) + B(t),\ t \in \mathbb{I}. \tag{2.3}$$

where

$$A := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & -a_{n-2} & -a_{n-1} \end{pmatrix}, \quad B(t) := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}.$$

If $Y : \mathbb{I} \to \mathbb{R}^n$ is a solution to (2.3), then its first coordinate Y_1 is a solution to (2.2).

Proof.
From the first $n-1$ lines of (2.3), we see that

$$\begin{aligned} Y_1'(t) &= Y_2(t) \\ Y_1''(t) &= Y_2'(t) = Y_3(t) \\ Y_1^{(3)}(t) &= Y_3'(t) = Y_4(t) \\ &\vdots \\ Y_1^{(n-1)}(t) &= Y_{n-1}'(t) = Y_n(t) \end{aligned}$$

Together with the last row, this yields

$$\begin{aligned} Y_1^{(n)}(t) &= Y_n'(t) \\ &= -a_0 Y_1(t) - a_1 \underbrace{Y_2(t)}_{Y_1'(t)} - a_2 \underbrace{Y_3(t)}_{Y_1''(t)} - \ldots - a_{n-1} \underbrace{Y_n(t)}_{Y_1^{(n-1)}(t)} + b(t), \end{aligned}$$

which means that Y_1 is a solution to the original system (2.2). □

2.3 Remarks/Examples

(i) *The theorem is based on the transformation*

$$y(t) \longleftrightarrow Y(t) = \big(y(t), y'(t), y''(t), \ldots, y^{(n-1)}(t)\big)$$

and it is easy to check that solutions of (2.2) *yield solutions of* (2.3) *by this transformation as well. The systems are therefore equivalent.*

(ii) *By Laplace expansion we see that the characteristic polynomial of A in* (2.3) *is*

$$\chi_A(\lambda) = \det(\lambda \mathrm{Id}_n - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots a_1\lambda^1 + a_0,$$

If we set this to zero we obtain the characteristic equation of the ODE (2.2)*, which has its name from this connection. The eigenvalues of A are thus used to setup the fundamental system of the homogeneous version of* (2.2)*.*

(iii) *If the coefficients a_k are functions, then the transformation holds as well. We simply have to replace A by*

$$A(t) := \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 1 \\ -a_0(t) & -a_1(t) & \ldots & -a_{n-2}(t) & -a_{n-1}(t) \end{pmatrix}.$$

(iv) The Bessel differential equation (Friedrich Bessel, 1784 – 1846*, German mathematician) is*

$$y''(t) + \frac{1}{t}y'(t) + \left(1 - \frac{\alpha^2}{t^2}\right) y(t) = 0, \qquad t > 0,$$

where $\alpha \in \mathbb{R}$ is a given constant. Its solutions are called Bessel functions which are important for many problems of wave propagation and static potentials (e.g. heat conduction in a cylindrical object, solutions to the radial Schrödinger equation for a free particle, FM synthesis, Kaiser windows or Bessel filters). We are dealing here with a 2nd order linear ODE with

$$a_1(t) = \frac{1}{t}, \;\; a_0(t) = 1 - \frac{\alpha^2}{t^2}, \;\; b(t) = 0.$$

So, in order to solve it, we can reduce it to a first order system of ODEs using the above transformation:

$$\begin{pmatrix} Y_1'(t) \\ Y_2'(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{\alpha^2}{t^2} - 1 & -\frac{1}{t} \end{pmatrix} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix}.$$

The first coordinate of the solutions to this system give the Bessel functions, i.e., the solutions to the Bessel ODE.

The last theorem showed us that an n-th order ODE with a one-dimensional solution can be solved using a first order ODE-system where the solution is n-dimensional. This embedding into a larger domain is also the idea for reducing an n-th order system of ODEs to a first order system.

2.4 Theorem
Let $A_k \in M_{m\times m}$ $(k = 0, \dots, n-1)$, $b : \mathbb{I} \to \mathbb{R}^m$ and consider the ODE(-system)

$$y^{(n)}(t) + A_{n-1}y^{(n-1)}(t) + A_{n-2}y^{(n-2)}(t) + \cdots + A_0 y(t) = b(t), \;\; t \in \mathbb{I}. \tag{2.4}$$

Further, consider the (first order) system of $n \cdot m$ ODEs

$$Y'(t) = AY(t) + B(t), \;\; t \in \mathbb{I}. \tag{2.5}$$

where

$$A := \begin{pmatrix} 0_m & \mathrm{Id}_m & 0_m & \dots & 0_m \\ \vdots & \ddots & \mathrm{Id}_m & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0_m \\ 0_m & \dots & \dots & 0_m & \mathrm{Id}_m \\ -A_0 & -A_1 & \dots & -A_{n-2} & -A_{n-1} \end{pmatrix}, \;\; B(t) := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}.$$

If $Y : \mathbb{I} \to \mathbb{R}^{nm}$ is a solution to (2.5)*, then its first m coordinates $(Y_1, \dots, Y_m)$ form a solution to* (2.4)*.*

2.5 Example
Consider the coupled vibratory system depicted in Figure VII.2.

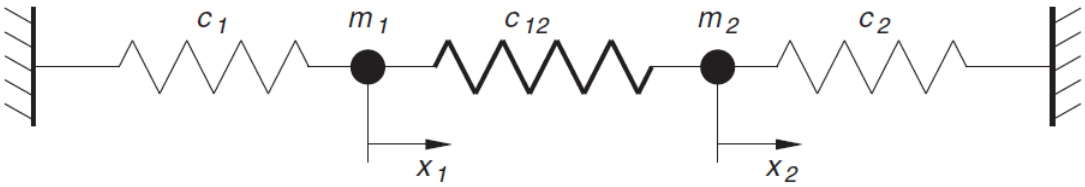


Figure VII.2: Model of vibratory system

Here, m_1, m_2 are masses, c_1, c_2 and c_{12} are spring constants and $x_1, x_2 : \mathbb{R}_+ \to \mathbb{R}$ are the horizontal amplitudes of the masses at time t. This system can be described by the second order linear ODE system

$$x_1''(t) + \frac{c_1}{m_1} x_1(t) + \frac{c_{12}}{m_1}(x_1(t) - x_2(t)) = 0$$

$$x_2''(t) + \frac{c_2}{m_2} x_2(t) + \frac{c_{12}}{m_2}(x_2(t) - x_1(t)) = 0$$

using Hook's and Newton's law, which can be written in matrix form as

$$\underbrace{\begin{pmatrix} x_1''(t) \\ x_2''(t) \end{pmatrix}}_{=y''(t)} + \underbrace{\begin{pmatrix} \frac{c_1+c_{12}}{m_1} & -\frac{c_{12}}{m_1} \\ -\frac{c_{12}}{m_2} & \frac{c_2+c_{12}}{m_2} \end{pmatrix}}_{=A_0} \underbrace{\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}}_{=y(t)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Using Theorem 2.4*, we can rewrite this system as the (equivalent) first order ODE system*

$$\begin{pmatrix} Y_1'(t) \\ Y_2'(t) \\ Y_3'(t) \\ Y_4'(t) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{c_1+c_{12}}{m_1} & \frac{c_{12}}{m_1} & 0 & 0 \\ \frac{c_{12}}{m_2} & -\frac{c_2+c_{12}}{m_2} & 0 & 0 \end{pmatrix} \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ Y_3(t) \\ Y_4(t) \end{pmatrix}.$$

The first two coordinates of a solution $Y = (Y_1, Y_2, Y_3, Y_4)$ yield the solution to the vibrating system.

As a result of this section we can focus on solving linear systems of ODEs of order one. The main tool is provided in the next section.

3 The exponential function

Linear (homogeneous) ODE systems for $n = 1$ are of the form

$$y'(t) = Ay(t), \;\; t \geq 0,$$

where A is a **real number** and it is well-known that the exponential function

$$y : \mathbb{R}_+ \to \mathbb{R}, \;\; t \mapsto y(t) = Ce^{tA}$$

is the solution to this ODE. As it will turn out, a multi-dimensional version of the exponential function will solve linear systems of ODEs. We recall the well-known properties of the one-dimensional exponential in the next discussion.

3.1 Discussion (1D-exponential function)
For $A \in \mathbb{R}$, the one-dimensional exponential function can be defined by

$$\exp : \mathbb{R} \to \mathbb{R}, \;\; t \mapsto exp(tA) = e^{tA} := \sum_{k=0}^{\infty} \frac{A^k}{k!} t^k.$$

Then, the following holds true:

(i) $\exp(\cdot)$ *is a power series with radius of convergence $\rho = \infty$, i.e., $e^{tA} \in \mathbb{R}$ for all $t \in \mathbb{R}$ (compare last semester).*

(ii) $\dfrac{d}{dt} e^{tA} = A \cdot e^{tA}$ *for all $A \in \mathbb{R}$ and $t \in \mathbb{R}$, because*

$$\left(\sum_{k=0}^{\infty} \frac{A^k}{k!} t^k \right)' = \sum_{k=1}^{\infty} \frac{A^k}{k!} k t^{k-1} = A \sum_{k=1}^{\infty} \frac{A^{k-1}}{(k-1)!} t^{k-1} = A \cdot exp(tA)$$

(iii) $e^{t(A+B)} = e^{tA} e^{tB}$ *and* $e^{(t+s)A} = e^{tA} e^{sA}$ *(functional equation).*

(iv) $e^{0 \cdot A} = 1$ *and* $\left(e^{tA}\right)^{-1} = e^{-tA}$, *i.e.,* $\exp(tA)$ *is always invertible.*

(v) $$e^{t(T^{-1}AT)} = \sum_{k=0}^{\infty} \frac{(T^{-1}AT) \cdot (T^{-1}AT) \cdots (T^{-1}AT)}{k!} t^k = T^{-1} \left(\sum_{k=0}^{\infty} \frac{A^k}{k!} t^k \right) T = T^{-1} e^{tA} T.$$

The named properties are precisely the ones that we wish to find in a generalization. When we replace the real number A by a matrix, we arrive at the matrix exponential function.

3.2 Definition
Let $A \in M_{n \times n}(\mathbb{R})$. We define the ***matrix exponential function*** *by*

$$\exp : \mathbb{R} \to M_{n \times n}(\mathbb{R}), \;\; t \mapsto \exp(tA) := \sum_{k=0}^{\infty} \frac{A^k}{k!} t^k,$$

where $A^0 := \mathrm{Id}_n$.

This exponential shares the same properties with the ordinary 1D-exponential observed in Discussion 3.1. They are summarized in the next theorem.

3.3 Theorem

(i) *The sequence (of matrices)* $\left(\sum_{k=0}^{N} \frac{A^k}{k!} t^k\right)_{N\in\mathbb{N}}$ *converges to a matrix in* $M_{n\times n}(\mathbb{R})$ *w.r.t. the maximum norm (for matrices) given by*

$$\|B\| := \max_{i,j=1,\dots n} |b_{ij}|.$$

Therefore $\exp(tA) \in M_{n\times n}(\mathbb{R})$ *for all* $t \in \mathbb{R}$ *and* $A \in M_{n\times n}(\mathbb{R})$.

(ii) *The function* $\exp(\cdot)$ *is differentiable w.r.t.* t *and we have*

$$\frac{d}{dt} e^{tA} = A \cdot e^{tA}.$$

(iii) $e^{t(A+B)} = e^{tA} \cdot e^{tB}$, *if* $AB = BA$. *In particular, the functional equation holds:*

$$\exp((t+s)A) = \exp(tA) \cdot \exp(tA).$$

(iv) $\exp(0 \cdot A) = \mathrm{Id}_n$ *and* $\exp(tA)^{-1} = \exp(-tA)$, *i.e.,* $\exp(tA)$ *is an invertible matrix for all* $t \in \mathbb{R}$ *and* $A \in M_{n\times n}(\mathbb{R})$.

(v) $\exp(t(TAT^{-1})) = T\exp(tA)T^{-1}$ *for an invertible (transformation) matrix* $T \in M_{n\times n}(\mathbb{R})$.

We immediately show some examples.

3.4 Examples

(i) *Let* D *be a diagonal matrix* $D = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. *We calculate the matrix exponential as*

$$\begin{aligned}
\exp(tD) &= \sum_{k=0}^{\infty} \frac{D^k}{k!} t^k \\
&= \sum_{k=0}^{\infty} \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}^k \frac{t^k}{k!} = \sum_{k=0}^{\infty} \begin{pmatrix} \lambda_1^k & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n^k \end{pmatrix} \frac{t^k}{k!} \\
&= \sum_{k=0}^{\infty} \begin{pmatrix} \frac{t^k}{k!}\lambda_1^k & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{t^k}{k!}\lambda_n^k \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^{\infty} \frac{t^k}{k!}\lambda_1^k & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sum_{k=0}^{\infty} \frac{t^k}{k!}\lambda_n^k \end{pmatrix} \\
&= \begin{pmatrix} e^{t\lambda_1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & e^{t\lambda_n} \end{pmatrix} = \mathrm{diag}(e^{t\lambda_1}, \dots, e^{t\lambda_n}).
\end{aligned}$$

One particular example is:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \quad \rightsquigarrow \quad \exp(tA) = \begin{pmatrix} e^t & 0 \\ 0 & e^{3t} \end{pmatrix}$$

(ii) *A matrix A is called diagonizable, if an invertible matrix T and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ exist such that*

$$A = TDT^{-1}.$$

A consequence of the famous spectral theorem is that all symmetric matrices are diagonizable! Moreover, the diagonalized version will hold the eigenvalues (according to multiplicities) on the diagonal and the transformation matrix T will hold the corresponding eigenvectors.
By Theorem 3.3(v)*, it is easy to calculate the matrix exponential of such a matrix:*

$$\exp(tA) = \exp\left(t\left(TDT^{-1}\right)\right) = T\exp(tD)T^{-1} = T\operatorname{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n})T^{-1}.$$

As a concrete example, we consider

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{=T} \cdot \underbrace{\begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}}_{=D} \cdot \underbrace{\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{=T^{-1}}$$

and calculate the matrix exponential as

$$\begin{aligned} exp(tA) &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{pmatrix} \cdot \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} e^{3t} & e^{3t} \\ e^{-t} & -e^{-t} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} e^{3t} + e^{-t} & e^{3t} - e^{-t} \\ e^{3t} - e^{-t} & e^{3t} + e^{-t} \end{pmatrix}. \end{aligned}$$

As a final result, we obtain that the matrix exponential serves to solve the initial value problem corresponding to a linear ODE system.

3.5 Corollary
The initial value problem corresponding to a linear homogeneous ODE-system with constant coefficients, i.e.,

$$\left\{ \begin{array}{rcl} y'(t) & = & Ay(t), \quad t \in \mathbb{R}_+, \\ y(0) & = & y_0, \end{array} \right. \tag{3.1}$$

where $A \in M_{n\times n}(\mathbb{R})$, $y_0 \in \mathbb{R}^n$ has the unique solution

$$\hat{y}(t) : \mathbb{R}_+ \to \mathbb{R}^n, \ \ t \mapsto \hat{y}(t) := \exp(tA)y_0.$$

Proof.
We only show that $\hat{y}$ is a solution of (3.1) here. By Theorem 3.3(ii) and (iv) we have

$$\begin{aligned} \hat{y}'(t) &= \frac{d}{dt}\exp(tA)y_0 = A\exp(tA)y_0 = A\hat{y}(t), \\ \hat{y}(0) &= \exp(0 \cdot A)y_0 = \mathrm{Id}_n y_0 = y_0, \end{aligned}$$

and consequently $\hat{y}$ is a solution to the initial value problem. □

3.6 Example
We consider the initial value problem

$$\begin{aligned} y_1'(t) &= y_1(t) + y_2(t), && t \geq 0, \\ y_2'(t) &= y_1(t) + y_2(t), && t \geq 0, \\ y_3'(t) &= y_3(t), && t \geq 0, \\ y_1(0) &= y_2(0) = y_3(0) = 1, \end{aligned}$$

which can be rewritten in matrix form by

$$\underbrace{\begin{pmatrix} y_1'(t) \\ y_2'(t) \\ y_3'(t) \end{pmatrix}}_{=y'(t)} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix}}_{=y(t)}$$

$$y_0 = (1,1,1)^\top.$$

The matrix A is symmetric with eigenvalues $\lambda_1 = 0$, $\lambda_2 = 2$, $\lambda_3 = 1$. Therefore, it is diagonizable using a transformation matrix T. In this case, the following equation holds true

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{=T} \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{=D} \underbrace{\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{=T^{-1}}.$$

In order to solve our system of ODEs, we need to calculate the matrix exponential:

$$\begin{aligned} \exp(tA) &= \exp\left(t\left(TDT^{-1}\right)\right) = T\exp(tD)T^{-1} \\ &= \begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{0t} & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{1t} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{e^{2t}+1}{2} & \frac{e^{2t}-1}{2} & 0 \\ \frac{e^{2t}-1}{2} & \frac{e^{2t}+1}{2} & 0 \\ 0 & 0 & e^t \end{pmatrix}. \end{aligned}$$

Consequently, according to Corollary 3.5*, the unique solution of the initial value problem considered is*

$$\hat{y} : \mathbb{R}_+ \to \mathbb{R}^3,\ t \mapsto \exp(tA)y_0 = \begin{pmatrix} \frac{e^{2t}+1}{2} & \frac{e^{2t}-1}{2} & 0 \\ \frac{e^{2t}-1}{2} & \frac{e^{2t}+1}{2} & 0 \\ 0 & 0 & e^t \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{2t} \\ e^t \end{pmatrix}.$$

4 Numerical methods for ODEs

In general, differential equations appearing in applications cannot be as easily solved by the presented method in the previous Section. For non-linear ODEs a closed form solution may not even be possible. With numerical methods, one can, however, still construct the solution at discrete points and interpolate the complete curve afterwards.
In this section we will show some famous methods applied in practice. The problem, we are trying to solve is always the following initial value problem corresponding to a first order ODE(-system)

$$\left\{ \begin{array}{rcl} y'(t) & = & f(t, y(t)), \quad t \in [a, b], \\ y(0) & = & y_0, \end{array} \right. \tag{4.1}$$

where $y : [a, b] \to \mathbb{R}^n$, $y_0 \in \mathbb{R}^n$ and $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$.

4.1 Discussion (Explicit Euler method)
Every numerical method starts with a discretization of the domain $[a, b]$, in our case we decompose it in N equidistant intervals of size $h = (b - a)/N$ and define

$$x_j = a + jh \ \ (j = 0, \ldots, N).$$

The construction of an approximate solution in the points x_j is the job of the numerical method.

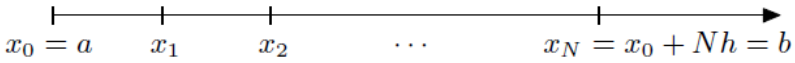


Figure VII.3: Discretization of $[a, b]$

To derive the explicit Euler method, we start at the point $x_0 = a$, where the function value $y(a) = y_0$ is already given. Then, we look at Taylor's formula to calculate $y(x_1)$ by

$$\begin{aligned} y(x_1) = y(a + h) &\approx y_1 = y(a) + hy'(a) \\ &= y(a) + hf(a, y(a)) = y_0 + hf(x_0, y_0), \end{aligned}$$

where we used the ODE(-system) (4.1). *We proceed in the same way to calculate $y(x_2)$, however, we don't have the exact function value $y(x_1)$ as we had $y(x_0)$ before, but only the approximation y_1. Since this is the best we have, we use it*

$$\begin{aligned} y(x_2) = y(x_1 + h) &\approx y(x_1) + hy'(x_1) = y(x_1) + hf(x_1, y(x_1)) \\ &\approx y_1 + hf(x_1, y_1). \end{aligned}$$

This leads to the explicit Euler method.

4.2 Algorithm (Explicit Euler method)
To solve the initial value problem (4.1)*, the explicit Euler method is executed as follows.*

(i) *Discretize the domain* $[a, b]$ *in* N *equidistant intervals of size* $h = (b - a)/N$ *and define*

$$x_j = a + jh \ \ (j = 0, \ldots, N).$$

(ii) *Calculate the approximate solution* $(y_0, \ldots, y_N)$ *of* (4.1) *at the discrete points* $(x_0, \ldots, x_N)$ *by*

$$y_{k+1} = y_k + hf(x_k, y_k).$$

Some remarks are necessary.

4.3 Remarks/Examples
(i) Consider the initial value problem

$$y'(t) = y(t), \ \ t \in [0, 1],$$
$$y(0) = 1.$$

Then, the explicit Euler method with step size $h = (1 - 0)/N$ *produces the steps*

$$\begin{aligned} y_1 &= y_0 + hy_0 = 1 + h \\ y_2 &= y_1 + hy_1 = (1 + h)y_1 = (1 + h)^2 \\ y_3 &= y_2 + hy_2 = (1 + h)y_2 = (1 + h)^3 \\ &\vdots \quad \vdots \end{aligned}$$

$$y_N = (1 + h)^N = \left(1 + \frac{1}{N}\right)^N \Rightarrow \lim_{N\to\infty} y_N = e.$$

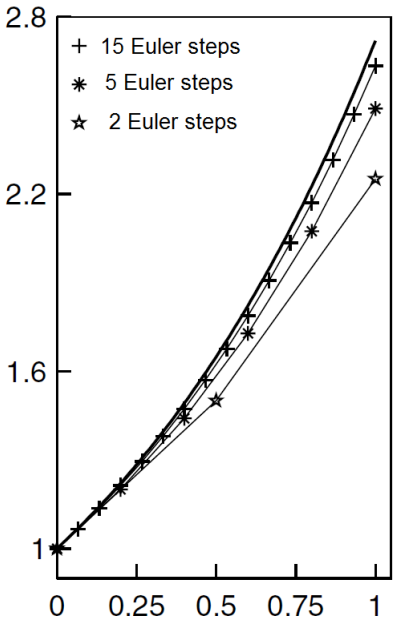


Figure VII.4: Explicit Euler for $y'(t) = y(t)$, y(0)=1

(ii) *A famous predator-prey model is given by the (non-linear) system*

$$x'(t) = (ay(t) - n) \cdot x(t), \quad t \in [a, b],$$
$$y'(t) = (d - cx(t)) \cdot y(t), \quad t \in [a, b],$$

*due to A.J. Lotka (*1880-1949*) and V. Volterra (*1860-1940*). We normalize all constants* $a = n = c = d = 1.0$ *and use the initial populations* $(2, 2)$ *to execute the explicit Euler method and derive the iteration procedure*

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + h \begin{pmatrix} (y_k - 1)x_k \\ y_k(1 - x_k) \end{pmatrix}.$$

The resulting orbit in the x-y-plane for different step sizes $h = \frac{1}{n}$ *is depicted in* Figure VII.5. *The correct orbit is periodic, which cannot even be observed when we use the (very small) step size* $h = \frac{1}{1000}$. *The nature of the problem does not seem to be correctly reflected by the explicit Euler method.*

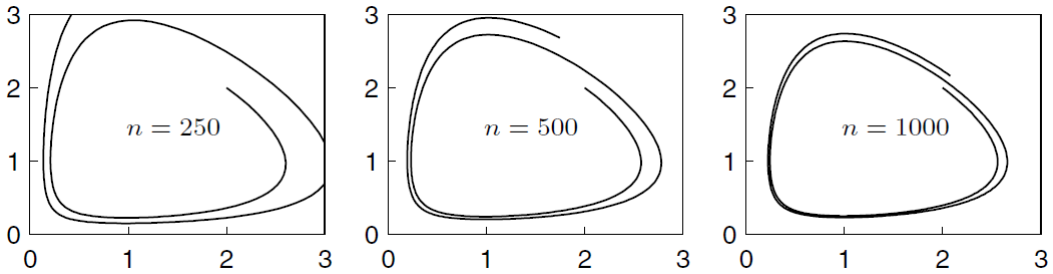


Figure VII.5: Explicit Euler applied to Lotka-Volterra model

(iii) *The explicit Euler method is not very accurate. To obtain a numerical solution with an acceptable accuracy, we have to use a very small step size h. A small step size h, however, implies a larger number of steps, thus more computing time. Therefore, it is desirable to develop methods that are more accurate than Euler's method.*
Commercial software packages implement methods of higher accuracy nowadays, like ***Runge-Kutta methods*** *or multiple-step methods. These methods also feature automatic error estimates and step size adaptions for high accuracy. We briefly discussed the very famous Runge-Kutta method of order 4 in the last semester. It can be found in* [Pap09] *or* [Mun06] *in more detail.*

Bibliography

[Kön00] K. Königsberger, *Analysis 2*, Springer-Verlag, 2000, 5. Auflage.

[Mun06] T. Munz, C.-D.; Westermann, *Numerische behandlung gewöhnlicher und partieller Differenzialgleichungenathematik - Ein interaktives Lehrbuch für Ingenieure*, Springer-Verlag, 2006, 3. Auflage.

[OO11] M. Oberguggenberger and A. Ostermann, *Analysis for Computer Scientists*, Foundations, Methods and Algorithms, Springer-Verlag, 2011, 1. Auflage.

[Pap09] L. Papula, *Mathematik für Ingenieure und Naturwissenschaftler, Band 2*, 12. Auflage ed., Vieweg+Teubner-Verlag, 2009.

[Str06] R. Stry, Y. und Schwenkert, *Mathematik kompakt für Ingenieure und Informatiker*, 2. Auflage ed., Springer-Verlag, 2006.

[Tes07] S. Teschl, G. und Teschl, *Mathematik für Informatiker*, 2. Auflage ed., Springer-Verlag, 2007.

[Wes08] T. Westermann, *Mathematik für Ingenieure - Ein anwendungsorientiertes Lehrbuch*, Springer-Verlag, 2008, 5. Auflage.