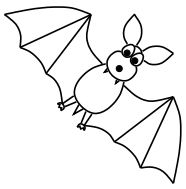
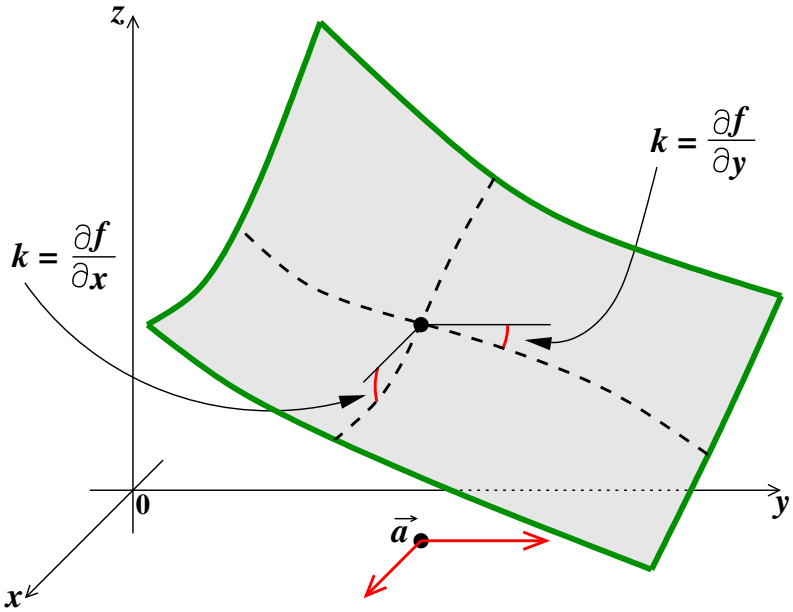


An Illustrated Introduction to Functions of More Variables

Petr Habala

habala@fel.cvut.cz
FEL, ČVUT Praha
CTU in Prague, FEE

2022



Introduction

The aim of these notes is to introduce important notions of the theory of functions of more variables so that the reader can understand them intuitively and appreciate the connection between mathematical formulas and geometry. It is therefore a complement, not a replacement, for traditional textbooks on multi-variable calculus that show theory and include proofs and practice problems, which is something that this text lacks entirely.

The notations is fairly standard, to be on the safe side we add an appendix with a summary of basic topological notions used when working in more-dimensional spaces. We also added an overview of basic analytic objects in the plane and space, because we will work with them a lot here.

Also the formatting is fairly standard. There is one peculiar convention used here, all Examples and Remarks are ended with the triangle you see below.
△

Contents

1. Introduction to functions of more variables

Functions of more variables are a natural generalization of functions of one variable. A function of one variable is a mapping (sort of a sending device) from one copy of real numbers (typically from some subset) to another copy. We obtain a function of more variables when we replace the starting one-dimensional set with a set of more dimensions.

People usually think of a formula when talking about functions (which is not always possible, but we usually meet them in this way), and this point of view also allows for a natural extension from formulas of the form $f(x) = x^2$ to formulas of the form $f(w, x, y, z) = (e^w + y)^x \cdot \sin(x - \pi z)$. This notation is very convenient when we work with specific functions, in particular when we know the number of variables and their names, but it is not suitable for general musings when we work with unknown number of variables. In such cases it pays to see the situation as if we worked with one variable, just this time we take that variable from a set of vectors. Instead of $f(x, y)$, $f(x, y, z)$, $f(u, v, w, x, y)$ and such we simply write $f(\vec{x})$.

In this notation, our work with functions of more variables is analogous to the case of one variable. The main ideas carry over, just calculations and procedures are sometimes a bit more complicated. All this is markedly easier if we can connect theoretical ideas with geometrical imagination, here some experience with basic objects in more-dimensional space (lines, planes) comes handy.

This is also the main aim of these notes, to show the geometry behind notions, calculations and procedures. We start with the definition of a function of more variables to put our reasoning on a firm footing.

Definition.
By a function of more variables we mean any mapping $f\colon D \mapsto \mathbb{R}$,
where $D = D(f)$ is some subset of $\mathbb{R}^n$.
If D is not explicitly given, then for the domain $D(f)$ we take the set of all $\vec{x} \in \mathbb{R}^n$ for which $f(\vec{x})$ makes sense.

Which brings us to the first topic.

1a. Domain

When a function is given by a formula, we determine its domain just like with functions of one variable: We ask ourselves what restrictions for the variables appear in that given formula.

Unlike the case of one variable, we need not try to express the resulting set in some standard way (union of intervals), since this is simply not possible in more dimensions; sets come in so many shapes that we are not able to write them down using one basic type of set (or several basic types). This means that we save some work, it is enough to write the answer as

$$D(f) = \{\vec{x} \in \mathbb{R}^n;\ \text{conditions for } \vec{x}\}.$$

On the other hand, there is one new thing: In more dimensions we can sometimes recognize what kind of object this set is.

Example: We determine $D(f)$ of the following functions:

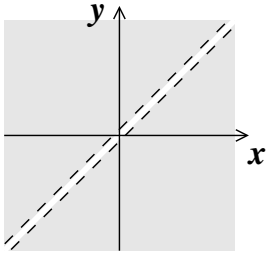
a) $f(x, y) = x^2 \sin(x + y)$: obviously $D(f) = \mathbb{R}^2$.

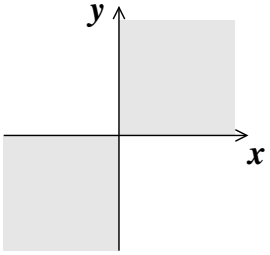
b) $f(x, y) = \sqrt{9 - x^2 - y^2}$: $D(f) = \{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 \le 3^2\}$,
it is a circle with radius 3 centered at the origin.

c) $f(x,y,z)=\sqrt{9-x^2-y^2-z^2}$: $D(f)=\{(x,y,z)\in\mathbb{R}^3;\ x^2+y^2+z^2\le 3^2\}$,
it is a ball with radius 3 centered at the origin.

d) $f(x,y)=\frac{e^{x+y}}{x-y}$: $D(f)=\{(x,y)\in\mathbb{R}^2;\ y\neq x\}$,
it is the plain with a line—the main diagonal—removed (see picture below).

e) $f(x,y)=\sqrt{x\cdot y}$: $D(f)=\{(x,y)\in\mathbb{R}^2;\ y\cdot x\ge 0\}$,
it is the first and third closed quadrant in the plane.

d) 

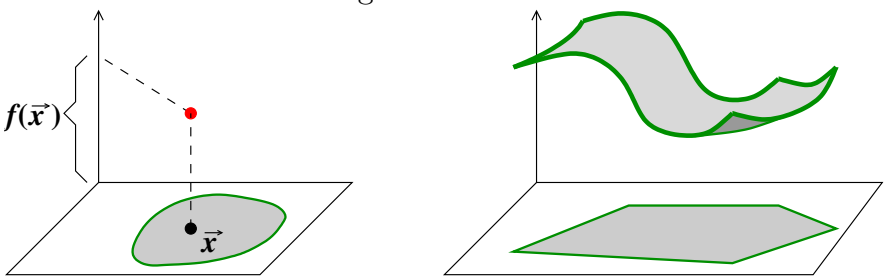
e) 

△

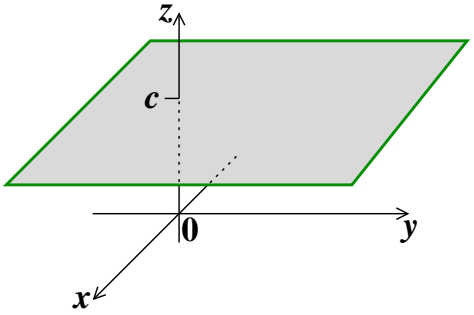
1b. Visualization of functions (sketching a graph)

We usually visualize functions of one variable through its graph. To generalize this notion to more dimensions is easy. In one variable, the graph of a function is the set of all points of the form $\big(x,f(x)\big)$ for $x\in D(f)$, so it requires two dimensions. If we take the variable of a function $f(\vec{x})$ from the space $\mathbb{R}^n$, then we will need one more dimension for the graph, and it is formed as the set of all points of the type $(x_1,x_2,\dots,x_n,f(\vec{x}))$, where $\vec{x}=(x_1,\dots,x_n)$.

In case of $n\ge 2$ we therefore need at least three dimensions, meaning that we jump out of our paper. We see that such a graph actually cannot be drawn. Still it is worthwhile to keep in mind the basic idea of the graph. On a "horizontal" representation of domain (precise or symbolic) we find the position of a certain multi-dimensional variable, and we draw a dot above it at elevation corresponding to the function value at the given variable.



If we do this at all points of the domain, the corresponding dots create a certain object that we can imagine to be a sort of wavy surface. This is what in fact happens with functions of two variables, a typical graph of such a function is some thin (essentially two-dimensional) object inside a three-dimensional space. With a bit of luck we can get a very suggestive picture of it using perspective, shading and similar methods. For instance, the graph of a constant function $f(x,y)=c$ will be the horizontal plane floating above the xy-plane at the appropriate elevation c.



Experience should suggest that there also are functions of two variables whose graphs are not that nice, sets of corresponding points in $\mathbb{R}^3$ could be rather wild, but we do not expect to find

such functions in applications. We like to work with functions of two variables precisely because we can usually sketch them.

If we have three or more variables, then we cannot even sketch a likeness of a graph, four-dimensional space is beyond our imagination. Therefore other tools were introduced that allow us to get a visual feeling for a function's behaviour. We will show here the most popular ones.

We will show how they work for functions of two variables, because then we can sketch the situations, but the ideas carry over also to higher dimensions.

There are also other ways to obtain sketches of graphs for functions of two variables, notably using appropriate programs (Maple, Mathematica and such), which is convenient and nice. However, it is useful to understand things well, after all, those programs sometimes fail and what is left is the good old brain and our knowledge.

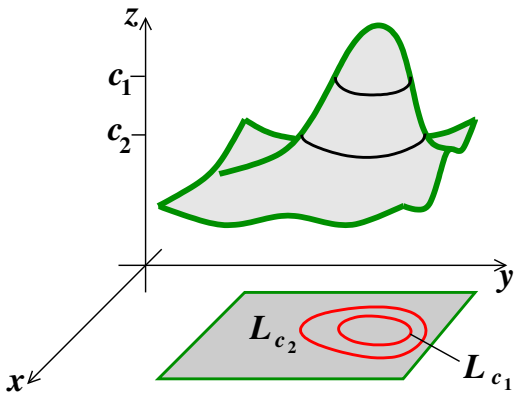
Level sets

This is one of the most powerful visualization methods, but this notion also has a wider use, for instance when investigating objects that were defined implicitly.

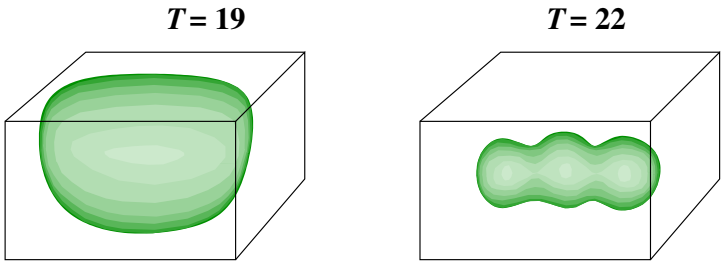
Definition.
Let $f\colon\ D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. For $c \in \mathbb{R}$ we define the corresponding **level sets** $L_c = \{\vec{x} \in D(f);\ f(\vec{x}) = c\}$.

From the definition we see that L_c lies in $D(f)$, so just n dimensions are enough to show level sets, an advantage compared to graphs.

How does it work? First we slice the graph at level c (we ask where the values of the function are equal to c) and then we check for which values of variable we get there, that is, we project that cut to the domain. If we imagine the graph to be a piece of land, then level sets are locations given by their geographical coordinates (that is, locations on the map) on which the elevation is precisely c. In other words, level sets are analogous to contours on a map. An experienced hiker can guess the shape of land just by looking at contours, similarly we can deduce a lot of useful information from level sets.



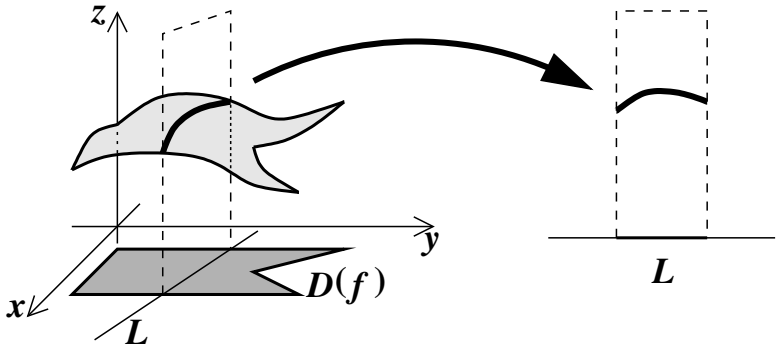
When working with functions of two variables, we traditionally say **level curves** instead of sets, and with three variables we often say **level surfaces**. It is a handy way of analyzing functions of three variables. For example, if we have a function $T(x, y, z)$ describing temperature at various places in a certain room, then for temperature c we see level sets (surfaces) L_c as "clouds" showing us where in the room that temperature is. These clouds are three-dimensional, we can use a perspective drawing to show them in 2D (on paper). Comparing such sketches for varying values of c we get a rather good understanding of behaviour of temperature, it is also possible to connect these pictures into an animation etc. The picture below shows two possible temperature "snapshots" of a classroom in February when it is being heated by three students sweating an exam.



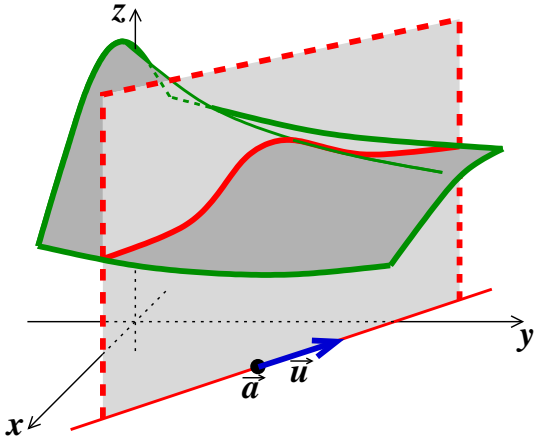
Slices

Slices are a simple but very powerful tool for exploring more-dimensional situations. However, they will be really simple for us only if we understand very well the interplay between formulas and geometric meaning. The basic idea is that we create a two-dimensional "slice" of a multi-dimensional situation by cutting it with a plane. Such a slice can be (hopefully) explored using tools for investigating (graphs of) functions of one variable.

The idea is simple. We have a graph of a function of n variables, that is, an object in $n+1$ dimensions. The variables are taken from the world $\mathbb{R}^n$ that we can symbolically picture as a horizontal object. When we choose a line in the domain and restrict our attention just to function values related to this line, then we actually get a flat object that we can carry over to $\mathbb{R}^2$ and capture with a function of one variable.



Formally we do it as follows. We choose some starting point $\vec{a}$ and consider a straight line l in $\mathbb{R}^n$ passing through this point in direction $\vec{u}$. Points of this line are given (if we assume uniform movement) by the parametric equation $\vec{x} = \vec{a} + t\vec{u}$, we can imagine that it is a description of a journey we take through the domain $D(f)$. At time t we are at a certain point from $D(f)$ and we see the function value $f(\vec{a} + t\vec{u})$ corresponding to this point. We thus obtain a mapping $t \mapsto f(\vec{a} + t\vec{u})$ describing the shape of the graph that we "see" above us while moving along the line l. This line therefore determines a slice that we obtain as an intersection of the graph of f and "vertical" plane erected over l.



The shape of this slice is obviously related to the function $\varphi(t) = f(\vec{a} + t\vec{u})$, which is a function of one variable and we can readily investigate it. Unfortunately this relationship is not as simple

as we would hope, that is, that drawing the graph of the function φ would already show us the shape of that slice. The problem lies in scale. In our two-dimensional picture of the slice, the mark for "1" (position at time $t = 1$) is at the place $\vec{a} + \vec{u}$, which does not necessarily need to be at the distance 1 from the point $\vec{a}$ in our original many-dimensional picture of the graph of f. The amount of distortion obviously depends on the size of the directional vector $\vec{u}$, which should not be surprising. The faster we go, the more the landscape around gets subjectively distorted (it get "squeezed", we pass things quicker).

Of course, we can simply restrict ourselves to directional vectors of magnitude 1, then the geometric information agrees. We will endeavor to do so, but it is not always possible, especially in applications (like physics). It is therefore useful to know how to handle general situations. We will encounter this problem again with derivatives.

Example: Consider the function $f(x, y) = \dfrac{3x^2 + 10x + 2y^2}{4x^2 + y^2}$. We will investigate the slice through its graph that we obtain by cutting it with the plane above the line $(x, y) = (2, 3) + t(-1, 1)$.

Interpretation: Function f indicates elevation, we stand at a place with GPS coordinates $(2, 3)$ and start off at the direction $(-1, 1)$. As we walk, we are interested in raise and fall of the land.

The description of the line yields the formulas $x = 2 - t$, $y = 3 + t$, substituting them into the function f gives the auxiliary function

$$\varphi :\ t \mapsto f(2 - t, 3 + t) = \frac{5t^2 - 10t + 50}{2t^2 - 10t + 25} = 1 + \frac{5}{t^2 - 2 + 5} = 1 + \frac{5}{(t - 1)^2 + 4}.$$

The last form shows that the graph of the function φ is shaped like a hill whose summit is at time $t = 1$.

This means that when we look at the corresponding cut through the graph of the function f, it will be shaped like a hill, but not precisely the same, because $\|\vec{u}\| = \sqrt{2}$, it is not a unit vector. However, only the horizontal scale of the picture is different, not the shape as such. Thus we can conclude that the slice is shaped like some hill and its highest point is at the point corresponding to time $t = 1$, that is, at the point $(1, 4)$.

The situation is therefore similar to the picture with slice above. Actually, the graph of the function is not right, but the point $\vec{a}$, vector $\vec{u}$ and line l are at the right place and the shape of the slice is essentially correct. By the way, the picture also shows that although wee see a hill (a local maximum) on the slice, it can easily happen that the function as such does not have any extreme there, we are just slicing through a side of a hill.

We remark that if we wanted to see the precise shape of the slice, then we would have to consider the directional vector $\frac{(-1,1)}{\|(-1,1)\|} = \frac{1}{\sqrt{2}}(-1, 1)$. The calculations are then analogous, just a bit less pleasant.

△

We prefer to work with lines parallel to coordinate axes, which simplifies calculations and the results can be easily interpreted. For instance, if we are in $\mathbb{R}^3$ at the point (x_0, y_0, z_0) and we want to move in the direction of the y-axis, then the other variables are left constant and the description of our movement simplifies significantly. Since we prefer directional vectors of magnitude 1, we end up with the pleasant description $t \mapsto (x_0, y_0 + t, z_0)$ of the line (and our movement). Similarly we use $t \mapsto (x_0 + t, y_0, z_0)$ to move in the x-direction etc.

In order to express this idea for a general setting we recall the standard coordinate vectors $\vec{e}_1 = (1, 0, \ldots, 0, 0)$, $\vec{e}_2 = (0, 1, \ldots, 0, 0)$ through $\vec{e}_n = (0, 0, \ldots, 0, 1)$, that is, we introduce the usual canonical basis of $\mathbb{R}^n$. Then the movement from a point $\vec{x}_0$ in the direction of the i-th coordinate can be expressed by the parametric formula $t \mapsto \vec{x}_0 + t\vec{e}_i$. Since $\|\vec{e}_i\| = 1$, shapes do not get distorted, so for instance when investigating the graph of the function $t \mapsto f(x_0 + t, y_0, z_0)$

we see exactly the shape of the slice of the graph of f above a line that is parallel with the x-axis and goes through (x_0, y_0, z_0).

In situations when we work with a known and relatively small number of variables (which is almost always except for general definitions and statements) we often further simplify notation by giving up on the otherwise good idea that the parametric description should place us at the original point for $t = 0$. Instead we use the chosen variable itself in place of a parameter, so we work with formulas $x \mapsto (x, y_0, z_0)$, $y \mapsto (x_0, y, z_0)$ etc., and we get to the original point by choosing $x = x_0$, $y = y_0$ etc. This is natural and efficient, in particular we directly obtain information about the influence of individual variables on the behaviour of our function.

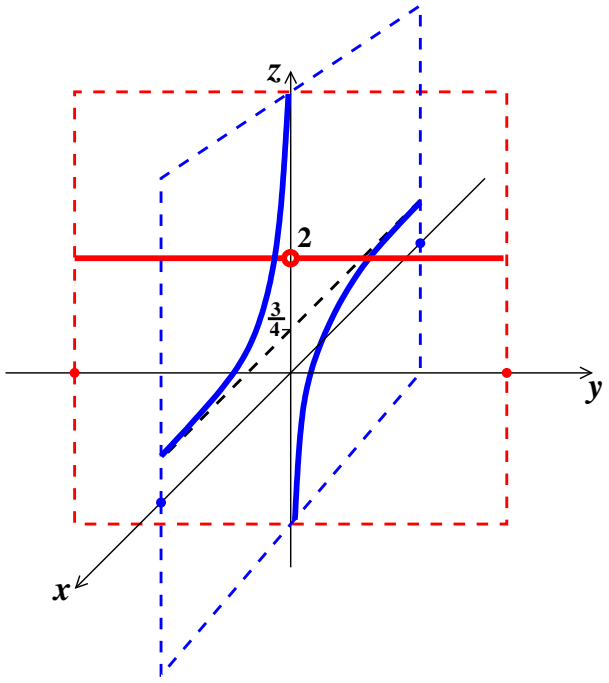
Example: Consider the function $f(x, y) = \dfrac{3x^2 + 10x + 2y^2}{4x^2 + y^2}$. What is the influence of individual variables around some point (x_0, y_0)? We obtain the answer through two functions of one variable,

$$x \mapsto f(x, y_0) = \frac{3x^2 + 10x + 2y_0^2}{4x^2 + y_0^2}$$
$$y \mapsto f(x_0, y) = \frac{3x_0^2 + 10x_0 + 2y^2}{4x_0^2 + y^2}.$$

For instance, if we wanted to know what is happening around the origin (where the function does not exist, but it exists elsewhere), then we would work with the original point $x_0 = 0$, $y_0 = 0$ and hence with functions

$$x \mapsto f(x, 0) = \frac{3x^2 + 10x}{4x^2} = \frac{3}{4} + \frac{5}{2x}$$
$$y \mapsto f(0, y) = \frac{2y^2}{y^2} = 2.$$

The first formula describes the shape of the graph of f above the x-axis, the second one describes the shape above the y-axis. We can express our findings by the following sketch:



This is not enough to give us a more definite idea of what the graph of f looks like, but we already suspect that it is going to be interesting near the origin.

△

This idea can be generalized as follows: We fix a certain number of variables and move with the others, which allows us to lower the number of dimensions we work with, depending on what we are

currently interested in. Take for the moment a function $T(x, y, z)$ of three variables, for instance a description of temperature in various places of a lecture room, and a certain point $\vec{a} = (x_0, y_0, z_0)$.

If we fix values $y = y_0$ and $z = z_0$, we get a one-dimensional object, that is, a line going from the point $\vec{a}$ in the direction of the x-axis and as we go along, we see the temperatures. It is a one-dimensional situation $x \mapsto T(x, y_0, z_0)$ that we easily investigate and draw.

If we fix only the variable $z = z_0$, then we have two degrees of freedom, that is, we move on the (horizontal) plane passing through the point $\vec{a}$ and perpendicular to the z-axis. Thus we obtain a function of two variables $(x, y) \mapsto T(x, y, z_0)$ whose graph we can (with a bit of luck) visualize using a perspective drawing (with shading, for instance).

In this way we get another tool for investigating functions. Still, most often we work with movement along a line and two-dimensional slices.

Standard shapes

When working with a function of one variable, we often think of its graph as an analytical object in the plane given by the equation $y = f(x)$, and we can recognize many of these objects. For instance, the graph of the function $f(x) = 1-2x$ is the object described by the equation $y = 1-2x$, that is, $2x + y = 1$, and we know that this determines a line. This trick sometimes also works rather well with functions of more variables.

For functions of n variables we obtain an object in $\mathbb{R}^{n+1}$ determined by the equation $y = f(\vec{x})$. Also here we can hope to recognize such an object. For instance, the function $f(x, y) = 2x + y - 5$ leads to the equation $z = f(x, y)$, that is, $2x + y - z = 5$, which defines a plane in $\mathbb{R}^3$.

Sometimes we have to reorganize the equation that we obtain, but then we have to be careful about not changing the resulting set. This is actually nothing new, we had to be careful already when working with functions of one variable. For instance, we know how the graph of the function $f(x) = \sqrt{x}$ looks like. If we rewrite the equation $y = \sqrt{x}$ as $x = y^2$, we immediately recognize this object as parabola, just the roles of variables are switched. Therefore this is a parabola that goes to the right, above and below the x-axis. However, the graph of our function $f(x) = \sqrt{x}$ is only the top half of this object, by squaring the equation we removed one restriction on values of y.

Such change in a set happens when we use some non-equivalent steps when rewriting our equation, the squaring that we used above is one popular case. Experience should suggest when it is time to pay extra attention.

In order to be successful in recognizing shapes we need to know basic geometric objects. The most popular are flat objects (lines, planes) and objects given by quadratic equations, which in two dimensions means the well-known family of conic sections (parabola, circle and, more generally, an ellipse). In more dimensions we get a richer family. It seem clear that one would have to be really lucky to hit one of the few known equations when choosing from infinitely many functions, but it actually happens more often than one would expect from purely probabilistic standpoint. And when it does happen, it is very pleasing.

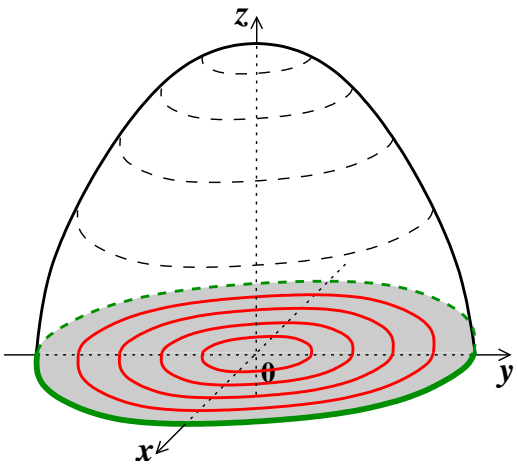
Example: Here we will showcase our methods on the function $f(x, y) = \sqrt{9 - x^2 - y^2}$.

Domain: $D(f) = \{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 \leq 3^2\}$. It is the circle of radius 3 with center $(0, 0)$. The graph will be above this circle as obviously $f(x, y) \geq 0$. We can also notice that $f(x, y) \leq 3$ and we can actually reach all values between 0 and 3, so the range is $R(f) = [0, 3]$.

We start with level curves. As we observed, the level curves will be interesting only for values between 0 and 3, the others are empty. So how does a level curve L_c look like for $c \in [0, 3]$?

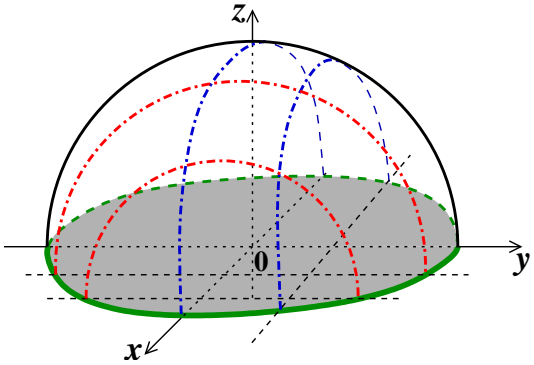
It is the set given by the relation $f(x, y) = c$, that is, $x^2 + y^2 = 9 - c^2$, this specifies the circle of radius $\sqrt{9 - c^2}$ in the domain. For larger c (larger values of the function) these circles get smaller, closer to the origin, and conversely, for $c = 0$ we get as the level curve the circle of radius 3, the boundary of the domain.

Conclusion: The function is equal to zero on the edge of the domain, largest at its center, hills look this way. On the picture we indicated with dashed lines several values in the graph and the corresponding level curves are in the domain.



Because the level curves are rotationally symmetric with respect to the origin (they are circles), we can deduce that the graph is also symmetric with respect to rotation about the z-axis.

Now we look at slices. If we fix $x = 0$, we are in effect asking how the graph looks like above the y-axis. We get $f(0, y) = \sqrt{9 - y^2}$, this has an upper half-circle as its graph. For other fixed $x = a$, the graphs of slices are smaller half-circles $f(a, y) = \sqrt{(9 - a^2) - y^2}$. So this is how cuts using vertical planes parallel to the y-axis look like.



It works symmetrically, so also vertical slices through the graph of f parallel to the x-axis are upper half-circles. It would seem that this graph is not just any hill, but a hill of spherical shape.

We will confirm this guess by recognizing the shape. The graph of f is given by the equation $z = \sqrt{9 - (x^2 + y^2)}$, which we readily rewrite as $x^2 + y^2 + z^2 = 3^2$ and we have the equation of sphere. Obviously the whole sphere cannot be the graph of our function (because it would offer two values for one point (x, y)), but some subset of it. Since $D(f)$ is the circle of radius 3 in the xy-plane, the graph must cover the whole expanse of that sphere, and from $f(x, y) = \sqrt{**} \geq 0$ it follows that we should look at its upper half, which is something that we already guessed from level curves and cuts.

Conclusion: The graph of f is the upper half-sphere (a dome).

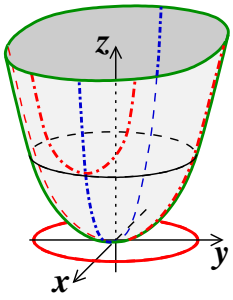
△

Example: Now we consider the function $f(x, y) = x^2 + y^2$. Obviously $D(f) = \mathbb{R}^2$.

Level curves: $x^2 + y^2 = c$, these are circles whose radii are increasing with larger values of the function. So the farther we are from the origin, the bigger the function, this looks like a pit. Since all level curves are invariant with respect to rotation (they look the same after rotating them about the vertical axis), the graph will also be like that.

Slices: If we choose $x = 0$, we get $f(0, y) = y^2$, this is a parabola. For other fixed x we get parabolas shifted up $f(x_0, y) = x_0^2 + y^2$, the shift grows larger the further we are from the origin. The vertices of these parabolas are themselves on a parabola. This is true also symmetrically, when we fix y.

When we put this together with our observation about rotational symmetry, the conclusion is clear. The graph is a paraboloid, that is, a rotated parabola.



△

Example: Consider the function $f(x, y) = y^2 - x^2$. We see that $D(f) = \mathbb{R}^2$.

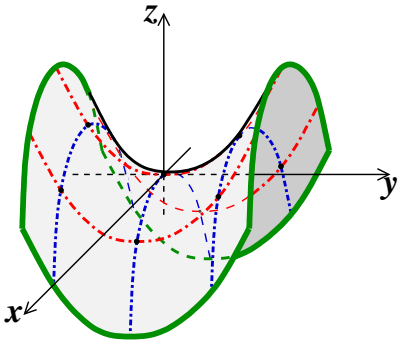
Level curves: $y^2 - x^2 = c$, these are hyperbolas of growing radii.

Slices: If we choose $x = 0$ (the slice above the y-axis), we get $f(0, y) = y^2$, the graph is the standard parabola. For other x we again get parabolas shifted down, $f(x_0, y) = y^2 - x_0^2$, the shift growing as we get further from the origin. Vertices of these parabolas lie on the parabola $z = -x^2$ oriented down.

If we choose $y = 0$ (the slice above the x-axis), we get $f(x, 0) = -x^2$, the graph is a parabola oriented down. For other y we again get parabolas opened down and shifted up, $f(x, y_0) = -x^2 + y_0^2$, the shift growing as we get further from the origin. Vertices of these parabolas lie on the parabola $z = y^2$.

Supplemental methods: There is no symmetry of rotation. The equation $z = y^2 - x^2$ is not some universally known shape. So not much help here, we have to make do with slices.

Conclusion: It is an interesting graph worth drawing.



Actually, this is a fairly popular shape called a saddle.

△

We conclude this chapter with a generalization of another useful notion.

Definition.
Let f: $D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
We say that f is **bounded** if there exists $K > 0$ such that $|f(\vec{x})| \leq K$ for all $\vec{x} \in D(f)$.
Let $M \subseteq D(f)$. We say that f is **bounded on** M if there exists $K > 0$ such that $|f(\vec{x})| \leq K$ for all $\vec{x} \in M$.

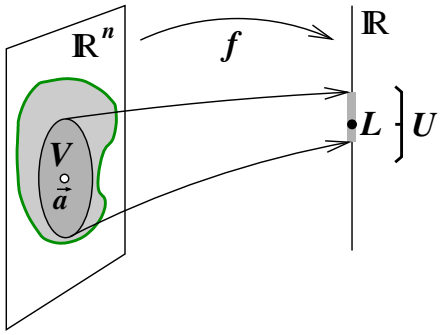
The graphical interpretation is easy. For a function of two variables, boundedness means that we can squeeze its graph between two horizontal planes. The situation in more dimensions is analogous, we just cannot visualize it that well.

2. Introduction to limits and continuity

The notion of limit has essentially the same definition in more dimensions as it does for one dimension:

Definition.
Let f: $D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
Let $\vec{a}$ be a point such that f is defined on some reduced neighborhood of $\vec{a}$, let $L \in \mathbb{R}^*$.
We say that L is the limit of f as $\vec{x}$ goes to $\vec{a}$, denoted $\lim_{\vec{x}\to\vec{a}} \big(f(\vec{x})\big) = L$, if for every neighborhood $U = U(L)$ of the value L there is a reduced neighborhood $V = V(\vec{a})$ of the point $\vec{a}$ such that $f[V] \subseteq U$.

To understand this notion, it is a good idea to visualize functions in the way that is used for general mappings, that is, using an "arrow-diagram" $D(f) \mapsto \mathbb{R}$.



What changed when we passed to more variables? The obvious modification is that neighborhoods in the domain are now more-dimensional. But they are still just reduced neighborhoods, so the value of f at the point $\vec{a}$ itself is irrelevant and may not even exist.

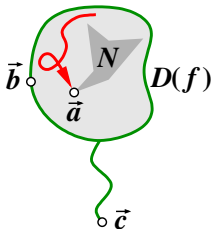
The change to higher dimension can be also seen in the classical epsilon-delta form of the condition:

• For every $\varepsilon > 0$ there is $\delta > 0$ such that $|f(\vec{x}) - L| < \varepsilon$ whenever $0 < \|\vec{x} - \vec{a}\| < \delta$.

We had to use a multi-dimensional notion of distance between vectors in the domain, it is the classical Euclidean distance.

There is one substantial difference compared to the world of one variable: There are actually two distinct and competing opinions on how a limit should look like for functions of more variables. We chose the version that is a direct generalization of the notion as it is defined for functions of one variable; it is sufficient in most cases and it is simple. Its disadvantage is that it does not allow us to ask about some interesting limits. For instance, the function $f(x, y) = \sqrt{x} + y$ exists on the half-plane $\{(x, y);\ x \geq 0\}$ and it would definitely be interesting to inquire about its limit at $(0, 13)$. Unfortunately, this point lies on the boundary of the domain and the function does not exist on any neighborhood of this point, reduced or not. The definition thus does not allow us to ask for the limit there. For some people this is a problem, and some authors therefore prefer a more general—and also more complicated—definition that allows calculating such a limit.

Here we adopted the point of view that it is handy to have a simple definition of the basic limit, and the more complicated situations were the reason for introducing one-sided limits in one dimension. We will do exactly the same thing here as well, however, in more dimensions it gets significantly more complicated. Indeed, having a point $\vec{a}$ in a more-dimensional space, it does not really make sense to approach it "from the left" or "from the right". We can actually approach it from infinitely many different directions, following many different curves, we can even approach it from within some sector or a cone or through some other interesting shape. Consider the following situation.



In the picture we see the domain of a certain function (probably not a common one). Perhaps we are interested in the limit at points $\vec{b}$ and $\vec{c}$. This cannot be done according to our definition (they are not surrounded by the domain), but it does make some sense. For instance, there is a "tail" in the domain leading to $\vec{c}$ that we can move along. The picture suggests that we can use it to get arbitrarily close to $\vec{c}$, and all the time we have the values of f available, so it makes sense to ask: Do these values approach some number (a limit)?

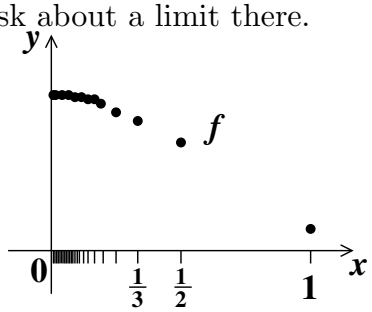
Similarly we can get to the point $\vec{b}$, it actually has a "full" set next to it offering many paths leading to $\vec{b}$, but that is not enough to make our definition applicable, we will need the new notion here as well. On the other hand, we can apply the definition to the point $\vec{a}$, but also here we may decide that for some reason it could be interesting to approach $\vec{a}$ along the outlined curve, or perhaps we would want to approach $\vec{a}$ while moving within the outlined set N. In such cases we actually work with the restriction of our function to a specific subset of the domain (the path or the set N).

If we want to approach a certain point $\vec{a}$ using some set on which f is defined, we have to make sure that this set allows us to get arbitrarily close to $\vec{a}$ in order that it makes sense to ask about a limit at $\vec{a}$. Mathematically, this requirement is captured in the notion of an accumulation point (see the appendix for basic topological notions).

Consider a weird function $f(x)$ whose domain is

$$\left\{\tfrac{1}{n};\ n \in \mathbb{N}\right\} = \left\{1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \dots\right\}.$$

We can get arbitrarily close to 0 using elements of this set, therefore 0 is an accumulation point of the domain and it makes sense to ask about a limit there.



In more dimensions it works similarly.

Definition.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
Let $\vec{a} \in \mathbb{R}^n$ and $N \subseteq D(f)$ be such that $\vec{a}$ is an accumulation point of N. Let $L \in \mathbb{R}^*$.
We say that L is a limit of f as $\vec{x}$ goes to $\vec{a}$ relative to N, denoted

$$\lim_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}} \big(f(\vec{x})\big) = L,$$

if for every neighborhood $U = U(L)$ of the value L there is a reduced neighborhood $V = V(\vec{a})$ of the point $\vec{a}$ such that $f[V \cap N] \subseteq U$.

Sometimes we also say "limit with respect to a set".

The set N could be some region, it could be a curve, or just a bunch of points. In the light of limitations discussed when we introduced the general notion of a limit, it may be interesting to take $N = D(f)$, because then we can actually inquire about limits on a boundary of the domain. In particular, going back to the example we discussed there, now we are entitled to ask about the limit $\lim\limits_{\substack{(x,y)\to(0,13)\\ x\geq 0}} \big(\sqrt{x} + y\big)$.

For functions of one variable it is true that if a limit at some point exists, then also one-sided limits there must yield the same answer. The same is true for our new notions. We also had two statements that go, in a way, in the opposite direction. Both statements have their analogues in more dimensions.

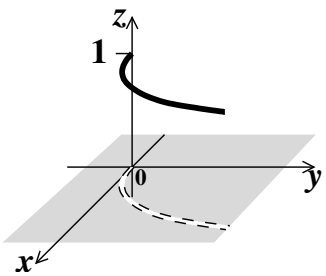
Theorem.
Let f: $D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
Let $\vec{a}$ be a point in the interior of $D(f)$, $L \in \mathbb{R}^*$.
(i) $\lim\limits_{\vec{x}\to\vec{a}} \big(f(\vec{x})\big) = L$ if and only if $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}} \big(f(\vec{x})\big) = L$ for all $N \subseteq D(f)$ such that $\vec{a}$ is an accumulation point of N.
(ii) (Heine's theorem) $\lim\limits_{\vec{x}\to\vec{a}} \big(f(\vec{x})\big) = L$ if and only if $\lim\limits_{k\to\infty} \big(f(\vec{x}(k))\big) = L$ for all sequences $\{\vec{x}(k)\} \subseteq D(f) - \{\vec{a}\}$ such that $\vec{x}(k) \to \vec{a}$.

This theorem is a very strong tool when investigating limits.

Example: Consider the function

$$f(x,y) = \begin{cases} 1, & y = x^2, x > 0; \\ 0, & \text{elsewhere.} \end{cases}$$

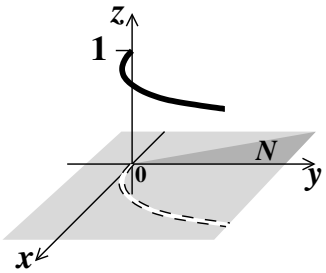
Its graph looks like a flat plain on the level 0, over which we can see a steep ridge shaped like a narrow path over a parabolic trace in the domain.



What can we say about the limit at the origin? If there was some limit L, then it would have to be able to approximate values of f near the origin with arbitrary precision. The picture shows quite clearly that no such value can exist, because regardless of how near the origin we decide to look, we can always find some points there where the function is 0 and points where the function is 1.

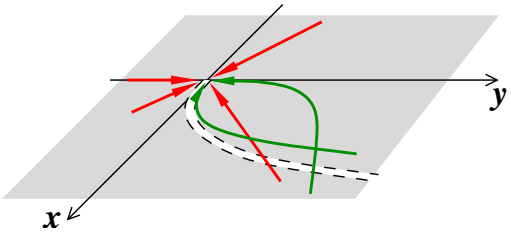
To make it more interesting we will look at limit respective to sets and curves. In the picture we see the outline of the wedge

$$N = \{(x,y) \in \mathbb{R}^2;\ x < 0, y > -x\}.$$



The origin $\vec{0}$ is an accumulation point for this set, so it makes sense to ask about the corresponding limit. Because the function is always zero on this set, consequently also the limit at $\vec{0}$ with respect to N must be zero.

Now we will try some sets N in the shape of a curve, namely we will check on straight lines going to the origin. Such rays can be coded using their slopes, so we are interested in sets of points of the type (x, kx) for $x > 0$, or for $x < 0$. Two special rays are not included in this type, namely when approaching the origin along the y-axis, so we will also consider the set of points $(0, y)$ for $y > 0$, or for $y < 0$. We see several of such sets in the picture, the arrows suggest that we are approaching the origin while moving within these sets. What values of the function do we encounter on the way?



It is obvious that arrows in three quadrants, including those that go along the x-axis and y-axis, do not meet the parabola in the domain at all. Consequently, f is zero everywhere on these sets and the corresponding relative limits are also zero. It remains to look at arrows that come from the first quadrant. Each of them actually intersects the parabola, but only once. Therefore, for each such ray, there must be some neighborhood of the origin on which this ray does not meet the parabola, it only "sees" values zero, and that is the value of the limit along this ray. We conclude that all limits with respect to straight lines going to the origin are equal to zero.

Interestingly enough, also most limits with respect to parabolic paths are equal to zero, because parabolas that are different from the one in the definition of the function either do not cross it at all, or cross it at most once, if we also allow for general parabolas of the form $y = ax(x - b)$ or $x = ay(y - b)$, see the green curves. One might start thinking that perhaps the function has limit zero at the origin.

There is just one "simple" path that shows that the limit at the origin has a problem: We have to go there along the key parabola $y = x^2$, $x > 0$. Then we encounter only function values 1, and that would be our limit with respect to this set N. We identified two sets so that the limits at the origin with respects to these sets are different, which proves that the limit as such (according to the first definition) does not exist. Note that if we were not lucky enough to think of this special path, then we would not be aware of the problem.

Actually, there are more possibilities, any curve that crosses our parabola arbitrarily close to the origin would do, for instance the path $y = x^2 + \sin\left(\frac{1}{x}\right)$. However, a typical limit investigator cannot not be expected to think of such a curve.

△

That "walking along curves" trick is the strongest and most popular tool for investigating limits and also a popular approach to proving that a limit does not exist. Unfortunately, it has its limitations. Clearly it is easy to replace the parabola with some unexpected monster as the basis for our function, and we get a function of two variables that convincingly pretends to have a limit equal to zero as you keep approaching the origin along pretty much any curve you can think of. Yet, it does not have a limit there. So this tool can fail fairly easily, but we do not really have anything better.

When investigating limits of functions of one variable, we often obtained our result by substituting the limit point. We needed the concept of continuity for that, and we can readily make it more general.

Definition.
Let f: $D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
Let $\vec{a} \in D(f)$. We say that the function f is continuous at $\vec{a}$ if
• $\vec{a}$ is an accumulation point of $D(f)$ and $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in D(f)}} \big(f(\vec{x})\big) = f(\vec{a})$,
• or $\vec{a}$ is an isolated point of $D(f)$.

The second condition is not universally agreed on by all authors. Here we declared automatic continuity for isolated points, because it simplifies some statements and it also makes sense logically. Continuity refers to the relationship of the function value at a point and values on its immediate surroundings. If the function exists at an isolated point, then there is no neighborhood on which the function would exist, and so the relationship cannot break down. On the other hand, some authors see this situation differently: Since the function does not exist around that point, then there can be no relationship. Their way to build the theory also works and I actually have a mild affinity for this approach, but I chose for this text the version that is preferred by textbooks used

at our school. Fortunately, in applications we rarely meet functions with isolated points in their domains, so we do not really have to worry which theory is being used.

It is just a little step passing from local continuity to a global notion.

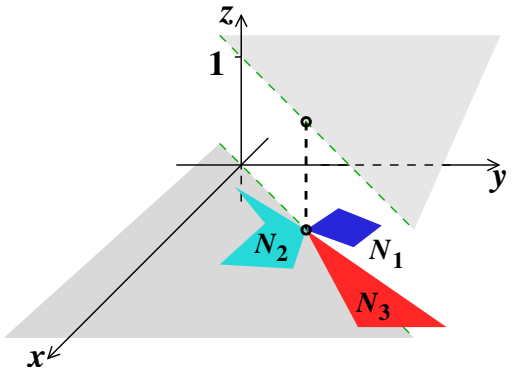
Definition.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
We say that the function f is continuous on a set $M \subseteq D(f)$ if its restriction to M is continuous at all points of M.
We say that this function is continuous if it is continuous at all points of its domain.

The notion of continuity in more dimensions behaves just like its one-dimensional inspiration. We again visualize continuous functions as those whose graphs (say, the sheets in three dimensions) are not "torn".

Example: Consider the function

$$f(x,y) = \begin{cases} 1, & y > x; \\ 0, & y \leq x. \end{cases}$$

Obviously, its domain is $\mathbb{R}^2$. The set given by the condition $y > x$ is the "infinite triangle" above the diagonal in the domain, there the graph of the function is just a horizontal "sail" on level 1. On the remaining part of the domain, the graph is a horizontal plain on the basic level 0.



From the picture it seems clear that this function is continuous on the sets $\{(x,y) \in \mathbb{R}^2;\ y > x\}$ and $\{(x,y) \in \mathbb{R}^2;\ y < x\}$ and discontinuous at all points on the line $y = x$.

Let's look closer at the point $(1,1)$. In the picture we outlined three sets N, we will inquire about limit relative to these sets.

The reader can hopefully see right away that the limit with respect to the blue set N_1 is 1, while the limit relative to the turquoise N_2 is 0. Because these two relative limits do not agree, according to one of the above theorems the limit at $(1,1)$ as such cannot exist, hence we also do not have continuity there. In a similar way we reason out that the limit with respect to the red set N_3 does not exist.

△

Key analytic theorems are true also for continuity in higher dimension. First, it is still true that elementary functions are continuous on their domains, which in particular includes polynomials. Second, it is still true that when we create a new function out of continuous functions using algebraic operations and/or composition, then the new function is continuous on its domain.

Thanks to this we get a substantial supply of continuous functions. This is crucial, because in this way we also get an easy way how to evaluate quite a lot of limits: We just substitute into a formula.

Example: Since the exponential is continuous and so are polynomials, the composed function $f(x,y) = e^{x^2y+y^2}$ is continuous on its domain, which obviously is $\mathbb{R}^2$. Thus we can easily evaluate limits at all points of $\mathbb{R}^2$, say,

$$\lim_{(x,y)\to(2,-4)} \left(e^{x^2y+y^2}\right) = e^{-16+16} = 1.$$

△

Of course, we cannot expect things to be so nice all the time; the list of indeterminate expressions is still valid, after all. This is where things get very difficult, because there is no multi-dimensional replacement for the l'Hospital rule, the most powerful tool for evaluating indeterminate limits in one variable. Practically speaking this means that while theoretically the notion of limit for functions of more variables is analogous to the case of one variable, when it comes to actual calculations it is a very different (and much harder) ball game.

The most difficult task is to prove that a certain L is really a limit. We will now look at some interesting blind alleys that still provide us with at least something. For instance, one may entertain the idea that perhaps we could approach $\vec{a}$ along slices (curves), which creates a one-dimensional situation where the l'Hospital rule is applicable. Unfortunately, it is not possible to confirm the existence of a limit in this way, because according to the Heine theorem we would have to check on all possible curves going to $\vec{a}$, which is practically impossible. As we saw, using limits along curves we can only (with a bit of luck) prove the non-existence of a limit.

There is also another idea that can get us to limits of one variable, namely the so-called "repeated limits". This refers to the process when we are considering a more-dimensional limit, but instead of approaching $\vec{a}$ with points $\vec{x}$ in the spatial sense, we just focus on individual coordinates one by one. The order in which we do this is not unique, as the case of two variables shows.

$$\lim_{(x,y)\to(a,b)} \big(f(x,y)\big) \implies \begin{cases} \lim\limits_{y\to a}\Big(\lim\limits_{x\to a}\big(f(x,y)\big)\Big) \\ \lim\limits_{x\to a}\Big(\lim\limits_{y\to b}\big(f(x,y)\big)\Big) \end{cases}$$

Let's start with the first repeated limit. It can be proved that if the limit $\lim\limits_{(x,y)\to(a,b)} \big(f(x,y)\big)$ (the standard one, in this context people also call it the "double limit") exists and is equal to L, and moreover, for all y_0 from some neighborhood of b the limits $\lim\limits_{x\to a}\big(f(x,y_0)\big)$ exist, then necessarily also the repeated limit $\lim\limits_{y\to a}\Big(\lim\limits_{x\to a}\big(f(x,y)\big)\Big)$ exists and is equal to L.

The analogous statement is also true for the second repeated limit. This means that if the two repeated limits come up different, then it actually proves that the investigated double limit does not exist.

Example: We investigate the limit of the function $f(x,y) = \dfrac{x^2}{x^2+y^2}$ at the point $(0,0)$.

This function exists on the set $\mathbb{R}^2 \setminus \{(0,0)\}$, in particular on a reduced neighborhood of $(0,0)$ and thus it makes sense to consider this limit. We try the repeated limits.

$$\lim_{y\to0}\Big(\lim_{x\to0}\Big(\frac{x^2}{x^2+y^2}\Big)\Big) = \lim_{y\to0}\Big(\frac{0}{0+y^2}\Big) = \lim_{y\to0}(0) = 0,$$

while

$$\lim_{x\to0}\Big(\lim_{y\to0}\Big(\frac{x^2}{x^2+y^2}\Big)\Big) = \lim_{x\to0}\Big(\frac{x^2}{x^2}\Big) = \lim_{x\to0}(1) = 1.$$

Since the outcomes differ, it follows that $\lim\limits_{(x,y)\to(0,0)}\Big(\dfrac{x^2}{x^2+y^2}\Big)$ does not exist.

△

Unfortunately, the implication that connects double and repeated limits is not in general true in the other direction. When the repeated limits yield the same answer, then it does not help us in determining the double limit.

Obviously, determining limits for functions of more variables is an adrenaline sport, not unlike evaluating integrals. There is no reliable algorithm for that, but at least there are some reasonably efficient approaches.

Investigating limits

• We try to substitute the limit point $\vec{a}$ into the given function f and if this leads to a result, then we obtain the value of the limit. Of course, the likelihood of this happening is very low when it comes to exam problems. In a typical case we get an indeterminate expression, which, unlike the case of functions of one variable, does not point to a straightforward procedure.

• If we suspect that the limit does not exist, we can confirm it by finding two distinct results when approaching $\vec{a}$ along various curves. Often we start by going to $\vec{a}$ along straight lines (rays), sometimes even just going along coordinate axes as it is easier. If we keep getting the same answer, we may in desperation try to walk along more complicated curves. In two dimensions it may also be helpful to try repeated limits. If all the attempts end up with the same answer, we may start getting suspicious that we actually found the limit, and we would need to prove that we are right.

• When we suspect that a limit exists, we need to confirm it. Unfortunately, there is no reasonably reliable procedure for that, we have to think of some individual approach. Sometimes we can cancel algebraically. Sometimes we get help from a comparison test, they work also in more dimensions. Sometimes a substitution helps, especially if it transforms our problem into a one-dimensional limit. Generally it is helpful if we can somehow incorporate into our musings the distance between $\vec{x}$ and $\vec{a}$, because it tends to zero during the limit process. And when everything else fails, we can always try to prove the validity of our answer using the definition.

Since proving validity of a limit answer is hard, limits are often covered just in passing in calculus courses and often they are not on the test at all.

We now look closer at some tools used in proving the existence of a limit. Regarding the comparison criterion, we will state the version that addresses the limit relative to some set; then the choice $N = \mathbb{R}^n$ yields comparison also for the standard limit.

Theorem. (The squeeze theorem)
Let $f\colon D(f) \mapsto \mathbb{R}$, $g\colon D(g) \mapsto \mathbb{R}$, and $h\colon D(h) \mapsto \mathbb{R}$ be functions,
where $D(f), D(g), D(h) \subseteq \mathbb{R}^n$.
Let $\vec{a} \in \mathbb{R}^n$ and $N \subseteq D(g)$ be such that $\vec{a}$ is an accumulation point of N. Let $L \in \mathbb{R}^*$.
Assume that $N \subseteq D(f) \cap D(h)$ and $f \le g \le h$ on N.
If $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}} \big(f(\vec{x})\big) = \lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}} \big(h(\vec{x})\big) = L$, then also $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}} \big(g(\vec{x})\big) = L$.

Regarding substitution, a function $f\colon \mathbb{R}^n \mapsto \mathbb{R}$ can be written as a composed function in several ways, so there are several possible theorems. We will look here at a situation that appears a lot, when $f = h(g)$ for some $g\colon \mathbb{R}^n \mapsto \mathbb{R}$ and $h\colon \mathbb{R} \mapsto \mathbb{R}$. Then we can calculate limit as follows:

$$\lim_{\vec{x}\to\vec{a}} \big(h(g(\vec{x}))\big) = \left| \begin{array}{l} y = g(\vec{x}) \\ b = \lim\limits_{\vec{x}\to\vec{a}} \big(g(\vec{x})\big) \end{array} \right| = \lim_{y\to b} \big(h(y)\big),$$

assuming that some assumptions are met. One popular condition is that h is continuous at b. Another possibility is to check that values $g(\vec{x})$ are distinct from b near $\vec{a}$. See the formal statement below in section Facts about limits.

We will now look at several limits to showcase various approaches.

Example: We investigate the limit of the function $f(x,y) = \dfrac{x^2y^2}{x^4+y^4}$ at the point $(0,0)$.

The functions exists on $\mathbb{R}^2 \setminus \{(0,0)\}$, so the limit makes sense.

We try to substitute the limit point to learn that it is not in the domain, and also that it yields the popular indeterminate expression $\frac{0}{0}$.

First we try to approach the given point along the coordinate axes. We move along the x-axis by choosing $y = 0$ and $x \to 0$, so we in fact do the limit $(x,0) \to (0,0)$.

$$\lim_{(x,0)\to(0,0)} \Big(\frac{x^2y^2}{x^4+y^4}\Big) = \lim_{x\to 0}\Big(\frac{0}{x^4+0}\Big) = \lim_{x\to 0}(0) = 0.$$

Similarly we go along the y-axis:

$$\lim_{(0,y)\to(0,0)} \Big(\frac{x^2y^2}{x^4+y^4}\Big) = \lim_{y\to 0}\Big(\frac{0}{0+y^4}\Big) = \lim_{y\to 0}(0) = 0.$$

Could the limit actually be zero? We will try to approach the origin along straight lines. A typical line through the origin is given by the formula $y = kx$, so we are interested in the limit process when $(x,kx) \to (0,0)$.

$$\lim_{(x,kx)\to(0,0)} \Big(\frac{x^2y^2}{x^4+y^4}\Big) = \lim_{x\to 0}\Big(\frac{k^2x^4}{x^4+k^4x^4}\Big) = \lim_{x\to 0}\Big(\frac{k^2}{1+k^4}\Big) = \frac{k^2}{1+k^4}.$$

We see that the outcome depends on k, that is, on which particular direction we took to approach the origin. Conclusion: The limit of f at $(0,0)$ does not exist.

Among other things it also implies that the function defined by the formula

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

is not continuous at the origin.

Out of curiosity we also try repeated limits.

$$\lim_{y\to 0}\Big(\lim_{x\to 0}\Big(\frac{x^2y^2}{x^4+y^4}\Big)\Big) = \lim_{y\to 0}\Big(\frac{0}{0+y^4}\Big) = \lim_{y\to 0}(0) = 0,$$

and also

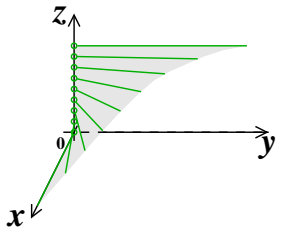
$$\lim_{x\to 0}\Big(\lim_{y\to 0}\Big(\frac{x^2y^2}{x^4+y^4}\Big)\Big) = \lim_{x\to 0}\Big(\frac{0}{x^4+0}\Big) = \lim_{x\to 0}(0) = 0.$$

So they would not make us aware of any problems with the limit.

△

How do we visualize the situation when the outcome of a limit depends on the direction from which we go? Probably the easiest example is that of the function $f(x,y) = \dfrac{y}{x}$ considered on the quadrant $D = \{(x,y) \in \mathbb{R}^2;\ x,y > 0\}$. When we look at the formula, we see that the value of the function only depends on the slope of the line on which the point (x,y) lies. Thus, for all points on the line given by $y = kx$ the value is $f(x,y) = k$. How does the graph look like? A bit like a piece of a spiral staircase where each step is a bit rotated relative to the previous one, the bottom step is in the direction of the x-axis and the top step is in the direction of y-axis. Then we have to smooth down the edges of stairs so that we can slide down better.

Since the slope goes to infinity for points very near to the y-axis, we prefer to draw the graph of the function $f(x,y) = \arctan\Big(\dfrac{y}{x}\Big)$. Then the values approach zero for points near the x-axis, while for points near the y-axis the values get close to $\frac{\pi}{2}$.



Example: We investigate the limit of the function $f(x,y)=\sqrt{2x^2+y^4}$ at the point $(0,0)$.

Since $(0,0)$ is in the domain of this obviously continuous function, we readily deduce by substituting that the limit is 0.

That was boring, let's try something a bit more complicated, but still with a happy ending:

$$\lim_{(x,y)\to(0,0)}\Big(\frac{x^4-y^4}{x^2+y^2}\Big)=\lim_{(x,y)\to(0,0)}\Big(\frac{(x^2+y^2)(x^2-y^2)}{x^2+y^2}\Big)=\lim_{(x,y)\to(0,0)}(x^2-y^2)=0.$$

The given function exists on $\mathbb{R}^2\setminus\{(0,0)\}$, that is, on some reduced neighborhood of the origin, so we had the right to ask for the limit there.

△

Example: We investigate the limit of the function $f(x,y)=\dfrac{x^2+y^4}{\sqrt{4x^2+y^2}}$ at the point $(0,0)$.

Attempting to substitute the given point we find that it is not in the domain; we also observe that that this attempt lead to the indeterminate expression $\frac{0}{0}$.

We try to approach $(0,0)$ along the coordinate axes:

$$\lim_{(x,0)\to(0,0)}\Big(\frac{x^2+y^4}{\sqrt{4x^2+y^2}}\Big)=\lim_{x\to0}\Big(\frac{x^2}{\sqrt{4x^2}}\Big)=\lim_{x\to0}\Big(\frac{x^2}{2|x|}\Big)=\lim_{x\to0}\Big(\frac{|x|}{2}\Big)=0.$$

Similarly,

$$\lim_{(y,0)\to(0,0)}\Big(\frac{x^2+y^4}{\sqrt{4x^2+y^2}}\Big)=\lim_{y\to0}\Big(\frac{y^4}{\sqrt{y^2}}\Big)=\lim_{y\to0}(|y|\cdot y^2)=0.$$

Now we try it along straight lines:

$$\lim_{(x,kx)\to(0,0)}\Big(\frac{x^2+y^4}{\sqrt{4x^2+y^2}}\Big)=\lim_{x\to0}\Big(\frac{x^2+k^4x^4}{\sqrt{4x^2+k^2x^2}}\Big)=\lim_{x\to0}\Big(\frac{x^2(1+k^4x^2)}{|x|\sqrt{4+k^2}}\Big)=\lim_{x\to0}\Big(\frac{|x|(1+k^4x^2)}{\sqrt{4+k^2}}\Big)=0.$$

What if we approach the origin along parabolas? We start with the simple ones, $y=kx^2$.

$$\lim_{(x,kx^2)\to(0,0)}\Big(\frac{x^2+y^4}{\sqrt{4x^2+y^2}}\Big)=\lim_{x\to0}\Big(\frac{x^2+k^4x^8}{\sqrt{4x^2+k^2x^4}}\Big)=\lim_{x\to0}\Big(\frac{|x|(1+k^4x^6)}{\sqrt{4+k^2x^2}}\Big)=0.$$

How about flipping their orientation, that is, using the curves $(ky^2,y)\to(0,0)$?

$$\lim_{(ky^2,y)\to(0,0)}\Big(\frac{x^2+y^4}{\sqrt{4x^2+y^2}}\Big)=\lim_{y\to0}\Big(\frac{k^2y^4+y^4}{\sqrt{4k^2y^4+y^2}}\Big)=\lim_{y\to0}\Big(\frac{y^4(1+k^2)}{|y|\sqrt{4k^2y^2+1}}\Big)$$
$$=\lim_{y\to0}\Big(\frac{|y|y^2(1+k^2)}{\sqrt{4k^2y^2+1}}\Big)=0.$$

If you enjoy it, you can try to approach the origin along the sine curves $(x,k\sin(x))$ and other curves according to your taste and imagination, but perhaps it is time to accept the suspicion that the limit actually exists and is equal to zero, and start focusing on proving this claim. It does not seem possible to somehow algebraically cancel the parts of expressions that create the indeterminate zeros. Also, no obvious substitution suggests itself, as the numerator and the denominator do not have a common basis. What else is left? Perhaps comparison. It would be nice to find some comparison that works with the distance from a point to the origin, because when $(x,y)\to(0,0)$, then necessarily $\sqrt{x^2+y^2}\to0$, perhaps we could use it somehow.

What estimates do we need? In order to enforce $f \to 0$, it is enough to make sure that $|f| \to 0$, that is, we use the absolute value version of the comparison test. We are therefore looking for an **upper** estimate for the function $|f|$. However, taking into account that obviously $f \geq 0$, it follows that it is enough to find a suitable upper estimate for f itself. This in turn means that we are looking for a lower estimate for the denominator, and preferably one that would involve that distance from the origin. We try this one:

$$\sqrt{4x^2+y^2} \geq \sqrt{x^2+y^2} \implies \frac{1}{\sqrt{4x^2+y^2}} \leq \frac{1}{\sqrt{x^2+y^2}}.$$

It was actually easy. The numerator will be a bit tougher due to that y^4. The key here is to realize that as $(x,y) \to (0,0)$, then eventually x, y become small, in particular we then should have $|y| < 1$. But then $y^4 < y^2$ and we can estimate

$$x^2+y^4 < x^2+y^2 = \sqrt{x^2+y^2}^2.$$

We put it together:

$$|f(x,y)| = \frac{x^2+y^4}{\sqrt{4x^2+y^2}} \leq \frac{\sqrt{x^2+y^2}^2}{\sqrt{x^2+y^2}} = \sqrt{x^2+y^2} = h(x,y).$$

Since $h(x,y) = \sqrt{x^2+y^2} \to 0$ at the origin, by the comparison test it follows that $f(x,y) \to 0$ at $(0,0)$.

△

A minor modification of the above approach shows that

$$\frac{\sqrt{4x^2+y^2}}{x^2+y^4} \geq \frac{1}{\sqrt{x^2+y^2}} \to \infty \qquad \text{as } (x,y) \to (0,0).$$

It is obvious that the estimates worked out only because this problem was carefully prepared, otherwise we would be in for some interesting times.

Example: We investigate the limit of the function $f(x,y) = \dfrac{x^2y}{x^4+y^2}$ at the point $(0,0)$.

Also this problem makes sense and leads to the limit type $\frac{0}{0}$. We go to the approach along lines right away.

$$\lim_{(x,kx)\to(0,0)} \Big(\frac{x^2y}{x^4+y^2}\Big) = \lim_{x\to 0} \Big(\frac{kx^3}{x^4+k^2x^2}\Big) = \lim_{x\to 0} \Big(\frac{kx}{x^2+k^2}\Big) = \frac{0}{0+k} = \frac{0}{k} = 0.$$

Really? What if $k = 0$? Then the calculation actually fails at the last step. However, when exploring the case $k = 0$, we already have this information right from the start and we can apply it already before doing the fatal last step.

$$\lim_{(x,0\cdot x)\to(0,0)} \Big(\frac{x^2y}{x^4+y^2}\Big) = \lim_{(x,0)\to(0,0)} \Big(\frac{x^2y}{x^4+y^2}\Big) = \lim_{x\to 0} \Big(\frac{0}{x^4+0}\Big) = \lim_{x\to 0}(0) = 0.$$

So it is the zero, but we had to justify it correctly.

Could it be that this limit is actually zero? To cover all bases we try a parabolic approach before we try to prove existence of the limit. Let's check on the simplest path, $y = x^2$.

$$\lim_{(x,x^2)\to(0,0)} \Big(\frac{x^2y}{x^4+y^2}\Big) = \lim_{x\to 0} \Big(\frac{x^4}{x^4+x^4}\Big) = \lim_{x\to 0} \Big(\frac{1}{2}\Big) = \frac{1}{2}.$$

We've got a different answer, great, we are done.

Conclusion: f does not have a limit at $(0,0)$.

△

Example: We investigate the limit of the function $f(x,y) = \dfrac{e^{x^2+y^2}-1}{x^2+y^2}$ at the point $(0,0)$.

Surprisingly enough, also this limit makes sense and leads to the expression $\frac{0}{0}$. However, now we face more than just polynomials, so we start carefully going along the axes.

$$\lim_{(x,0)\to(0,0)}\Big(\frac{e^{x^2+y^2}-1}{x^2+y^2}\Big)=\lim_{x\to 0}\Big(\frac{e^{x^2}-1}{x^2}\Big)\overset{\substack{\frac{0}{0}\\ \text{l'H}}}{=}\lim_{x\to 0}\Big(\frac{2x\,e^{x^2}}{2x}\Big)=\lim_{x\to 0}(e^{x^2})=1.$$

That one-dimensional limit could have been also handled through a Taylor expansion, using the substitution $t=x^2$, and some people actually simply remember the answer for this type of limit.

By symmetry, the limit along the y-axis would work out exactly the same. How about straight lines?

$$\lim_{(x,kx)\to(0,0)}\Big(\frac{e^{x^2+y^2}-1}{x^2+y^2}\Big)=\lim_{x\to 0}\Big(\frac{e^{x^2(1+k^2)}-1}{x^2(1+k^2)}\Big)=1$$

by a similar calculation.

If you want, try approaching the origin along parabolas, but perhaps it is time to ask whether the limit actually could be 1. How would we prove it? One possible approach uses Taylor expansion, but it is not clear whether we are ready for this with more variables. Regarding comparison, there is no apparent one that would help. However, we can notice that the variables always appear in the form of the same expression, which suggests substitution.

$$\lim_{(x,y)\to(0,0)}\Big(\frac{e^{x^2+y^2}-1}{x^2+y^2}\Big)=\left|\begin{array}{c} r=x^2+y^2 \\ (x,y)\to(0,0)\implies r\to 0\end{array}\right|=\lim_{r\to 0}\Big(\frac{e^r-1}{r}\Big)\overset{\substack{\frac{0}{0}\\ \text{l'H}}}{=}\lim_{r\to 0}(e^r)=1.$$

△

Facts about limit and continuity

For the sake of completeness we will recall here some key statements.

Theorem. (on limit and operations)
Let f: $D(f)\mapsto\mathbb{R}$ and g: $D(g)\mapsto\mathbb{R}$ be functions, where $D(f),D(g)\subseteq\mathbb{R}^n$.
Assume that f,g are defined on some reduced neighborhood of a certain point $\vec{a}$. Then the following are true:
(i) $\lim\limits_{\vec{x}\to\vec{a}}\big((f+g)(\vec{x})\big)=\lim\limits_{\vec{x}\to\vec{a}}\big(f(\vec{x})\big)+\lim\limits_{\vec{x}\to\vec{a}}\big(g(\vec{x})\big)$,
(ii) $\lim\limits_{\vec{x}\to\vec{a}}\big((f-g)(\vec{x})\big)=\lim\limits_{\vec{x}\to\vec{a}}\big(f(\vec{x})\big)-\lim\limits_{\vec{x}\to\vec{a}}\big(g(\vec{x})\big)$,
(iii) $\lim\limits_{\vec{x}\to\vec{a}}\big((f\cdot g)(\vec{x})\big)=\lim\limits_{\vec{x}\to\vec{a}}\big(f(\vec{x})\big)\cdot\lim\limits_{\vec{x}\to\vec{a}}\big(g(\vec{x})\big)$,
(iv) $\lim\limits_{\vec{x}\to\vec{a}}\Big(\frac{f}{g}(\vec{x})\Big)=\dfrac{\lim\limits_{\vec{x}\to\vec{a}}(f(\vec{x}))}{\lim\limits_{\vec{x}\to\vec{a}}(g(\vec{x}))}$,
(v) $\lim\limits_{\vec{x}\to\vec{a}}\big((f^g)(\vec{x})\big)=\lim\limits_{\vec{x}\to\vec{a}}\big(f(\vec{x})\big)^{\lim\limits_{\vec{x}\to\vec{a}}(g(\vec{x}))}$,
assuming that the right-hand sides make sense.

Theorem. (on limit of a composition)
Let g: $D(g)\mapsto\mathbb{R}$ be a function, where $D(g)\subseteq\mathbb{R}^n$.
Assume that g is defined on some reduced neighborhood of a certain point $\vec{a}$, that $b=\lim\limits_{\vec{x}\to\vec{a}}\big(g(\vec{x})\big)$ exists and that there is a reduced neighborhood $P\subseteq D(g)$ of $\vec{a}$ such that $g(\vec{x})\neq b$ for $\vec{x}\in P$.
Assume further that h is a real function defined on some reduced neighborhood of b. Then

$$\lim_{\vec{x}\to\vec{a}}\big((h\circ g)(\vec{x})\big)=\lim_{t\to b}\big(h(t)\big).$$

Sometimes it is convenient to rewrite the conclusion as $\lim_{\vec{x}\to\vec{a}}\big(h(g(\vec{x}))\big) = h\big[\lim_{\vec{x}\to\vec{a}}\big(g(\vec{x})\big)\big]$. This works as long as the expression on the right makes sense. In calculations it often happens that h is continuous at b which is not an isolated point of $D(h)$, then we no longer need to worry about the condition $g(\vec{x}) \neq b$.

The Heine theorem, and in general the idea of investigating limits by passing to subsets, can be actually used in more generality when investigating a limit relative to a certain set.

Theorem.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$.
Let $\vec{a} \in \mathbb{R}^n$ and $N \subseteq D(f)$ be such that $\vec{a}$ is an accumulation point of N, let $L \in \mathbb{R}^*$.
(i) $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}}\big(f(\vec{x})\big) = L$ if and only if $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in M}}\big(f(\vec{x})\big) = L$ for all $M \subseteq N$ such that $\vec{a}$ is an accumulation point of M.
(ii) $\lim\limits_{\substack{\vec{x}\to\vec{a}\\ \vec{x}\in N}}\big(f(\vec{x})\big) = L$ if and only if $\lim\limits_{k\to\infty}\big(f(\vec{x}(k))\big) = L$ for all sequences $\{\vec{x}(k)\} \subseteq N - \{\vec{a}\}$ such that $\vec{x}(k) \to \vec{a}$.

In more dimensions we also have the theorem linking limit and boundedness.

Theorem.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. Let $\vec{a} \in \mathbb{R}^n$ and $N \subseteq D(f)$ be such that $\vec{a}$ is an accumulation point of N.
If the limit of f at $\vec{a}$ relative to N exists and is finite, then there is some neighborhood U of $\vec{a}$ such that f is bounded on $U \cap N$.

Now some facts about continuity.

Fact.
Functions created out of elementary functions using algebraic operations and compositions are continuous on their domains.

Theorem. (Extreme value theorem)
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. If M is a bounded closed subset of $D(f)$ and f is continuous on M, then f attains its minimum and maximum on M, that is, there are $\vec{x}_{\min}, \vec{x}_{\max} \in M$ such that $f(\vec{x}_{\min}) = \inf\{f(\vec{x});\ \vec{x} \in M\}$ and $f(\vec{x}_{\max}) = \sup\{f(\vec{x});\ \vec{x} \in M\}$.

Theorem. (Intermediate value theorem)
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. If f is continuous, then for every connected set $M \subseteq D(f)$ also the image $f[M]$ is connected.

Advanced corner. In many applications we appreciate the property of continuous functions that when we are at some point $\vec{a}$ and we move by a bit, then also f changes just a little. Sometimes we need to know how large this little is, and this can be troublesome, even in case of one variable. For instance, we know that the functions $\arctan(x)$, $\frac{1}{x}$, and x^2 are continuous. Now imagine that we are at some point a and we move by $\Delta x = 0.1$.

The function $\arctan(x)$ does not care where a is, after such a small change in variable it simply cannot change its value by more than 0.1. However, this is not true about the other two. Looking at the graph of $y = \frac{1}{x}$ we see that if we take a close to the origin, then shifting by Δx could result in a huge change in function value, and there is no upper bound for this change. The closer to 0 we get with a, the larger the change in function value resulting from the shift by 0.1, and we can achieve arbitrarily large changes. The function $y = x^2$ has a similar problem near infinity.

However, it is possible to show that, say, on the interval $[0, 13]$ the function x^2 no longer causes troubles, a shift by 0.1 cannot make a larger change than 2.6. Similarly, $\frac{1}{x}$ is well-behaved on the interval $[1, \infty)$.

Functions whose change can be controlled deserve a special name.

Definition.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. Let $M \subseteq D(f)$.
We say that f is **uniformly continuous** on M if for every $\varepsilon > 0$ there is $\delta > 0$ such that $\|f(\vec{x}) - f(\vec{y})\| < \varepsilon$ for all $\vec{x}, \vec{y} \in M$ satisfying $\|\vec{x} - \vec{y}\| < \delta$.

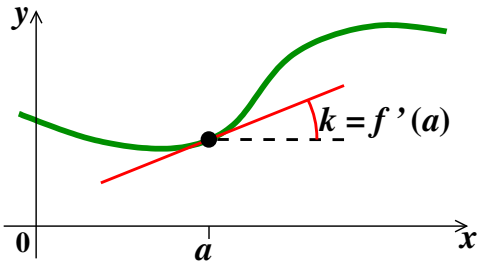
Some functions are uniformly continuous on their whole domain, for instance the aforementioned $\arctan(x)$, also $\sin(x)$ or $\frac{1}{x^2+1}$. However, this has nothing to do with boundedness, for instance $\sin(x^2)$ is bounded but not uniformly continuous on $\mathbb{R}$. Conversely, the functions $\sqrt{x}\sin\left(\sqrt{x}\right)$ or simply $13x$ are not bounded on $[0, \infty)$, but they are uniformly continuous there.

Many useful functions are not uniformly continuous on $\mathbb{R}$ (say, e^x or powers), but according to the following statement we can get to uniform continuity by passing to a subset, which is often enough.

Theorem.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. If M is a bounded and closed subset of $D(f)$ and f is continuous on M, then f is uniformly continuous on M.

3. Introduction to derivatives

We start by recalling the major interpretations of derivative for a function of one variable. In applications we heavily use the fact that the derivative $f'(a)$ tells us the rate of change of the function f (how fast its value changes) when the variable passes through the chosen point a. The geometric interpretation is that $f'(a)$ gives the slope of the tangent line to the graph of f at the point a, which characterizes the inclination of the graph at the point $\big(a, f(a)\big)$.

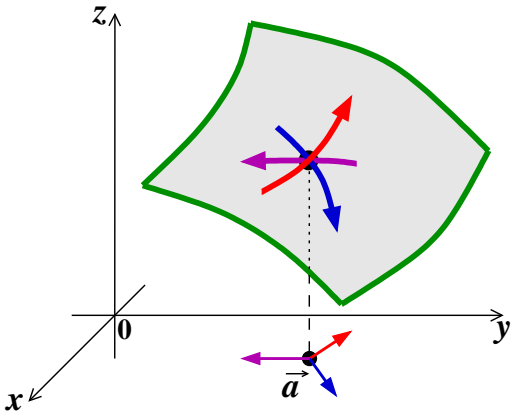


The third useful point of view is that using derivative we can approximate values of the function on some neighborhood of a using the tangent line, the formula is (in two versions)

$$f(x) \approx f(a) + f'(a) \cdot (x - a), \qquad f(a + h) \approx f(a) + f'(a)h.$$

3a. Derivative in higher dimension

It is clear that these ideas would have to be heavily modified if we want to apply them to functions of more variables. We see it in the picture of a function of two variables: When we choose a point $\vec{a} \in D(f)$, then the rate of change of the graph when passing through $\vec{a}$ depends which way we go, and unlike the case of one variable, here we significantly more possible directions and thus also answers.



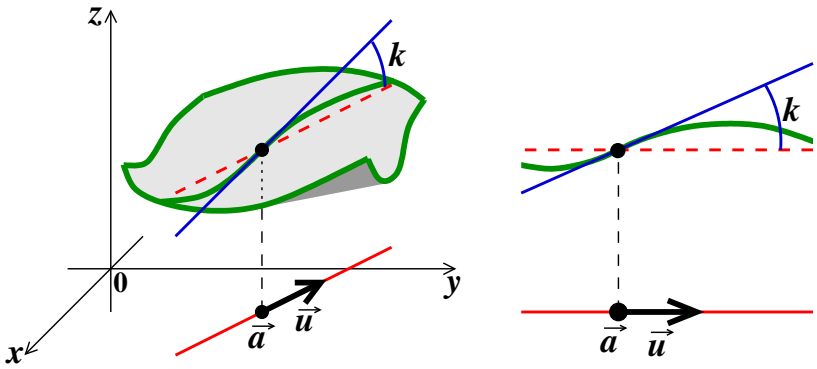
However, these observations bring us to one important notion. If somebody does tell us in which direction to go from $\vec{a}$, then the question how fast the graph grows does make sense, and it also makes sense to attach a line as the tangent and use it to approximate function values (in that chosen direction).

Example: Consider the function $f(x, y) = x^2 + y^2$, we are at the point $\vec{a} = (1, 2)$. What happens when we start off in the direction $\vec{u} = (h, k)$?

We move along the line given by the parametric equation $(x, y) = (1, 2) + t(h, k)$, along the way we meet values

$$\varphi(t) = f(1 + th, 2 + tk) = (1 + th)^2 + (2 + tk)^2 = (h^2 + k^2)t^2 + (2h + 4k)t + 5.$$

This is just another example of slicing, as we saw it already in the first chapter. We cut the graph of f with a vertical plane above the line given by the formula $t \mapsto \vec{a} + t\vec{u}$ and expect to get a two-dimensional situation that we can describe using one variable.



Indeed, exactly such description is provided by our function φ. We can differentiate it; we visit the point $\vec{a}$ at time $t = 0$ and at that moment, the value of φ changes at the rate

$$\varphi'(0) = 2t(h^2 + k^2) + (2h + 4k)\big|_{t=0} = 2h + 4k.$$

△

How can we interpret this result? For instance, if we start off from the point $\vec{a}$ in the direction $\vec{v} = (-1, 1)$, then differentiating $\varphi(t) = 2t^2 + 2t + 5$ at time $t = 0$ we get number 2. What does it mean? It is the rate at which we see (subjectively) the value of f changing while passing through the point $\vec{a}$. However, this observed rate depends on how fast we move, that is, on the magnitude of the directional vector $\vec{v}$. Therefore this number 2 does not carry any geometric information that we like to have when working with tangent lines and approximations.

If we want the observed and geometric properties to agree, then we need to use directional vectors of magnitude 1. In our case we would use the vector $\vec{u} = \frac{(-1,1)}{\|(-1,1)\|} = \frac{1}{\sqrt{2}}(-1, 1)$. Repeating the calculations above we find that in this direction, the graph is changing at the rate $\frac{1}{\sqrt{2}}(-2+4) = \sqrt{2}$.

This information now has a geometric meaning as well, for instance it gives the slope of a "directional tangent line", that is, the tangent line that we would construct using the real slice through the graph at $\vec{a}$.

Understandably, we prefer to work with directional vectors of magnitude 1 in mathematics. However, we will develop a general theory, because directional vectors of arbitrary magnitude can be useful in applications, for instance in physics (velocity vectors).

Remark: Using the tangent line we are now able to approximate values of the function in the direction $\vec{u}$. For the function φ we have $\varphi(t) \approx \varphi(0) + \varphi'(0)t$. If we use this with the normalized vector $\vec{u}$ and pass back to f, we get the formula

$$f\Big((1, 2) + \frac{1}{\sqrt{2}}(-1, 1)t\Big) \approx 5 + \sqrt{2}\,t, \quad \text{that is,} \quad f\Big(1 - \frac{1}{\sqrt{2}}t, 2 + \frac{1}{\sqrt{2}}t\Big) \approx 5 + \sqrt{2}\,t.$$

Substitution $s = \frac{1}{\sqrt{2}}t$ leads to an equivalent but more pleasant formula

$$f(1 - s, 2 + s) \approx 5 + 2s.$$

It is interesting that we can get this nicer form directly if we use φ with the original directional vector $\vec{v}$. So if we are not interested in precise shape of graph, just going for approximation, then we need not normalize. In any case, the conclusion is that if we want to move from $\vec{a}$ only in that particular direction, then we can approximate values of the function (for small t) using the formula

$$f(1 - t, 2 + t) \approx f(1, 2) + 2t.$$

△

The rate of change of f at $\vec{a}$ was found as $\varphi'(0)$. What were we actually calculating?

$$\varphi'(0) = \lim_{t\to 0}\Big(\frac{\varphi(t) - \varphi(0)}{t}\Big) = \lim_{h\to 0}\Big(\frac{f(\vec{a} + t\vec{u}) - f(\vec{a})}{t}\Big).$$

So we can get to this information directly, without the auxiliary function φ. That last limit hopefully looks familiar to the reader, and we are not surprised to see it in the following definition.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Let $\vec{u}$ be a vector from $\mathbb{R}^n$.
We say that the function f is **differentiable** at point $\vec{a}$ in direction $\vec{u}$ if the limit $\lim_{t\to 0}\left(\frac{f(\vec{a}+t\vec{u})-f(\vec{a})}{t}\right)$ exists and is finite.
Then we define the **(directional) derivative** of f at point $\vec{a}$ in direction $\vec{u}$ as

$$D_{\vec{u}}f(\vec{a}) = \lim_{t\to 0}\Big(\frac{f(\vec{a}+t\vec{u})-f(\vec{a})}{t}\Big).$$

By the way, when choosing a direction we obviously need a non-zero vector $\vec{u}$. We did not put this restriction in the definition, but there is no harm done. For $\vec{u} = \vec{0}$ the definition yields $D_{\vec{0}}f = 0$, which will not be of any use in applications, but it does not spoil various theorems.

Now we can express one of the results in the above example as $D_{(-1,1)}f(1,2) = 2$. This example also suggests that perhaps it makes no difference at which stage of the calculation we do the normalization of our directional vector. The following statement confirms it.

Fact.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Let $\vec{u}$ be a vector from $\mathbb{R}^n$ and assume that f is differentiable at $\vec{a}$ in the direction $\vec{u}$.
Then f is also differentiable in the direction $\lambda\vec{u}$ for any $\lambda \in \mathbb{R}$ and we have

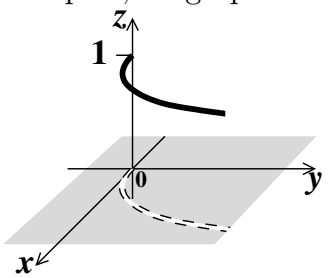
$$D_{\lambda\vec{u}}f(\vec{a}) = \lambda D_{\vec{u}}f(\vec{a}).$$

One can actually prove much more. If we fix some direction $\vec{u}$, then the corresponding directional derivative behaves just like the usual derivative for functions of one variable, in particular we still have the rules for calculation like the product rule and such. Even the Mean value theorem is still valid, see the chapter More on derivative.

However, investigating derivative in just one direction cannot be expected to yield sufficient information about the behaviour of our function. Common sense suggests that if we want to understand this behaviour, we should consider derivatives in all directions. Surprisingly enough, it turns out that this is still not enough. We will show one interesting complication.

Example: Consider the function $f(x,y) = \begin{cases} 1, & y = x^2, x > 0; \\ 0, & \text{elsewhere.} \end{cases}$

We have already met it in the limits chapter, its graph looks like this:



We found that the limit at the origin $(0,0)$ does not exist, now we check on directional derivatives. Take an arbitrary vector $\vec{u} \neq \vec{0}$. As we observed already in our previous encounter with this function, the line passing through the origin in direction $\vec{u}$ either does not meet the parabola $y = x^2$ at all, or intersects it just once. Either way it means that there is some neighborhood of

the origin so that as we approach the origin along our line, then we only see values zero for our function. This means that we perceive the function as being constant, hence $D_{\vec{u}}f(0,0) = 0$.

In particular, we see that the directional derivatives exist at $(0,0)$ in all directions and are even the same, yet the function f is not continuous at $(0,0)$.

△

This shows that some analogue of the classical statement "differentiable hence continuous" does not work for directional derivatives. Thus they do not constitute a proper generalization of the notion of derivative. However, this does not mean that they would not be very useful, as we will shortly see.

In the first chapter we saw that the most convenient slices are those parallel with coordinate axes, because then we do not need to introduce a new parameter, we work with coordinate functions $x \mapsto (x, y_0, z_0, \dots)$, $y \mapsto (x_0, y, z_0, \dots)$ and so on, and directional vectors are of norm one so things are as good as they can get.

So what do we get when we differentiate in the direction $\vec{e}_1$, that is, along the x-axis? We work with the parametric formula $x \mapsto (x, y_0, z_0, \dots)$, obtaining the function $\varphi(x) = f(x, y_0, z_0, \dots)$ that we want to differentiate with respect to x. We see that we do not really need any new function, it is enough to fix the other variables in f and differentiate by x in the usual way.

Example: We return to the function $f(x,y) = x^2 + y^2$, we are interested in derivative in the y-direction at the point $(1,2)$.

First we try it by definition. We move along the parametric line $t \mapsto (1,2)+t(0,1)$ with directional vector $\vec{u} = (0,1)$, giving rise to the function $\varphi(t) = 1^2 + (2+t)^2 = t^2 + 4t + 5$. Then our previous calculations lead to $D_{\vec{u}}f(1,2) = \varphi'(0) = 4$.

Alternative approach: We take the function $f(x,y)$, substitute 1 for x and differentiate the resulting formula $f(1,y) = 1^2 + y^2$ "with respect to y" in the usual way: $[1+y^2]' = 2y$. Finally we substitute $y = 2$ and obtain the same result.

△

In such an easy way we can obtain derivative at arbitrary general point $\vec{a} = (x_0, y_0)$, for instance in the direction of the x-axis we find the derivative by differentiating the function $x^2 + y_0^2$, where y_0 is now some constant (unknown, but it is a fixed number). Since the derivative of the constant y_0^2 (regardless of its value) is zero, we obtain $[x^2 + y_0^2]' = 2x$, therefore the derivative at (x_0, y_0) in the x-direction is $2x_0$.

In real life calculations we do not write those subscript zeros, we simply say that the derivative of $f(x,y) = x^2 + y^2$ in the direction x is $2x$ and it is understood that this applies to arbitrary point (x,y). Similarly, derivative in the direction of the y-axis is $2y$. And that's the whole secret.

Because these derivatives are so easy to derive and the axial directions are the most important, it is no surprise that this whole idea has a special name.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Consider unit vectors $\vec{e}_i$ in axial directions, $\vec{e}_1 = (1,0,0,\dots,0)$, $\vec{e}_2 = (0,1,0,\dots,0)$, $\dots$, $\vec{e}_n = (0,0,0.\dots,1)$.
For $i = 1,\dots,n$ we define the **partial derivative** of f with respect to x_i as $\dfrac{\partial f}{\partial x_i}(\vec{a}) = D_{\vec{e}_i}f(\vec{a})$, if this exists.

Alternative notations: $\frac{\partial f}{\partial x_i}(\vec{a}) = \frac{\partial}{\partial x_i}[f](\vec{a}) = D_i f(\vec{a}) = f'_{x_i}(\vec{a}) = f_{x_i}(\vec{a})$.

In calculations we differentiate with respect to the given variable simply by imagining that the other variables (and expressions that they create) are constant and we differentiate by the given variable following the usual rules.

Example: We find all partial derivatives of the function $f(x,y,z) = x^2y + \sin(y^3 + 2z)$. We will start slowly and with details.

We find the partial derivative with respect to x by imagining that y and z are some particular numbers, for instance $\sqrt{2}$ and π, then $\sin(\sqrt{2}^3 + 2\pi)$ is also a number, that is, a constant. Differentiation thus yields

$$[x^2\sqrt{2} + \sin(\sqrt{2}^3 + 2\pi)]' = 2x \cdot \sqrt{2} + 0.$$

Of course, this is not what we wanted, but this auxiliary calculation allows us to trace the fate that befell the variables y and z when treated as constants. In particular, in the first term the $\sqrt{2}$ representing some chosen value for y remained, because of its role as a multiplicative constant. Exactly the same role will be played in the product x^2y by y when we imagine that it is constant. So if we take "y" and "z" as constants, the same reasoning as before leads us to the result

$$f_x(x,y) = \frac{\partial f}{\partial x}(x,y) = 2x \cdot y + 0 = 2xy.$$

Similarly, to get partial derivative with respect to y we imagine that instead of x and z there are constants, say $\sqrt{5}$ and π, and we calculate

$$[\sqrt{5}^2 y + \sin(y^3 + 2\pi)]' = \sqrt{5}^2 + \cos(y^3 + 2\pi) \cdot 3y^2,$$

that is,

$$f_y(x,y) = \frac{\partial f}{\partial y}(x,y) = x^2 + \cos(y^3 + 2z) \cdot 3y^2.$$

Of course, an experienced derivator (as in "terminator") does not write numbers, just learns to pretend that there are constants at the right places and analyze the resulting expression. We still owe you the derivative by z, for that we take x and y to be constants, then also the whole term x^2y is constant. Therefore

$$f_z = \frac{\partial f}{\partial z} = 0 + \cos(y^3 + 2z) \cdot 2.$$

Now we did not write (x,y) next to the derivative symbol, people often leave it out.

△

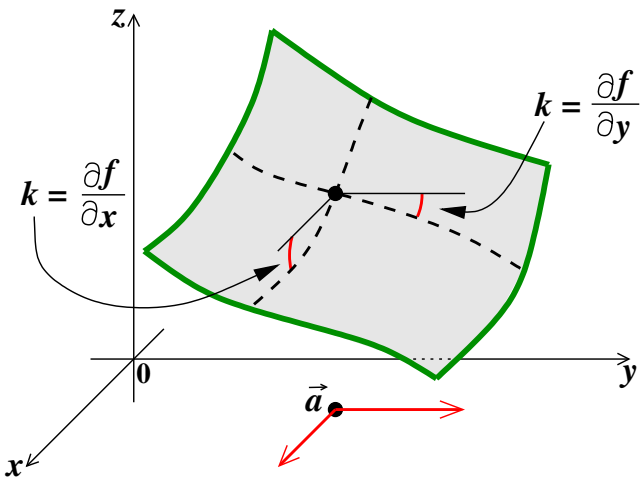
The convenient subscript notation f_x definitely looks preferable, but it is not as flexible as the notation with the curved shape that surprisingly does not have a short name (well, you can call it the "partial derivative mark"). The symbol $\frac{\partial}{\partial x_i}$ can be used just like the usual derivative mark to indicate derivative of a particular expression, but we write it on the left. For instance, in the last calculation above we can indicate application of the chain rule as follows:

$$\frac{\partial}{\partial z}[\cos(y^3 + 2z)] = \cos(y^3 + 2z) \cdot \frac{\partial}{\partial z}[y^3 + 2z] = 2\cos(y^3 + 2z).$$

We will therefore mostly use this more complicated notation.

3b. The meaning of partial derivatives

We already know that partial derivatives tell us how fast the function changes (grows, falls) in key directions.



It might seem that in other directions, a function is free to do whatever it wants, and that is true.

Example: Consider $f(x,y) = \dfrac{x^2y^2}{x^4+y^4}$ for $(x,y) \neq (0,0)$ and $f(0,0) = 0$.

First we find derivatives at points $(x,y) \neq (0,0)$ to get some practice. There we always find some neighborhood on which our function is defined by a specific formula, so we can apply the classical rules.

$$\frac{\partial f}{\partial x} = \frac{2xy^2(x^4+y^4) - x^2y^2 \cdot 4x^3}{(x^4+y^4)^2} = \frac{2xy^6 - 2x^5y^2}{(x^4+y^4)^2},$$
$$\frac{\partial f}{\partial y} = \frac{2x^2y(x^4+y^4) - x^2y^2 \cdot 4y^3}{(x^4+y^4)^2} = \frac{2x^6y - 2x^2y^5}{(x^4+y^4)^2}.$$

However, the really interesting things are happening at the point $(0,0)$. The function is not defined by that algebraic formula there, so we have to find derivatives using the definition:

$$\frac{\partial f}{\partial x}(0,0) = \lim_{x\to 0}\Big(\frac{f(x,0) - f(0,0)}{x}\Big) = \lim_{x\to 0}\Big(\frac{0-0}{x}\Big) = 0.$$

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

So the partial derivatives at the origin do exist, but surprisingly, derivatives at other directions do not. In order to see that, we choose $k \neq 0$ and attempt to move away from the origin along a line with equation $y = kx$ for $x > 0$, using the directional vector, say, $\vec{u} = (1,k)$. Then we encounter function values $f(x,kx) = \frac{k^4}{1+k^4}$ for $x \neq 0$, while $f(0,0) = 0$. The directional derivative therefore is

$$D_{\vec{u}} f(0,0) = \lim_{h\to 0}\Big(\frac{f(h,kh) - f(0,0)}{h}\Big) = \lim_{h\to 0}\Big(\frac{k^4}{1+k^4}\cdot\frac{1}{h}\Big) = \frac{1}{0}.$$

So this directional derivative does not exist.

△

This shows that in general, knowing just partial derivatives is not sufficient to learn about other directions. However, if we prevent the graph of the function from having "sharp bends", then this freedom of behaviour is lost. It is perhaps surprising that relatively mild assumptions on the function guarantee that its raise and fall in all directions is completely determined by its behaviour in coordinate directions.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. If there exists some neighborhood of $\vec{a}$ on which partial derivatives $\frac{\partial f}{\partial x_i}(\vec{x})$ exist for all $i = 1, \dots, n$ and they are continuous at $\vec{a}$, then f has directional derivatives at $\vec{a}$ in all directions and for every $\vec{u}$ the following is true:

$$D_{\vec{u}} f(\vec{a}) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a}) \cdot u_i.$$

The requirement on continuity of derivatives is often satisfied, essentially every function given by a formula fits in, and for such functions we can deduce derivative in arbitrary direction purely from knowing partial derivatives. This means that the condition of continuous derivatives has a rather large impact.

For convenient manipulation we usually store all partial derivatives in one packet.

Definition.
Let f be function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. If all partial derivatives $\frac{\partial f}{\partial x_i}(\vec{a})$ for $= 1, \dots, n$ exist, then we define the **gradient** of f at $\vec{a}$ as the vector

$$\nabla f(\vec{a}) = \Big(\frac{\partial f}{\partial x_1}(\vec{a}), \dots, \frac{\partial f}{\partial x_n}(\vec{a}) \Big).$$

Alternative notations: $\nabla f(\vec{a}) = \vec{\nabla} f(\vec{a}) = \operatorname{grad}(f)(\vec{a})$.

The symbol "nabla" ∇ is universally used, the version $\vec{\nabla}$ is sometimes read as "del". The notation $\operatorname{grad}(f)$ is nice in that it is very upfront about what it means, but it is too long for my taste. I will therefore stick with nablas.

It is worth noting that gradient is a vector from $\mathbb{R}^n$, that is, we see it as an object from the function's domain; on a symbolic graph we see it within the horizontal representation of $D(f)$ (see picture below).

For a function with continuous derivatives (in other words, for most functions that we normally meet) we can express the conclusion of the above theorem in an elegant way using dot product,

$$D_{\vec{u}} f(\vec{a}) = \nabla f(\vec{a}) \bullet \vec{u}.$$

Example: Consider $f(x, y) = x^2 + y^2$ again. We already found its partial derivatives, so we readily write the gradient $\nabla f(x, y) = (2x, 2y)$.

At the point $\vec{a} = (1, 2)$ then $\nabla f(1, 2) = (2, 4)$ and the new formula gives

$$D_{(h,k)} f(1, 2) = \nabla f(1, 2) \bullet (h, k) = 2h + 4k,$$

exactly as we calculated it before by the definition.

△

The gradient has interesting theoretical properties, for instance we can calculate it using familiar rules for operations, or we have the familiar proposition that when a sufficiently smooth function has gradient equal to zero on some region, then it must be constant there. However, we will leave this to the chapter More on derivative, here we are interested in what useful information gradient can provide. It turns out that it is one of key notions for investigating functions.

Gradient and slope

Imagine that we are at a point $\vec{a}$, sitting on the graph and looking around. Depending on in which direction we look, the graph raises or falls. The rate at which it changes is given by the

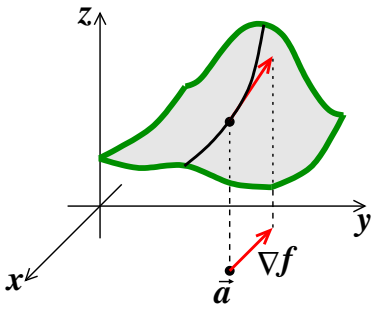
directional derivative. In other words, it is given by the expression $\nabla f(\vec{a}) \bullet \vec{u}$, where $\vec{u}$ are unit vectors.

According to a well-known formula,

$$\nabla f(\vec{a}) \bullet \vec{u} = \|\nabla f(\vec{a})\| \cdot \|\vec{u}\| \cos(\alpha) = \|\nabla f(\vec{a})\| \cos(\alpha),$$

here α is the angle between the vectors $\nabla f(\vec{a})$ and $\vec{u}$.

We are interested in extreme values of this directional derivative. We see that we will climb fastest if we start off so that $\cos(\alpha) = 1$, which happens for $\alpha = 0$, that is, in the direction of the gradient. Conversely, the steepest fall happens when $\cos(\alpha) = -1$, in exactly the opposite direction.

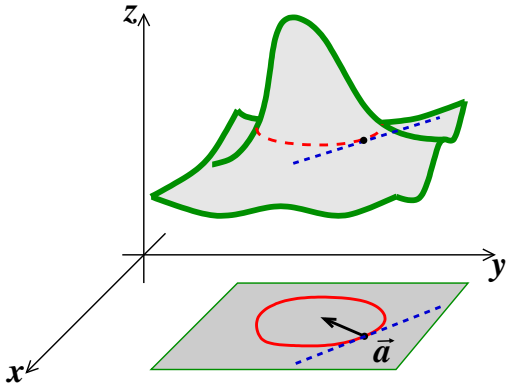


Fact.
Let f be a function that has continuous first partial derivatives on some neighborhood of a point $\vec{a}$. Then the gradient $\nabla f(\vec{a})$ is the direction of steepest growth of the function at $\vec{a}$, the function increases at the rate $\|\nabla f(\vec{a})\|$ there. The vector $-\nabla f(\vec{a})$ is the direction of the steepest descent at $\vec{a}$.

Gradient and level sets

We are still at a point $\vec{a}$ and sitting on the graph. This location on the graph has a certain level (elevation), namely the level $c = f(\vec{a})$, and the point $\vec{a}$ then belongs to the corresponding level set (which is situated in the domain). If we keep walking along this level curve, then our elevation does not change. Assume that the level curve does not break sharply at the point $\vec{a}$ (it is differentiable as a curve, therefore smooth), so it makes sense to talk about a tangent line to this level curve at $\vec{a}$. Consider the directional vector $\vec{u}$ of this tangent line.

If we start off from the point $\vec{a}$ in the direction $\vec{u}$, that is, along the tangent line, then for a tiny while we almost do not stray away from our level curve, and therefore also the function value will essentially stay the same. This means that $D_{\vec{u}} f(\vec{a}) = 0$, that is, $\|\nabla f(\vec{a})\| \cos(\alpha) = 0$, meaning that $\alpha = \frac{\pi}{2}$.



We just reasoned that the direction of the tangent line to the level curve at $\vec{a}$ must be perpendicular to the gradient there. In other words, the direction in which we would walk to stay on the same elevation is at all times perpendicular to the direction of the steepest climb at that point. We confirm it.

Fact.
If a function f has continuous first partial derivatives on some neighborhood of a point $\vec{a}$, then the gradient $\nabla f(\vec{a})$ is perpendicular to the level set passing through $\vec{a}$.

This is very useful. Many objects can be represented as level sets for suitable functions, and then the gradient allows one to easily obtain normal vectors to such an object, which is helpful in many applications.

Example: Consider the ellipse given by the equation $\frac{x^2}{6} + \frac{y^2}{3} = 1$, we want to find its tangent line at the point $(2, 1)$.

One possible approach is through graphs. The given point lies on the upper half of the ellipse, where it can be viewed (after solving the formula for y) as the graph of the function $f(x) = \sqrt{3 - \frac{x^2}{2}}$. To find the tangent line at $x = 2$ we need the derivative $f'(x) = \frac{-x}{2\sqrt{3-\frac{x^2}{2}}}$, slope of the tangent line is therefore $k = f'(2) = -1$. We obtain the line $y - 1 = -(x - 2)$, that is, $x + y = 3$.

Alternative approach: We rewrite the given equation into a more pleasant form $x^2 + 2y^2 = 6$ and decide to see it as the level curve of the function $F(x, y) = x^2 + 2y^2$ corresponding to $c = 6$. We find the gradient at $(2, 1)$: $\nabla F = (2x, 4y)$, therefore $\nabla F(2, 1) = (4, 4)$.

This vector is perpendicular to the level set passing through $(1, 2)$, therefore also to the ellipse, therefore also to its tangent line. The equation of the line perpendicular to $(4, 4)$ and passing through the point $(2, 1)$ is $4(x - 2) + 4(y - 1) = 0$, that is, $x + y = 3$.

△

Gradient and tangents, approximation

Intuitively, tangent line to a graph of a function $f(x)$ is an object of the same nature (the graph is a curved line, the tangent line is a straight line) that touches and merges with the graph at the given point. This suggests how we will want to adopt this idea to the case of more variables. We visualize a graph of a function of two variables, so we see some two-dimensional sheet floating in a three-dimensional space, and it seems natural to touch it with a plane, that is, a flat two-dimensional object.

Similarly, for a function of three variables we would look for tangent three-dimensional spaces (they look "flat" when placed in the four-dimensional space where the graph lives) and so on. In general, by a **hyperplane** in $\mathbb{R}^{n+1}$ we mean an n-dimensional affine linear space. We are looking for special ones.

How do we find it? By leaving the world of geometry for the moment and turning to analytical approach. We know that with functions of one variable, the tangent line at a is the line that better than any other line approximates behaviour of f around a.

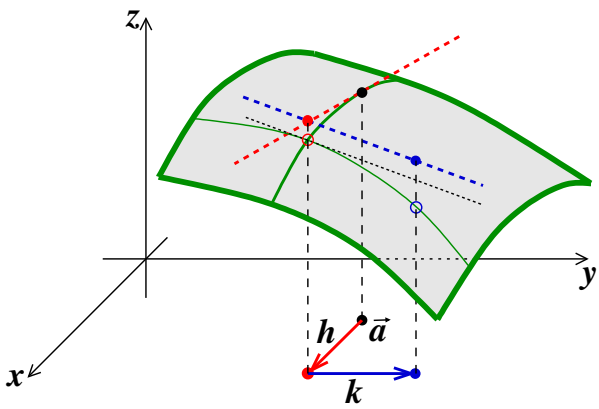
$$f(a + h) \approx f(a) + f'(a)h.$$

How could we best approximate values of a function $f(x, y)$ around a point $\vec{a} = (a, b)$? Assume that we move a tiny bit away from this point, namely by a vector $\vec{u} = (h, k)$. How much does the function change?

Instead of one "diagonal" movement we can arrive at the place $(a+h, b+k)$ by first moving by h along the x-axis and then by k in the direction of y-axis. But now the first movement is a one-dimensional affair, we change only one variable, and we know how to estimate the corresponding change in function, we use differentiation in the appropriate direction:

$$f(a+h,b) \approx f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})h.$$

We see it in the picture marked in red, instead of the value in circle we take the one marked with a full dot.



From the point $(a+h, b)$ we now move in the direction of y-axis by k and similarly estimate

$$f(a+h,b+k) \approx f(a+h,b) + \frac{\partial f}{\partial y}(a+h,b)k.$$

We put it together:

$$f(a+h,b+k) \approx f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})h + \frac{\partial f}{\partial y}(a+h,b)k.$$

The fact that in the last term we do not take derivative at $\vec{a}$ is unpleasant. However, if we assume that this derivative exists on some neighborhood of $\vec{a}$ and is continuous at $\vec{a}$, then for very small values of h we can disregard this shift by h, hoping that the value of $\frac{\partial f}{\partial y}$ would not change much. Then we get

$$f(a+h,b+k) \approx f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})h + \frac{\partial f}{\partial y}(\vec{a})k.$$

We obtained an approximating formula that is linear, and therefore it describes a plane in three dimensions. To see this we use the usual substitution $x = a+h$, $y = b+k$ and obtain the other popular form of approximating formula

$$f(x,y) \approx f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})(x-a) + \frac{\partial f}{\partial y}(\vec{a})(y-b).$$

The expression on the right defines a function whose graph can be described by the formula

$$z = f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})(x-a) + \frac{\partial f}{\partial y}(\vec{a})(y-b)$$

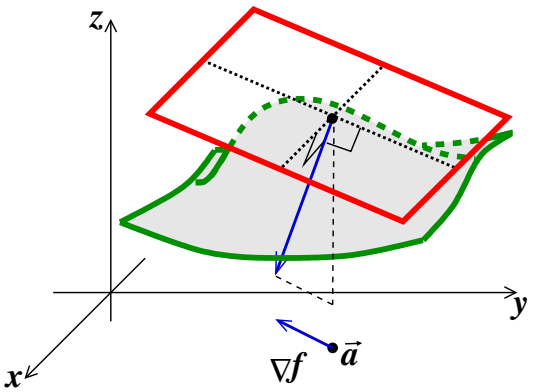
and we recognize this as the equation of a plane, our desired tangent plane. We can simplify it as

$$\frac{\partial f}{\partial x}(\vec{a})x + \frac{\partial f}{\partial y}(\vec{a})y - z = -f(\vec{a}) + \frac{\partial f}{\partial x}(\vec{a})a + \frac{\partial f}{\partial y}(\vec{a})b,$$

that is,

$$\frac{\partial f}{\partial x}(\vec{a})x + \frac{\partial f}{\partial y}(\vec{a})y - z = d$$

for an appropriate d and see that this plane is determined by the normal vector $\left(\frac{\partial f}{\partial x}(\vec{a}), \frac{\partial f}{\partial x}(\vec{a}), -1\right)$



In principle this also makes sense on an intuitive level. Imagine that we place a tangent plane on the top of that hill in the graph. Then it would have to be horizontal, and its normal vector should be vertical. We take it pointing down and with natural magnitude 1, so we have the normal vector $(0, 0, -1)$. If we want to attach a tangent plane to a point on a slope, then we have to tilt this horizontal plane, and it seem obvious that we should tilt the plane, and hence the attached normal vector, in the direction in which the slope goes up. And that is exactly the direction of the gradient (that we can see in the domain). The steeper the slope, the more we have to tilt, but steeper slope also means larger gradient; it all fits.

Similar reasoning works in more dimensions, we have the approximating formula

$$f(\vec{a}+\vec{h}) \approx f(\vec{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a}) h_i = f(\vec{a}) + \nabla f(\vec{a}) \bullet \vec{h}$$

or the version

$$f(\vec{x}) \approx f(\vec{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a})(x_i - a_i) = f(\vec{a}) + \nabla f(\vec{a}) \bullet (\vec{x} - \vec{a})$$

that leads to the equation of the tangent hyperplane

$$z = f(\vec{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a})(x_i - a_i).$$

We confirm our observations with an official statement.

Fact.
Let a function f have continuous first derivatives on some neighborhood of a point $\vec{a}$. If we extend the vector $\nabla f(\vec{a})$ by one coordinate, namely we add -1 as the $(n+1)$th coordinate, we obtain a vector from $\mathbb{R}^{n+1}$ that is perpendicular to the tangent hyperplane to the graph of f at the point $\vec{a}$.

Example: Consider $f(x, y) = x^2 + y^2$ and the point $(1, 2)$. We find the tangent plane to the graph of f at the corresponding point.

We have already found $\nabla f(1, 2) = (2, 4)$. We can therefore take $\vec{n} = (2, 4, -1)$ as a normal vector to the graph.

Through which point should the plane go? Since $f(1, 2) = 5$, the point is $(1, 2, 5)$. We have a point and normal vector, the equation of the plane follows easily:

$$0 = \vec{n} \bullet \big((x, y, z) - (1, 2, 5)\big) = 2(x-1) + 4(y-2) - (z-5) \implies 2x + 4y - z = 5.$$

Alternative: Plane perpendicular to the vector $(2, 4, -1)$ has equation $2x + 4y - z + d = 0$. Substituting in the point $(1, 2, 5)$ we get $d = -5$, hence $2x + 4y - z - 5 = 0$ is the equation.

Another alternative: The graph is given by the equation $z = x^2+y^2$. Rewriting it as $x^2+y^2-z = 0$ we can treat it as the level surface of the function $F(x, y, z) = x^2 + y^2 - z$ corresponding to the

value $c = 0$. We easily find $\nabla F = (2x, 2y, -1)$ and we know that the vector $\nabla F(1, 2, 5) = (2, 4, -1)$ is perpendicular to this level surface, therefore also perpendicular to the graph and in particular to the desired tangent plane. Its equation is therefore

$$2(x - 1) + 4(y - 2) - (z - 5) = 0$$

and we are done.

Conclusion: The tangent plane to the graph of f at the point given by $\vec{a} = (1, 2)$ has the equation $2x + 4y - z = 5$.

We use this example to review other uses of gradient.

The function grows fastest when we leave the point $(1, 2)$ in the direction $\nabla f(1, 2) = (2, 4)$, that is, in the direction $(1, 2)$ (every positive multiple of a vector has the same direction), the rate of growth then will be $\|\nabla f(1, 2)\| = \|(2, 4)\| = \sqrt{20} = 2\sqrt{5}$.

The point $(1, 2) \in D(f)$ lies on the level curve $f(1, 2) = 5$, that is, on the circle given by the equation $y^2 + x^2 = 5$. At the point $(1, 2)$ the vector $\nabla f(1, 2) = (2, 4)$ is perpendicular to this curve, which allows us to easily write the equation of the tangent line to this circle:

$$0 = \nabla f(1, 2) \bullet ((x, y) - (1, 2)) = 2(x - 1) + 4(y - 2) \implies 2x + 4y = 10.$$

Using the normal direction $(1, 2)$ and the popular trick we can obtain a vector from $\mathbb{R}^2$ tangent to the circle, for instance $(2, -1)$.

△

3c. Partial derivatives of higher order

Just like with functions of one variable, also functions of more variables can be differentiated more times (if they allow us). For instance, imagine that we have a function f, directional vector $\vec{u}$, and we find that $D_{\vec{u}} f$ exists on some neighborhood of a point $\vec{a}$. Then this directional derivative in fact creates a function $\vec{x} \mapsto D_{\vec{u}} f(\vec{x})$ on a neighborhood of $\vec{a}$ and as a such it can be differentiated at $\vec{a}$, for instance in the direction $\vec{v}$. We obtain $D_{\vec{v}}[D_{\vec{u}} f] = D_{\vec{v}} D_{\vec{u}} f$. Note the order of differential operators in the expression on the right. They are applied right-to-left, so differentiation in direction $\vec{u}$ comes first.

One could develop a general theory in this way, but we will naturally focus on partial derivatives. We can apply them repeatedly (at least for sufficiently nice functions), and quite obviously we have a choice what derivatives we want to apply. A function of two variables has $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ and both these derivatives can be in turn differentiated by x or by y, obtaining four distinct partial derivatives of order two, for instance the two below. We will show first a detailed record of the procedure and then the standard condensed notation:

$$\frac{\partial}{\partial x}\Big[\frac{\partial}{\partial x}[f]\Big] = \frac{\partial}{\partial x}\Big[\frac{\partial f}{\partial x}\Big] = \frac{\partial^2 f}{\partial x^2}, \qquad \frac{\partial}{\partial x}\Big[\frac{\partial}{\partial y}[f]\Big] = \frac{\partial}{\partial x}\Big[\frac{\partial f}{\partial y}\Big] = \frac{\partial^2 f}{\partial x \partial y}.$$

Note the order of differentiation, the symbols in the denominator are read right to left, so we start with differentiation by the rightmost variable. For instance, in the partial derivative of third order $\frac{\partial^3 f}{\partial x \partial y \partial x}$ we would first differentiate f with respect to x, then the result is differentiated by y and this by x again, whereas to obtain $\frac{\partial^3 f}{\partial x^2 \partial y}$ we would first differentiate with respect to y and then twice by x.

Definition.
Consider a function f defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Let $i_1, i_2, \dots, i_m \in \{1, 2 \dots, n\}$ are some indices of variables. We define the corresponding **partial derivative of order** m of the function f by induction as follows:

$$\frac{\partial^m f}{\partial x_{i_m} \partial x_{i_{m-1}} \cdots \partial x_{i_2} \partial x_{i_1}} = \frac{\partial}{\partial x_{i_m}} \Big[\frac{\partial^{m-1} f}{\partial x_{i_{m-1}} \cdots \partial x_{i_2} \partial x_{i_1}} \Big],$$

assuming that all derivatives that are needed exist.
If all the coordinate indices i_k are not the same, then we call this derivative a **mixed derivative**.

To make things more interesting, for such repeated differentiation we also have the handy subscript notation, but it reverses the order and we naturally apply left-to-right: $[f_x]_y = f_{xy} = \frac{\partial^2 f}{\partial y \partial x}$. If you happen to be unsure about the order, it may help to expand the compact notation into the repeated derivation form and that should suggest the right way. For instance,

$$f_{x^2zy} = [[[f_x]_x]_z]_y = \frac{\partial}{\partial y} \Big[\frac{\partial}{\partial z} \Big[\frac{\partial}{\partial x} \Big[\frac{\partial}{\partial x} [f] \Big] \Big] \Big] = \frac{\partial^4 f}{\partial y \partial z \partial x^2}.$$

With both notations we first differentiate by x twice, then by z and finally by y.

In general, a function of n variables has n^m distinct partial derivatives of order m. It is easy to find examples where the order really matters, but if the reader does so, it turns out that those examples are not exactly friendly. In other words, for nice functions the actual order does not matter, which greatly simplifies life.

Theorem.
If a function f has all partial derivatives of order m on some neighborhood of a point $\vec{a}$ and they are all continuous at $\vec{a}$, then the order of differentiation makes no difference when calculating derivatives up to the order m.

What does this mean? In general there are quite a lot partial derivatives. For instance, if we work with a function of three variables, then we are looking at $3^4 = 81$ partial derivatives of the fourth order. However, if the derivatives of order four are continuous (for instance if our function is given by an algebraic expression built using elementary functions), then it is enough to find just 10 derivatives of third order instead of 27 and 15 instead of 81 for order four. Actually, this savings is more of a theoretical nature, since we rarely need higher than second derivative in applications, but it is a nice thing to have anyway.

Just like derivatives of order one, the higher ones can be also collected into packets.

Definition.
Assume that a function f has all derivatives of order two at a point $\vec{a}$. Then we define its **Hess matrix** at $\vec{a}$ as

$$H(\vec{a}) = \Big(\frac{\partial^2 f}{\partial x_j \partial x_i} \Big)_{i,j=1,\dots,n} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Practically speaking, we differentiate the function f by its first variable, this derivative is then repeatedly differentiated once more, by all available variables, and the results create the first row

of the matrix. We can actually see $\nabla\left(\frac{\partial f}{\partial x_1}\right)$ in the first row, similarly we create the others. On the diagonal we see non-mixed derivatives $\frac{\partial^2 f}{\partial x_i^2}$, mixed derivatives are off the diagonal.

If the function is reasonable and the order of differentiation does not matter, then the Hess matrix is symmetric, so it takes roughly just half the work. Actually, for functions of two variables it may be a good idea to calculate the mixed derivatives independently, since it is not so much extra work and their match serves as a validity check.

To collect derivatives of the third order we would need a three-dimensional matrix, which brings us to tensors, a topic that we definitely do not want to explore here. In many (most?) applications we can do with the first two derivatives, we settle for them here as well.

Because functions with continuous partial derivatives are very convenient, we often restrict our statements to them, and to facilitate this we introduce a handy notation.

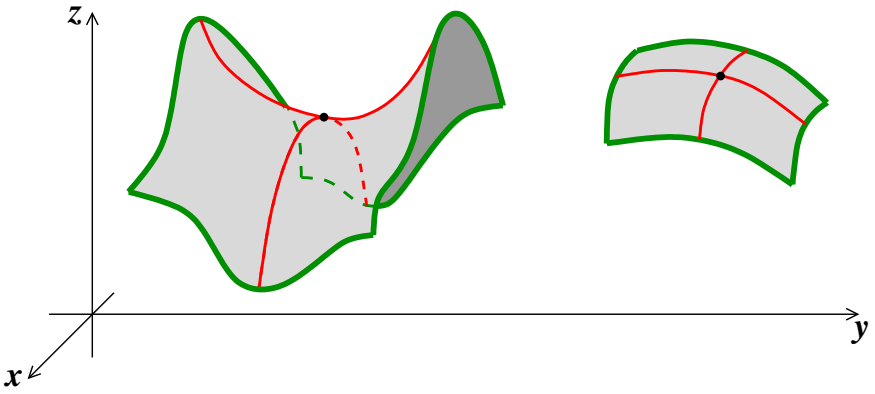
Definition.
Let G be an open set in $\mathbb{R}^n$. For $m \in \mathbb{N}$ the symbol $C^m(G)$ denotes the set of all functions defined (at least) on G that have all partial derivatives of order m at all points of G and these partial derivatives are continuous on G.

Just like with functions of one variable, also here it is true that if all partial derivatives of order m of some function are continuous, then also all partial derivatives of lower orders must be continuous.

3d. The meaning of higher order derivatives

For functions of one variable we usually offer interpretation of first two derivatives. The first derivative provides information about the slope of a tangent line or the rate of change of the function, the second derivative tells us about how much is the graph curved measured as concavity. We already talked about first order partial derivatives, so now we look at interpretation of partial derivatives of the second order.

We start with the easier case, partial derivatives that are not mixed, because when we repeatedly differentiate with respect to the same variable, then we can ignore the other directions. In effect, we work with a slice of our situation that can be treated as a function of one variable. If we slice a function's graph in the direction of the x-axis, then $\frac{\partial^2 f}{\partial x^2}$ determines concavity of the cut, just like we are used to, similar information comes from $\frac{\partial^2 f}{\partial y^2}$, $\frac{\partial^2 f}{\partial z^2}$, etc. This tells us something about the shape of the graph.



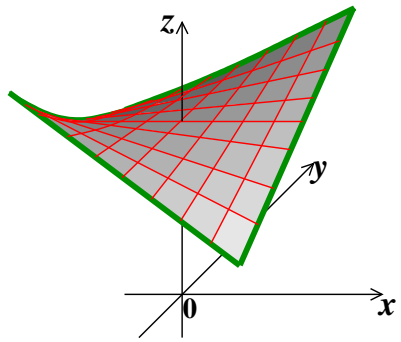
On the left we see the case when non-mixed second partial derivatives have opposite signs at some point, so our function is concave up in one direction and concave down in the other direction there. We can form a fairly good picture of the situation. On the right we see a situation when we have concavity down in both directions.

That second case inspires a question analogous to the one we had for first order derivatives. There we eventualy found that when we know the gradient of a smooth function, that is, we know

its rate of growth in coordinate directions, then it already determines growth in other directions. Is it analogously true that when we know non-mixed second derivatives of a smooth function, that is, concavity in coordinate directions, then it already determines concavity in other slices? The answer is in the negative, to know about concavity of a graph we need information about all partial derivatives of order two. Which brings us to mixed derivatives.

We first look at the derivative $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right]$ and assume that it is positive. We have a function $\frac{\partial f}{\partial x}$ and we differentiate it by y, so we are moving in the y direction, looking at its values. We assumed that this second derivative is positive, which tells us that the function $\frac{\partial f}{\partial x}$ grows as we move in the direction of the y-axis. Geometrically speaking, we are moving in the direction of the y-axis and observe slopes of tangents in the perpendicular direction x. We find that these slopes grow, that is, the tangent lines are turning upwards as we go.

Can you imagine a situation when you move in the y-direction and tangent lines taken in direction x are turning towards faster growth? Such a graph must be twisted. We will show it on a picture where we look at behaviour at the origin. To see the shape better, we rotated the graph so that the x-axis goes to the right, as we are used to when drawing tangent lines. But then the y-axis must necessarily go away from us.



Note that slices of the graph in the x and y directions are straight lines. This is possible (for instance the simple function $f(x,y) = xy$ has this property) and the important thing is that for the function in the picture we have $\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial y^2} = 0$. In other words, the partial derivatives $\frac{\partial^2 f}{\partial x^2}$ and $\frac{\partial^2 f}{\partial y^2}$ do not influence the shape of the graph, so we can directly observe the influence of the mixed second derivative.

If the function is sufficiently smooth, then we should have $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$, so we should get the same picture also when interpreting the expression $\frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right] > 0$: When we move in the x-direction, the slopes of tangents taken in the y-direction are increasing, lines are turning up. The picture fits well, larger slopes of "y-tangents" means steeper growth in the direction of the y-axis, that is, away from us.

To conclude, the meaning of the second mixed derivative is the direction and measure of the graph's twist. Deformation of a graph of this sort is the reason why concavity information in axial directions does not determine the whole shape. When investigating a graph, we have to compare (in a mathematical way) the convexity influence of non-mixed derivatives and the twisting action as indicated by the mixed derivative. Obviously, this awaits us in the section on local extrema. Here we sum up our exploration by saying that information that we deduce from the second derivative in case of one variable is, in case of more variables, encoded in the Hess matrix, where all entries (all derivatives of the second order) play a role of equal importance.

4. Introduction to local and global extrema

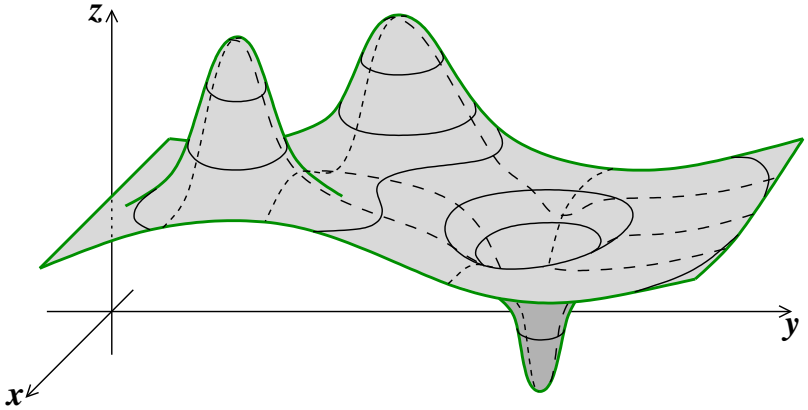
Both local and global extrema can be directly adopted to more variables. Although definitions look the same, we encounter substantial differences when actually investigating extrema.

4a. Local extrema

We just copy the usual definition of a local extreme, replacing points with vectors.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$.
We say that f has a **local maximum** at $\vec{a}$, or that $f(\vec{a})$ is a local maximum, if there exists a neighborhood U of $\vec{a}$ such that $f(\vec{a}) \geq f(\vec{x})$ for all $x \in U$.
We say that f has a **local minimum** at $\vec{a}$, or that $f(\vec{a})$ is a local minimum, if there exists a neighborhood U of $\vec{a}$ such that $f(\vec{a}) \leq f(\vec{x})$ for all $x \in U$.

The picture below for the case of two variables shows two local maxima on the left and a local minimum on the right.



This is how we also imagine these notions for more dimensions. A local maximum has the property that if we slice the graph through that point in any direction (thus passing to the situation of one variable), then on this slice we have a local maximum in the usual meaning on that notion. Analogous property is true for every local minimum.

There is one more interesting kind of behaviour, we see it in the picture between the two hills. If we cut the graph there with a vertical plane leading from one hilltop to another, then on this slice we see a local minimum in the valley. However, if we cut the graph using a perpendicular vertical plane passing between the two summits, then in the valley we see a local maximum on the slice. Such points are called saddles or saddle points and we encounter them when investigating extrema, so they are usually counted among points to explore when a question asks about extrema.

How do we find those local extrema? The procedure is similar to investigating local extrema for functions of one variable. Roughly speaking, first we find candidates using the first derivative, then we classify them using the second derivative.

If we cut the graph by an arbitrary vertical plane through some local extreme, we also get an extreme on the slice, so the derivative in that direction must be equal to zero. This in particular applies to slices parallel to coordinate axes, so partial derivatives must be zero. In other words, the gradient must be a zero vector.

Another reasoning: At a local extreme, the tangent plane must be horizontal, so its normal vector must be vertical. Now for its normal vector we can take the vector $\left(\frac{\partial f}{\partial x_1}(\vec{a}), \ldots, \frac{\partial f}{\partial x_n}(\vec{a}), -1\right)$, and it is vertical exactly if $\frac{\partial f}{\partial x_i}(\vec{a}) = 0$ for all i, that is, $\nabla f(\vec{a}) = \vec{0}$.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. If f has a local extreme at $\vec{a}$, then $\nabla f(\vec{a}) = \vec{0}$ if this gradient at $\vec{a}$ exists.

Points where $\nabla f(\vec{x}) = \vec{0}$ are called **stationary points**. With a bit of luck we can find them by solving the system of n equations $\frac{\partial f}{\partial x_i}(\vec{x}) = 0$ for n unknowns $x_1, \dots, x_n$.

As usual the statement does not work in the opposite direction, not every stationary point is a local extreme. Just recall saddle points that are stationary points but not extrema. The practical impact of this theorem is therefore as follows: If we search for local extrema among stationary points (and points without gradient), then we will not leave any out.

When we gather all candidates (points with zero or no gradient), then we need to classify them. This is not easy for points without gradient, so to make our life easier we will from now on focus only on smooth functions in this chapter. Then the only source of candidates is stationary points, and for those there is a fairly reliable test known as the Sylvester criterion. It is easier to remember if we can understand it intuitively, so we will try to see where it comes from.

Imagine that a function f has a local maximum (hill) at $\vec{a}$. If we slice the graph in the direction of the x-axis, then on this slice we again see a hill, so the function $x \mapsto f(x, a_2, a_3, \dots)$ also has a local maximum at a_1. Therefore, its second derivative, which is actually $\frac{\partial^2 f}{\partial x^2}(\vec{a})$, must be negative (or zero). Since there is nothing special about the x-coordinate, we conclude that if a function f has a local maximum at $\vec{a}$, then $\frac{\partial^2 f}{\partial {x_i}^2}(\vec{a}) \leq 0$ for all i. To put it another way, and disregarding the rare case of zeros, we can expect that at a point of local maximum, the Hess matrix should have negative numbers on its diagonal.

Similarly, for a local minimum we expect $\frac{\partial^2 f}{\partial {x_i}^2}(\vec{a}) > 0$ for all i (or zeros, rarely), and thus we have positive numbers on the diagonal of the Hess matrix. If we look at both cases together, we see that existence of a local extreme means that non-mixed second partial derivatives (that is, diagonal terms on the Hess matrix) must have the same sign. It is interesting to ask what situation would lead to the case when we see different signs on the diagonal of the Hess matrix. That would mean that on some slices through the graph we have concavity up while on others we have concavity down, in other words, we have a saddle there.

It seems that we are developing a nice test for recognizing local extrema and saddles based on signs of the diagonal terms of the Hess matrix, but unfortunately, it is not so simple. We also have to take into account twistiness of the graph, in other words, the mixed derivatives also have their say. To make things easier we will now look closer at the case of two variables.

There we can express a match of signs in an elegant way using a product. We now know that local extrema must satisfy $\frac{\partial^2 f}{\partial x^2}(\vec{a})\,\frac{\partial^2 f}{\partial y^2}(\vec{a}) \geq 0$, while saddles must satisfy $\frac{\partial^2 f}{\partial x^2}(\vec{a})\,\frac{\partial^2 f}{\partial y^2}(\vec{a}) \leq 0$. We are actually looking at the product of the diagonal of the Hess matrix

$$H(x,y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(x,y) & \frac{\partial^2 f}{\partial y \partial x}(x,y) \\ \frac{\partial^2 f}{\partial x \partial y}(x,y) & \frac{\partial^2 f}{\partial y^2}(x,y) \end{pmatrix}.$$

If we want this to work in the opposite direction, then we have to incorporate into this expression also the other terms of this matrix. Do we know some mathematical formula that features all terms of a matrix and includes the product of its diagonal? Yes we do, the determinant, and remarkably enough, it is a good guess. One can prove that for stationary points of functions of two variables, the sign of $\det(H)$ can recognize local extrema and saddles.

If we do have a local extreme, we need to tell apart maxima and minima. But this is easy, for a confirmed local extreme it is enough to check on the shape of one slice, for instance by checking on the sign of $\frac{\partial^2 f}{\partial x^2}(\vec{a})$. We obtain the following algorithm.

Algorithm. (Investigating local extrema for $f(x,y)$)
1. By solving the equation $\nabla f(x,y) = \vec{0}$, that is, the system

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

we find stationary points.
2. For each stationary point (a,b) we find the corresponding Hess matrix $H = H(a,b)$.
• If $\det(H) < 0$, then there is a saddle at (a,b).
• If $\det(H) > 0$, then there is a local extreme at (a,b).
It is a local maximum if $\frac{\partial^2 f}{\partial x^2}(a,b) < 0$, it is a local minimum if $\frac{\partial^2 f}{\partial x^2}(a,b) > 0$.
△

This specific algorithm is not just an inspiration for what will come soon, but it can be useful in itself for students, because many of them will encounter only functions of two variables and this special version is then rather handy. It should be noted that when zeros appear in key moments then this algorithm fails, we know nothing and more advanced methods have to be used. This is a topic beyond this introduction.

If we want to generalize this procedure to more variables, we have to unify the two tests. First we looked at the determinant of the whole matrix, then we decided based on $\frac{\partial^2 f}{\partial x^2}(a,b)$, but that is actually also a determinant, namely determinant of the upper left 1×1 submatrix. We thus work with the same objects, subdeterminants of the Hess matrix, which can be easily generalized.

To make sense of conditions that we will introduce, consider for the moment a simpler situation when mixed derivatives are all zero, so the Hess matrix is diagonal and determinants can be easily evaluated.

$$\begin{array}{c} \\ \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \\ \end{array}\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & 0 & 0 & 0 & \cdots \\ 0 & \frac{\partial^2 f}{\partial x_2^2} & 0 & 0 & \cdots \\ 0 & 0 & \frac{\partial^2 f}{\partial x_3^2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

How will such a matrix behave at a local maximum? We expect $\frac{\partial^2 f}{\partial x_i^2}(\vec{a}) < 0$ for all i. We will now look at upper left subdeterminants $i \times i$, these are traditionally denoted Δ_i.

• The first subdeterminant is given by the upper left entry of H, so $\Delta_1 = \frac{\partial^2 f}{\partial x_1^2}(\vec{a}) < 0$.

• The second subdeterminant is given by the 2×2 matrix in the upper left corner of H, so $\Delta_2 = \frac{\partial^2 f}{\partial x_1^2}(\vec{a}) \cdot \frac{\partial^2 f}{\partial x_2^2}(\vec{a}) > 0$.

• The third subdeterminant is given by the 3×3 matrix in the upper left corner of H, hence $\Delta_3 = \frac{\partial^2 f}{\partial x_1^2}(\vec{a}) \cdot \frac{\partial^2 f}{\partial x_2^2}(\vec{a}) \cdot \frac{\partial^2 f}{\partial x_3^2}(\vec{a}) < 0$.

The reader can surely work out how this should go on. For local maxima the signs alternate, the general formula is $\operatorname{sign}(\Delta_i) = (-1)^i$. This can be also written as $(-1)^i \Delta_i > 0$.

For local minima, all the partial derivatives on the diagonal are positive and hence also all subdeterminants come up positive.

If there is some other progression of signs, then we do not have a maximum or minimum, and if some of the determinants are zero, then the whole procedure fails and we do not know what is going on at $\vec{a}$.

The conclusions deduced for a diagonal H, that is, for the case of mixed derivatives equal to zero, are true in general.

Theorem. (Sylvester criterion)
Let f be defined and have continuous second order partial derivatives on some neighborhood of a point $\vec{a}$ that is stationary for f, that is, $\nabla f(\vec{a}) = 0$. Let H be the Hess matrix of f at $\vec{a}$, let Δ_i be its upper left subdeterminants, $i = 1, \dots, n$.
If $\Delta_i > 0$ for all i, then $f(\vec{a})$ is a local minimum.
If $\Delta_1 < 0$, $\Delta_2 > 0$, $\Delta_3 < 0$, and so on up to $(-1)^n \Delta_n > 0$, then $f(\vec{a})$ is a local maximum.

We will show a more substantial justification in section 7d after we develop proper tools.

Algorithm. (Investigating local extrema for $f(\vec{x})$)
1. By solving the equation $\nabla f(\vec{x}) = \vec{0}$, that is, the system

$$\begin{aligned} \frac{\partial f}{\partial x_1}(x_1, \dots, x_n) &= 0 \\ &\vdots \\ \frac{\partial f}{\partial x_n}(x_1, \dots, x_n) &= 0 \end{aligned}$$

we find stationary points.
2. For each stationary point $\vec{a}$ we find the corresponding Hess matrix $H = H(\vec{a})$ and evaluate subdeterminants Δ_i, that is, determinants of upper left submatrices of size $i \times i$.
• If $\Delta_i > 0$ for all i, then there is a local minimum at $\vec{a}$.
• If the signs alternate $\Delta_1 < 0$, $\Delta_2 > 0$, $\Delta_3 < 0$, ... , then there is a local maximum at $\vec{a}$.
△

Example: We find and classify local extrema of the function $f(x, y, z) = 2xy^2 - 4xy + x^2 + z^2 - 2z$.
First we find stationary points.

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2y^2 - 4y + 2x = 0 \\ \frac{\partial f}{\partial y} &= 4xy - 4x = 0 \\ \frac{\partial f}{\partial z} &= 2z - 2 = 0. \end{aligned}$$

It is a system of three equations of three variables; this sounds hopeful, but the equations are not linear, so the nice theory is of no use. How do we solve general systems?

We start by noticing that the third equation is independent on the others, so definitely $z = 1$. What next? The most reliable method is by elimination, we keep expressing certain variables from equations and substituting them into others, thus reducing the number of equations and unknowns. Here we could use the first equation to find $x = 2y - y^2$ and substitute this into the second equation, creating an equation of third degree with unknown y, this can be handled by smart factoring with a bit of luck (try it). However, this looks a bit like an adventure, it is good to know some alternatives.

We focus on the second equation that we rewrite as $4x(y - 1) = 0$. If we can create a product on one side and zero on the other, we are in luck. In this particular case we see that there are two possibilities, $x = 0$ and $y = 1$.

The case $y = 1$ changes the first equation into $-2 + 2x = 0$, that is, $x = 1$ and we have the first stationary point $(1, 1, 1)$.

The case $x = 0$ changes the first equation into $2y^2 - 4y = 0$, that is, $y(y-2) = 0$, and there are two solutions, $y = 0$ and $y = 2$. Thus we get two more stationary points, $(0, 0, 1)$ and $(0, 2, 1)$.

Now we have to investigate all three stationary points, so we need the Hess matrix. We prepare the second partial derivatives, thanks to the symmetry it is enough to calculate six of them:

$$\frac{\partial^2 f}{\partial x^2} = 2, \qquad \frac{\partial^2 f}{\partial y \partial x} = 4y - 4, \qquad \frac{\partial^2 f}{\partial z \partial x} = 0,$$
$$\frac{\partial^2 f}{\partial y^2} = 4x, \qquad \frac{\partial^2 f}{\partial z \partial y} = 0, \qquad \frac{\partial^2 f}{\partial z^2} = 2,$$

The Hess matrix is

$$H(x, y) = \begin{pmatrix} 2 & 4y-4 & 0 \\ 4y-4 & 4x & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Here we go:

Point $(1, 1, 1)$: $H = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, hence $\Delta_1 = 2$, $\Delta_2 = \det \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} = 8$ and $\Delta_3 = \det(H) = 16$. Signs go $+$, $+$, $+$, therefore $f(1, 1, 1) = -2$ is a local minimum.

Point $(0, 0, 1)$: $H = \begin{pmatrix} 2 & -4 & 0 \\ -4 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, hence $\Delta_1 = 2$, $\Delta_2 = \det \begin{pmatrix} 2 & -4 \\ -4 & 0 \end{pmatrix} = -16$ and $\Delta_3 = \det(H) = -32$. Signs go $+$, $-$, $-$, therefore there is no local extreme at $f(0, 0, 1) = -1$. Judging by signs it would seem that in some slices we have a maximum and in some a minimum, in case of two variables we would call it a saddle.

Point $(0, 2, 1)$: $H = \begin{pmatrix} 2 & 4 & 0 \\ 4 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, hence $\Delta_1 = 2$, $\Delta_2 = \det \begin{pmatrix} 2 & 4 \\ 4 & 0 \end{pmatrix} = -16$ and $\Delta_3 = \det(H) = -32$. Signs go $+$, $-$, $-$, therefore $f(0, 2, 1) = 3$ is not a local extreme, see the previous point.

△

Example: We investigate local extrema of the function $f(x, y) = xy\, e^{x - y^2/2}$.

First we find stationary points.

$$\frac{\partial f}{\partial x} = y\, e^{x-y^2/2} + xy\, e^{x-y^2/2} \cdot 1 = (y + xy) e^{x-y^2/2} = 0$$
$$\frac{\partial f}{\partial y} = x\, e^{x-y^2/2} + xy\, e^{x-y^2/2} \cdot (-y) = (x - xy^2) e^{x-y^2/2} = 0.$$

Since the exponential is always positive, we can divide the equations by it and solve the equations $(1 + x)y = 0$ and $x(1 - y^2) = 0$ instead. We rewrote the equations to the advantageous form of a product and the first one yields two possibilities. If $y = 0$, then from the second equation we have $x = 0$. If $x = -1$, then from the second equation we have $y = \pm 1$. We found stationary points $(0, 0)$, $(-1, -1)$, $(-1, 1)$.

We prepare the second partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = y\, e^{x-y^2/2} + (y + xy) e^{x-y^2/2} \cdot 1 = (x + 2) y\, e^{x-y^2/2},$$
$$\frac{\partial^2 f}{\partial x \partial y} = (1 + x) e^{x-y^2/2} + (y + xy) e^{x-y^2/2} (-y) = (x + 1)(1 - y^2) e^{x-y^2/2},$$
$$\frac{\partial^2 f}{\partial y^2} = -2xy\, e^{x-y^2/2} + (x - xy^2) e^{x-y^2/2} (-y) = xy(y^2 - 3) e^{x-y^2/2}.$$

The Hess matrix is

$$H(x,y) = \begin{pmatrix} (x+2)y\, e^{x-y^2/2} & (x+1)(1-y^2)e^{x-y^2/2} \\ (x+1)(1-y^2)e^{x-y^2/2} & xy(y^2-3)e^{x-y^2/2} \end{pmatrix}.$$

The term $e^{x-y^2/2}$ is always positive, so we factor it out of all entries, it will not influence the signs of determinants. It suffices to use the matrix

$$\widehat{H}(x,y) = \begin{pmatrix} (x+2)y & (x+1)(1-y^2) \\ (x+1)(1-y^2) & xy(y^2-3) \end{pmatrix}.$$

Since we have a function of two variables, we use the first algorithm where we first check on Δ_2.

Point $(0,0)$: $H = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, hence $\Delta_2 = -1 < 0$ and $f(0,0) = 0$ is a saddle.

Point $(-1,1)$: $H = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, hence $\Delta_2 = 2 > 0$ and we have a local extreme. Since $\Delta_1 = 1 > 0$, $f(-1,1) = -e^{-3/2}$ is a local minimum.

Point $(-1,-1)$: $H = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}$, hence $\Delta_2 = 2 > 0$ and we have a local extreme. Since $\Delta_1 = -1 < 0$, $f(-1,-1) = e^{-3/2}$ is a local maximum.

△

4b. Global extrema

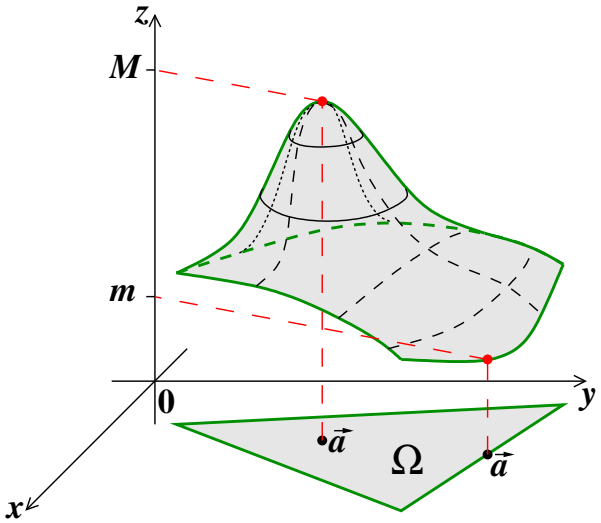
Definition.
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. Let $\Omega \subseteq D(f)$.
We say that M is a **(global) maximum of f on Ω**, denoted $M = \max\limits_{\vec{x}\in\Omega}\big(f(\vec{x})\big)$, if

- $f(\vec{x}) \le M$ for all $\vec{x} \in \Omega$, and
- $f(\vec{x}_0) = M$ for some $\vec{x}_0 \in \Omega$.

We say that m is a **(global) minimum of f on Ω**, denoted $m = \min\limits_{\vec{x}\in\Omega}\big(f(\vec{x})\big)$, if

- $f(\vec{x}) \ge m$ for all $\vec{x} \in \Omega$, and
- $f(\vec{x}_0) = m$ for some $\vec{x}_0 \in \Omega$.

Sometimes we just write $\max\limits_{\Omega}(f)$ and $\min\limits_{\Omega}(f)$. If some point $\vec{a} \in \Omega$ satisfies $f(\vec{a}) = \max\limits_{\Omega}(f)$, then we say that f **attains its maximum at** $\vec{a}$, similarly we attain minimum. The following picture shows a typical situation.



It also explains why the following statement should be true.

Definition.
Let f: $D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. Let $\Omega \subseteq D(f)$.
If f attains its minimum or maximum over Ω at $\vec{a}$, then f has a local extreme at $\vec{a}$ or $\vec{a}$ belongs to the boundary $\partial\Omega$ of the set Ω.

The statement is exactly the same as it was for functions of one variable, and it is again the starting point for an algorithm for determining global extrema.

The main steps:

1. We identify candidates.
2. We compare the values of f at these candidates, then choose the largest and the smallest.

The second step should be easy, let's have a look at the first one. According to the theorem, we should worry about local extrema and points on the boundary. We know how to find local extrema, and to make things even easier, we are not really interested in classification. If we add an undeserving candidate, then it will get ruled out in the second (comparison) step anyway. So instead of adding exactly the points of local extrema to our list, we simply add all candidates for local extrema. In other words, we just take all points from Ω where the gradient is zero or does not exist. Formally we should take only points from the interior of Ω, but again, we need not worry, adding a few points from the boundary does not spoil anything (just adds work in the second step).

Regarding the boundary, there things get complicated. When we were investigating global extrema for functions of one variable, then Ω was usually an interval. Its boundary is just the endpoints, so we simply added them to our list of candidates (if they were not infinite). With functions of more variables the situation is completely different, as a typical set in $\mathbb{R}^n$ has infinite boundary. For instance, the boundary of a square in $\mathbb{R}^2$ is its perimeter, and the boundary of a ball in $\mathbb{R}^3$ is its surface, a sphere.

We definitely do not want to add infinitely many points to our list of candidates, but fortunately we do not have to. Take some value attainable on a boundary. If it is not the largest or the smallest from values on the boundary, then it also cannot be largest or smallest over the whole set, and therefore it is pointless to consider it. Therefore, when we are looking for candidates for global extrema from a boundary, then it is sufficient to consider only those that are extreme with respect to this boundary.

This means that when considering the boundary, we actually pass to a new subproblem, namely we are looking for global extrema (or rather, for points suspected of providing them) of the given function, but only on the boundary of the original set. In a typical case, this boundary has one dimension less compared to the original set, and as a set of lower dimension it also has some interior and boundary, so we apply the general algorithm to this new subproblem: We gather stationary points from the interior and check on the boundary (of boundary).

For instance, the boundary of a square consists of four sides. Each side is a one-dimensional object, a segment, and as such it has an interior and a boundary, namely the endpoints. It can happen that this boundary of boundary is infinite again, then we investigate global extrema on it, passing to a subproblem of a subproblem. Since we decrease dimension each time, sooner or later this process has to stop.

Example: Imagine some function $f(x, y, z)$ of three variables. We are looking for its extrema on the set Ω which is the upper half-ball of radius 2, formally,

$$\Omega = \{(x, y, z) \in \mathbb{R}^3;\ z \geq 0 \text{ and } x^2 + y^2 + z^2 \leq 2^2\}.$$

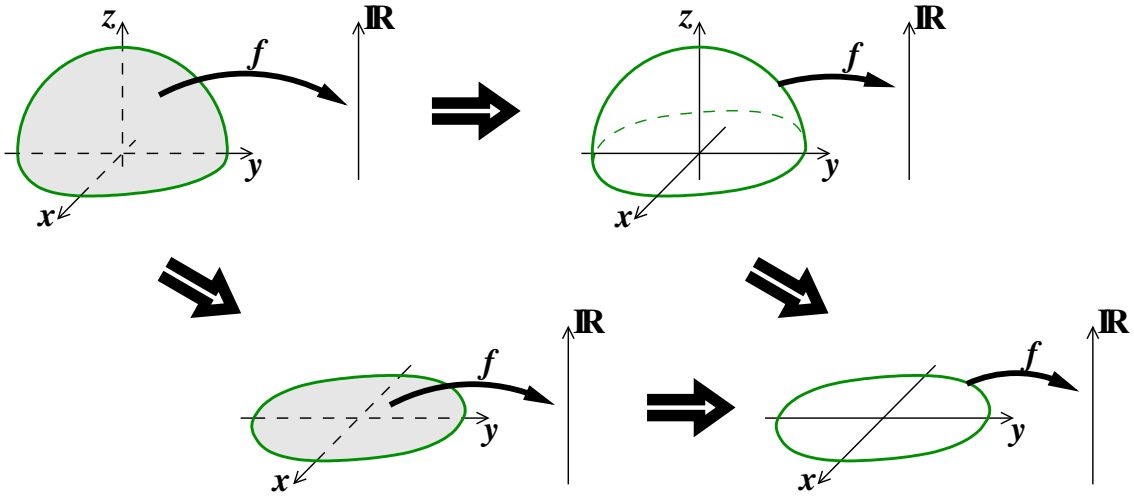
When looking for candidates, we first find points suspicious of being local extrema, and choose those from Ω. Then we look at the boundary, which is the surface of the half-ball.

It consists of two parts, the base and the upper half-sphere. Both objects are essentially two-dimensional and we look for global extrema of the given function on each them, that is, we will explore two subproblems.

Regarding the base, it is the disc of radius 2 in the xy-plane, so it is a true two-dimensional object. It has an interior on which we would look for local extrema, and a boundary, namely its perimeter. This perimeter is infinite, so we pass to another subproblem: We investigate global extrema on the perimeter of the base. It is a circle, which is essentially a one-dimensional object (it can be described using one parameter). It is special, because as a one-dimensional object it does not have a boundary (or rather, the boundary is empty), and all points of a circle are its interior points. So it is enough to identify local extrema on this circle.

If the reader feels troubled by those things with dimensions and empty boundary, do not worry, we will see this properly when we return to this example later and describe these objects mathematically.

The second major subproblem, investigation of the function on the upper half-sphere, is similar. It has an interior, the open upper half-sphere, where we look for local extrema. Then we have to check on the boundary, but this is the same circle that we already looked at while investigating the base.

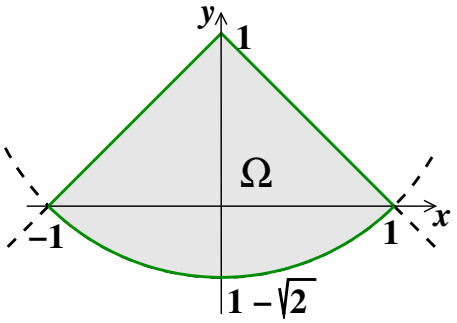


△

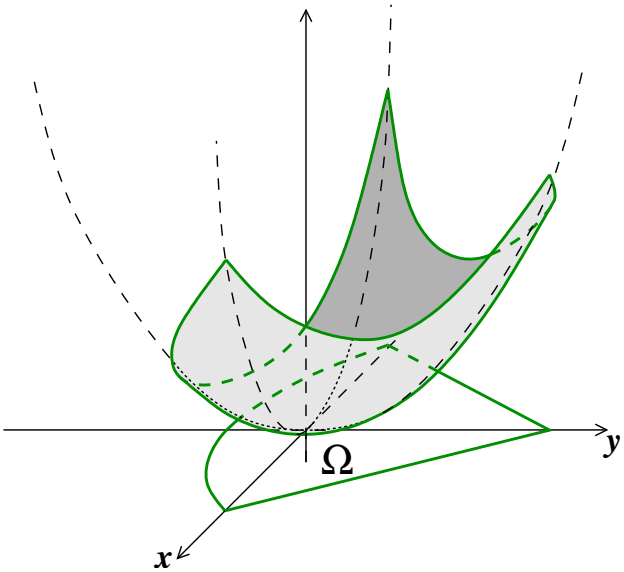
How do we actually investigate functions on sets of smaller dimension? We parametrize those sets (describe them mathematically) and the number of parameters that are needed determine dimension. These parameters then also appear as new variables in our function, so we in fact work with auxiliary functions with smaller number of variables. We will show how it works now.

Example: We find global extrema of the function $f(x,y) = x^2 + y^2$ on the set Ω delimited by curves $y = 1 - |x|$ and $x^2 + (y-1)^2 = 2$ that contains the origin.

First we identify the region Ω. We draw the curves, the first is shaped like a chevron, the other is the circle centered at $(0,1)$ with radius $\sqrt{2}$. They divide the plane into four regions, we choose the one that includes the origin.



We find global extrema of f with respect to this set. We will cheat a bit now and show how the situation actually looks like. We are interested in the piece of a rotational paraboloid that lies over our set.



Normally we would not know this, so we will now pretend that we do not see this picture and solve everything just using mathematics. However, it will be useful for the reader to compare individual steps with this sketch.

1. First we collect the candidates.

a) Local extrema: The equation $\nabla f = \vec{0}$ reads $(2x, 2y) = (0, 0)$, hence $x = 0$ and $y = 0$. The point $(0, 0)$ is in the region Ω and thus it is a valid candidate.

b) Extrema on the boundary: We need to investigate three sections of the border.

We start with the segment in the upper-right quadrant. It is a part of the line given by the formula $y = 1 - x$ and thus it is natural to describe it parametrically as $x \mapsto (x, 1 - x)$ for $x \in [0, 1]$. This parametrization confirms that this segment is essentially one-dimensional. Using the formula $y = 1 - x$ we eliminate y in f and obtain an auxiliary function ϕ of one variable describing the behaviour of f on this segment of boundary.

$$\phi(x) = f(x, 1 - x) = x^2 + (1 - x)^2 = 2x^2 - 2x + 1.$$

In this way we transformed the problem of finding global extrema of f over the segment into a problem of finding global extrema of $\phi(x)$ over the set $[0, 1]$. We know how to solve this.

First we find critical points from that interval.

$$\phi'(x) = 0 \implies 4x - 2 = 0 \implies x = \tfrac{1}{2}.$$

This x belongs to the investigated interval, therefore it is valid. We thus add to the main list of candidates the point from the boundary segment (returning to two-dimensional situation) that corresponds to this x: $y = 1 - \frac{1}{2} = \frac{1}{2}$, so we add the point $\left(\frac{1}{2}, \frac{1}{2}\right)$.

We also have to check on the boundary. For an interval this is easy, the boundary is just the endpoints $x = 0$ and $x = 1$, so we add corresponding two-dimensional points $(0, 1)$ and $(1, 0)$ to our list of candidates. These are actually the endpoints of that oblique segment on the boundary of Ω that we work on here, so things fit nicely.

Note that the point $\left(\frac{1}{2}, \frac{1}{2}\right)$ is not a local extreme of the function f itself. But when investigating the boundary segment, we restricted the function just to this set, and relative to this set it is a local minimum, as we can see in the picture.

In the same way we handle the part of the boundary of Ω given by the formula $y = 1 + x$ for $x \in [-1, 0]$. By investigating the function

$$\phi(x) = f(x, 1 + x) = 2x^2 + 2x + 1$$

we find the suspected local extreme $x = -\frac{1}{2}$ and add the point $\left(-\frac{1}{2}, \frac{1}{2}\right)$ to our list of candidates. Considering endpoints we arrive at $(-1, 0)$ and $(0, 1)$; we add the first to the list, the second is already there.

The last part of the boundary is the arc. The formula $x^2 + (y - 1)^2 = 2$ offers two ways to get

rid of one variable, we will try the approach $y = 1 - \sqrt{2 - x^2}$ for $x \in [-1, 1]$. We are interested in the auxiliary function

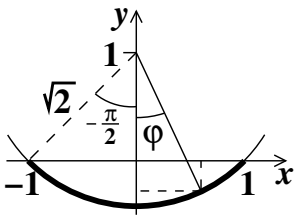
$$\phi(x) = f\big(x, 1 - \sqrt{2 - x^2}\,\big) = x^2 + 1 + (2 - x^2) - 2\sqrt{2 - x^2} = 3 - 2\sqrt{2 - x^2}.$$

What could be the extrema of this function on $[-1, 1]$? The endpoints of this interval lead to points that we already have in our list. It remains to check on local extrema.

$$\phi'(x) = 0 \implies \frac{2x}{\sqrt{2 - x^2}} = 0 \implies x = 0.$$

We add the last point $\big(0, 1 - \sqrt{2}\,\big)$ to our list.

Alternative for the arc: Since it is a part of a circle, polar coordinates may be a good idea. We will use a special version adapted to our situation.



$$x = \sqrt{2}\sin(\varphi),$$
$$y = 1 - \sqrt{2}\cos(\varphi).$$

The third part of the boundary is obtained by taking $\varphi \in \big[-\frac{\pi}{4}, \frac{\pi}{4}\big]$. We get the auxiliary function and its derivative

$$\phi(x) = f\big(\sqrt{2}\sin(\varphi), 1 - \sqrt{2}\cos(\varphi)\big) = 3 - 2\sqrt{2}\cos(\varphi), \qquad \phi'(x) = 2\sqrt{2}\sin(\varphi).$$

Candidates for extrema are endpoints $\varphi = \pm\frac{\pi}{4}$ leading to $(\pm 1, 0)$ that we already have, and zero point $\varphi = 0$ of the derivative, which supplies the point $\big(0, 1 - \sqrt{2}\,\big)$, confirming our previous work.

2. We now compare values.

$f(0,0) = 0$, $f(0,1) = 1$, $f(1,0) = 1$, $f(-1,0) = 1$, $f\big(\frac{1}{2}, \frac{1}{2}\big) = \frac{1}{2}$, $f\big(-\frac{1}{2}, \frac{1}{2}\big) = \frac{1}{2}$, $f\big(0, 1 - \sqrt{2}\,\big) = 3 - 2\sqrt{2} \approx 0.17$.

We found that $\min\limits_{\Omega}(f) = 0 = f(0,0)$ and $\max\limits_{\Omega}(f) = 1 = f(0,1) = f(\pm 1, 0)$.

△

What do we actually mean by a "local extreme with respect to a set"? Consider a function $f(\vec{x})$ with $D(f) \subseteq \mathbb{R}^n$ and a set $M \subseteq D(f)$. A point $\vec{a} \in M$ is a "local maximum with respect to M" if there is a neighborhood U of $\vec{a}$ such that $f(\vec{x}) \le f(\vec{a})$ whenever $\vec{x} \in U \cap M$. Relative local minimum can be defined in analogous way. We did not make an official definition, because people usually focus on a more specific setting.

Namely, we are interested in situation when the set M is described by an equation or several equations. We can think of a circle in the plane, or a plane in a three-dimensional space. The equations that define the set M are called "constraints". In order to unify notation we recall that every (algebraic) equation can be rewritten in the form $g(x, y, \dots) = 0$. The set M is then uniquely determined by a constraint function g, or constraint functions g_j. Mathematically,

$$M = \{\vec{x} \in D;\ g_1(\vec{x}) = 0, \dots, g_p(\vec{x}) = 0\},$$

where D is some set on which all constraint functions exist. Very often there is just one constraint:

$$M = \{\vec{x} \in D;\ g(\vec{x}) = 0\}.$$

What can we expect of such a set? In general such sets can be very wild, but in applications it usually works like this: Our requirement that the equality $g = 0$ be true takes away one degree of freedom from the set, so we can expect that the set M will essentially have dimension $n - 1$. This among other things means that we should be able to describe it using $n - 1$ parameters.

For instance, the constraint function $g(x, y) = 169 - x^2 - y^2$ defined on the two-dimensional space $\mathbb{R}^2$ leads to the constraint $169 - x^2 - y^2 = 0$, that is, $x^2 + y^2 = 169$. So the resulting set is the circle of radius 13 centered at the origin, which is essentially a one-dimensional object. We can tell by observing that just one parameter is enough to determine the position of a point on this circle (for instance an angle).

However, it may not always work this way. For instance, the function $g(x) = x - \sqrt{x^2}$ is defined

on $\mathbb{R}$, which is a one-dimensional world, and sets up the set of all points satisfying $x = |x|$, so $M = [0, \infty)$. This set is also one-dimensional, so we did not drop any dimension here.

On the other hand, the function $g(x, y) = (x-1)^2 + (y-13)^2$ exists on $\mathbb{R}^2$ and defines the set of all points satisfying the constraint $(x-1)^2 + (y-13)^2 = 0$. This can be satisfied by only one point, namely $(1, 13)$. This set is not one-dimensional but is of dimension zero, so our requirement took away not one, but two dimensions from the original set.

The good news is that examples like these were thought up by mischievous mathematicians, in application the one-dimension loss usually works.

It gets interesting when we have more constraint functions g_j. When we insist on simultaneous validity of all constraints $g_j = 0$, then the resulting set M is actually the intersection $\bigcap M_j$ of sets M_j defined by individual constraint functions g_j. In a typical case, each of the sets M_j has dimension $n-1$ (assuming that everything takes place in the space $\mathbb{R}^n$). Then one would expect that with every new intersection we loose one dimension, so with p constrains the resulting intersection, that is, the set M, could have dimension $n-p$. Actually, the reader most likely encountered this kind of behaviour in linear algebra. A linear function $g = a_1x_1 + \cdots + a_nx_n + d$ leads to an equation that defines a hyperplane in $\mathbb{R}^n$, that is, an object of dimension $n-1$. By intersecting non-parallel hyperplanes we successively remove dimensions, which allows us to get all the way to straight lines, that is, to one-dimensional objects. For instance, a straight line in $\mathbb{R}^3$ is determined by two equations.

However, in linear algebra we also learned that things need not work this way; for instance, intersecting two parallel but distinct hyperplanes yields an empty intersection. With more interesting shapes things can get even more complicated, which is unpleasant for theory. We do want to work with the situation that the resulting set M has dimension $n-p$, and therefore we will have to incorporate one technical assumption on constraint functions g_j into the statement that comes soon.

Before we get there, let us state the problem that we want to address. We have a function $f(\vec{x})$ and a set M determined by some constraints. We are looking for local extrema relative to this set. Such local extrema are called **constrained extrema**. In the example above we approached this problem using elimination of variables (or parametric description of a set), but that can lead to complicated calculations if the constraints are not friendly. The following theorem offers an alternative approach.

Theorem. (Lagrange multipliers)
Let D be an open set in $\mathbb{R}^n$ and $f \in C^1(D)$.
Consider constraint functions $g_1, \dots, g_p \in C^1(D)$, denote

$$M = \{\vec{x} \in D;\ g_1(\vec{x}) = 0, \dots, g_p(\vec{x}) = 0\}$$

and assume that the vectors $\nabla g_1(\vec{a}), \dots, \nabla g_p(\vec{a})$ are linearly independent for all points $\vec{a} \in M$.
If $\vec{x}_0$ is a local extreme of the function f with respect to the set M, then there exists numbers $\lambda_1, \dots, \lambda_p$ such that

$$\nabla f(\vec{x}_0) = \sum_{j=1}^{p} \lambda_j \nabla g_j(\vec{x}_0).$$

The condition on gradients of g_j makes sure that these constraints create a reasonable set. The numbers λ_j are called Lagrange multipliers. From the point of view of exploring local extrema they are just working parameters and we usually get right of them right away using elimination. However, it is worth noting that they can have a practical meaning in physics. There they work

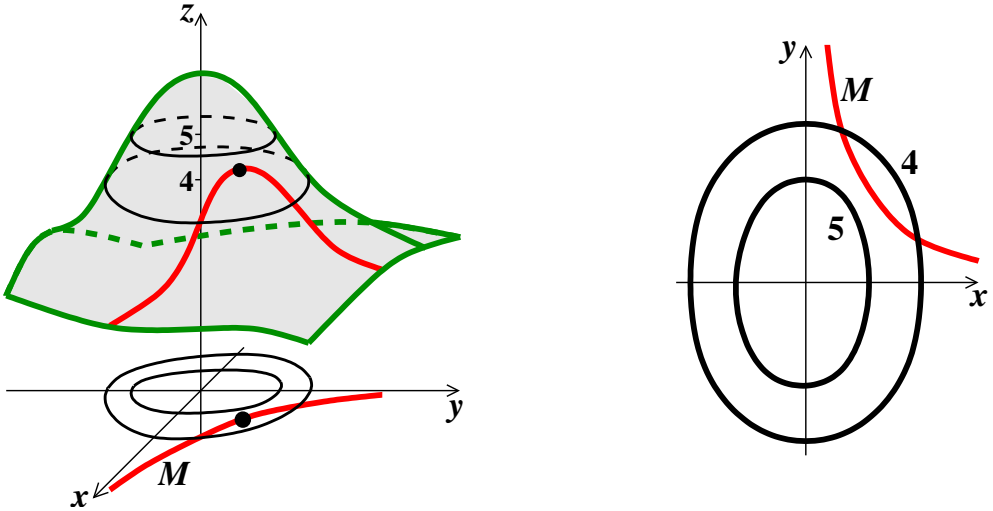
with the so-called Lagrange function

$$L = f - \sum_{j=1}^{p} \lambda_j g_j$$

and the conclusion of the theorem actually says that $\nabla L = \vec{0}$.

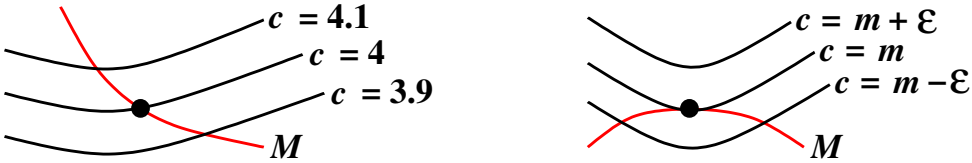
Why should something like this be true? We can actually make a good visual argument for the case of one constraint.

In the picture we see the graph of some function f, but we are only interested in values that are attained on the set M outlined in the xy-plane, that is, in the domain of f. We outlined these values in the graph of the function in red. We also marked our guess for the maximal value that can be reached on M, both in the graph and in the set M itself, but this is really just to satisfy our curiosity. What is more important is the black curves that show the values 4 and 5 on the graph of f. They allow us to conclude that the maximal value on M is between 4 and 5. However, this is not the right place to look, because graphs are not easy to work with. We thus move our attention to the domain, where we naturally pass to level curves corresponding to the values 4 and 5 for our function f. Can we tell something by looking at the domain only?



We see right away that the function can never reach the value 5 within the set M, as the set M has empty intersection with the set of all $\vec{x}$ that yield value 5. We also see that we can reach value 4, we even have two possible points $\vec{x}$ to do so. However, the key question is whether we can recognize from this picture that 4 is not the largest possible value, therefore it is not what we are seeking. One may think that the existence of more intersections will play important role, but it is not so. The important thing is that at the intersection, the level curve L_4 and the set M intersect at an angle (measured as the angle between their tangent lines).

This is related to the fact that level sets L_c behave in a continuous way with respect to the value of c, assuming that the investigated function f is continuous. This means that if we change c just a little, then also the set L_c in the space $\mathbb{R}^n$ changes just a little. A good visualization is that as we turn a knob that controls c gradually from value 4 to value 5, then the corresponding level sets change gradually from L_4 to L_5, a bit like the popular morphing. What does it mean for our situation? If the set M and a level set L_4 intersect at a non-zero angle, then for a small change of constant c the new level sets will still intersect the set M. This then implies that the value $c = 4$ cannot be maximal nor minimal.

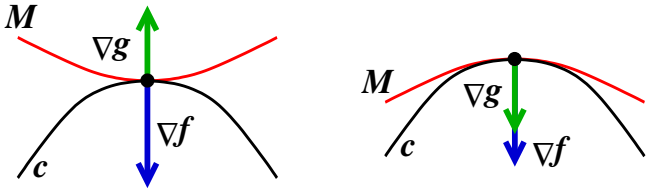


So how do we recognize that we found a maximum or minimum? The corresponding level set

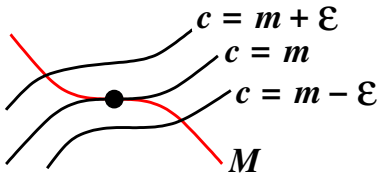
must intersect the set M is such a way that their tangent lines agree. We are getting near, so it is time to state this in a way that allows also for higher dimension: If we have a function defined on $\mathbb{R}^3$, then level sets are surfaces in 3D, just like the set M. If they are not to intersect at an angle, then we need to require that their tangent planes agree. Similarly we work in higher dimensions.

So how do we find in general that tangent hyperplanes of two $(n-1)$-dimensional surfaces are equal? The easiest way is to check on their normal vectors, these then have to be parallel. We are almost there. How do we find normal vectors to such surfaces? For the level set we can use a statement from chapter 3: The normal vector is simply the gradient ∇f. How about the set M?

It is a set of points that satisfy $g(\vec{x}) = 0$, so it is actually also a level set, this time for the function g. A normal vector for the set M can thus be obtained as ∇g. We reached the following conclusion: In order for a function f to have a maximum or minimum with value c with respect to a set M at some point $\vec{a}$, then the set M must intersect the level set L_c at $\vec{a}$ and the vectors $\nabla f(\vec{a})$, $\nabla g(\vec{a})$ must be parallel; that is, there must be some non-zero constant λ so that $\nabla f(\vec{a}) = \lambda \nabla g(\vec{a})$. And that's exactly what the Lagrange multiplier theorem says.



We remark for the sake of completeness that this theorem is only an implication, just like the popular theorem on critical points for local extrema. It can happen that $\nabla f(\vec{a})$ and $\nabla g(\vec{a})$ are parallel at some $\vec{a}$, and yet there is no constrained local extreme there.



This is no problem in applications, because we are just gathering candidates anyway, so if we put in an extra one, we will then discard it in the second, comparative step anyway. The main point of this theorem is that when we use it to find candidates, then we know that we did not leave any out.

Hopefully the theorem now makes sense mathematically, another good question is whether it makes sense from practical point of view. Here comes an example.

Example: We return to the problem of finding extrema of the function $f(x,y) = x^2+y^2$. Namely, we will revisit the stage where we were looking for extrema of f relative to the set given by the formula $x^2 + (y-1)^2 = 2$. In the first solution we used this constraint to eliminate one variable and investigated a one-dimensional situation.

Now we try Lagrange multipliers. We rewrite the constraint as $x^2 + (y-1)^2 - 2 = 0$. We thus have the constraint function $g(x,y) = x^2 + (y-1)^2 - 2$ and write down the equations arising from the Lagrange condition and the constraint.

$$\begin{aligned} \nabla f(x,y) &= \lambda \nabla g(x,y) \\ g(x,y) &= 0 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} 2x &= 2x\lambda \\ 2y &= 2(y-1)\lambda \\ x^2 + (y-1)^2 &= 2. \end{aligned}$$

We obtained three equations for three unknowns, which sounds hopeful, but unfortunately they are not linear. Fortunately, there are some interesting openings, in particular, the first equation just begs to be simplified by cancelling. We get $1 = \lambda$. The second equation then reads $y = y-1$, which does not have a solution. This is a blind alley, but it is not the end of the road. The cancelling can be done only under the assumption that $x \neq 0$, so we also have to check on the other possibility.

If $x = 0$, then the first equation is automatically true. Substituting into the constraint we get $(y-1)^2 = 2$. Then $y = 1 \pm \sqrt{2}$, but it is obvious from the specification of the problem that we are only interested in values $y \leq 0$, so we take just $y = 1 - \sqrt{2}$. Given that we do not really need λ

one might think that our work is done, but it is necessary to check that also the second equation can be made true. We easily find that $\lambda = \frac{\sqrt{2}-1}{\sqrt{2}}$ will do the trick.

We obtained a suspicious point $\big(0, 1-\sqrt{2}\,\big)$ just like in the previous attempt. Was it easier now? This is for the reader to decide.

△

Now we are ready to compare strategies. Consider some function f with n variables and p independent constraint functions g_j that also have n variables, where $p < n$. Our previous (intuitive) strategy was based on the idea that with a bit of luck, we can use the constraints to express some variables as functions of other variables. Substituting these into f then creates an auxiliary function ϕ of $n-p$ variables. Then we look for its extrema, which among other things leads to the condition $\nabla\phi = \vec{0}$, that is, a system of $n-p$ equations with $n-p$ unknowns.

If we follow the approach from the new theorem, then we should keep the original variables and add λ_j on the top of it, which gives altogether $n+p$ unknowns. How many equations do we have? The equality $\nabla f = \sum \lambda_j \nabla g_j$ compares vectors, which means that we have equality for all components of these vectors. We thus get n equations, plus we still have the constraints $g_j = 0$, so we get $n+p$ equations.

$$\begin{aligned}
\frac{\partial f}{\partial x_1}(x_1,\dots,x_n) &= \lambda_1 \frac{\partial g_1}{\partial x_1}(x_1,\dots,x_n) + \dots + \lambda_p \frac{\partial g_p}{\partial x_1}(x_1,\dots,x_n)\\
&\vdots\\
\frac{\partial f}{\partial x_n}(x_1,\dots,x_n) &= \lambda_1 \frac{\partial g_1}{\partial x_n}(x_1,\dots,x_n) + \dots + \lambda_p \frac{\partial g_p}{\partial x_n}(x_1,\dots,x_n)\\
g_1(x_1,\dots,x_n) &= 0\\
&\vdots\\
g_p(x_1,\dots,x_n) &= 0.
\end{aligned}$$

The good news is that this number of equations fits the number of unknowns, the bad news is that now we have $2p$ more equations and $2p$ more unknowns compared to the previous approach. Why would anyone want to go this new way? It depends on how complicated the formulas obtained when expressing some variables from constraints $g_j = 0$ are, and how well (or badly) they fit into f. If we end up with unpleasant expressions, then the Lagrange approach could help, and in fact it often does. This approach is popular in applications, and it can even provide us with some general results and proofs of some theorems.

Example: We find the distance between a point $P = (x_0, y_0, z_0)$ and a plane q: $ax+by+cz+d = 0$.

This distance can be found as the distance between P and the point $Q = (x, y, z)$ from the plane q that is closest to P. We are thus looking for the minimum of the function

$$(x,y,z) \mapsto \operatorname{dist}\big(P,(x,y,z)\big) = \sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}$$

for (x, y, z) from the given plane. Because square root is an increasing function, it is enough to identify the minimum of the function

$$f(x,y,z) = (x-x_0)^2+(y-y_0)^2+(z-z_0)^2,$$

and that with respect to the set given by the constraint $ax+by+cz+d = 0$. So we will work with the constraint function $g(x, y, z) = ax+by+cz+d$.

According to the theorem, the desired point is hiding among solutions of the system created by the Lagrange condition and the given constraint,

$$\begin{aligned}
\nabla f &= \lambda \nabla g\\
g &= 0,
\end{aligned}$$

that is,

$$\begin{aligned} 2(x-x_0) &= \lambda \cdot a \\ 2(y-y_0) &= \lambda \cdot b \\ 2(z-z_0) &= \lambda \cdot c \\ ax+by+cz+d &= 0. \end{aligned}$$

The equations are linear, but we leave the standard approach through Gaussian elimination or Cramer rule to curious readers and focus on approaches typical for Lagrange multipliers.

Since we do not really need λ, we get rid of it using elimination. Assume that $a \neq 0$ (at least one of the coefficients a, b, c is not zero, otherwise we do not have an equation of a plane). Then from the first equation $\lambda = \frac{2}{a}(x-x_0)$ and the first three equations reduce to

$$\begin{aligned} 2(y-y_0) &= b\frac{2}{a}(x-x_0) \\ 2(z-z_0) &= c\frac{2}{a}(x-x_0) \end{aligned} \quad \Longrightarrow \quad \begin{aligned} y &= y_0 + \frac{b}{a}(x-x_0) \\ z &= z_0 + \frac{c}{a}(x-x_0). \end{aligned}$$

We substitute into the constraint:

$$\begin{aligned} & ax + b\big(y_0 + \tfrac{b}{a}(x-x_0)\big) + c\big(z_0 + \tfrac{c}{a}(x-x_0)\big) + d = 0 \\ \Longrightarrow\ & x = \frac{-ad + (a^2+b^2+c^2)x_0 - a(ax_0+by_0+cz_0)}{a^2+b^2+c^2} \\ \Longrightarrow\ & x = x_0 - a\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2} \\ \Longrightarrow\ & y = y_0 - b\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2} \\ & z = z_0 - c\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2}. \end{aligned}$$

This is the closest point Q in the plane q to the point P. Before we determine the distance, we explore another interesting possibility for solving the system. From the first three equations we will deduce formulas for x, y, z depending on the Lagrange multiplier:

$$\begin{aligned} x &= x_0 + \tfrac{a}{2}\lambda \\ y &= y_0 + \tfrac{b}{2}\lambda \\ z &= z_0 + \tfrac{c}{2}\lambda. \end{aligned}$$

Now we substitute into the constraint and obtain

$$\begin{aligned} & a\big(x_0 + \tfrac{a}{2}\lambda\big) + b\big(y_0 + \tfrac{b}{2}\lambda\big) + c\big(z_0 + \tfrac{c}{2}\lambda\big) + d = 0 \\ \Longrightarrow\ & \lambda = -2\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2}. \end{aligned}$$

Hence

$$\begin{aligned} x &= x_0 - a\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2} \\ y &= y_0 - b\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2} \\ z &= z_0 - c\frac{ax_0+by_0+cz_0+d}{a^2+b^2+c^2}. \end{aligned}$$

This was perhaps easier.

Both approaches were successful and necessarily lead to the same point Q. While calculating the

distance it is useful to use λ at first.

$$\begin{aligned}\operatorname{dist}(P,Q) &= \sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}\\ &= \sqrt{\Big(\frac{a}{2}\lambda\Big)^2+\Big(\frac{b}{2}\lambda\Big)^2+\Big(\frac{c}{2}\lambda\Big)^2} = \sqrt{a^2+b^2+c^2}\,\frac{1}{2}|\lambda|\\ &= \frac{|ax_0+by_0+cz_0+d|}{\sqrt{a^2+b^2+c^2}}.\end{aligned}$$

This is the standard formula for the distance between a point and a plane: We substitute the point into the equation, apply absolute value and divide by the magnitude of the normal vector associated with the plane equation.

Nitpicking question: How do we know that this distance is the minimum and not the maximum? A sophisticated answer would point out that the function f is concave up. However, we are not on that level, so we try a different argument. When we take points Q from the plane that in some way go to infinity, then also their distance from P will increase to infinity, so the above number cannot be a maximum. It still could be a local maximum. However, because this function is continuous, bounded from below and goes to infinity on the "outside", then it must have a local minimum. Since we found only one suspicious point, this must be it. It is not precise mathematics, but a complete answer would be rather long.

△

This example confirmed that Lagrange multipliers can provide general formulas, and also outlined the two most common strategies for solving equations that we get using Lagrange multipliers.

• A very popular approach is to use the equations that came from the gradient equality to eliminate parameters λ_j (we do not care about them anyway) and in this way find some relationships between variables. Using these and also the original constraints we then find the suspicious points.

• Sometimes it is possible to use the first equations (those from the gradient) to express all variables using the parameter λ. After substituting into the constraint we obtain the value for λ and then also for the variables. However, we have to get lucky for this to work.

If none of these approaches work, then we simply have to try something different, inspired by the actual equations that we face.

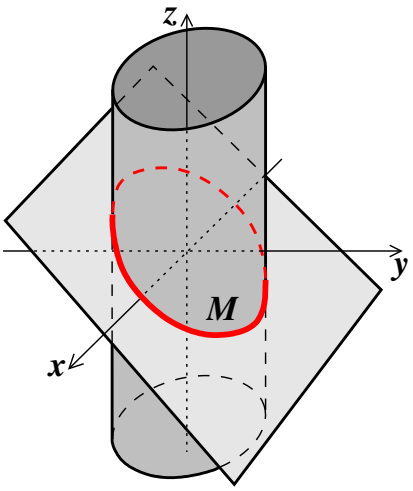
We met all the key tools and now we are ready to put together an algorithm for finding global extrema. It will not be entirely complete and we will use recursion.

Algorithm. (Investigating global extrema of $f(\vec{x})$ on Ω)
1. We prepare the list of candidates $\{x_1,\dots,x_N\}$.
a) We find points from Ω at which $\nabla f=\vec{0}$ or ∇f does not exist.
b) We find points from the boundary $\partial\Omega$ at which f could attain global extrema with respect to $\partial\Omega$. To this end we again apply the basic algorithm, that is, steps a) and b), to the set $\partial\Omega$.
Local extrema from the interior of $\partial\Omega$ can be identified using Lagrange multipliers if this set is given by constraints, the other common approach is to parametrize this set.
If the boundary of the set $\partial\Omega$ is infinite, then we investigate it by recursively applying the step b) to it. This reduction should in a typical case end up with investigating a function of one variable over a bounded closed interval, where the boundary is just its endpoints that we add to the list.
2. We compare values $f(x_1),\dots,f(x_N)$ and choose the largest and smallest among them.
△

Now it is time for a representative problem.

Example: We find global extrema of the function $f(x,y,z) = xyz$ relative to the constraints $x^2+y^2=3$ and $y+z=0$.

Just out of curiosity we first reason out on what set M we are finding the extremes. The equation $x^2+y^2=3$ would define a circle with radius $\sqrt{3}$ in $\mathbb{R}^2$, but we are working in $\mathbb{R}^3$ here. This condition does not restrict z at all, so the circle can move in the z direction freely and creates an infinite cylinder.



The equation $y+z=0$ defines a straight line in the yz-plane. It passes through the origin and decreases, in fact it is the secondary diagonal $z=-y$. This line then moves freely in the x- direction and creates a plane. A good visualization is that we take the xz plane and rotate it about the x-axis by 45 degrees.

When we intersect the cylinder and the plane, we get an ellipse, which is essentially a one-dimensional object. This fits, our constraints took away two degrees of freedom from the original three-dimensional space. We noted that it need not always work like this, here we got lucky.

This ellipse does not have any endpoints as a one-dimensional object, so global extrema on it must be attained at points of local extrema. We identify them using the Lagrange multipliers theorem. First we check that its assumptions are satisfied. We have two constraint functions, $g(x,y,z)=x^2+y^2-3$ and $h(x,y,z)=y+z$, that define the set M. The theorem asks us to check that on this set, the gradients of g and h are linearly independent. Let's see:

$$\nabla g = (2x, 2y, 0),$$
$$\nabla h = (0, 1, 1).$$

The third coordinate shows that we cannot obtain the vector ∇h as a multiple of the vector ∇g, so we have the independence. We also notice that all functions taking part in the problem have continuous partial derivatives of order one, so we are entitled to use the theorem. We introduce two Lagrange multipliers, λ and μ (I do not feel like writing indices), and write the equations.

$$\begin{aligned} \nabla f &= \lambda \nabla g + \mu \nabla h \\ g &= 0 \\ h &= 0 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} yz &= \lambda \cdot 2x + \mu \cdot 0 \\ xz &= \lambda \cdot 2y + \mu \cdot 1 \\ xy &= \lambda \cdot 0 + \mu \cdot 1 \\ x^2+y^2 &= 3 \\ y+z &= 0. \end{aligned}$$

We will try the recommended procedure and use the first three equations to eliminate the multipliers.

$$\begin{aligned} yz &= 2\lambda x \\ xz &= 2\lambda y + \mu \\ xy &= \mu \end{aligned} \quad \Longrightarrow \quad \begin{aligned} yz &= 2\lambda x \\ xz &= 2\lambda y + xy. \end{aligned}$$

Now we would like to express λ from the first equation, but first we have to ask ourselves a question: What if $x=0$? Then this equation would imply that $yz=0$, but the second constraint $y+z=0$ gives $y=-z$, which would lead to $y=z=0$. We get the point $(0,0,0)$, but it does not fulfill the first constraint, hence it is not a valid candidate.

So we continue under the assumption that $x \neq 0$. Then we express 2λ from the first equation and substitute into the second:

$$\begin{aligned} yz &= 2\lambda x \\ xz &= 2\lambda y + xy \end{aligned} \quad \Longrightarrow \quad xz = \frac{yz}{x} y + xy \quad \Longrightarrow \quad x^2 z = y^2 z + x^2 y.$$

Elimination of parameters gave us a connection between the variables, now we put it together with

the two given constraints. From the second one we substitute $z = -y$:

$$-x^2y = -y^3 + x^2y \implies y^3 = 2x^2y.$$

Is it possible to have $y = 0$? Then also $z = 0$ and the first constraint yields $x = \pm\sqrt{3}$. We have the first two candidates.

If $y \neq 0$, then we cancel and obtain $y^2 = 2x^2$. We substitute into the first constraint and obtain

$$x^2 + 2x^2 = 3 \implies x = \pm 1,\ y = \pm\sqrt{2}.$$

We get four more points, where the third coordinate is determined using $z = -y$.

There are no more cases to explore, the list is complete. Now we compare values for candidates.

$f\big(\pm\sqrt{3}, 0, 0\big) = 0$,

$f\big(1, \sqrt{2}, -\sqrt{2}\big) = -2$, $f\big(1, -\sqrt{2}, \sqrt{2}\big) = -2$, $f\big(-1, \sqrt{2}, -\sqrt{2}\big) = 2$, $f\big(-1, -\sqrt{2}, \sqrt{2}\big) = 2$.

Conclusion: $\max\limits_{\Omega}(f) = 2 = f\big(-1, \sqrt{2}, -\sqrt{2}\big) = f\big(-1, -\sqrt{2}, \sqrt{2}\big)$

and $\min\limits_{\Omega}(f) = -2 = f\big(1, \sqrt{2}, -\sqrt{2}\big) = f\big(1, -\sqrt{2}, \sqrt{2}\big)$.

An alternative: What if we did not know the Lagrange multiplier theorem? Then we could use the constraints to reduce the number of variables. If we decide to use x as a parameter, then $y = \pm\sqrt{3 - x^2}$, obviously $-\sqrt{3} \leq x \leq \sqrt{3}$.

We start with the version $y = \sqrt{3 - x^2}$. Then $z = -y = -\sqrt{3 - x^2}$ and we have to investigate the auxiliary function

$$\phi(x) = f\big(x, \sqrt{3 - x^2}, -\sqrt{3 - x^2}\big) = x\sqrt{3 - x^2} \cdot \Big(-\sqrt{3 - x^2}\Big) = -x(3 - x^2) = x^3 - 3x.$$

We need to identify global extrema on $\big[-\sqrt{3}, \sqrt{3}\,\big]$. Right away we add to our list the endpoints $x = \pm\sqrt{3}$, returning to the original setting we actually add points $\big(\pm\sqrt{3}, 0, 0\big)$.

Now we look for local extrema:

$$\phi'(x) = 0 \implies 3x^2 - 3 = 0 \implies x = \pm 1.$$

We obtain the points $\big(1, \sqrt{2}, -\sqrt{2}\big)$ and $\big(-1, \sqrt{2}, -\sqrt{2}\big)$.

Similarly we work out the version $y = -\sqrt{3 - x^2}$, $z = -y = \sqrt{3 - x^2}$ which yields the points $\big(1, -\sqrt{2}, \sqrt{2}\big)$ and $\big(-1, -\sqrt{2}, \sqrt{2}\big)$.

So this approach is also possible and quite pleasant.

$\triangle$

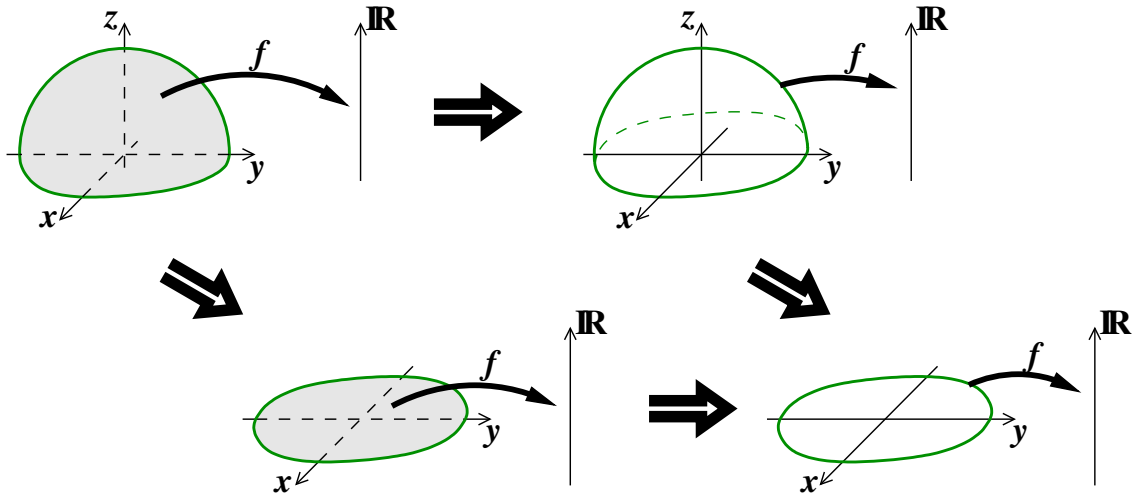
This example shows how we work with two constraints, but perhaps it was not the best advertisement for Lagrange multipliers. We will therefore conclude this chapter with a more substantial example.

Example: We find global extrema of the function $f(x, y, z) = xy + (z - 1)^2$ on the set

$$\Omega = \{(x, y, z) \in \mathbb{R}^3;\ z \geq 0 \text{ a } x^2 + y^2 + z^2 \leq 2^2\},$$

that is, on the upper half-ball (dome) of radius 2.

We already thought of the steps needed to solve this problem, including pictures, so now we do it properly.



1. Collecting candidates.

a) Local extrema: The set Ω has a non-empty interior, that is, it is a fully n-dimensional set, therefore local extrema relative to this set correspond to the usual local extrema. The equation $\nabla f = \vec{0}$ leads to the system

$$\begin{aligned} y &= 0 \\ x &= 0 \implies x = 0, y = 0, z = 1. \\ 2(z-1) &= 0 \end{aligned}$$

The point $(0, 0, 1)$ lies in our dome, so it is a valid candidate.

b) Extrema on the boundary.

The boundary of a half-ball is its surface and consists of two parts, the upper half-sphere and the circular base.

b1) We now look for extrema of the function f on the upper half-sphere. It is actually a new subproblem that we solve in the usual way, that is, we will investigate the behaviour of f inside this set and on its boundary.

Local extrema: We are looking for extrema of the function $f(x, y, z) = xy + (z-1)^2$ with respect to the constraint $x^2+y^2+z^2-4 = 0$, that is, with the constraint function $g(x, y, z) = x^2+y^2+z^2-4$. We will use Lagrange multipliers. We set up the appropriate equations:

$$\begin{aligned} \nabla f(x,y,z) &= \lambda \nabla g(x,y,z) \\ g(x,y,z) &= 0 \end{aligned} \implies \begin{aligned} y &= 2x\lambda \\ x &= 2y\lambda \\ 2(z-1) &= 2z\lambda \\ x^2+y^2+z^2 &= 4. \end{aligned}$$

Is it possible to use the popular strategy and eliminate the parameter λ that we do not want anyway? We would like to express it from the first equation, but we need $x \neq 0$ for that. So let's start by exploring the case $x = 0$.

The first equation then yields $y = 0$, the constraint then gives $z = \pm 2$, but we are only interested in $z \geq 0$. We also check that $\lambda = \frac{1}{2}$ validates the third equation and we have the first suspicious point $(0, 0, 2)$ for our list.

If $x \neq 0$, then from the first equation we get $\lambda = \frac{y}{2x}$. We substitute into the second equation and obtain $x = \frac{2y^2}{2x}$, that is, $x^2 = y^2$. So $y = \pm x$ and consequently $\lambda = \pm\frac{1}{2}$.

Case $y = x$ and $\lambda = \frac{1}{2}$: Then the third equation reads $2(z-1) = z$, that is, $z = 2$. From the constraint $x^2 + x^2 + 2^2 = 4$ we then get $x = y = 0$, we already included this point in our list.

Case $y = -x$ and $\lambda = -\frac{1}{2}$: Then the third equation reads $2(z-1) = -z$, that is, $z = \frac{2}{3}$. Substituting this and $y = -x$ into the constraint we get $2x^2 + \frac{4}{9} = 4$, so $x = \pm\frac{4}{3}$. We obtain suspicious points $\left(\frac{4}{3}, -\frac{4}{3}, \frac{2}{3}\right)$ and $\left(-\frac{4}{3}, \frac{4}{3}, \frac{2}{3}\right)$.

Now we should check on the boundary of the upper half-sphere. It is actually the circle in the

xy-plane with radius 2. If we want to work formally, we can see it as a curve in $\mathbb{R}^3$ given by constraints $x^2+y^2=4$ and $z=0$, that is, by constraint functions $g(x,y,z)=x^2+y^2-4$ and $h(x,y,z)=z$. We set up Lagrange equations:

$$\begin{array}{rcl} \nabla f(x,y,z)=\lambda\nabla g(x,y,z)+\mu\nabla h(x,y,z) & & y=2x\lambda+0\cdot\mu \\ g(x,y,z)=0 & \Longrightarrow & 2(z-1)=0\cdot\lambda+1\cdot\mu \\ h(x,y,z)=0 & & x^2+y^2=4 \\ & & z=0 \end{array} \quad \begin{array}{r} y=2x\lambda \\ x=2y\lambda \\ \Longrightarrow\ 2(z-1)=\mu \\ x^2+y^2=4 \\ z=0. \end{array}$$

From $z=0$ we immediately get $\mu=2$, which we do not need, the important thing is that the third equation is satisfied. We like to eliminate parameters, can we express λ from the first equation? What if $x=0$? Then this equation also yields $y=0$, but these values do not satisfy the constraint equation of the circle, so this cannot happen in our system. We can therefore rewrite the first equation as $\lambda=\frac{y}{2x}$. Substituting into the second we get $x=\frac{2y^2}{2x}$, that is, $x^2=y^2$, that is, $x=\pm y$. We substitute into the constraint and obtain $2x^2=4$, so $x=\pm\sqrt{2}$. We obtain four more points into our list, namely $\big(\pm\sqrt{2},\pm\sqrt{2},0\big)$.

The reader may have noticed that we did not really use the variable z, as it was in fact just a two-dimensional problem. Indeed, we could have introduced the auxiliary function

$$\phi(x,y)=f(x,y,0)=xy+1$$

and look for its extrema with respect to the constraint $x^2+y^2=4$. Then we would work with the constraint function $g(x,y)=x^2+y^2-4$ and with the system

$$\begin{array}{rcl} \nabla\phi(x,y)=\lambda\nabla g(x,y) & & y=2x\lambda \\ g(x,y)=0 & \Longrightarrow & x=2y\lambda \\ & & x^2+y^2=4. \end{array}$$

These are exactly the equations that produced the points $\big(\pm\sqrt{2},\pm\sqrt{2},0\big)$.

b2) Now we look for extrema of the function f on the base, that is, on a disc in the xy-plane given by the condition $x^2+y^2\le 4$. It is essentially a two-dimensional problem, so we introduce the auxiliary function

$$\phi(x,y)=f(x,y,0)=xy+1.$$

We are looking for its extrema on a disc. Again, this is a subproblem that we solve as an independent problem, so we check on two sources of candidates: inside and on the boundary.

The boundary of our disc is the circle that we already investigated, so this part is done. It remains to find local extrema of ϕ in the disc. By introducing the auxiliary function we got rid of the constraint, so this is a classical local extreme problem and we just find stationary points:

$$\nabla\phi(x,y)=\vec{0}\ \Longrightarrow\ \begin{array}{r} y=0 \\ x=0 \end{array}\ \Longrightarrow\ x=0,\ y=0.$$

We obtained another (and last) suspicious point $(0,0,0)$.

Note that it was not possible to use Lagrange multipliers to handle the disc, because it is given by an inequality, not by an equation.

2. Comparison of candidates.

We found nine suspicious points altogether. The function has the following values at these points: $f(0,0,1)=0$, $f(0,0,0)=1$, $f\big(\sqrt{2},\sqrt{2},0\big)=3$, $f\big(-\sqrt{2},\sqrt{2},0\big)=-1$, $f\big(\sqrt{2},-\sqrt{2},0\big)=-1$, $f\big(-\sqrt{2},-\sqrt{2},0\big)=3$, $f(0,0,2)=1$, $f\big(\frac{4}{3},-\frac{4}{3},\frac{2}{3}\big)=-\frac{5}{3}$, $f\big(-\frac{4}{3},\frac{4}{3},\frac{2}{3}\big)=-\frac{5}{3}$.

Comparing them we find that

$$\min_{\Omega}(f) = -\tfrac{5}{3} = f\big(\tfrac{4}{3}, -\tfrac{4}{3}, \tfrac{2}{3}\big) = f\big(-\tfrac{4}{3}, \tfrac{4}{3}, \tfrac{2}{3}\big),$$
$$\max_{\Omega}(f) = 3 = f\big(\sqrt{2}, \sqrt{2}, 0\big) = f\big(-\sqrt{2}, -\sqrt{2}, 0\big).$$

△

We close this chapter with an interesting question: How would it go if instead of using Lagrange multiplicators we just followed a common sense approach, that is, if we used the constraints to reduce the number of variables and similar tricks?

We start with the base, because there we work with a pleasant two-dimensional function

$$\phi(x, y) = f(x, y, 0) = xy + 1.$$

We already found the local extrema, but how about the boundary? It is given by the condition $x^2 + y^2 = 4$ that offers two immediate approaches to reduce the number of variables.

One possibility is to split the circle into two parts, one given by the equation $y = \sqrt{4 - x^2}$ and the other given by the equation $y = -\sqrt{4 - x^2}$ for $x \in [-2, 2]$. The behaviour of ϕ and thus of f on the first part of the circle is described by the auxiliary function

$$\psi(x) = f\big(x, \sqrt{4 - x^2}, 0\big) = x\sqrt{4 - x^2} + 1.$$

We are looking for extrema of ψ on the set $[-2, 2]$, so it is a new optimization problem.

First we find local extrema.

$$\psi'(x) = 0 \implies \sqrt{4 - x^2} + \frac{-x^2}{\sqrt{4 - x^2}} = 0 \implies x = \pm\sqrt{2}.$$

We obtained points $\big(\sqrt{2}, \sqrt{2}, 0\big)$ and $\big(-\sqrt{2}, \sqrt{2}, 0\big)$ for our list of suspects.

It remains to check on the boundary of the set $[-2, 2]$. This boundary consists of two values, namely $x = \pm 2$, and we add the corresponding three-dimensional points into our list: $(2, 0, 0)$ and $(-2, 0, 0)$.

Similarly we work out the other part of the circle given by $y = -\sqrt{4 - x^2}$, the endpoints agree and we get two more candidates, $\big(\sqrt{2}, -\sqrt{2}, 0\big)$ and $\big(-\sqrt{2}, -\sqrt{2}, 0\big)$.

We see that we found the four points identified by the Lagrange approach, but now we also have extra two points $(\pm 2, 0, 0)$. These are artefacts of parametrization and result in more work in the comparison stage.

It is also possible to use different two parts of the circle given by the formulas $x = \pm\sqrt{4 - y^2}$. This would also lead to the candidates $\big(\pm\sqrt{2}, \pm\sqrt{2}, 0\big)$ and additionally we would get the endpoints $(0, \pm 2, 0)$. So also here we would have extra two points.

What is unpleasant about this approach is the necessity to do everything twice, because we were unable to express the circle with just one formula. This motivates us to try some better parametrization, namely polar coordinates. Recall that we are looking for extrema of the function $\phi(x, y) = xy+1$ on a certain circle of radius 2. We try the parametrization $x = 2\cos(\varphi)$, $y = 2\sin(\varphi)$ for $\varphi \in [0, 2\pi]$. We investigate the values of the auxiliary function

$$\psi(\varphi) = f(2\cos(\varphi), 2\sin(\varphi), 0) = 4\cos(\varphi)\sin(\varphi) + 1 = 2\sin(2\varphi) + 1.$$

One source of suspicious points is the endpoints for φ, that is, 0 and 2π. Both values add the same point into our list, namely $(2, 0, 0)$. It is an artifact of parametrization again.

The question of local extrema is more interesting.

$$\psi'(\varphi) = 0 \implies 4\cos(2\varphi) = 0 \implies \varphi = \frac{\pi}{4} + k\frac{\pi}{2},\ k \in \mathbb{Z}.$$

We obtain four values, $\varphi = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$, and after substituting into appropriate formulas we get the good old suspicious points $\big(\pm\sqrt{2}, \pm\sqrt{2}, 0\big)$.

In my view the Lagrange multipliers worked better, in particular because it did not produce those extra unnecessary points, but the reader is welcome to differ.

Now it is time to check on local extrema on the upper half-sphere. We will show three approaches.

1. Looking at the equation $x^2+y^2+z^2=4$ we may get the idea that it would be nice to express $z=\sqrt{4-x^2+y^2}$. Then we should look for extrema of the auxiliary function

$$\phi(x,y)=f\big(x,y,\sqrt{4-x^2-y^2}\,\big)=xy+(\sqrt{4-x^2-y^2}-1)^2$$

on the set of points (x,y) satisfying $x^2+y^2\le 2$. Here we go:

$$\nabla\phi=\vec{0}\implies \begin{array}{l} y-2x\dfrac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}=0\\ x-2y\dfrac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}=0\end{array}\implies \begin{array}{l} y=2x\dfrac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}\\ x=2y\dfrac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}\end{array}$$

When we substitute x from the second equation into the first, we get

$$y=4y\Big(\frac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}\Big)^2.$$

If $y=0$, then also $x=0$ and we are getting the point $(0,0,2)$ for our list of candidates.

If $y\neq 0$, then we can cancel and obtain

$$1=4\Big(\frac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}\Big)^2\implies \frac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}=\pm\frac12.$$

The case of $+\frac12$: The first equation from the gradient condition then yields $y=x$. We also have

$$\frac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}=\frac12\implies \sqrt{4-x^2-y^2}=2\implies x^2+y^2=0.$$

This has only one solution, $x=y=0$. Then $z=2$ and we already have this point in our list.

The case of $-\frac12$: The first equation from the gradient condition then yields $y=-x$. We also have

$$\frac{\sqrt{4-x^2-y^2}-1}{\sqrt{4-x^2-y^2}}=-\frac12\implies \sqrt{4-x^2-y^2}=\frac23.$$

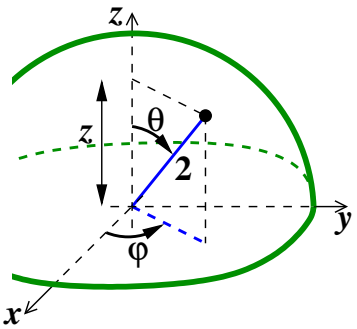
This tells us that $z=\frac23$. When we put in $y=-x$, we get

$$\sqrt{4-2x^2}=\frac23\implies x=\pm\sqrt{2-\frac29}=\pm\frac43.$$

We get two more suspicious points, namely $\big(\frac43,-\frac43,\frac23\big)$ and $\big(-\frac43,\frac43,\frac23\big)$.

We obtained the same points as with the Lagrange approach, in my view the calculations were a bit more friendly then.

2. A more experienced solver might recall that the sphere can be described rather conveniently using polar coordinates, or more precisely, spherical coordinates. Given a point on a sphere, we use θ for the polar angle describing how much this point leans away from the z-axis. This determines the z-coordinate and also the circle on which the point lies. We fix its position on this circle by an azimuth φ.



$$\begin{aligned} x&=2\cos(\varphi)\sin(\theta)\\ y&=2\sin(\varphi)\sin(\theta)\\ z&=2\cos(\theta).\end{aligned}$$

Then we work with the auxiliary function

$$\phi(\varphi,\theta)=4\cos(\varphi)\sin(\varphi)\sin^2(\theta)+(2\cos(\theta)-1)^2=2\sin(2\varphi)\sin^2(\theta)+(2\cos(\theta)-1)^2.$$

We are looking for extrema for $0\le\varphi\le 2\pi$ and $0\le\theta\le\frac{\pi}{2}$. We focus on local extrema here, the

condition is

$$\nabla\phi = \vec{0} \implies \begin{array}{r} 4\cos(2\varphi)\sin^2(\theta) = 0 \\ 4\sin(2\varphi)\sin(\theta)\cos(\theta) - 4(2\cos(\theta)-1)\sin(\theta) = 0. \end{array}$$

The first equations offers two possibilities. One is $\sin(\theta) = 0$. Given the range of this variable, the only possible value is $\theta = 0$, that is, $z = 2$. Then $x = y = 0$, we get the point $(0, 0, 2)$.

The second possibility is $\sin(\theta) \neq 0$, which means that we can cancel it in the second equation and the first yields $\cos(2\varphi) = 0$. Then $2\varphi = \frac{\pi}{2} + 2k\pi$, that is, $\varphi = \frac{\pi}{4} + k\pi$.

The case $\varphi = \frac{\pi}{4}$ or $\varphi = \frac{5\pi}{4}$: Then $\sin(2\varphi) = 1$, so the second equations says that

$$\cos(\theta) - (2\cos(\theta) - 1) = 0 \implies \cos(\theta) = 1.$$

We get $z = 2$, we have already been there.

Case $\varphi = \frac{3\pi}{4}$ or $\varphi = \frac{7\pi}{4}$: Then $\sin(2\varphi) = -1$, so the second equations says that

$$-\cos(\theta) - (2\cos(\theta) - 1) = 0 \implies \cos(\theta) = \frac{1}{3}.$$

We get $z = \frac{2}{3}$. What else do we know? From $\cos(\theta) = \frac{1}{3}$ we get $\sin(\theta) = \sqrt{\frac{8}{9}} = \frac{2}{3}\sqrt{2}$. Again two subcases:

If $\varphi = \frac{3\pi}{4}$, then

$$x = 2 \cdot \frac{2}{3}\sqrt{2} \cdot \left(-\frac{\sqrt{2}}{2}\right) = -\frac{4}{3},$$
$$y = 2 \cdot \frac{2}{3}\sqrt{2} \cdot \frac{\sqrt{2}}{2} = \frac{4}{3}.$$

We have a suspicious point $\left(-\frac{4}{3}, \frac{4}{3}, \frac{2}{3}\right)$.

If $\varphi = \frac{7\pi}{4}$, then

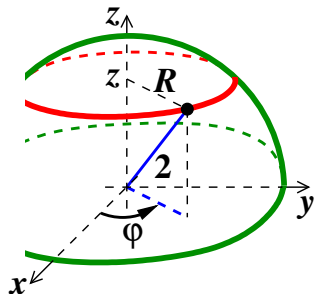
$$x = 2 \cdot \frac{2}{3}\sqrt{2} \cdot \frac{\sqrt{2}}{2} = \frac{4}{3},$$
$$y = 2 \cdot \frac{2}{3}\sqrt{2} \cdot \left(-\frac{\sqrt{2}}{2}\right) = -\frac{4}{3}.$$

We have a suspicious point $\left(\frac{4}{3}, -\frac{4}{3}, \frac{2}{3}\right)$.

So we obtained the same four points, this time the calculations were perhaps a bit more involved.

3. It seems that the calculations were made less pleasant by a composed function in the expression $(z-1)^2$. We can avoid it if we use z as one of the parameters. In this way people come to cylindrical coordinates, but these only work with three-dimensional objects. Here we have to find our own way.

If we fix some value for z, then it uniquely determines a certain circle on the upper half-sphere, namely the one at elevation z. Its radius is $R = \sqrt{4 - z^2}$. A specific point on this circle can be determined, say, by an angle, which is suitable for our circular situation. In this way we arrive at a non-standard parametrization



$$x = \sqrt{4 - z^2}\cos(\varphi)$$
$$y = \sqrt{4 - z^2}\sin(\varphi)$$
$$z = z,$$

where $z \in [0, 2]$ and $\varphi \in [0, 2\pi]$. We are looking for local extrema of the auxiliary function

$$\phi(\varphi, z) = (4 - z^2)\cos(\varphi)\sin(\varphi) + (z-1)^2 = \tfrac{1}{2}(4 - z^2)\sin(2\varphi) + (z-1)^2.$$

Gradient supplies equations:

$$\nabla\phi = \vec{0} \implies \begin{array}{c}(4-z^2)\cos(2\varphi) = 0,\\ -z\sin(2\varphi) + 2(z-1) = 0.\end{array}$$

In the first equation we may have $4 - z^2 = 0$, the parametric formulas then yield $x = y = 0$ and we have the candidate $(0, 0, 2)$.

Otherwise we must have $\cos(2\varphi) = 0$. We now look at cases just like in the previous solution.

The case $\varphi = \frac{\pi}{4}$ or $\varphi = \frac{5\pi}{4}$: Then $\sin(2\varphi) = 1$, so the second equations says that

$$-z + 2(z-1) = 0 \implies z = 2.$$

We have been here before.

The case $\varphi = \frac{3\pi}{4}$ or $\varphi = \frac{7\pi}{4}$: Then $\sin(2\varphi) = -1$, so the second equations says that

$$z + 2(z-1) = 0 \implies z = \frac{2}{3}.$$

We work out the other variables. For $\varphi = \frac{3\pi}{4}$ we have

$$x = \sqrt{4 - \frac{4}{9}} \cdot \left(-\frac{\sqrt{2}}{2}\right) = -\frac{4}{3},$$
$$y = \sqrt{4 - \frac{4}{9}} \cdot \frac{\sqrt{2}}{2} = \frac{4}{3}.$$

We get the suspicious point $\left(-\frac{4}{3}, \frac{4}{3}, \frac{2}{3}\right)$.

For $\varphi = \frac{7\pi}{4}$ then

$$x = \sqrt{4 - \frac{4}{9}} \cdot \frac{\sqrt{2}}{2} = \frac{4}{3},$$
$$y = \sqrt{4 - \frac{4}{9}} \cdot \left(-\frac{\sqrt{2}}{2}\right) = -\frac{4}{3}.$$

We get the suspicious point $\left(\frac{4}{3}, -\frac{4}{3}, \frac{2}{3}\right)$.

It was similar to the polar coordinates approach, but perhaps the calculations were somewhat more friendly.

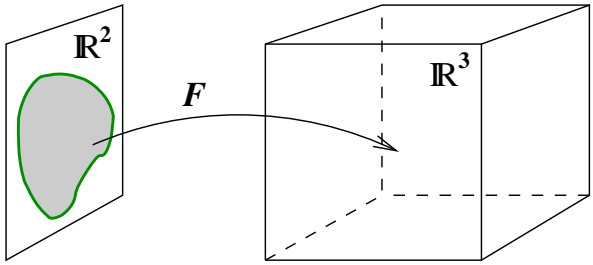
△

This completes our survey of basic approaches to investigation of global and constrained extrema. We note that we will return to parametric surfaces and curves again in section 5a from a theoretical point of view and in section 6d practically, while coordinates will be revisited in section 7b.

5. Introduction to vector functions

We introduced functions of more variables as a natural generalization, when we replaced one-dimensional domain with a multi-dimensional one. It is equally natural to do the same with the range.

A **vector function** is an arbitrary mapping $F\colon D \mapsto \mathbb{R}^m$, where $D \subseteq \mathbb{R}^n$. Sometimes people also say **vector field**, we will see the motivation in section 6c.

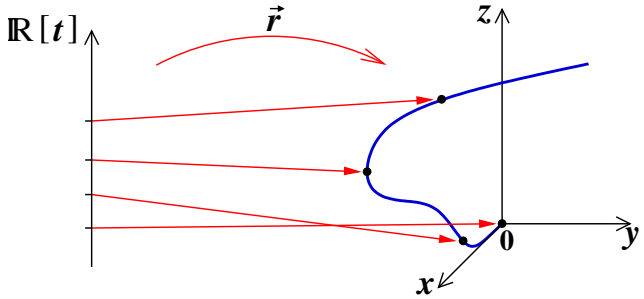


We have already met an example of a vector function. Given a function of more variables $f(\vec{x})$, its gradient also features the variable $\vec{x}$ and its values are vectors, hence it is a vector function. For instance, given the function $f(x,y) = xy^2$, we obtain the gradient $\nabla f(x,y) = (y^2, 2xy)$, it is a function $\mathbb{R}^2 \mapsto \mathbb{R}^2$.

Since the values of a vector function F are vectors, it would be helpful to denote such function $\vec{F}$ to remind the reader. But mathematicians never bothered to do so, and thus we will also write F here to be in tune with other sources. Another important convention is that we can write values of such a vector function using components, that is, we write $F(\vec{x}) = \big(F_1(\vec{x}), \dots, F_m(\vec{x})\big)$. Typically we work with Cartesian coordinates in the range, but it is not necessary. The important thing is that the behaviour of the function F is closely related to properties of its components F_j. In many situations it is natural to work with just the components (see transformation of variables below), then we actually do not need the notion of a vector function. However, there are situations when vector functions provide a natural and efficient framework.

5a. Parametric curves

In the mathematical world, people commonly meet vector functions in the form of parametric curves and surfaces. Imagine an airplane flying in the air. We introduce some natural Cartesian system, say, with the origin at the airport and axes pointing east and north. The actual position of the airplane is then given as a three-dimensional vector traditionally denoted $\vec{r}$ that depends on time, so it is a function $\vec{r} = \vec{r}(t)$, that is, a vector function $\mathbb{R} \mapsto \mathbb{R}^3$.



The path that the airplane followed, more precisely, the set of points through which the airplane passed, and even more precisely, the set of values of the vector function $\vec{r}(t)$ is called a parametric curve. In this case it is an object that is, by its nature, one-dimensional (we can only move in one direction within it, there or back along the path). However, it is placed in a three-dimensional space. In many situations we can recognize such an object and give it a name. For instance, if the airplane was circling, then the resulting curve would be a circle in $\mathbb{R}^3$.

The function $\vec{r}(t)$ is called a parametrization of this curve. It carries information about its shape, but also about the process that created it, for instance the velocity of movement along this curve and such, so parametric curves are very useful in applications. The reader surely reasoned out that

the same curve can be parametrized in many ways; for instance, we can consider a certain stretch of a railway (for simplicity we imagine that it is one point wide, so it is a curve), and then every train passing along creates its own parametrization of this curve.

Generally speaking, every vector function of the type $\mathbb{R} \mapsto \mathbb{R}^n$ can be interpreted as a parametric curve, a mathematical description of an object that is essentially one-dimensional but placed in the space $\mathbb{R}^n$. Since the focus of this chapter is on vector functions, we will use the general notation $F(t)$ rather than $\vec{r}(t)$ here.

An interesting mathematical question is to recognize which curve is given by a given vector function. Conversely, sometimes we have a certain mathematical object and we would like to describe it. It may be possible to do it using equations (say, a line or a circle in a plane), but it may be preferable to do it using a parametric curve, then we say that we parametrized this object.

Example: Consider the vector function $F(t) = (a_1 t + b_1, a_2 t + b_2)$, where a_i, b_i are fixed parameters. It is a function $\mathbb{R} \mapsto \mathbb{R}^2$. Assuming the standard axis names in the range, we in fact have a description of a movement in the form

$$x = a_1 t + b_1$$
$$y = a_2 t + b_2.$$

This form is used quite a lot, often people work directly with coordinates without referring to the notion of a vector function.

We can imagine that this is a recording of a bug's movement on a table, with t being time. If we dipped its belly in ink, it would leave a trace behind, some mathematical object in the plane. What object is it? The classical approach is to eliminate the parameter.

$$t = \frac{1}{a_1}(x - b_1) \implies y = a_2 \frac{1}{a_1}(x - b_1) + b_2 \implies y = \frac{a_2}{a_1} x + \left(b_2 - \frac{a_2}{a_1} b_1\right).$$

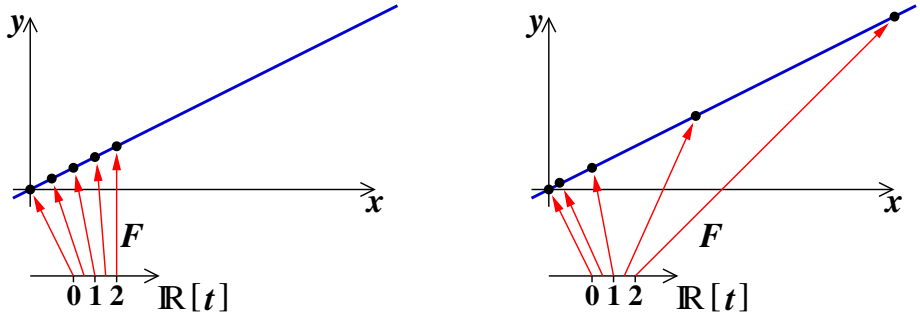
We just found out that the bug crawled along a straight line. Mathematically speaking, the given function is a parametrization of a line. We see right away one advantage of a parametric description. The formula $y = ax + b$ cannot express a vertical line (and thus it is not a universal formula for lines in the plane), while the parametric description handles this easily by taking $a_1 = 0$.

The same line can be described parametrically in other ways, sometimes it makes no difference, but sometimes the right choice can significantly simplify work. For instance, the line given by a certain formula $y = ax + b$ can be parametrized as $F(t) = (t, at + b)$, that is,

$$x = t$$
$$y = at + b,$$

but also as $F(t) = (13t + 7, 13at + 7a + b)$.

One could play even more. For instance, now we know how to deduce that $F(t) = (2t, t)$ is a line corresponding to the equation $y = \frac{1}{2}x$, but the same line can be obtained also by $F(t) = (2t^3, t^3)$. Geometrically it is the same object, but in physics these would be two very different processes, the second bug is accelerating massively.



Now imagine this accelerating bug, moving along a straight line road, but this road is not directly on the ground; it is raised on pylons two meters above ground due to marshy terrain. Then we

would record the position as $F(t) = (2t^3, t^3, 2)$, so it would be a one-dimensional parametric curve in the 3D space.
△

Example: The curve determined by the vector function $F(t) = \big(\cos(t), \sin(t)\big)$, that is,

$$x = \cos(t)$$
$$y = \sin(t)$$

can also be identified by eliminating t. Probably the easiest way here is to square both equations and then add them. We find that the curve is the unit circle $x^2+y^2 = 1$, the parametric description says more, however. We can imagine, say, a dog on a leash attached to a spike in the ground, running around at a constant speed in the positive direction (counterclockwise). If the dog runs in the other direction and twice as fast, it creates the parametrization $F(t) = \big(\cos(-2t), \sin(-2t)\big)$ of the same circle.

The parametric curve $F(t) = (\cos(t^2), \sin(t^2))$ also describes the unit circle, but in this case the speed of rotation keeps increasing. Now we will look at some interesting modifications.

$F(t) = (t\cos(t), t\sin(t))$ for $t \in [1, \infty)$ creates a spiral, we run around while increasing the radius.

$F(t) = \big(\frac{1}{t}\cos(t), \frac{1}{t}\sin(t)\big)$ for $t \in [1, \infty)$ is also a spiral, but this time it is winding in towards the origin.

$F(t) = (\cos(t), \sin(t), t)$ is a journey in three dimensions. When projected into the xy plane, it is the familiar unit circle. However, now we also gradually increase our elevation. The resulting curve is therefore a raising spiral, we can imagine a glider that caught a good thermal, so now it circles and climbs.

We close this off with something different. Consider $F(t) = (t^2, 2t, \sin(t))$. The first two coordinates show us the projection of a bug's path onto the xy plane. We eliminate t from $x = t^2$, $y = 2t$ and obtain $x = \frac{1}{4}y^2$. The bug therefore follows the parabola open to the right, in the x direction. The elevation of the bug over the xy plane is given by the third coordinate, which is $\sin(t)$. We see that the bug takes turns flying above and below that parabola. Mathematics cannot tell us why.
△

For the sake of completeness we recall some classical **parametric curves**:

A circle in $\mathbb{R}^2$: $x = R\cos(t)$, $y = R\sin(t)$ for $t \in [0, 2\pi)$ or another interval of sufficient length, or perhaps for $t \in \mathbb{R}$, we can run around the circle repeatedly.

An ellipse in $\mathbb{R}^2$: $x = A\cos(t)$, $y = B\sin(t)$ for $t \in [0, 2\pi)$ or another interval of sufficient length, or perhaps for $t \in \mathbb{R}$.

A straight line in $\mathbb{R}^n$ passing through the point $\vec{a}$ in direction $\vec{u}$ is given by the function $F(t) = \vec{a} + t\vec{u}$ for $t \in \mathbb{R}$. Here we see one advantage of a parametric approach. It is not possible to specify a line in three dimensions by just one algebraic equation, two are needed for that (line is an intersection of two hyperplanes), while just one formula is enough in the parametric description.

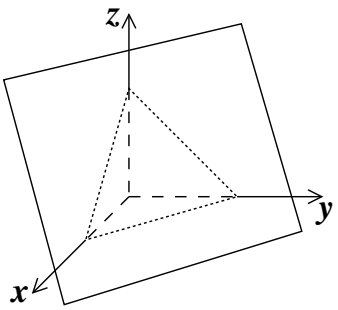
Parametric surfaces is another interesting type of object. These are objects that are, by its nature, two-dimensional, but they are positioned in a space of higher dimension. They can thus be described using vector functions of type $\mathbb{R}^2 \mapsto \mathbb{R}^n$ for $n \geq 3$. We can imagine a bed sheet floating in a breeze.

A useful physics point of view is to interpret it as a description of the behaviour of some device that moves in three dimensions, but not arbitrarily, so thanks to some mechanical limitations we are able to control it using just two controls. For instance, a portal crane (or overhead/bridge crane) is a carriage riding on a straight segment of rails, often suspended under the ceiling of a factory hall, with hook suspended from it on a cable. The operator can ride the carriage along the rail and lower or raise the hook, that is, there are two controls. The set of all possible positions of the hook is a vertical flat rectangle, an essentially two-dimensional object in three-dimensional

space. If the rail is curved, then the set of all possible positions of the hook is no longer flat, it is a "bent" rectangle, still an essentially two-dimensional object.

Example: Consider the plane in $\mathbb{R}^3$ that intersects all three axes at position 1, that is, it passes through the points $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. It is given by the equation $x+y+z=1$.

If we express $z = 1-x-y$, we can interpret it as a description of how high above (or below) we see that plane as we stand in the xy plane at the position (x,y). Using this we easily create a parametric description of our plane:



$$\begin{aligned} x &= x \\ y &= y \\ z &= 1-x-y. \end{aligned}$$

It will look better if we use different names for parameters:

$$\begin{aligned} x &= s \\ y &= t \\ z &= 1-s-t. \end{aligned}$$

In other words, this is a vector function $F(s,t) = (s,t,1-s-t)$. As usual, there are also other parametrizations, for instance $F(u,v) = (u+v, u-v, 1-2u)$. We could confirm the match by eliminating parameters u, v from the equations

$$\begin{aligned} x &= u+v \\ y &= u-v \\ z &= 1-2u \end{aligned}$$

and obtaining $x+y+z=1$.

△

We have actually spent quite some time with parametrization of surfaces and curves in section 4b, especially at the end, and we will use it heavily in section 6d.

Generally speaking, every vector function $F\colon \mathbb{R}^n \mapsto \mathbb{R}^m$, where $n < m$, can be interpreted as a description of some n-dimensional object embedded in an m-dimensional space, we just cannot really visualize it for $m > 3$.

The case $m = n$ has a different useful interpretation, namely a change of coordinates. For instance, if we want to pass between Cartesian and polar coordinates in a two-dimensional space, we can in one direction use pleasant formulas

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

This is actually a vector function $F(r,\varphi) = \big(r\cos(\varphi), r\sin(\varphi)\big)$ from $\mathbb{R}^2$ to $\mathbb{R}^2$.

In applications people often work with just the individual formulas $x = x(r,\varphi)$, $y = y(r,\varphi)$ for new coordinates, but the vector function approach can sometimes simplify situation. Change of coordinates is addressed theoretically in section 7b and used in section 6d.

5b. Basic analytic notions

Limit and continuity can be easily generalized, we just consider multi-dimensional neighborhoods in the range space.

Definition.
Let F be a vector function with values in $\mathbb{R}^m$ defined on some reduced neighborhood of a point $\vec{a}$, let $\vec{L} \in \mathbb{R}^m$.
We say that $\vec{L}$ is a limit of F as $\vec{x}$ goes to $\vec{a}$, denoted $\lim_{\vec{x}\to\vec{a}} \big(F(\vec{x})\big) = \vec{L}$, if for every neighborhood U of $\vec{L}$ there is some reduced neighborhood V of $\vec{a}$ such that $F[V] \subseteq U$.

A vector function can be written as $F = (F_1, \dots, F_m)$. It is easy to prove that $\lim_{\vec{x}\to\vec{a}} \big(F(\vec{x})\big) = \vec{L}$ if and only if $\lim_{\vec{x}\to\vec{a}} \big(F_j(\vec{x})\big) = L_j$ for all $j = 1, \dots, m$.

It would therefore be possible to define the limit in this way, "coordinatewise". We will now use this approach and proclaim that a vector function F is continuous at a certain point, on a certain set or simply continuous (which means on its domain) if this is true for all its components F_j.

As usual, continuous vector functions are nice, for instance their graphs are not "torn" and many statements involving continuous functions can be generalized. We now list briefly some popular facts.

Theorem. (Heine theorem)
Let F be a vector function with values in $\mathbb{R}^m$ defined on some reduced neighborhood D of a point $\vec{a}$, let $\vec{L} \in \mathbb{R}^m$.
Then $\lim_{\vec{x}\to\vec{a}} \big(F(\vec{x})\big) = \vec{L}$ if and only if $\lim_{k\to\infty} \big(F(\vec{x}(k))\big) = \vec{L}$ for all sequences $\{\vec{x}(k)\} \subseteq D - \{\vec{a}\}$ satisfying $\vec{x}(k) \to \vec{a}$.

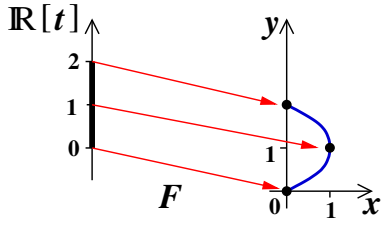
Theorem.
Let F be a vector function with values in $\mathbb{R}^m$ defined on some reduced neighborhood of a point $\vec{a}$, let $\vec{L} \in \mathbb{R}^m$.
If the limit of F at $\vec{a}$ exists and is finite, then F is bounded on some reduced neighborhood of $\vec{a}$.

Theorem.
Let $F\colon D \mapsto \mathbb{R}^m$ be a function, where $D \subseteq \mathbb{R}^n$ is a bounded closed set.
(i) If f is continuous on D, then $F[D]$ is a bounded closed set.
(ii) If F is continuous and one-on-one on D, then also the inverse function $F_{-1}\colon F[D] \mapsto D$ is continuous.

However, we should not get carried away, for instance it is not possible to generalize the Intermediate value theorem.

Example: Consider the parametric curve $F(t) = (2t - t^2, t)$ for $t \in [0, 2]$.

It is a hill in the xy plane. We easily find $F(0) = (0, 0)$ and $F(2) = (0, 2)$, so the endpoints of this curve are on the y-axis. But it is definitely not true that the function would attain also the values between these two points. This would mean the segment between the points $(0, 0)$ and $(0, 2)$, and the function cannot get there, although it is continuous.



However, what works is that the domain $[0, 2]$ as a connected set gets mapped onto a curve that is also connected.
△

Theorem.
Let F: $D(F) \mapsto \mathbb{R}^m$ be a function, where $D(F) \subseteq \mathbb{R}^n$. If F is continuous, then for every connected set $M \subseteq D(F)$ also the set $F[M]$ is connected.

We can also introduce uniform continuity for vector functions.

Definition.
Let F: $D(F) \mapsto \mathbb{R}^m$ be a function, where $D(F) \subseteq \mathbb{R}^n$. Let M be a subset of $D(F)$.
We say that F is **uniformly continuous** on M if for every $\varepsilon > 0$ there is some $\delta > 0$ so that $\|F(\vec{x}) - F(\vec{y})\| < \varepsilon$ for all $\vec{x}, \vec{y} \in M$ satisfying $\|\vec{x} - \vec{y}\| < \delta$.

Theorem.
Let F: $D \mapsto \mathbb{R}^m$ be a function, where $D \subseteq \mathbb{R}^n$.
If F is continuous on D and D is bounded and closed, then F is uniformly continuous on D.

Just like with multi-variable functions, exploration of derivatives for vector functions also starts with a directional derivative.

Definition.
Let F be a vector function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Let $\vec{u}$ be a vector from $\mathbb{R}^n$.
We say that the function F is **differentiable** at point $\vec{a}$ in direction $\vec{u}$ if the limit $\lim\limits_{t\to 0}\big(\frac{F(\vec{a}+t\vec{u})-F(\vec{a})}{t}\big)$ exists and is finite.
Then we define the **(directional) derivative** of F at point $\vec{a}$ in direction $\vec{u}$ as

$$D_{\vec{u}}F(\vec{a}) = \lim_{t\to 0}\Big(\frac{F(\vec{a}+t\vec{u}) - F(\vec{a})}{t}\Big).$$

We actually just replaced all symbols f with F in the definition of directional derivative. However, now the expressions in limits are vectors, and thus also the directional derivative F is a vector from the target space $\mathbb{R}^m$ of the function F.

Since all the operations in the formula in the definition—the subtraction, the division by a scalar, the limit process—work coordinatewise, we conclude that we are actually applying the directional derivative process to individual components of $F = (F_1, \dots, F_m)$. Consequently, the directional derivative $D_{\vec{u}}F(\vec{a})$ exists exactly if the directional derivatives $D_{\vec{u}}F_j(\vec{a})$ exist for all $j = 1, \dots, m$, and then we can write

$$D_{\vec{u}}F(\vec{a}) = \big(D_{\vec{u}}F_1(\vec{a}), \dots, D_{\vec{u}}F_m(\vec{a})\big).$$

As usual, a special role among directional derivatives is played by derivatives in coordinate directions. That is, we are interested in $\dfrac{\partial F_j}{\partial x_i}$, where $i = 1, \dots, n$ and $j = 1, \dots, m$.

For functions of more variables were wrapped partial derivatives in an object called gradient that proved itself quite useful. For vector functions we have something similar.

Definition.
Let $F = (F_1, \dots, F_m)$: $D(F) \mapsto \mathbb{R}^m$ be a vector function, where $D(F) \subseteq \mathbb{R}^n$.
We define its **Jacobi matrix**

$$J_F(\vec{a}) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\vec{a}) & \dots & \frac{\partial F_1}{\partial x_n}(\vec{a}) \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1}(\vec{a}) & \dots & \frac{\partial F_m}{\partial x_n}(\vec{a}) \end{pmatrix}$$

at points where these partial derivatives exist.
For the case $m = n$ we also define the **jacobian** as

$$\Delta_F(\vec{a}) = \det(J_F(\vec{a})).$$

We create a Jacobi matrix by putting all partial derivatives of F_1 into the first row, so in fact we make it out of the gradient of F_1. Similarly we fill in the other rows, so the matrix has m rows and n columns. Symbolically,

$$J_F = \begin{pmatrix} \nabla F_1 \\ \vdots \\ \nabla F_m \end{pmatrix}.$$

As expected, the relationship between gradients and directional derivatives also extends to vector functions. In the formula we will need to multiply the Jacobi matrix by a vector, which thus needs to be a column vector, so we use transpositions.

$$D_{\vec{u}}F(\vec{a})^T = J_F(\vec{a})\vec{u}^T.$$

The Jacobi matrix and jacobian play a very important role in investigation of parametric surfaces and also in substitution in integral.

Since we will often work with functions having derivatives, we extend the convenient spaces of smooth functions also to vector functions. For a region G the symbol $[C(G)]^m$ denotes the set of all vector functions $G \mapsto \mathbb{R}^m$ that are continuous on G, and $[C^1(G)]^m$ is the set of all vector functions $G \mapsto \mathbb{R}^m$ that have all their partial derivatives of order one (so the Jacobi matrix exists) continuous on G.

The notation is rather intuitive, given that for a general set M the symbol M^m denotes the set of all vectors of dimension m whose components are from M. Here in the set $[C(G)]^m$ we find vector functions, that is, functions that can be written as $(G_1, \dots, G_m)$, and these belong to $[C(G)]^m$ exactly if the individual component functions G_j come from $C(G)$. For $[C^1(G)]^m$ it works analogously.

We now look at some interesting interpretations of derivatives of a vector function. We start with parametric curves. Consider a parametric curve $F(t)$: $\mathbb{R} \mapsto \mathbb{R}^m$. Then $n = 1$, so instead of partial derivatives we differentiate in the usual way. The Jacobi matrix is then a column vector

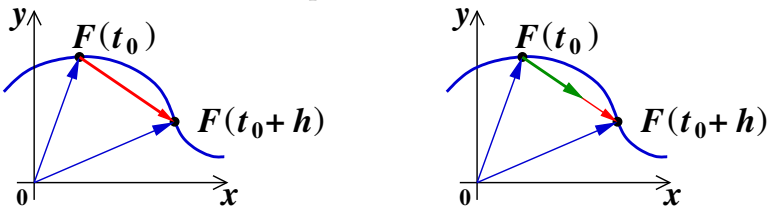
$$J_F(t) = \begin{pmatrix} F_1'(t) \\ \vdots \\ F_n'(t) \end{pmatrix}.$$

People in applications often prefer to use the row vector $F'(t) = \big(F_1'(t), \dots, F_m'(t)\big)$ and call it the derivative of F. It is easy to show that

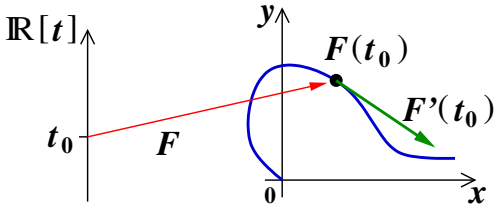
$$F'(t_0) = \lim_{h \to 0} \Big(\frac{F(t_0 + h) - F(t_0)}{h} \Big).$$

This hints at the information that this derivative supplies. Let's look at the fraction inside. We imagine that $F(t)$ is the positional vector at time t. So we are moving along a curve, the difference $F(t_0 + h) - F(t_0)$ shows how our position changed between times t_0 and $t_0 + h$. This displacement is also a vector, we see it in red in the picture on the left. It clearly shows that this displacement

need not correspond to the distance that we travelled along the curve. However, if h is very small then the displacement is a very good approximation of the distance covered. We divide by time in the fraction, then it shows the average velocity (green arrow in the picture on the right). This average velocity has the same direction as the displacement, which for small h should agree very well with the direction of the curve at that position.



Since we can expect that for reasonable movements, this fraction will approximate actual velocity at time t_0 the better the smaller h we take, we get the following interpretation: The derivative $F'(t_0)$ (if it exists) yields the actual velocity of the movement along the curve at time t_0. This velocity is a vector that is tangent to the given curve at the point $F(t_0)$.



In this context we work with the notion of **smooth curves**, these are parametric curves $F(t)$ defined on an interval on which they are continuous and have a continuous derivative such that $F'(t) \neq 0$ on the interior of that interval.

Often it is useful to have the normal vector to a curve (for instance when talking about centrifugal force). We find it using the formula

$$\left[\frac{F'(t)}{\|F'(t)\|}\right]'.$$

It follows from the following interesting fact that is then applied to the vector function $T = F'$:

Fact.
If a vector function $T(t)$ is differentiable on a neighborhood of t_0 and its magnitude $\|T(t)\|$ is constant on this neighborhood, then the vector $T'(t_0)$ is perpendicular to $T(t_0)$.

We will now look at another interesting application of derivative. Consider a differentiable vector function $F(t)$ of one variable and choose a point a. We are interested in approximating the vector $F(a+h)$. Because the vector function F is differentiable, we can differentiate also its components $F_j(t)$, these are in fact good old functions. We can thus proceed as follows:

$$\begin{aligned} F(a+h) &= \big(F_1(a+h), \dots, F_m(a+h)\big) \\ &\approx \big(F_1(a) + F_1'(a)h, \dots, F_m(a) + F_m'(a)h\big) \\ &= \big(F_1(a), \dots, F_m(a)\big) + \big(F_1'(a), \dots, F_m'(a)\big)h \\ &= F(a) + F'(a)h. \end{aligned}$$

This is not really surprising. We are interested in how our position F in space $\mathbb{R}^m$ changes when the variable t moves a bit from $t = a$. For smooth functions F, that is, for smooth curves we should be able to approximate this movement by moving along the tangent line, and we already know that the direction of a tangent line to a curve given by function F is exactly $F'(a)$.

This result suggests that also for functions of more variables there could be something like approximation using a Taylor polynomial. This is easy to deduce for vector functions of one variable by working individually with its components, just like we did above. For functions of more variables the Taylor polynomial gets more complicated and we leave it to chapter 7. Significant applications of derivatives can be found in section 6d.

5c. Differential operators

An operator is a mapping that is expected to have some pleasant algebraic properties, for instance the popular linearity. A differential operator is an operator that is based on differentiation. So far we have met two important differential operators. For functions of one variable we have the gradient that creates vector functions based on functions, and for vector functions we have the Jacobi matrix which can be seen as a process that creates matrix functions out of vector functions. Both operators conform to a number of rules for which there is no room in a text called "Illustrated introduction", so we refer the reader to the chapter More on derivative; let us just mention here that both are linear, which is also very practical.

In this section we introduce two more differential operators. Both play an important role in applications, especially in physics, and they also appear in several key theorems in integral calculus. Both operators require vector functions with $m = n$, that is, we restrict ourselves to transformations, although it will be useful to focus on a different interpretation now.

The right interpretation for this section is that a vector function captures the flow of some liquid. In the case of $n = 3$ we can imagine a river bed, then the domain D is the set of positions where there is some water, and $F(\vec{a})$ shows the direction and speed of a small particle of water located at $\vec{a}$. This is then called a velocity field. Since such a function F does not feature time as variable, it follows that we are talking here about a steady flow.

In the case $n = 2$ we may imagine water flowing on our table in a very thin layer (again we need a steady flow, so it flows in the same way all the time). We will mostly focus on this case because we can draw it, with the understanding that our observations will apply to higher dimension where we usually use then.

However, it need not be exactly water that is flowing, it could be also air or another similar medium, including heat or magnetic field. These popular settings explain why vector functions are often called vector fields in applications.

For the sake of completeness we note that there is a close connection between functions of more variables and vector fields. Imagine a function of three variables that records the temperature at different places in a room. Physics (or common sense) tell us that if there are two neighboring places with different temperatures, then there will be some heat transfer happening, that is, there will be a flow. The intensity of this flow is related to how much the temperatures differ, and it is exactly this information that gradient provides. However, the gradient gives the direction of the maximal growth, while heat flows in exactly the opposite direction (from warmth to cold). This suggests that if a function T of more variables describes the distribution of temperature, then the vector function $-\nabla T$ describes the flow of heat.

This principle can be found in many areas of physics. Not every vector field arose as a gradient, but some did, and then they share some special properties, we call them conservative fields. However, this is moving us someplace else, so we go back to the topic and introduce the first differential operator of this section.

Definition.
Let $F = (F_1, \dots, F_n)$: $D(F) \mapsto \mathbb{R}^n$ be a vector function, where $D(F) \subseteq \mathbb{R}^n$. We define its **divergence** as

$$\operatorname{div}(F)(\vec{a}) = \sum_{i=1}^{n} \frac{\partial F_i}{\partial x_i}(\vec{a}) = \frac{\partial F_1}{\partial x_1}(\vec{a}) + \frac{\partial F_2}{\partial x_2}(\vec{a}) + \cdots + \frac{\partial F_n}{\partial x_n}(\vec{a})$$

at points $\vec{a} \in D$ where the appropriate partial derivatives exist.

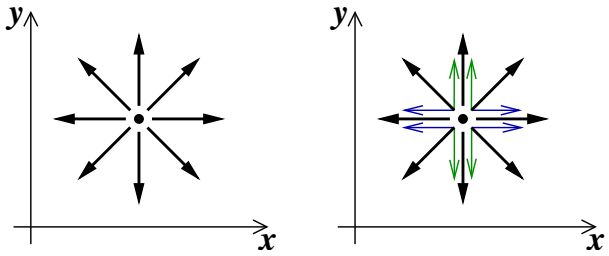
We are in fact adding numbers from the diagonal of the Jacobi matrix, which is a known mathematical operation called trace. The fans of linear algebra could therefore write $\operatorname{div}(F) = \operatorname{tr}(J_F)$, but usually we prefer different notation, see the chapter More on derivative.

What information does the divergence of a vector field provide? The standard interpretation is that if $\operatorname{div}(F)(\vec{a}) > 0$, then there is inflow of the medium at this place, it is a so-called **source**. For instance, if F describes the flow of heat, then we could expect some heating device at that place. Conversely, $\operatorname{div}(F)(\vec{a}) < 0$ means that the medium is disappearing there, it is a **sink**.

We can simply accept this, but a curious reader might wonder about several things: How does the formula recognize source from a sink? And how come that the other derivatives are missing completely, can they really do without knowing, say, $\frac{\partial F_1}{\partial x_2}$? (That is my favourite question.)

It is actually possible to deduce everything by a moderately complicated calculation involving limit of a certain integral and then we would trust it, but the aim of this illustrated introduction is to provide the reader with some intuitive understanding that makes the interpretations believable. Can it be done for the divergence?

Many authors showcase two very specific situations for a two-dimensional flow to convince readers that the definition provides the right answers.

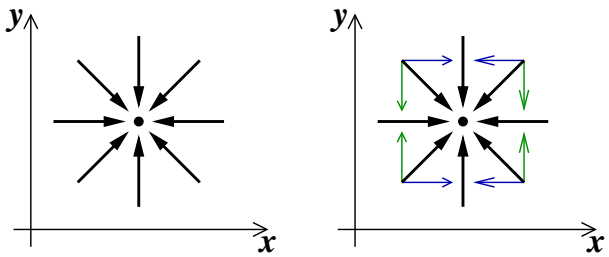


On the left we see a typical source, the flow is indicated by black arrows. Say, there is a thin stream of water falling on our table from the ceiling because the neighbor above us overfilled his bathtub. From the point of view of the thin, two-dimensional layer of water on the table it seems as if there was new water magically appearing out of nowhere at that place.

As the water falls on the table, it spreads away from this point, so the velocity vectors in the picture should make sense. If we look at the components of velocity vectors in the x-direction (we marked them in blue for oblique vectors in the picture on the right, it is actually the component F_1 of the vector function F), we can notice that for the vectors on the left of the point $\vec{a}$ they point to the left, so they are negative, while for the points to the right from $\vec{a}$ these components are positive. In other words, as we move from the left to the right across the point $\vec{a}$, then the horizontal component increases (goes from negative to positive); it is so even if we move horizontally a bit above or below the point $\vec{a}$. We thus conclude that the first component F_1 of our vector function is increasing around $\vec{a}$ as we move in the x-direction, that is, $\frac{\partial F_1}{\partial x} > 0$.

Similarly, if we move up in the vertical direction, then the green components F_2 of vectors F increase in the y-direction (from negative they become positive), so $\frac{\partial F_2}{\partial y} > 0$. We get

$$\operatorname{div}(F) = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} > 0.$$



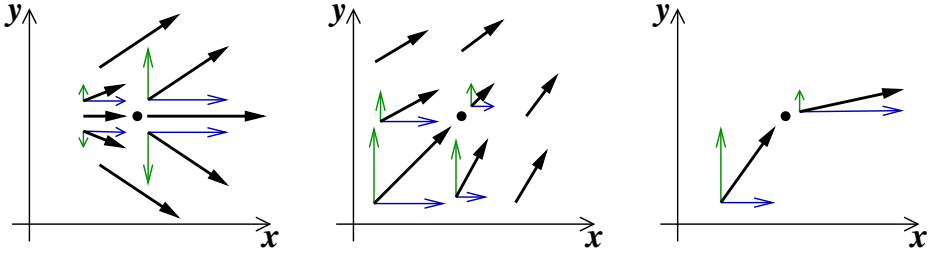
Now in the left picture we see a typical sink, for instance if we have some water on our desk and we decide to get rid of it by drilling a small hole at a point $\vec{a}$. Again, the directions of flow are natural and this time they flip from positive to negative as we move left to right; thus the first component F_1 of the vector function F is decreasing, similarly the second component F_2 decreases

when we move from down up. Conclusion: The two partial derivatives are negative and we have

$$\operatorname{div}(F) = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} < 0.$$

You can find this explanation at many places and if you are happy with it, then the mission is completed, but I was not quite convinced. The argument shows that divergence yields the right sign with a sink and a source, but it does not explain why we have to recognize sinks and sources by just this formula. In both cases the direction of vector F changes abruptly when passing through the point $\vec{a}$, and it is not clear why this has to be checked on using derivative whose purpose is to show a different information (rate of change).

So let's have a look at situations that are less extreme.



In the first picture we see a river, we watch it from above and we also see a place where some factory pours its waste (properly cleaned, of course). This is, therefore, an example of a source. When we check on how the blue first component F_1 of the function F changes as we move left to right, and how the green second component F_2 changes as we move upward, we observe that they both increase, so it sounds plausible that in this case also $\operatorname{div}(F)(\vec{a}) > 0$.

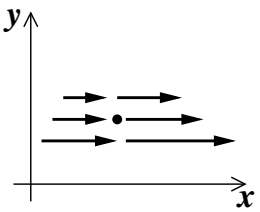
In the second picture we see the river flowing diagonally so that we finally have a proper example, but at the point $\vec{a}$ somebody placed one end of a hyperspace tunnel that sucks off water, but not too much so that something is left (it seems ETs are no longer content with abducting people). Now it seems that the blue component F_1 is getting smaller as we move left to right, just like the green component F_2 when moving upward, so the partial derivatives $\frac{\partial F_1}{\partial x}$, $\frac{\partial F_2}{\partial y}$ should be negative and we get $\operatorname{div}(F)(\vec{a}) < 0$.

In both cases the components were changing in magnitude, not in orientation, so it is beginning to look that partial derivatives are indeed the right tool.

For completeness we look at the third picture. There is no adding or drawing of water happening there, just the river changes direction. We therefore expect that $\operatorname{div}(F) = 0$. In the picture we see that the blue component F_1 increases as we move left to right, while the green component F_2 decreases when moving upward. The appropriate partial derivatives therefore have opposite signs. Could it be that they cancel each other out? We cannot really decide from a mere sketch, this would require a more thorough approach.

These explanations were better, but we still do not know why the formula for $n = 2$ does not feature the other two partial derivatives $\frac{\partial F_1}{\partial y}$ and $\frac{\partial F_2}{\partial x}$. In this context it is interesting to revisit the above examples and notice that for the component F_1, the sign of $\frac{\partial F_1}{\partial y}$ always matches the sign of $\frac{\partial F_1}{\partial x}$, similar observation holds for the component F_2. Perhaps it is not necessary to consider the two partial derivatives because they would provide the same information that we already have. But what should we think about this picture?

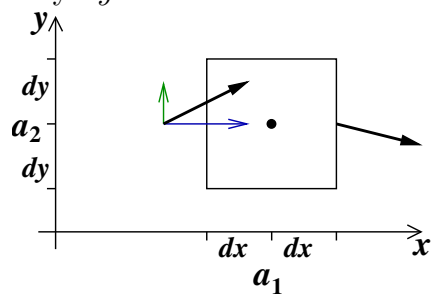
To make our life easier, the component F_2 is always zero, everything happens only in the direction x. Looking left to right we readily observe that $\frac{\partial F_1}{\partial x} > 0$ around the point $\vec{a}$, but we also have $\frac{\partial F_1}{\partial y} < 0$ around $\vec{a}$. What do we do now? Can we really ignore the second observation and simply say that there is a source at the point $\vec{a}$? Or is perhaps such a picture impossible?



My conclusion is that the pictures have shown the divergence as defined to be capable of providing the right information in many cases, which may be enough for many readers. However, if we want

to understand it properly, then we need to take another approach. To this end we will now show an intuitive reasoning, which means that we will cheat, but only moderately and our story will have the right moral at the end.

We take a point $\vec{a}$ and we will investigate how much media it produces. This will determine whether it is a source, a sink or the third possibility when nothing is lost, nothing is gained. To facilitate our calculations we will make a tiny square about this point that exceeds the point by dx to the left and to the right, and by dy in the vertical direction.



We will make an inventory of the medium passing through this square. We start by investigating the horizontal sides. How much media flows through the left side during some time interval of length dt? This depends on the velocity of the flow passing through this side, and this need not be the same everywhere. Fortunately we have a really small square, even infinitely small, so we can assume that the velocity of inflow is the same along the left side. We will use the value in the middle of that side, that is, at $(a_1 - dx, a_2)$. We also notice that the vertical component of this velocity vector does not cause any inflow or outflow of medium through this side, everything is decided by the first component. The inflow velocity is therefore $F_1(a_1 - dx, a_2)$.

The total volume of the medium that entered the square is then this velocity times the size of the side times the time, that is, $F_1(a_1 - dx, a_2)(2dy)dt$. In the same way we figure out the amount of the medium that left the square through the right side, we get $F_1(a_1 + dx, a_2)(2dy)dt$. Subtracting these two gives us the overall balance through horizontal sides; because we want to obtain a positive number when the square adds to the medium, we use the formula

$$F_1(a_1 + dx, a_2)(2dy)dt - F_1(a_1 - dx, a_2)(2dy)dt.$$

However, there are two good questions. First, what if the flow goes right to left? Then the components F_1 are negative and the reader can easily deduce that our formula still provides a positive number when there is more outflow than inflow, which is exactly what we want. Similarly we get the right sign if there was inflow through both sides (the case when $F_1(a_1 - dx, a_2) > 0$ and $F_1(a_1 + dx, a_2) < 0$) or if both sides were leaking medium. Our formula is therefore universal.

The last two cases bring us to the second question that is fundamental: What if some of the medium that enters the left side leaves this square through some of the horizontal sides? The answer is that we do not have to worry, because then it enters into the calculations regarding horizontal sides.

Before we get to it, we will simplify the formula that we obtained. Because dx is very small, we can dare to replace both function values with linear Taylor approximations. Note that in both cases we only change the first coordinate.

$$\begin{aligned} &F_1(a_1 + dx, a_2)(2dy)dt - F_1(a_1 - dx, a_2)(2dy)dt \\ &\quad = \Big(F_1(a_1, a_2) + \frac{\partial F_1}{\partial x}(a_1, a_2) \cdot dx\Big)(2dy)dt - \Big(F_1(a_1, a_2) + \frac{\partial F_1}{\partial x}(a_1, a_2) \cdot (-dx)\Big)(2dy)dt \\ &\quad = \frac{\partial F_1}{\partial x}(\vec{a})(2dx)(2dy)dt. \end{aligned}$$

This reasoning finally made it clear why we do not have to worry about the change of F_1 in the vertical direction represented by $\frac{\partial F_1}{\partial y}$, just in the horizontal one.

Now we follow exactly the same steps to summarize the flow through the horizontal sides. This

time we will not care about the horizontal component of flow, only the vertical component F_2 counts.

$$\begin{aligned} &F_2(a_1, a_2 + dy)(2dx)dt - F_2(a_1, a_2 - dy)(2dx)dt \\ &\quad = \Big(F_2(a_1, a_2) + \frac{\partial F_2}{\partial y}(a_1, a_2) \cdot dy\Big)(2d2)dt - \Big(F_2(a_1, a_2) + \frac{\partial F_2}{\partial y}(a_1, a_2) \cdot (-dy)\Big)(2dx)dt \\ &\quad = \frac{\partial F_2}{\partial y}(\vec{a})(2dy)(2dx)dt. \end{aligned}$$

When we add the two partial balances, we get a number that tells us how much media the square added or took away during the time dt.

$$\begin{aligned} \frac{\partial F_1}{\partial x}(\vec{a})(2dx)(2dy)dt + \frac{\partial F_2}{\partial y}(\vec{a})(2dy)(2dx)dt &= \Big(\frac{\partial F_1}{\partial x}(\vec{a}) + \frac{\partial F_2}{\partial y}(\vec{a})\Big)(2dy)(2dx)dt \\ &= \operatorname{div}(F)(\vec{a}) \cdot dA \cdot dt. \end{aligned}$$

In the last step we replaced the product $(2dx)(2dy)$ with the area of the investigated square. If we wanted the contribution of $\vec{a}$ as such, then we would have to divide by that area dA and time dt and we obtain the divergence F. This calculation has shown that the definition of divergence was done the right way to provide the information that we want to see.

It is worth noting that local inflow or outflow of a medium need not be caused by some outside influence (heating device, water tap, magnet), but it can be also caused by change in the density of the medium. For instance, if we heat a gas at some place, then it starts expanding, effectively becoming a source, but pays for it by lowered density.

On the other hand, if the medium is denser at some place, then it means that more came in compared to the outflow. In this way we derive the following intuitive rule for medium without outside inflow/outflow: $\operatorname{div}(F) > 0$ means expansion, that is, lowering of density, while $\operatorname{div}(F) < 0$ means gathering of the medium, that is, increase in density.

When a perfectly incompressible fluid flows, then there is no density change possible, and if there is no outside influence on the amount of the medium then necessarily $\operatorname{div}(F) = 0$. This condition plays an important role in the mechanics of liquid flows.

Now we look at the second operator that is often used in applications, namely the curl of a vector field (that is, of a vector function). There is a universal definition but it is rather impractical, so everybody works with another formula; unfortunately, this formula has specific versions depending on the dimension.

Definition.
Let $F = (F_1, F_2)$: $D(F) \mapsto \mathbb{R}^2$ be a vector function, where $D(F) \subseteq \mathbb{R}^2$.
We define its **curl** as

$$\operatorname{curl}(F)(\vec{a}) = \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})$$

at points $\vec{a} \in D(F)$ where the appropriate partial derivatives exist.

Let $F = (F_1, F_2, F_3)$: $D(F) \mapsto \mathbb{R}^3$ be a vector function, where $D(F) \subseteq \mathbb{R}^3$.
We define its **curl** as

$$\operatorname{curl}(F)(\vec{a}) = \Big(\frac{\partial F_3}{\partial y}(\vec{a}) - \frac{\partial F_2}{\partial z}(\vec{a}), \frac{\partial F_1}{\partial z}(\vec{a}) - \frac{\partial F_3}{\partial x}(\vec{a}), \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})\Big)$$

at points $\vec{a} \in D(F)$ where the appropriate partial derivatives exist.

The curl is sometimes called the rotation and denoted $\operatorname{rot}(F)$; we will see shortly why.

What information does this operator provide? The standard answer is that the curl describes the tendency of the vector field to curve at the given point. What does it mean?

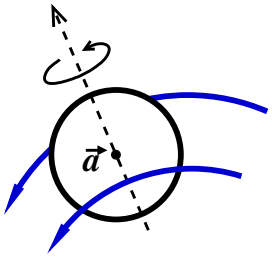
Consider a flowing river. In order to judge its turning effect at a point $\vec{a}$ we place a small (virtual) ball there in such a way that its location is fixed (say, using a magnetic field), but it can rotate freely. If the medium flows in a straight way there, then the friction forces on the ball are the same on all sides from the point of view of this flow (above, below, on the left, on the right), and thus the ball has no reason to rotate. On the other hand, if the medium changes the direction of its flow at $\vec{a}$, then it causes an imbalance of friction forces on different sides of the ball and it begins to rotate. Intuitively, if the flow bends to one side, then on the outer side of this bend the medium has to flow faster, and thus it pushes this side of the ball more.

Now we understand what we are trying to recognize, and we need to address the problem of capturing a specific rotation of a ball mathematically. First, imagine a rotating disc in two dimensions. What data do we need to describe its rotation completely? We need to know the direction of this rotation and also describe its intensity somehow. For that people use either the angular velocity or angular acceleration, depending on the context. Here we are interested in the strength of the influence of the medium on the ball, that is, in the force that acts on the ball, which means that we will want to see the angular acceleration. These two pieces of data, angular acceleration and its direction, can be captured using one number. By a remarkable coincidence, the curl of a two-dimensional vector field is defined as a number, so this fits.

We remind the reader that in physics, a positive direction of rotation is taken to be the counter-clockwise direction when viewed from "above", that is, as if we also added the z-axis and looked down from its tip.

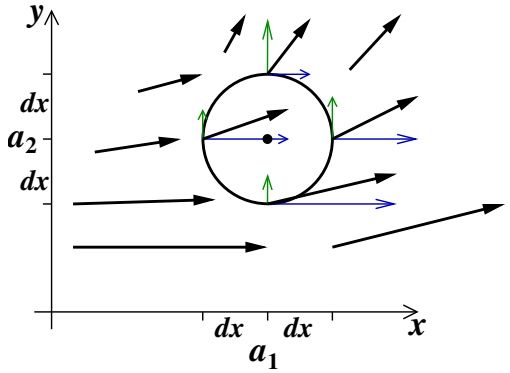
Now we imagine a three-dimensional ball with center at a point $\vec{a}$. How do we capture its rotation? One important piece of information is the axis of its rotation. How do we know that there is such a thing? This is actually simple: if we cannot find two points on opposite ends of the ball that stay in place, then this movement is not a rotation but some other type of movement.

In order to store the axis of rotation, that is, a certain direction, we need a vector. We need to decide on its orientation, and the convention is to orient the vector in such a way that we see the ball rotating in positive direction, that is, counter-clockwise, when looking down from the tip of this vector. It remains to choose the magnitude of this vector, so we use this opportunity to encode the angular acceleration and we have all the information stored in one vector. We check that curl was defined as a vector for a three-dimensional vector field, so things fit together well also in this case.



Now we know what we are looking for, and it is time to ask how come that the formulas from the definition provide the right information.

We start with the case of two dimensions. We put a ball of radius dx at the location $\vec{a}$ inside a two-dimensional vector field, so it is in fact a disc. We will investigate the situation. On the picture the flow bends to the left, that is, in the positive way, so we would expect the disc to rotate in the positive direction. We will determine how strong the forces of the flow are at various places of the disc.



We start with the bottom point. First we recall some basic physics. The intensity of rotational

influence is called the torque, or the moment of the force. This is easiest to calculate when the acting force is perpendicular to the displacement, that is, to the segment connecting the point where the force acts and the center of rotation. Then we just multiply the magnitude of the force with the distance and we get the torque.

When the force is not perpendicular (which is the typical case and we can also see it at the bottom of the disc), then we need to decompose the force into two components. One component is radial, it leads towards or away from the axis of rotation, and does not contribute to the torque. The other component is perpendicular to the displacement and we use it to calculate the torque.

In the picture we can see this decomposition marked in green and blue, and as it happens, these are exactly the components F_2 and F_1 of our vector function; this is the reason why we started with the bottom point. Now we easily find the torque at this point, it is $F_1(a_1, a_2 - dx) \cdot dx$. The sign fits, force acting to the right has positive component F_1 and it pushes the disc to spin in the positive direction.

At the top point we consider only the component F_1 again. However, this time positive value of F_1 leads to rotation in the negative direction, therefore the contribution to torque on the top is $-F_1(a_1, a_2 + dx)dx$. The total contribution of the bottom and top point is therefore

$$F_1(a_1, a_2 - dx)dx - F_1(a_1, a_2 + dx)dx.$$

Does the sign make sense? In our case the force at the bottom is larger than the one on the top, so we expect to see tendency of the disc to spin counter-clockwise. The difference is also positive here in the formula, so this fits. You can draw other situations, with different directions of vectors F, and convince yourself that the difference always gives the right direction of spin.

We simplify this expression using linear approximation and obtain

$$\begin{aligned} &F_1(a_1, a_2 - dx)dx - F_1(a_1, a_2 + dx)dx \\ &\quad = \Big(F_1(a_1, a_2) + \frac{\partial F_1}{\partial y}(a_1, a_2) \cdot (-dx)\Big)dx - \Big(F_1(a_1, a_2) + \frac{\partial F_1}{\partial y}(a_1, a_2) \cdot dx\Big)dx \\ &\quad = -2\frac{\partial F_1}{\partial y}(\vec{a})(dx)^2. \end{aligned}$$

In an analogous way we now derive the total torque contribution of the points on the right and on the left, obtaining

$$\begin{aligned} &F_2(a_1 + dx, a_2)dx - F_2(a_1 - dx, a_2)dx \\ &\quad = \Big([F_2(a_1, a_2) + \frac{\partial F_2}{\partial x}(a_1, a_2) \cdot dx\Big)dx - \Big(F_2(a_1, a_2) + \frac{\partial F_2}{\partial x}(a_1, a_2) \cdot (-dx)\Big)dx \\ &\quad = 2\frac{\partial F_2}{\partial x}(\vec{a})(dx)^2. \end{aligned}$$

The total torque is then

$$-2\frac{\partial F_1}{\partial y}(\vec{a})(dx)^2 + 2\frac{\partial F_2}{\partial x}(\vec{a})(dx)^2 = \mathrm{curl}(F) \cdot 2(dx)^2.$$

To find angular acceleration we need to divide by the momentum of inertia, this causes the term $(dx)^2$ to disappear and we see that $\mathrm{curl}(F)$ really describes the rotational influence of the flowing medium.

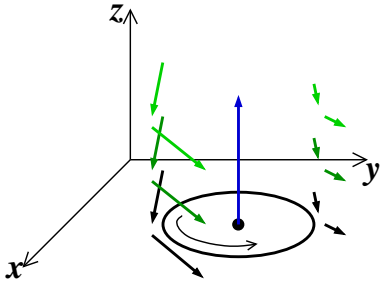
This explanation seems quite convincing, just like the one for divergence. Unfortunately, a closer look reveals some significant gaps. For instance, we have the traditional question here why the derivatives $\frac{\partial F_1}{\partial x}$ and $\frac{\partial F_2}{\partial y}$ are not needed. They did not enter our calculations above, but that was because we looked only at four carefully chosen points. If we looked at any other point, for instance the one in the southeast, then the components F_1 and F_2 would no longer be radial and perpendicular. Instead, we would have to calculate appropriate projections of F, which would be much more complicated and the two missing derivatives would appear in our formulas.

To see that we eventually arrive at the expression from the definition we would have to do the

thorough analysis that physics people do, which is not what we want in this illustrated introduction. Unfortunately, unlike the divergence, I am not aware of any accessible pictorial explanation for the curl that would explain it to complete satisfaction, but the above story gives at least some insight.

Accepting the two-dimensional explanation, we will now build on it to see the meaning of the three-dimensional definition.

We start by returning to our two-dimensional story, but we put it into a three-dimensional setting. We have a table on which something flows in a thin layer, and we put in a small disc on a pin. We see it spinning, and suddenly we realize that we in fact have three dimensions. Formally, we can turn our two-dimensional vector function $F(x,y) = (F_1(x,y), F_2(x,y))$ into a three-dimensional function $F(x,y,z) = (F_1(x,y), F_2(x,y), 0)$, where we add a third variable into the function F but it is not used. This sets up the situation when the things happening on the table are extended upward and downward uniformly, and the disc becomes a vertical cylinder.



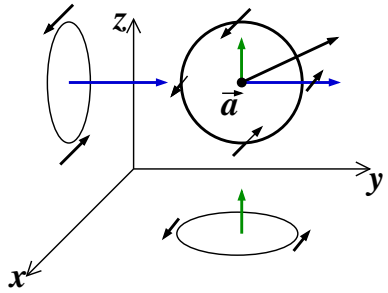
The rotation of the disc (that is, the cylinder) is still happening in the xy-plane, so the axis of rotation is parallel with the z-axis. In our picture the disc is spinning in the direction from the x-axis toward the y-axis, so it is the positive direction from the point of view of z. The vector of rotation must therefore point up. Its magnitude is determined by the torque, which we calculated as a two-dimensional curl. The vector we want is therefore

$$\big(0, 0, \operatorname{curl}(F)\big) = \Big(0, 0, \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})\Big).$$

You may notice that the third coordinate is exactly the same as in the definition of the three-dimensional curl.

The important thing is that the step from two to three dimensions can be also reversed. If we have a ball spinning in three dimensions, we can cut this situation with a plane passing through the center of this ball, preferably one parallel with some coordinate axes. Then the physical processes will be preserved, in particular, the the torque that we find by analyzing the slice corresponds to the appropriate component of the three-dimensional vector of angular acceleration.

In the next picture we have a ball spinning in three dimensions, propelled by a vector field $F(x,y,z) = (F_1, F_2, F_3)$. To simplify our situation, imagine that the axis of rotation and also the vector of angular acceleration are parallel to the yz-plane, so we do not have to worry about the x-component. This means that the ball spins as if it wanted to roll towards us out of the picture. This could be caused by water spinning around the ball; above the ball it flows towards us while below the ball it flows away from us, which we attempted to suggest with arrows.



In front of the ball the water flows down, but not quite because then the axis of rotation would point horizontally to the right. It is actually tilted up a bit, so the water in front of the ball flows down and also a bit to the right. This then means that the flow also has a component to the left and to the right of the ball, but weaker compared to above and below the ball.

If we slice this situation with a plane through $\vec{a} = (a_1, a_2, a_3)$ parallel with the xy-plane, we obtain a two-dimensional situation whose picture was, for the sake of clarity, shifted down to the actual xy-plane. We also carried over some typical forces that are in the original picture visible to

the left and to the right of the ball. In this slice we only see how the ball turns around its vertical axis, which explains why we were only interested in forces on the left and on the right, while the forces above and below the ball do not contribute to this rotation visible in the plane xy.

In fact on this slice we work with the vector function

$$F(x,y) = (F_1(x,y,a_3), F_2(x,y,a_3)).$$

We can calculate its torque with respect to vertical axis exactly as we did above, obtaining

$$\Big(0, 0, \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})\Big).$$

We see this vector in green and its magnitude exactly fits the third component of three-dimensional curl.

If we slice this situation with the plane parallel with the xz-plane, (again we move the picture of this slice to the actual xz-plane, including some typical vectors above and below the ball), then we in fact work with the auxiliary function

$$F(x,z) = (F_1(x,a_2,z), F_3(x,a_2,z)).$$

Playing with the two-dimensional situation we arrive at the blue vector pointing in the y-direction, whose magnitude is the torque $\frac{\partial F_3}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial z}(\vec{a})$. However, here we have to be careful. This torque goes from the first working axis to the second. Previously this meant from x to y, which is the positive rotation. However, now it goes from x to z, which is from the point of view of the y-axis a negative rotation. In order to obtain a positive rotation we have to change sign. We therefore take the torque around the y-axis as $\frac{\partial F_1}{\partial z}(\vec{a}) - \frac{\partial F_3}{\partial x}(\vec{a})$. We get the vector

$$\Big(0, \frac{\partial F_1}{\partial z}(\vec{a}) - \frac{\partial F_3}{\partial x}(\vec{a}), 0\Big).$$

If we add the green and blue vector, then taking into account our special situation where the x-component does not play role, we should obtain the actual torque vector of the original three-dimensional situation. In a general case we would also slice the ball with the plane parallel to the yz-plane. The auxiliary function

$$F(y,z) = (F_2(a_1,y,z), F_3(a_1,y,z))$$

would yield the torque $\frac{\partial F_3}{\partial y}(\vec{a}) - \frac{\partial F_2}{\partial z}(\vec{a})$. Does it have the right sign? Positive means rotation from the y-axis to z, which is positive from the point of view of x, this fits. So this number becomes the first component of the torque vector pointing in the x-direction.

In this way we get vectors

$$\Big(0, 0, \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})\Big)$$
$$\Big(0, \frac{\partial F_1}{\partial z}(\vec{a}) - \frac{\partial F_3}{\partial x}(\vec{a}), 0\Big)$$
$$\Big(\frac{\partial F_3}{\partial y}(\vec{a}) - \frac{\partial F_2}{\partial z}(\vec{a}), 0, 0\Big).$$

When we compose them into one vector, we get exactly the formula that we have in the definition of the three-dimensional curl:

$$\Big(\frac{\partial F_3}{\partial y}(\vec{a}) - \frac{\partial F_2}{\partial z}(\vec{a}), \frac{\partial F_1}{\partial z}(\vec{a}) - \frac{\partial F_3}{\partial x}(\vec{a}), \frac{\partial F_2}{\partial x}(\vec{a}) - \frac{\partial F_1}{\partial y}(\vec{a})\Big).$$

So much for the meaning of the curl of a vector field.

For some identities valid for differential operators see section 7a. Integral theorems featuring divergence and curl are explained in section 6c.

6. Introduction to integrals

We start by returning to definite integral for functions of one variable. The expression $\int_a^b f(x)\,dx$ is understood as the "area under the graph". There is an interpretation that is not formally correct, but it conveys the right idea: We somehow "add" values of the function at points x, and we add them as areas of very thin columns or rectangles. The key notion here is dx. Formally we can interpret it as the differential of variable x, which some authors emphasize by writing $\mathrm{d}x$.

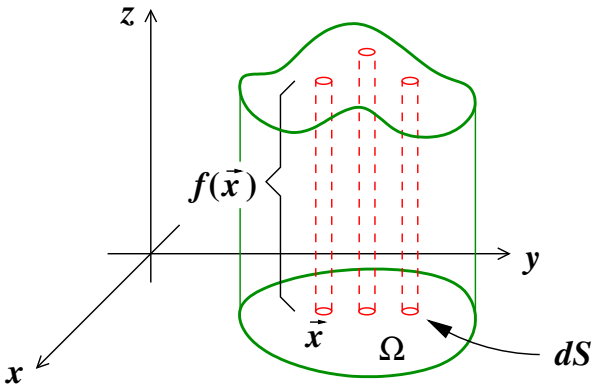
However, here we will use the the original idea; in previous centuries, dx has been understood as an infinitely small piece of the x-axis, also called an infinitesimal. We are definitely leaving the world of rigorous mathematics here. To make things even more obscure, the symbol dx is also used to denote the length of such a small piece. Fortunately, the appropriate interpretation of dx is usually clear from the context.

The integration over an interval I is usually visualized as a process where we divide the interval I into really tiny parts of length dx, and then approximate the region under the graph as a union of thin rectangles. Each of them has area $f(x)\cdot dx$ and we add them using integration. Here we will make use of a slightly different story. When integrating, we go through all points of the interval I. For each such point x we include the contribution of the appropriate function value $f(x)$ by taking that infinitely small piece of the x-axis of length dx centred around x and erect a column above it of height $f(x)$. Its area $f(x)\cdot dx$ is then added to the total.

Now we extend this idea to functions of more variables. When we have some set Ω in $\mathbb{R}^n$ and a function f on it, then we also can "add" values over Ω. In the space $\mathbb{R}^n$ we have infinitesimals that we will denote $\mathrm{d}[\vec{x}]$ if the variable is $\vec{x}$, but typically we use specific notation depending on dimension. In $\mathbb{R}^2$ we have planar infinitesimals, we can imagine infinitely small rectangles or circles, and we denote them dS. We also write $dS[\vec{x}]$, $dS[x,y]$ and so on when we want to emphasize the variables. These infinitesimals have some area that we will call dS.

Similarly, in $\mathbb{R}^3$ we have 3D infinitesimals, we can think of cubes or balls and denote them dV or $dV[x,y,z]$ etc. Again, we also use dV for their volume. In general we use n-dimensional volume for infinitesimals $\mathrm{d}[\vec{x}]$ in $\mathbb{R}^n$.

We can easily imagine the situation in $\mathbb{R}^2$, that is, when integrating a function of two variables. Following the inspiration from one variable, we want to find the volume of the solid whose base is Ω and its top part coincides with the graph of the function f. We find it as follows. For every point (x,y) in Ω we take a suitable infinitesimal around it, we can imagine a small disc. This will be the base for a vertical column (in this case a cylinder) of height $f(x,y)$. Its volume is $f(x,y)\cdot dS$. When we "add" these volumes, we obtain the integral of f over Ω, denoted $\iint_\Omega f(x,y)\,dS$.



For a function of three variables and a set Ω we would, for each $\vec{x}\in\Omega$, take a 3D-infinitesimal dV and erect a four-dimensional "column" of height $f(\vec{x})$ over it. By adding all corresponding

4D-volumes $f(\vec{x}) \cdot dV$ we get $\iiint\limits_{\Omega} f(\vec{x})\, dV$, which is the four-dimensional volume of the object that is situated between the base Ω and the graph of f, which is where our imagination gives up. But the idea is hopefully clear.

In this way we arrive at the general notion of an integral $\int\limits_{\Omega}\cdots\int f(\vec{x})\mathrm{d}[\vec{x}]$. It should be noted that the actual shape of infinitesimals is not crucial, they just have to be simple enough so that we can determine their area, volume etc. However, the ability to decompose a given set Ω into suitable elementary parts is not guaranteed. When integrating a function of one variable over an interval, we only had to worry about the quality of the function (some are integrable, for instance the continuous ones, some are not). Once we pass to more variables, the ability to integrate depends on both the function and the set Ω being reasonable. Precise statements (especially specification of reasonability for sets) are complicated and definitely beyond an illustrated introduction, more information can be found in chapter 8. As we will soon see, we normally integrate only over sets of specific types.

What happens if we integrate the constant function $f = 1$ over some set? Then all those "columns" have height 1, so their $n+1$-dimensional volumes are equal to the n-dimensional volumes of the bases $\mathrm{d}[\vec{x}]$. When we "add" them through integration, we in fact add volumes of the little bases into which we split the set Ω, and therefore the integral yields the n-dimensional volume of Ω. This is obvious in one dimension where "volume" (meaning size) corresponds to length. The integral $\int\limits_a^b 1\, dx$ gives $b-a$, that is, the length of the segment $[a,b]$. In two dimensions the integral provides the area, in three dimensions the volume and things work analogously for higher dimension.

This can be handy, for instance when we want to find the average of some function with respect to some set Ω. The formula for a function of one variable is well-known and we easily generalize it to functions of more variables:

• If a function $f(\vec{x})$ can be integrated over a set Ω, then the average is given by the formula

$$\operatorname*{Ave}_{\Omega}(f) = \frac{\int\limits_{\Omega}\cdots\int f(\vec{x})\,\mathrm{d}[\vec{x}]}{\int\limits_{\Omega}\cdots\int 1\,\mathrm{d}[\vec{x}]}.$$

The main idea of a more-dimensional integral seems clear, now we will address the problem of evaluating such integrals.

Before we start, we have to clarify terminology. The standard meaning of the term "region" in analysis is that it denotes a bounded open connected set. Unfortunately, when it comes to integration, people use "region" for the set over which we integrate, regardless of its properties. In fact, integral is often understood to be done over closed sets. To avoid this issue I tried to talk chiefly about sets, but I was not totally successful as there were occasions where I felt that the context just calls for the word region and I did not fight the urge. Fortunately, it usually does not matter whether we integrate over a set, its closure or its interior, so we do not have to worry about it too much.

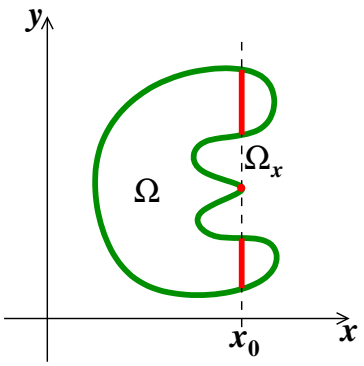
6a. Two-dimensional integral

Integral in which we integrate a function of two variables over a set $\Omega \subseteq \mathbb{R}^2$ is called the **double integral**,

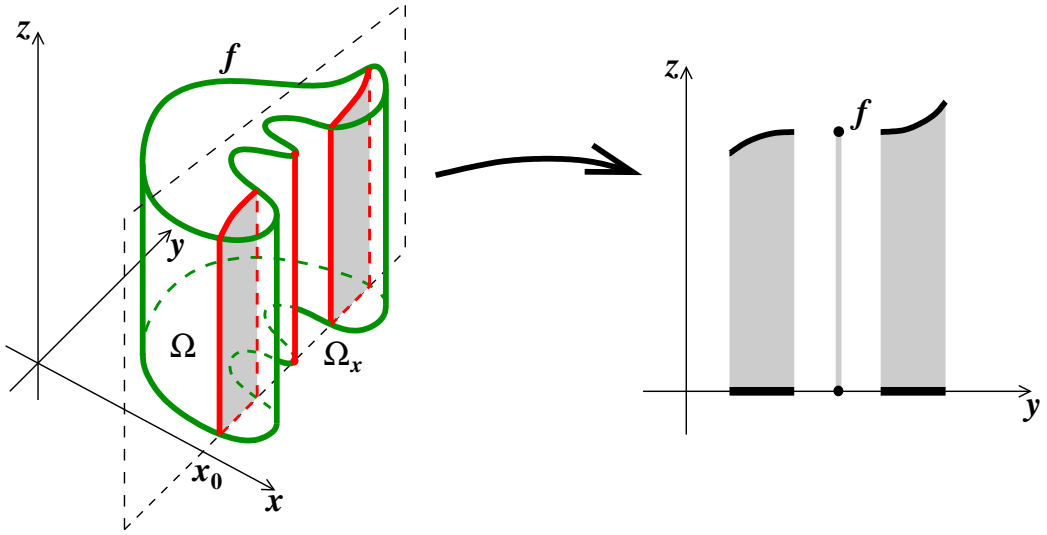
$$\iint_\Omega f(x,y)dS.$$

We want to develop a procedure for evaluating it. We have to sum up volumes of "columns" over all points of the set Ω. The key idea of multi-dimensional integration is that in order to account for all points in Ω, we organize them into smaller groups and handle these one-by-one. The trick is to create those groups in such a way that they have lower dimension.

How do we do it? We choose a basic direction, say, along the y-axis, which in the xy plane means vertical direction. A specific line in this direction is determined by fixing some specific value $x = x_0$ for the variable x. Assume that this line intersects the set Ω, call this intersection Ω_{x_0}. We will call this set a slice, and depending on the shape of Ω, these slices could be rather wild. This will later force us to restrict our attention only to sets of reasonable shape.



The equation $x = x_0$ can be also considered in $\mathbb{R}^3$, where it is the equation of a plane perpendicular to the x-axis, that is, parallel to the yz-plane. This plane intersects the solid under the graph of f and creates a slice of it.



Note that now we have two kinds of slices, slices Ω_x through Ω and also slices through the solid under the graph. Hopefully we will always make the distinction clear. Fortunately we only work with Ω when it comes to actual calculations, so we will not have to worry for long.

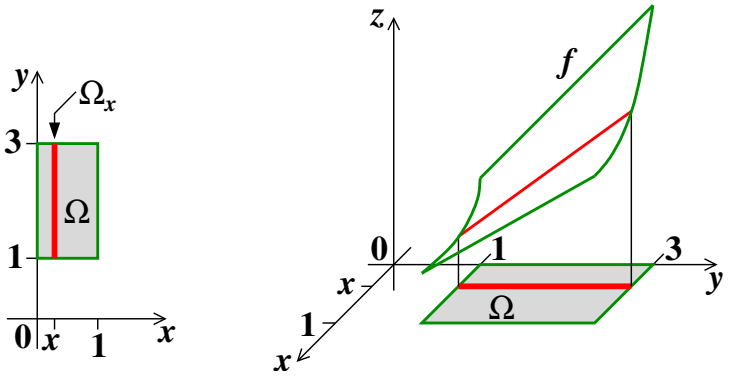
When we look at this slice through the solid, we actually see the graph of the function of one variable $y \mapsto f(x_0, y)$, where y is taken from points of the slice Ω_{x_0}. If the set Ω_{x_0} is reasonable (preferably an interval), then we can integrate the function $f(x_0, y)$ there (with respect to y, obviously) and obtain the area of that slice through the solid under the graph. Note that this is a definite integration with y, so the variable y will disappear in the process, leaving only x, which is to be expected as the areas of slices will most likely differ depending on the choice of x_0.

Example: Consider the function $f(x,y) = y\,e^{-x}$ on the rectangle $[0,1] \times [1,3]$, that is, on the set

$$\Omega = \{(x,y) \in \mathbb{R}^2;\ 0 \le x \le 1 \text{ and } 1 \le y \le 3\}.$$

We choose $x_0 = 0.3$. This determines a slice $\Omega_{0.3}$ through the set Ω, it is a segment. We show a picture. Usually we work with a picture of the set Ω, but here to add some context we make an

exception and also show a three-dimensional picture with the graph of f.



On this slice values of y change between 1 and 3, and for these we have the function

$$y \mapsto f(x_0, y) = f(0.3, y) = y\, e^{-0.3}$$

that we can integrate.

When integrating, the term $e^{-0.3}$ is a constant, hence we easily obtain

$$\int_1^3 y\, e^{-0.3}\, dy = e^{-0.3} \int_1^3 y\, dy = e^{-0.3} \left[\frac{1}{2}y^2\right]_1^3 = e^{-0.3}\left[\frac{9}{2} - \frac{1}{2}\right] = 4e^{-0.3}.$$

This is the area of the corresponding slice of the solid under the graph. As expected, it does not feature the variable y. The outcome depends on the choice of x_0, for instance for $x_0 = 0$ we would get

$$\int_1^3 y\, e^0\, dy = e^0 \int_1^3 y\, dy = 1 \cdot \left[\frac{1}{2}y^2\right]_1^3 = 4.$$

It is useful to observe that for a general x between 0 and 1 the appropriate integral evaluates to

$$\int_1^3 y\, e^{-x}\, dy = e^{-x} \int_1^3 y\, dy = e^{-x} \left[\frac{1}{2}y^2\right]_1^3 = e^{-x}\left[\frac{9}{2} - \frac{1}{2}\right] = 4e^{-x}.$$

It is a bit like partial integration. We thus obtain information about areas of all slices.

It is not always possible to factor out the variable we do not use. This should not be a problem, we just pretend that some terms are constant and work with them as usual. Thus we could have evaluated our integral also in this way,

$$\int_1^3 y\, e^{-x}\, dy = \left[\frac{1}{2}y^2\, e^{-x}\right]_{y=1}^{y=3} = \frac{9}{2}e^{-x} - \frac{1}{2}e^{-x} = 4e^{-x}.$$

Since there are two variables in the antiderivative, to be on the safe side we reminded ourselves where to substitute the limits.

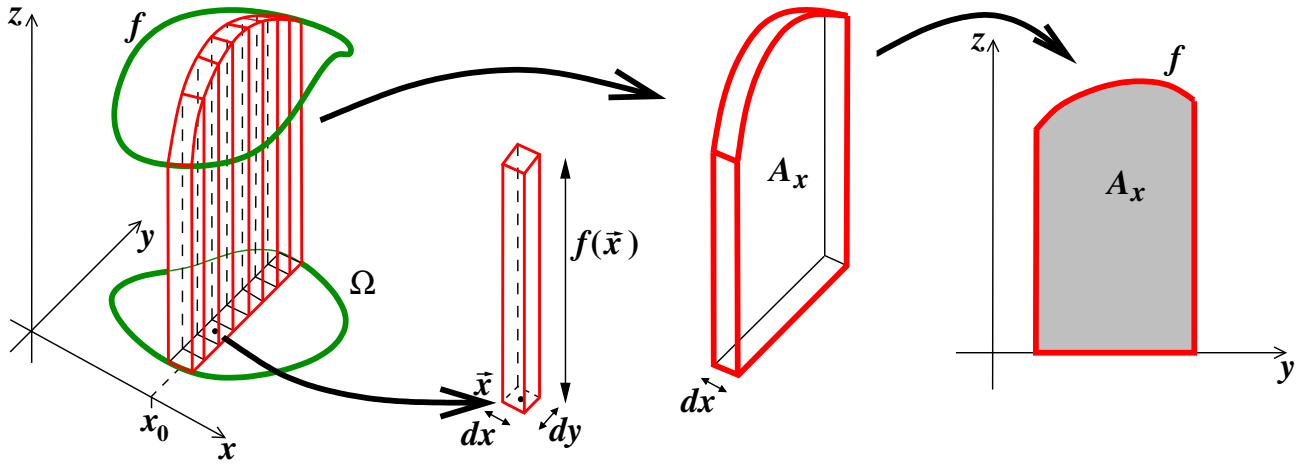
△

So it does make sense to find the areas of slices through a solid by integration. We introduce the notation

$$\int_{\Omega_x} f(x, y)\, dy$$

for such integrals. It is just symbolic; we cannot take it literally because for that Ω_x would have to be some set in $\mathbb{R}$. But Ω_{x_0} is actually a subset of $\mathbb{R}^2$ whose points all share the common first coordinate x_0. We understand this integral notation as follows: We substitute that x_0 into the function f for x and then integrate over all y such that points (x_0, y) are in Ω_x.

What is the use of such integration over slices Ω_x? Recall that our eventual aim is to add contributions $f(x,y)\cdot dS$ of all points from the set Ω, but right now we are thinking of taking only contributions from points (x,y) on a specific slice Ω_{x_0} of Ω. It will be helpful to assume that dS are infinitely small squares $dx\times dy$. When integrating over the slice we are actually joining volumes of vertical narrow boxes that correspond to points of Ω_{x_0}. They stand next to each other, they share common width dx, and therefore they create a sort of board (or a slab) with an irregular upper edge (which is given by the function f). The volume of this board is obtained by multiplying the area of its side by the thickness dx.



We actually calculated area of such a slice in the example above. The area of a particular slice obviously depends on the choice of x, so it makes sense to denote it A_x, and we saw exactly this in that example. Common sense tells us that the volume of the solid is the sum of volumes of the slabs, that is, we want to "add" volumes $dV = A_x\cdot dx$. Infinitesimals are added using integration, so the volume of the solid (that is, the double integral) can be rewritten as

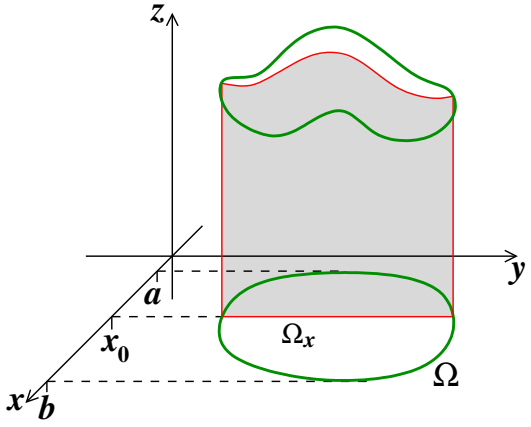
$$\iint_\Omega f\,dS = \int A_x\,dx.$$

This points to a useful informal principle:

- The volume of a solid can be obtained by summing up the areas of all its parallel slices.

This is true also in other dimensions, for instance the area of a planar set can be found by adding lengths of its parallel slices, or the four-dimensional volume of a four-dimensional set can be obtained by adding the volumes of its slices, which are three-dimensional solids.

We need to add limits for the integral on the right. If the set Ω is bounded, then majority of its slices Ω_x will be empty. Then also the areas A_x of the slices through the solid are zero and there is no need to include them in the integral. Thus it should be possible to find some range $[a,b]$ for the variable x so that Ω_x are non-empty only for x from this range; then it makes sense to integrate there.



Recalling that the area A_x can be found as an integral over a slice

$$A_x = \int_{\Omega_x} f(x,y)\,dy$$

we arrive at a key identity.

$$\iint_\Omega f(x,y)\,dS = \int_a^b \int_{\Omega_x} f(x,y)\,dy\,dx.$$

The integral on the right is called the **repeated integral**. The process of finding some repeated integral for a given double integral is called **setting up a repeated integral**.

How is such an integral evaluated? It is actually an integral from an integral, and they are "nested", which means that we interpret it from the outside towards the inside: the integrals at the beginning are taken left to right, while differentials at the end are taken right to left. This could be emphasized by parentheses, but usually people are too lazy to do it.

$$\iint_\Omega f(x,y)\,dS = \int_a^b \Big(\int_{\Omega_x} f(x,y)\,dy \Big)\,dx.$$

We see the **outer integral** that works with variable x and is related to integration limits a, b. It is applied to the expression inbetween, which is another integral. Obviously it makes no sense to try this integration with x until we figure out what is actually being integrated. Accordingly, we always start evaluation of a repeated integral by evaluating the **inner integral** $\displaystyle\int_{\Omega_x} f(x,y)\,dy$.

As we saw, the outcome of such integral is some expression featuring x, and we thus have a chance to evaluate the outer integral $\displaystyle\int_a^b \cdots\,dx$.

The three-dimensional pictures with graph of f helped us to understand the motivation behind this procedure, but for practical calculations it is better to work just with Ω. We read the above formula as follows: The integration over Ω can be replaced by integration over slices Ω_x (of lower dimension) and the outer integral makes sure that we take all slices that contribute.

This reduction of dimension is the key principle for evaluation of multi-dimensional integrals:

• We change integration over a set Ω into integration over its slices that have lower dimension (by one). If this new dimension is still too high, we repeat the slicing process.

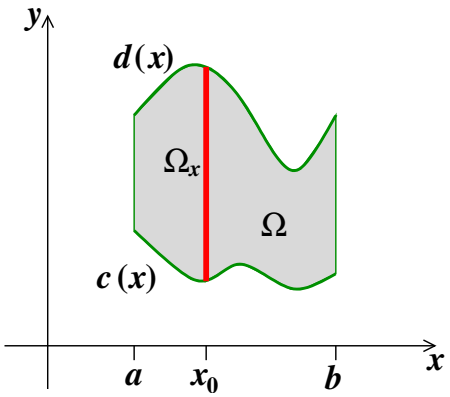
This general principle suggests that we could slice also in different directions that the one we have been using here. Oblique slices are possible in principle, but hard to handle. However, there is a natural alternative that is as pleasant as slicing by fixing x. We can create slices parallel to the x-axis, these are created by choosing a value for the variable y. If we denote the corresponding slice through the set Ω as Ω_y, we get the formula

$$\iint_\Omega f(x,y)\,dS = \int_c^d \int_{\Omega_y} f(x,y)\,dx\,dy.$$

This is also evaluated from the inside, so this time we start by integrating with respect to x.

We will now turn these abstract ideas to actual calculations, which should help the reader in digesting them. This means that we will have to start worrying about the shape of the integrating domain Ω. If Ω is too wild, then also the slices Ω_x or Ω_y could be wild and it can easily happen that integration fails. Therefore we restrict our calculations to integral domains of reasonable shape (type).

We start with the situation when we slice in direction y, so a particular slice is chosen by fixing $x = x_0$. In an ideal case the slice Ω_x through the set Ω has the form of a segment, which means that in order to go through all points of Ω_x we let values of y change between two specific numbers. These numbers, which are integral limits for the inner integral with respect to y, will in general depend on the selected slice, that is, on x. Let's denote them $c(x)$ and $d(x)$. In other words, we are interested in the case when there are some functions $c(x)$ and $d(x)$ such that the slice corresponding to some x is the segment $[c(x), d(x)]$. Sets Ω that work this way have a specific shape: These are regions squeezed between graphs of two functions on an interval:



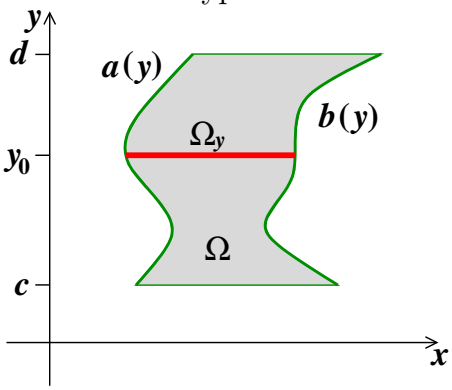
Some people call them sets of type I. Mathematically, we are interested in sets that can be expressed as follows:

$$\Omega = \{(x, y) \in \mathbb{R}^2;\ a \le x \le b \ \text{ and } \ c(x) \le y \le d(x)\}.$$

For such sets, integration over slices Ω_x is straightforward and we get the following repeated integral:

$$\iint_{\Omega} f(x, y)\, dS = \int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx.$$

The other direction of slicing works with fixing y and slices Ω_y that set up the integral domain for the inside integral with respect to the variable x. Here we again want the slices Ω_y to be intervals with reasonable endpoints, which leads to the type of set that some people call sets of type II.



Formally, these are sets of the form

$$\Omega = \{(x, y) \in \mathbb{R}^2;\ c \le y \le d \ \text{ and } \ a(y) \le x \le b(y)\}.$$

Integration over such sets can be rewritten as the following repeated integral:

$$\iint_{\Omega} f(x, y)\, dS = \int_c^d \int_{a(y)}^{b(y)} f(x, y)\, dx\, dy.$$

Note one common feature in both reworkings of the integral, a feature that is valid in general. When a repeated integral is set up properly, the following must be true (among other things): The outer integral has only numbers for its limits. The inner integral can take numbers or expressions with a variable for its limits, but this variable must be more to the outside in the list of differentials, that is, the one that will be integrated with later.

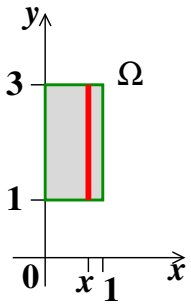
Some sets fit both specifications, then we can choose in which direction to cut. In that case we usually decide for direction that offers more pleasant integration.

Example: We will integrate the function $f(x,y) = y\,e^{-x}$ over the rectangle $[0,1]\times[1,3]$, that is, over the set

$$\Omega = \{(x,y)\in\mathbb{R}^2;\ 0\le x\le 1 \text{ and } 1\le y\le 3\}.$$

We have a choice which variable to use for integration first, because in a rectangle, slicing in both directions works equally well.

We decide to fix x and first integrate with y (that is, create slices parallel to the y-axis), because we prepared ground for it in the previous example.



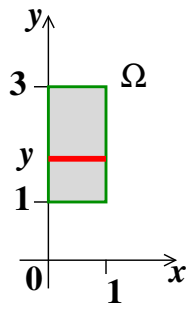
We have already evaluated the integral over one slice as

$$\int_1^3 y\,e^{-x}\,dy = \Big[e^{-x}\frac{1}{2}y^2\Big]_{y=1}^{y=3} = 4e^{-x}.$$

Now we "add" these integrals over slices using x, we get

$$\iint_\Omega y\,e^{-x}\,dS = \int_0^1\int_1^3 y\,e^{-x}\,dy\,dx = \int_0^1 4e^{-x}\,dx = \Big[-4e^{-x}\Big]_0^1 = 4-4e^{-1}.$$

Now we try the other possible slicing. How do horizontal slices, that is, in the direction of the x-axis look like?



Slices are non-empty pro y between 1 and 3, so we get the decomposition

$$\iint_\Omega y\,e^{-x}\,dS = \int_1^3\int_{\Omega_y} y\,e^{-x}\,dx\,dy.$$

We can deduce from the picture that on every slice, x changes between 0 and 1. This determines the inner integral.

$$\iint_\Omega y\,e^{-x}\,dS = \int_1^3\int_0^1 y\,e^{-x}\,dx\,dy.$$

We evaluate from inside:

$$\iint_\Omega y\,e^{-x}\,dS = \int_1^3\int_0^1 y\,e^{-x}\,dx\,dy = \int_1^3\Big[-y\,e^{-x}\Big]_{x=0}^{x=1}dy = \int_1^3 y-e^{-1}y\,dy$$

$$= \int_1^3 (1-e^{-1})y\,dy = \Big[(1-e^{-1})\tfrac{1}{2}y^2\Big]_1^3 = 4(1-e^{-1}).$$

When substituting into the inner integral we reminded ourselves what the working variable was at the time, that is, where should we substitute the integrating limits.

△

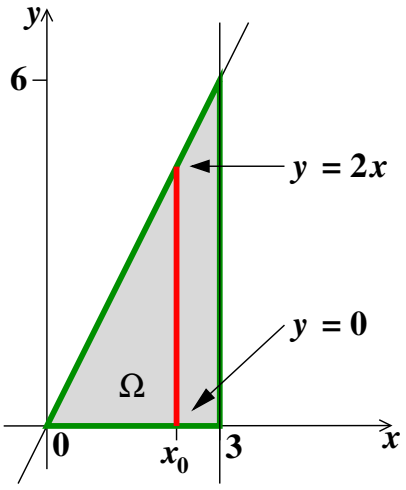
Compare the two repeated integrals that we created for the given double integral:

$$\iint_\Omega y\,e^{-x}\,dS = \int_0^1\int_1^3 y\,e^{-x}\,dy\,dx = \int_1^3\int_0^1 y\,e^{-x}\,dx\,dy.$$

It looks as if we just switched the integrals and the corresponding differentials. However, it is this simple only when integrating over rectangles, our favourite integrating domains. In other cases, the so-called "change of order of integration" is more complicated.

Example: We will integrate the function $f(x,y) = e^{x^2}$ over the bounded region Ω determined by the curves $y = 2x$, $x = 3$, and $y = 0$.

It always pays off to draw a picture. The given lines split the plane into seven regions, but only one of them is bounded: the triangle with vertices $(0,0)$, $(3,0)$, and $(3,6)$.



Vertical slicing (in direction of the y-axis) will surely work. We see that we obtain a meaningful slice only by choosing x between 0 and 3. The resulting repeated integral will therefore have the outside integral $\int_0^3 \cdots dx$.

On one particular slice, the variable y moves between the values $y = 0$ and $y = 2x$. After all, this corresponds to the formal description of the set in the form

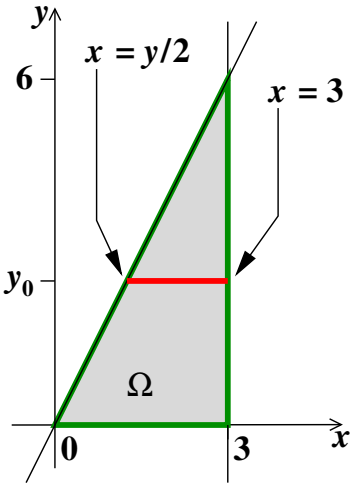
$$\Omega = \{(x,y) \in \mathbb{R}^2;\ 0 \le x \le 3 \text{ and } 1 \le y \le 2x\}.$$

This settled the shape of the inside integral over a typical slice. We can evaluate.

$$\iint_\Omega e^{x^2}\,dS = \int_0^3 \int_0^{2x} e^{x^2}\,dy\,dx = \int_0^3 \Big[ye^{x^2}\Big]_{y=0}^{y=2x} dx = \int_0^3 2x\,e^{x^2}\,dx$$

$$= \left| \begin{array}{c} w = x^2 \\ dw = 2x\,dx \\ x=0 \implies w=0 \\ x=3 \implies w=9 \end{array} \right| = \int_0^9 e^w\,dw = \Big[e^w\Big]_0^9 = e^9 - 1.$$

Now it is time to try horizontal slices.



Location of a particular slice is determined by choosing y from the range between 0 through 6, this makes it clear how the outside integral will go. The slice (on which x is the working variable) is a segment with its right end at the level $x = 3$, the left end lies on the curve given by the formula $y = 2x$ and we need to know about x. We therefore see that when moving along such a slice, x goes from $\frac{y}{2}$ to 3. We obtain

$$\iint_\Omega e^{x^2}\,dS = \int_0^6 \int_{y/2}^3 e^{x^2}\,dx\,dy.$$

We start with the inside integral $\int_{y/2}^3 e^{x^2}\,dx$. And we run into trouble right away, the antiderivative to e^{x^2} cannot be expressed by a closed algebraic formula, so this is the end of the road. △

We see that our choice of the slicing direction, that is, our choice of the order of integration can have a big influence on the evaluation that follows. In less extreme cases it may influence the level of complexity of our evaluation.

Although one of our attempts did not work out, it was still a useful exercise. Both ways of rewriting the integral were entirely correct and show that when we change the order of integration, the integrating limits may also change.

$$\iint_\Omega e^{x^2}\,dS = \int_0^3 \int_0^{2x} e^{x^2}\,dy\,dx = \int_0^6 \int_{y/2}^3 e^{x^2}\,dx\,dy.$$

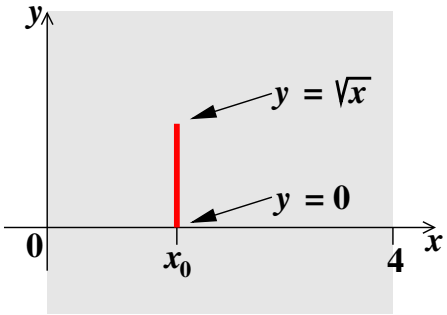
This is typical: When changing the order, we have to rework the limits for integrals. Sometimes we are given a repeated integral and we are asked to change the order of integration (one reason

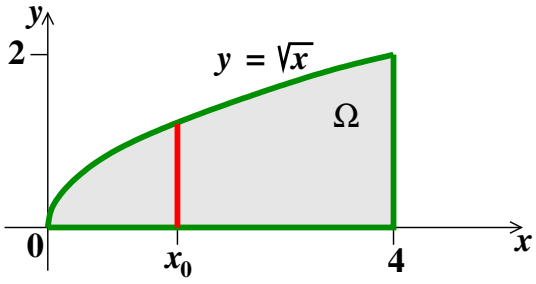
may be that we are not able to evaluate the integral as given). In such a case we first need to reconstruct the shape of the integrating domain Ω using the known integrating limits (some sort of a reverse engineering), then we apply slices in the other direction to this domain.

Example: Change the order of integration in the integral

$$\int_0^4 \int_0^{\sqrt{x}} f \, dy \, dx.$$

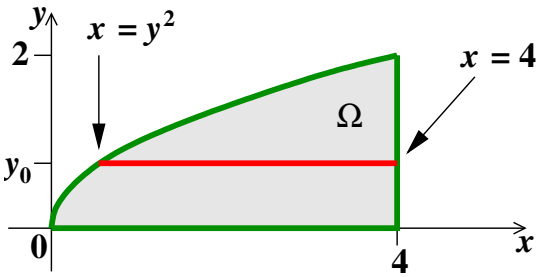
First we need to identify the domain of the integral Ω. The outer integral shows limits for x, which tells us two things: We take slices perpendicular to the x-axis, that is, vertical slices, and they are only relevant for x between 0 and 4. This means that the set Ω must lie within the infinite vertical strip determined by $0 \le x \le 4$.





Now we take some $x \in [0, 4]$ and look at the range of y allowed for the corresponding slice. We see that the values start at $y = 0$ and stop at $y = \sqrt{x}$. In other words, the lower edge of the set is given by the formula $y = 0$ and the upper edge is given by $y = \sqrt{x}$.

We are ready for slicing in perpendicular direction. We take a specific horizontal slice by fixing y from $[0, 2]$. On this particular slice, the leftmost value of x is given by $y = \sqrt{x}$, so x changes between $x = y^2$ and 4. We can write the desired integral.



$$\int_0^4 \int_0^{\sqrt{x}} f \, dy \, dx = \int_0^2 \int_{y^2}^4 f \, dx \, dy.$$

△

Note that such a straightforward change is not always possible, because with some sets, slicing in a certain direction does not lead to one resulting integral. We will see this in the next example.

An inquisitive reader has surely already thought of the fact that not all sets Ω fall into the two types we covered. In such a case it is usually possible to split this set into subsets that are of the right type, then we set up an integral for each of them. This is possible thanks to the additivity of integral with respect to integrating domain. The reader knows the formula

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

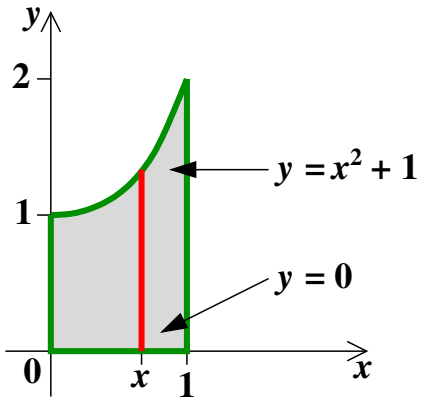
which for $a < b < c$ says that we can divide the integrating interval into two parts. These should not overlap, on the other hand it is definitely not true that intervals $[a, b]$ and $[b, c]$ would have an empty intersection. We need to allow two sets to share some of their boundary, which can be done mathematically by relaxing the condition on disjointness to not involve boundaries. We get the following principle.

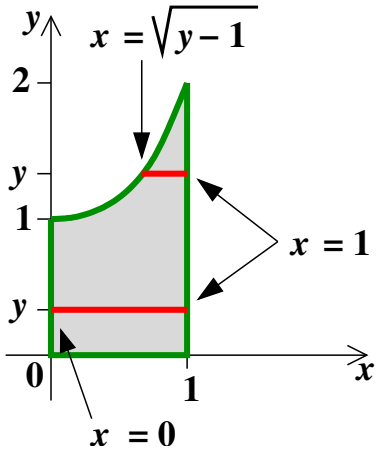
• Consider a set Ω that we split into subsets $\Omega_1, \dots, \Omega_m$ so that $\Omega = \Omega_1 \cup \cdots \cup \Omega_m$ and the interiors of sets Ω_j, Ω_k are disjoint for any $j \neq k$. Then

$$\iint_{\Omega} f(\vec{x})\,\mathrm{d}[\vec{x}] = \sum_{k=1}^{m} \iint_{\Omega_k} f(\vec{x})\,\mathrm{d}[\vec{x}].$$

Example: Let Ω be the region between $y = x^2 + 1$ and the x-axis above $[0, 1]$. We want to find $\iint_{\Omega} 8x(x^2 - y + 1)^3\,dS$.

First we draw Ω and analyze our slicing options.





The shape of this set just calls for vertical slicing, where we first integrate with respect to y. Indeed, we see that if we tried slicing in the x-direction, then we would not get a universal formula for left endpoints of slices.

We therefore prefer vertical slices determined by choosing x, obviously from the interval $[0, 1]$, and inside we integrate over a slice where y has limits that we see in the picture. We get the repeated integral

$$\iint_{\Omega} 8x(x^2 - y + 1)^3\,dS = \int_0^1 \int_0^{x^2+1} 8x(x^2 - y + 1)^3\,dy\,dx.$$

How do we handle the inner integral? When integrating the function $8x(x^2 - y + 1)^3$ by y, then $8x$ is actually a constant and we can factor it out of the integral. When integrating the expression $(x^2 - y + 1)^3$ with respect to y, then also the term $x^2 + 1$ is treated as a constant, so in fact we integrate an expression of the form $(-y + a)^3$ that an experienced integrator works out right away, those more careful can try a suitable substitution:

$$\begin{aligned} 8x \int (x^2 - y + 1)^3\,dy &= \left| \begin{array}{l} w = -y + x^2 + 1 \\ dw = \frac{\partial}{\partial y}[-y + x^2 + 1]\,dy = -dy \end{array} \right| \\ &= -8x \int w^3\,dw = -2xw^4 + C = -2x(x^2 - y + 1)^4 + C. \end{aligned}$$

We get

$$\begin{aligned} \iint_{\Omega} 8x(x^2 - y + 1)^3\,dS &= \int_0^1 \int_0^{x^2+1} 8x(x^2 - y + 1)^3\,dy\,dx = \int_0^1 \Big[-2x(x^2 - y + 1)^4\Big]_{y=0}^{y=x^2+1} dx \\ &= \int_0^1 0 - (-2x(x^2 + 1)^4)\,dx = \int_0^1 2x(x^2 + 1)^4\,dx = \left| \begin{array}{l} w = x^2 + 1 \\ dw = 2x\,dx \end{array} \right| \end{aligned}$$

$$= \Big[\frac{1}{5}(x^2+1)^5\Big]_0^1 = \frac{2^5}{5} - \frac{1}{5} = \frac{31}{5}.$$

What if for some reason we do want slices in the x-direction? While all the right endpoints of slices are given by the same formula $x = 1$, the left endpoint depends on where we slice. For slices given by $y \in [0,1]$, the left endpoints are $x = 0$, so we will handle this part of set Ω (lower half) in a separate integral. In the upper half, that is, for slices determined by $y \in [1,2]$, the left endpoints for slices lie on the curve $y = x^2 + 1$. Therefore the starting value for x on such a slice is $\sqrt{y-1}$. We thus set up our integration as a sum of two integrals.

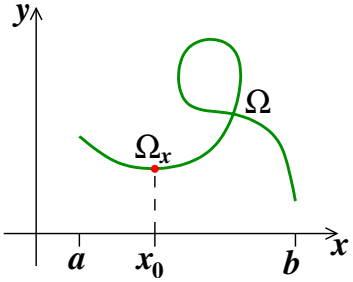
$$\begin{aligned}
\iint_\Omega 8x(x^2-y+1)^3\,dS &= \int_0^1\int_0^1 8x(x^2-y+1)^3\,dx\,dy + \int_1^2\int_{\sqrt{y-1}}^1 8x(x^2-y+1)^3\,dx\,dy \\
&= \left|\begin{array}{l} w = x^2-y+1 \\ dw = \frac{\partial}{\partial x}[x^2-y+1]\,dx = 2x\,dx \end{array}\right| \\
&= \int_0^1 \Big[(x^2-y+1)^4\Big]_{x=0}^{x=1} dy + \int_1^2 \Big[(x^2-y+1)^4\Big]_{x=\sqrt{y-1}}^{x=1} dy \\
&= \int_0^1 (2-y)^4 - (1-y)^4\,dy + \int_1^2 (2-y)^4 - 0\,dy \\
&= \Big[-\frac{1}{5}(2-y)^5 + \frac{1}{5}(1-y)^5\Big]_0^1 + \Big[-\frac{1}{5}(2-y)^5\Big]_1^2 = \frac{31}{5}.
\end{aligned}$$

A sigh of relief, we've got the same answer.

△

Remark: We conclude this section by addressing an interesting situation. We all know the formula $\int_a^a f(x)\,dx = 0$. What does it tell us? We are evaluating a one-dimensional integral over an interval that is actually just one point, so it is a set of dimension zero. The integrating interval forms the base for the region under the graph of f whose area we want to find, and if this base has length zero, then of course the area is also zero.

The same principle is valid also for double integral. It can happen that a set Ω lies in $\mathbb{R}^2$, but its actual dimension is lower. For a typical example take a segment. We can only move there and back in one direction on it, so it is one-dimensional. We can also say that it only offers one degree of freedom for movement on it. Another significant indication is that a segment has empty interior as a set in $\mathbb{R}^2$. True, we are used to the fact that the interior of an interval $[a,b]$ is the interval (a,b), but this is true only in one dimension. In two dimensions, a point is an interior point of some set only if it lies in it along with some of its neighborhoods, which in two dimensions means some disc. There are no such points in a segment in a plane. The same argument applies to all reasonable curves in plane.



When we have such a set with an empty interior, then the slices Ω_x (or Ω_y) are typically just a

point or several points. Consequently, the one dimensional integrals $\int_{\Omega_x}$ are automatically zero, and the whole double integral is zero.

In conclusion, when we attempt to integrate using a double integral over a set that is not sufficiently "substantial", for instance when its two-dimensional interior is empty, then the integral automatically yields zero. Typical sets of this sort are curves, for instance segments, arcs, circles and such.

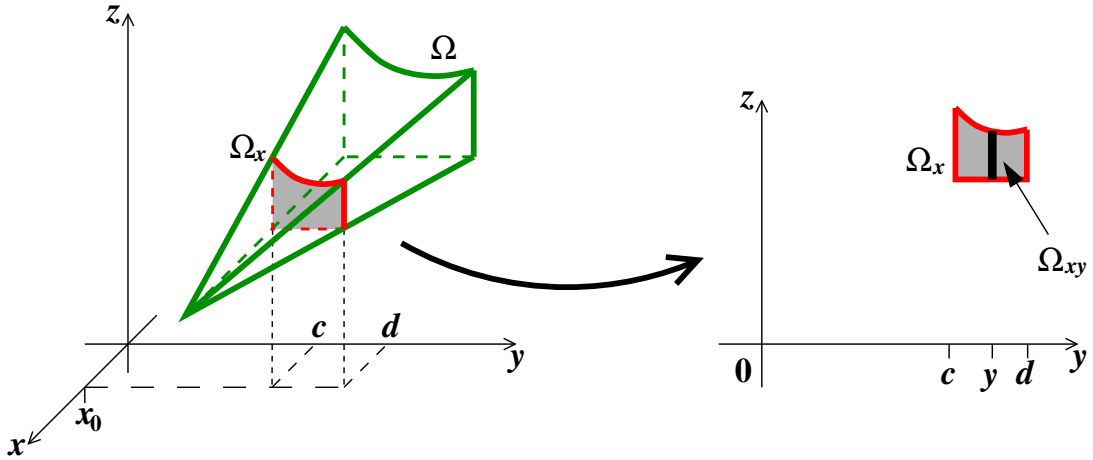
△

6b. Three-dimensional integral

A triple integral $\iiint_\Omega f(x,y,z)\,dV$ can be evaluated in an analogous way. We need to go through all points of the solid Ω, and we simplify this by arranging them into slices. In other words, we reduce the dimension of Ω one by one.

When we fix the value of one variable, for instance x as x_0, then it selects a plane in the three-dimensional space that is perpendicular to the x-axis. This plane then slices the solid Ω and creates a two-dimensional slice, call it Ω_x. For a reasonable solid Ω these slices are non-empty only for x from some range $[a,b]$, which determines the first stage of our setup. The outer integral makes sure that we go through all relevant slices. The inner integral then handles a particular slice, which is a two-dimensional set, so the inner integral will be two-dimensional. We move over this slice by changing y and z. We can write

$$\iiint_\Omega f(x,y,z)\,dV = \int_a^b \iint_{\Omega_x} f(x,y,z)\,dS[y,z]\,dx.$$



The next step is working out the inner integral over a two-dimensional set. We use the usual approach, that is, we slice the slice Ω_x. For instance, with a bit of luck we can make reasonable slices through Ω_x in the z direction, that is, by fixing y. We can call them Ω_{xy}. The range for y now depends on the shape of a particular slice Ω_x, so for the next integral we consider y between certain values $c(x)$ and $d(x)$. We have

$$\iint_{\Omega_x} f(x,y,z)\,dS[y,z] = \int_{c(x)}^{d(x)} \int_{\Omega_{xy}} f(x,y,z)\,dz\,dy.$$

This reduces the double integral to integration over a one-dimensional slice, with a bit of luck it

will be a segment whose endpoints depend on y, which in turn depends on x. We thus obtain

$$\iiint_\Omega f(x,y,z)\,dV = \int_a^b \int_{c(x)}^{d(x)} \int_{g(x,y)}^{h(x,y)} f(x,y,z)\,dz\,dy\,dx.$$

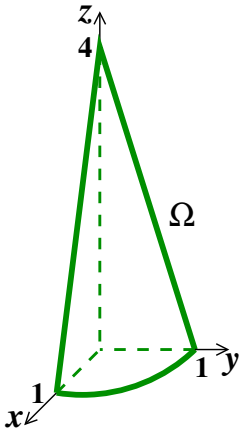
Of course, we do not know that exactly this order will work for a particular shape of Ω. There are six possible orders of decomposition, at least one might work well.

It is clear that in order to succeed in setting up such an integral, one needs good spatial imagination and be quite familiar with analytical handling of geometric objects.

Example: We will find the average of the function $f(x,y,z) = 8xy(4-z)$ on the set

$$\Omega = \left\{(x,y,z) \in \mathbb{R}^3;\ x,y \geq 0,\ 0 \leq z \leq 4 - 4\sqrt{x^2+y^2}\right\}.$$

First we need to figure out what kind of solid this set represents. We start with the restrictions $x,y,z \geq 0$, our solid will therefore be situated in the first quadrant. The condition on z tells us that Ω is actually the solid between two surfaces. The lower one is the horizontal plane $z = 0$, it will form the base of the solid. The upper surface is given by the equation $z = 4 - 4\sqrt{x^2+y^2}$. What surface is it?



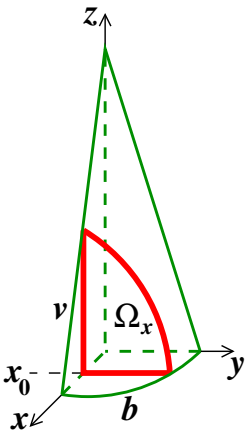
It is transformation of one of the essential equations. The equation $z = \sqrt{x^2+y^2}$ determines a cone (the elevation at points (x,y) is given by its distance from the origin), with its vertex at the origin and opening up; its axis coincides with the z-axis.

When we introduce a multiplicative constant, we get the equation $z = 4\sqrt{x^2+y^2}$ that speeds up the raise of the cone four times, so it will be markedly sharper; the incline of its sides is no longer $1:1$, but $4:1$. The change of sign $z = -4\sqrt{x^2+y^2}$ flips this cone downward, and adding four shifts it up.

So we get a cone with its vertex up at $(0,0,4)$ and sides falling steeply down. This cone delimits our solid from above, but we also have the plane $z = 0$ delimiting from below. Where do these two intersect?

$$0 = 4 - 4\sqrt{x^2+y^2} \implies x^2+y^2 = 1.$$

It follows that the condition $0 \leq z \leq 4 - 4\sqrt{x^2+y^2}$ determines the cone with vertex at $(0,0,4)$ and whose base is the unit circle in the xy plane centered at $(0,0,0)$. From this cone we cut out the quarter that lies in the main quadrant.

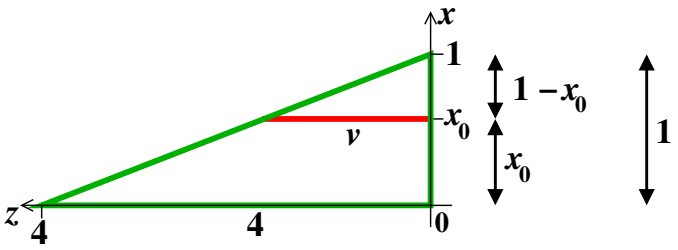


We need to integrate over this set, so we will now consider possible slicings. In the general introduction we started with x, so we will try it here. It is obvious that it only makes sense to take x from the range $[0,1]$. Fixing a certain x from this range will create a slice Ω_x parallel to the yz-plane. We get the first reduction of our integral.

$$\iiint_\Omega f(x,y,z)\,dV = \int_0^1 \iint_{\Omega_x} f(x,y,z)\,dS\,dx.$$

The slice Ω_x is two-dimensional, so the integral over this set has to be further reduced. How does this set look like?

What we see right away is that this shape has straight horizontal and vertical sides, that is, these are straight segments. How long are they? The height v can be deduced by a close look at the vertical triangular side of the quarter-cone that lies in the xz-plane, we put it on its side to save room.



Similarity of triangles yields

$$\frac{v}{1-x} = \frac{4}{1} \implies v = 4(1-x).$$

The size of the base b can be deduced from information about the base of the cone. It is given by the circle, hence $b = \sqrt{1-x^2}$. These two formulas also provide us with reasonable values for z, respectively y for the next step when we will cut this slice.

These next-generation cuts will be segments. For their precise specification we need to know the precise shape of the "hypotenuse", which is unfortunately not a real hypotenuse, as vertical slices through cones off their axes are not triangles. This third side is given by the equation of the cone where we take x as a fixed parameter, corresponding to our choice $x = x_0$ that created the slice. Thus the variables y, z on the third side of the slice are related by the formula

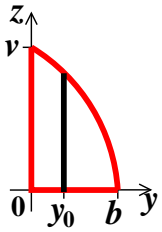
$$z = 4 - 4\sqrt{x_0^2 + y^2},$$

that is,

$$(4-z)^2 - y^2 = x_0^2.$$

It follows that the shape of the third side is hyperbolic.

We are ready to set up the integral for integration over Ω_x, we will try vertical slicing. This means that we choose some $y = y_0$ between 0 and $b = \sqrt{1-x_0^2}$, which determines a vertical segment on which z changes between 0 and $4 - 4\sqrt{x_0^2 + y_0^2}$. This means that our triple integral changes into the following repeated integral:



$$\iiint_\Omega f(x,y,z)\,dV = \int_0^1 \int_0^{\sqrt{1-x^2}} \int_0^{4-4\sqrt{x^2+y^2}} f(x,y,z)\,dz\,dy\,dx.$$

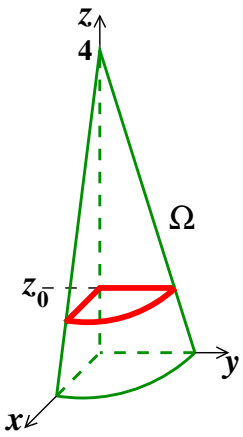
To find the average we need to integrate the given function. We start with the inner integral and work our way out.

$$\begin{aligned}
\iiint_\Omega 8xy(4-z)\,dV &= \int_0^1 \int_0^{\sqrt{1-x^2}} \int_0^{4-4\sqrt{x^2+y^2}} 8xy(4-z)\,dz\,dy\,dx = \left| \begin{matrix} w = 4-z \\ dw = -dz \end{matrix} \right| \\
&= \int_0^1 \int_0^{\sqrt{1-x^2}} \Big[-4xy(4-z)^2\Big]_{z=0}^{z=4-4\sqrt{x^2+y^2}}\,dy\,dx \\
&= \int_0^1 \int_0^{\sqrt{1-x^2}} 4^3xy - 4^3xy(x^2+y^2)\,dy\,dx = \left| \begin{matrix} w = x^2+y^2 \\ dw = 2y\,dy \end{matrix} \right| \\
&= \int_0^1 \Big[2\cdot 4^2xy^2 - 4^2x(x^2+y^2)^2\Big]_{y=0}^{y=\sqrt{1-x^2}}\,dx
\end{aligned}$$

$$= \int_0^1 2 \cdot 4^2 x(1-x^2) - 4^2 x + 4^2 x^5 \, dx = 4^2 \int_0^1 x - 2x^3 + x^5 \, dx$$
$$= 4^2 \Big[\frac{1}{2}x^2 - \frac{1}{2}x^4 + \frac{1}{6}x^6 \Big]_0^1 = 4^2 \frac{1}{6} = \frac{8}{3}.$$

Alternative: Symmetry makes it clear that trying to take slices perpendicular to the y-axis, parallel to the xz plane would lead to slices of the same shape as before. Also x and y in the function f are interchangeable, so all the calculations would be the same, just roles of x and y would be switched. To see something different we will try horizontal slicing.

So we fix some value $z = z_0$, this makes sense for the range $[0, 4]$. This creates a slice that should have the shape of a quarter-circle.



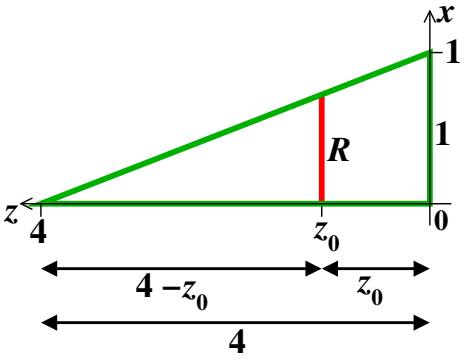
We confirm this by taking the equation of the conic surface

$$z = 4 - 4\sqrt{x^2 + y^2}$$

and use $z = z_0$ as a parameter. We get

$$x^2 + y^2 = \Big(1 - \frac{1}{4}z_0\Big)^2.$$

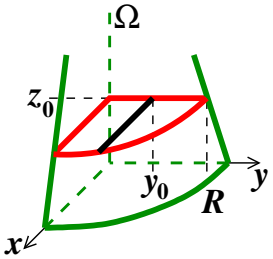
This is the shape of the concave side of our slice and it is a circle indeed, its radius is $R = 1 - \frac{1}{4}z_0$. The latter can be independently confirmed using similarity of triangles on the left vertical side of our quarter-cone.



We get

$$\frac{R}{4-z} = \frac{1}{4} \implies R = \frac{1}{4}(4-z).$$

Now take one such quarter-circle. We will decompose it into segments. If we fix some value $y = y_0$ between 0 and $R = \frac{1}{4}(4-z)$, it determines a cut shaped like a segment, parallel with the x-axis with values for x going between 0 and $\sqrt{R^2 - y_0^2}$, that is, $\sqrt{\frac{1}{4^2}(4-z_0)^2 - y_0^2}$. In this way we set up the following transform of a triple integral to repeated integral.



$$\iiint_\Omega f(x,y,z)\, dV = \int_0^4 \int_0^{(z-4)/4} \int_0^{\sqrt{(4-z)^2/16 - y^2}} f(x,y,z)\, dx\, dy\, dz.$$

Again we integrate the given function.

$$\iiint_\Omega 8xy(z-4)\, dV = \int_0^4 \int_0^{(z-4)/4} \int_0^{\sqrt{(4-z)^2/16 - y^2}} 8xy(z-4)\, dx\, dy\, dz$$

$$= \int_0^4 \int_0^{(z-4)/4} \Big[4x^2 y(z-4)\Big]_{x=0}^{x=\sqrt{(4-z)^2/16-y^2}} \, dy\, dz$$
$$= \int_0^4 \int_0^{(z-4)/4} 4\Big(\frac{1}{16}(4-z)^2 - y^2\Big) y(4-z)\, dy\, dz.$$

Here we see two possible approaches. We could multiply out the first group and then integrate y and y^3 separately. An interesting alternative is to use substitution, because the whole term $\frac{1}{16}(4-z)^2$ is taken as constant now. If we denote $w = \frac{1}{16}(4-z)^2 - y^2$, we get $-2y\,dy = dw$ and proceed as follows:

$$\int 4\Big(\frac{1}{16}(4-z)^2 - y^2\Big) y(4-z)\, dy = \int -2w(4-z)dw = -w^2(4-z) = -\Big(\frac{1}{16}(4-z)^2 - y^2\Big)^2 (4-z).$$

We thus have

$$\iiint_\Omega 8xy(z-4)\, dV = \int_0^4 \int_0^{(z-4)/4} 4\Big(\frac{1}{16}(4-z)^2 - y^2\Big) y(4-z)\, dy\, dz$$
$$= \int_0^4 \Big[-\Big(\frac{1}{16}(4-z)^2 - y^2\Big)^2 (4-z)\Big]_{y=0}^{y=(z-4)/4} dz$$
$$= \int_0^4 0 - \Big[-\Big(\frac{1}{16}(4-z)^2 - 0\Big)^2 (4-z)\Big]\, dz$$
$$= \int_0^4 \frac{1}{4^4}(4-z)^5\, dz = \Big[\frac{1}{-6\cdot 4^4}(4-z)^6\Big]_0^4 = \frac{16}{6} = \frac{8}{3}.$$

We confirmed the result.

In order to find the average we also need to know the volume of Ω. In general, it would be possible to integrate the function 1 over Ω, we now have two orders of integration prepared. However, there is an interesting alternative.

It is based on the the interpretation of slicing that volume is the sum of areas of slices. This points us to the horizontal slices that are quarter-circles. Formally, if we denote their areas as A_z, we can write

$$\mathrm{vol}(\Omega) = \iiint_\Omega 1\, dV = \int_0^4 \iint_{\Omega_z} 1\, dS\, dz = \int_0^4 A_z\, dz.$$

We found that the quarter-circle corresponding to fixed $z = z_0$ has radius $R = \frac{1}{4}(4-z)$, hence $A_z = \frac{1}{4}\pi\big(\frac{1}{4}(4-z)\big)^2$. Thus

$$\mathrm{vol}(\Omega) = \int_0^4 \pi \frac{1}{4^3}(4-z)^2 dz = \Big[-\pi\frac{1}{3\cdot 4^3}(4-z)^3\Big]_0^4 = \frac{\pi}{3}.$$

Actually, we could have just taken the standard formula for the volume of a cone and then divide by four, obtaining the same result. But we would not have learned anything interesting in that way.

Now we are ready to find the average:

$$\operatorname*{Ave}_{\Omega}\big(8xy(4-z)\big) = \frac{\iiint_{\Omega} 8xy(4-z)\,dV}{\operatorname{vol}(\Omega)} = \frac{8}{\pi}.$$

△

This example shows that the basic idea of integration in more dimensions is not that difficult, but decomposing the given solid to suitable slices could be challenging if our grasp of three-dimensional geometry is not up to scratch.

Remark: We conclude this section by repeating the closing remark of the previous section. If we try to integrate some function $f(x,y,z)$ over a set in $\mathbb{R}^3$ that is not truly three-dimensional, for instance it has an empty interior in $\mathbb{R}^3$, then the integral is automatically zero.
△

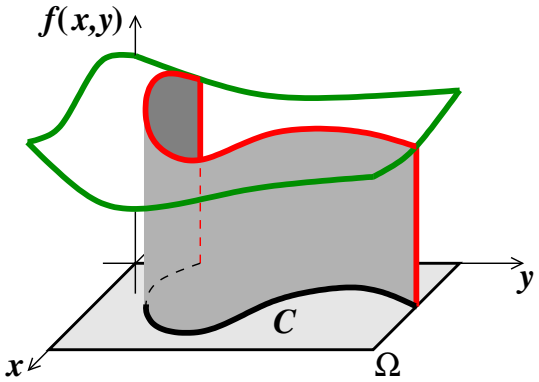
6c. Line and surface integrals

Imagine the heating plate of a stove, with some stable distribution of temperature $f(x,y)$ on it. If we represent this plate by a set Ω, then we can easily find the average temperature on our stove using double integration over Ω.

Now imagine that a bug (with asbestos slippers) crawled across this plate, and the path that it took creates some curve C in the plane. The bug is naturally curious what is the average temperature that it experienced on this trip. Since C is most likely not a straight line but it twists and curves, the bug cannot use the usual one-dimensional integral. On the other hand, such curve is essentially a one-dimensional object, so if the bug decides to evaluate the double integral of temperature over the curve C, then it automatically yields zero.

However, bug's question about average temperature is entirely legitimate and shows that we need to develop yet another type of integral. How should it work? The situation is as follows.

We have a function $f(x,y)$ defined on a planar set Ω. But we are interested in values of f only on some curve C. We intend to "add" them using integration, so we are in fact looking for the area of the slice determined by the curve C in the solid under the graph of f. This slice is an object that is essentially two-dimensional, so the notion of area is appropriate, and it is not flat but curved. We have to keep both things in mind when developing the new integral.



When we take some point $\vec{x}$ on the curve C, then it does not make much sense to consider some planar infinitesimal around it, because this would lead to a 3D-column with volume while we need a flat object with area. Instead we consider a "length infinitesimal" in the direction of the curve, that is, an infinitely short part of this curve around $\vec{x}$ of length ds. This will be the base over which we erect a rectangle of height $f(\vec{x})$. As a flat rectangle it has an area, namely $f(\vec{x})\cdot ds$. Summing up these contributions using integral we arrive at a number that the bug is interested in, the area of the curved slice.

This number is called the **line integral** (sometimes also **curve integral**) and we denote it

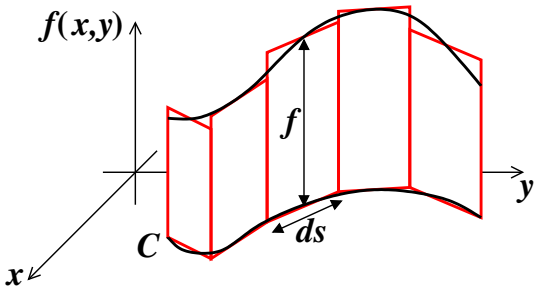
$$\int_C f(\vec{x})\,ds.$$

There is a special notation

$$\oint_C f(\vec{x})\,ds$$

for line integrals over closed curves (the end connects to the start).

To appreciate better its meaning we apply the story that is usually given for ordinary definite integral. We can imagine that we divided the curve C into very small parts of length ds. Because such a part is very tiny, we can assume that the curve does not have time to bend much, so it is almost a straight segment, and the function f does not have time to change its value much, so it can be approximated by its value, say, in the middle of the segment. In this way we approximate the shaped area above the curve by flat rectangular panels.



Adding areas of these panels we get an approximation for the area of the slice, an approximation that can be made better and better as we decrease the widths of these rectangles.

We note that in the notation for a line integral above we used a vector $\vec{x}$ instead of the coordinates (x, y). There is a reason for that. Our ideas above did not depend on the curve being on a flat plate, the main ingredient was the ability to split it into small parts whose lengths were known. The idea of line integral is therefore applicable to curves in arbitrary space $\mathbb{R}^n$ and a function of n variables.

For instance, imagine a fly flying through a classroom where we know the temperature $f(x, y, z)$ at all points of the room. If the fly wanted to know the average temperature over the path it took (which is a one-dimensional curve), it would again split it into small parts of length ds and each would contribute $f(x, y, z) \cdot ds$ to the total.

Once we have the line integral established, we can find the **length of a curve** by integrating number 1 (as a constant function) over it, because by adding contributions $1 \cdot ds$ we are in fact adding lengths of those parts. The average of a function on a curve can be found by integrating the function over the curve and then dividing this by the length of the curve.

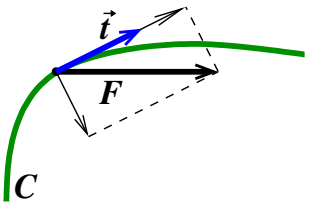
An important application of line integrals is related to vector functions. Actions of differentiation and integration are typically applied to vector functions by applying them to individual coordinates. That would also be the interpretation of the line integral $\int_C F(\vec{x})\,ds$, the result would be a vector again. However, in applications people usually need something else.

Imagine that a vector function F describes an acting force that causes some object to move along a path C. What work was done? If the force is constant and acts in the direction of a straight path, then it is enough to work with its magnitude $\|F\|$ and use the formula known from elementary school stating that works equals force times displacement: $W = \|F\| \cdot s$. If the magnitude of the force changes but it still acts in the direction of movement along a straight path, then we can split the path into small parts ds; on each of them the force does not have time to change much and thus we can use integration to add local contributions $\|F(x)\|ds$. This approach will work using

line integral also in cases when the path is curved in a plane or in space, assuming that the force acts in the appropriate direction.

However, this is not guaranteed in general. Then we need to decompose the force into two components, one in the direction of the movement and the other perpendicular to it. Work is done only by the tangent component.

How do we find it? We need to know the unit tangent vector $\vec{t}$ that points in the direction of movement (there is also the unit tangent vector pointing in the opposite direction that we do not want). Linear algebra tells us that the magnitude of the tangent component of the vector F can be obtained as the dot product $F \bullet \vec{t}$. The corresponding contribution to work is therefore $F \bullet \vec{t}\, ds$.



If we denote by $\vec{t}(\vec{x})$ the unit tangent vector to the curve C at the a point $\vec{x}$ pointing in direction of travel, then the total work is given by the line integral

$$\int_C F(\vec{x}) \bullet \vec{t}(\vec{x})\, ds.$$

This is the usual way of integrating vector functions along curves in applications. The term $\vec{t}(\vec{x})ds$ incorporates the information about the length of the infinitely small piece of the curve and also its direction, so it can be considered an "oriented infinitesimal". Many authors use the notation $d\vec{s} = \vec{t}(\vec{x})ds$ for it. When properly introduced, this can be seen as "oriented differential". This type of line integral for vector functions is therefore often written as

$$\int_C F(\vec{x}) \bullet d\vec{s}.$$

It should be noted that having a curve C does not yet allow us to set up this type of integral, because a curve by itself does not have a direction of travel associated with it. Obviously, we use this integral only in situation when more information is available, in particular the direction of travel must be somehow determined.

One popular interpretation of vector functions is to see them as vector fields. Imagine that F describes local velocity of a flowing liquid or gas. When we look at some point $\vec{x}$ of a curve C, then the decomposition of F to tangent and normal components tells us how much of the media passes along the curve at this point, and how much crosses the curve. When we add local information about the flow along the curve C using the line integral, we get another useful interpretation: The integral

$$\int_C F(\vec{x}) \bullet d\vec{s}$$

determines the total **flow along the curve**, also called flow circulation (especially if the curve C is closed).

For the case $F = (F_1, F_2)$: $\mathbb{R}^2 \mapsto \mathbb{R}^2$ we sometimes see flow along a curve stated in the form

$$\int_C F_2 dx + F_2 dy.$$

In order to decipher this we would need to talk about parametrization, so this will wait for the next section.

The flow along the curve tells us about the component of flow that does not influence how much of the medium is to the left and to the right of the curve. Sometimes we want to know that information. Then we have to work with the other component of the decomposition of F, the one that is perpendicular to the curve. To find its magnitude we need to know the unit normal vector $\vec{n}(\vec{x})$ of the curve. There are more candidates again, and context should tell us which orientation

is proper for our situation. The amount of flow across the curve at a point $\vec{x}$ is then obtained as $F(\vec{x}) \bullet \vec{n}(\vec{x})$. Integrating this we obtain

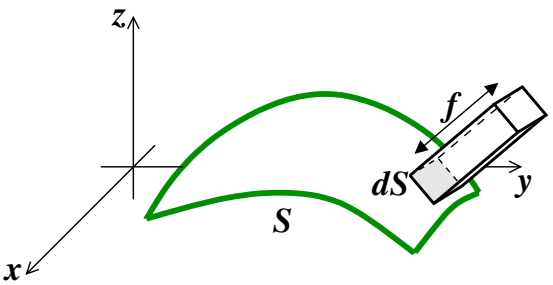
$$\int_C F(\vec{x}) \bullet \vec{n}(\vec{x})\, ds.$$

This **flow across the curve** tell us by how much the balance of medium on the left and on the right changed. There is no special notation for the element $\vec{n}(\vec{x})\, ds$.

The idea of line integral can be readily generalized. Given a function of n variables, we may be interested in its behaviour over some set whose actual dimension is smaller. Such a general theory is possible, but we do not really need it in typical applications. People therefore typically focus on two important cases, when the set is essentially one dimensional (we just worked it out) and when it is two-dimensional. Such sets are called surfaces.

Consider a room with recorded local temperature $f(x, y, z)$. We place a balloon there and want to know what is the average temperature on its surface. That is also a good question, and obviously the line integral does not help. We will follow the usual approach.

Consider some surface S in $\mathbb{R}^n$ for $n \geq 3$. For a point $\vec{x}$ on this surface we take an infinitely small square around it whose area is dS. Over this square base we erect a "column" of height $f(\vec{x})$, creating a three-dimensional object in this way whose volume we easily find.



We then "add" volumes of these "columns" via integration. In this way we obtain the **surface integral**, denoted

$$\iint_S f(\vec{x})\, dS.$$

If this surface is closed, (think of surface of that balloon), then we use the notation

$$\oiint_S f(\vec{x})\, dS.$$

As usual, this formula can provide us with the **area of a surface**, we just integrate value 1 over it.

What is the typical application of surface integrals to vector functions? Again, the definition is interpreted as integration by coordinates, but in applications we prefer another approach denoted

$$\iint_S F(\vec{x}) \bullet d\vec{S}.$$

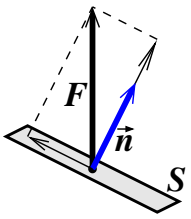
What do we mean by this? Consider some infinitely small piece of the surface S around a point $\vec{x}$. Then dS stores its area and $\vec{x}$ provides the location, but to know the precise position we would also need to know the tilt, which is for deformed surfaces determined as the tilt of its tangent plane at $\vec{x}$. We addressed the problem of recording tilt for length infinitesimals and we used a tangent vector to capture the direction. However, this only works for one-dimensional objects. If we wanted to capture the tilt of a plane in $\mathbb{R}^3$ using tangent vectors, then we would need two of them, which is not convenient for our situation. If we want to capture the tilt with just one vector, then we have to use the normal vector, that is, a vector perpendicular to the surface. The column

in the above picture shows the right direction.

Accordingly, for that infinitely small part of the surface at $\vec{x}$ we find a unit normal vector, call it $\vec{n}(\vec{x})$, and then $d\vec{S} = \vec{n}(\vec{x}) \cdot dS$. We can see it as "oriented area infinitesimal". Note that there are always two candidates for the normal vector pointing in opposite directions. It is assumed that when we use $d\vec{S}$, then there is some additional information that allows us to determine the proper orientation of $\vec{n}$. In order for key theorems to work, this orientation should be consistent, namely $\vec{n}(\vec{x})$ should be continuous as a function of $\vec{x}$.

Interestingly, this it is not always possible. For instance, it is not possible to assign continuous orientation to the famous Mobius strip, which is problematic as continuity is one of basic assumptions for integration. However, one does not expect troubles of this sort in practical applications.

Assume then that thanks to some information we determined unit normal vectors $\vec{n}(\vec{x})$. Then the dot product $F \bullet \vec{n}$ provides us with the magnitude of the normal component of F. If F is a vector field that describes the flow of a liquid or gas, then the tangent component of F describes particles that move parallel to the surface of S, while the normal component tells us what proportion of flow actually crosses the surface.



A typical application of a surface integral to vector functions, called **flux through a surface** or **flux across a surface**, therefore has the form

$$\iint_S F(\vec{x}) \bullet d\vec{S} = \iint_S F(\vec{x}) \bullet \vec{n}(\vec{x})\, dS.$$

A reader may be concerned now about the fact that the length differential $d\vec{s}$ works with the tangent direction while the area differential $d\vec{S}$ works with the normal direction, therefore their dot product with a vector function results in different information (flow along versus flow across). One has to get used to this and there is some inner logic, namely that in both cases we store the information about tilt of ds or dS using one vector.

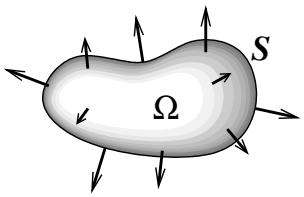
Both the line integral and the surface integral can be defined only for curves and surfaces that are reasonable, which typically means that we are able to describe them parametrically using differentiable vector functions. We will look at this closer in the next section where we address the problem of actually evaluating these integrals, here we will look at two interesting correspondences between ordinary multi-dimensional integrals and line/surface integrals that are useful in applications.

Consider some vector function $F(x, y, z)$. We will now interpret it as a vector field, for instance the recording of the velocity at different locations of a water flowing through a river. The flow is steady, that is, it does not change over time. Consider some region, that is, an imaginary solid Ω positioned in the river bed. Because it is just imaginary, the water flows freely through it. We want to know whether this region as a whole adds or takes away water.

There are two possible approaches. We know that at every point of the river, divergence provides the corresponding local information (source, sink). If we add all these local balances throughout the whole region, we obtain its overall contribution:

$$\iiint_\Omega \operatorname{div}(f)\, dV.$$

However, we should be able to obtain the same information in another way. Namely, we will look at how much water is flowing in or out through the boundary S of this solid. The boundary is a two-dimensional object, that is, a surface. If we can determine how much water is flowing in or our at every point, then we will be able to add all these numbers using surface integral. This will require some work.



As we saw above, the flow through some tiny part of surface around a point $\vec{x}$ can be obtained as $F \bullet \vec{n}(\vec{x})$, but we need to clarify the problem of orientation. If the surface S is the boundary of a reasonable three-dimensional solid, then there are two recognizable directions, namely in and out. We expect here that a positive sign for the total balance means that medium is produced, which means an outflow.

This will happen if we always take normal vectors pointing out of the region. These are called **outer normals**. We will therefore introduce the convention that in these considerations we always choose such normal vectors.

Having settled the problem of orientation, we can recall out earlier musings on "flux across a surface" and observe that the amount of medium that the region Ω produces is also given by the integral

$$\oiint_S F(\vec{x}) \bullet d\vec{S}.$$

Both approaches provide the same information, which means that the two resulting integrals must be equal. We just deduced an important statement known as **divergence theorem**, or the **Gauss-Ostrogradsky theorem**:

$$\iiint_\Omega \operatorname{div}(F)(\vec{x})\, dV = \oiint_S F(\vec{x}) \bullet d\vec{S}$$

or

$$\iiint_\Omega \operatorname{div}(F)(\vec{x})\, dV = \oiint_S F(\vec{x}) \bullet \vec{n}(\vec{x})\, dS.$$

As the notation suggests, the surface S will be closed for reasonable regions Ω.

As expected, this equality is true only assuming that the region Ω and its surface are sufficiently nice and the vector function F is continuously differentiable.

The divergence theorem also applies to other dimensions. Imagine a very thin plate isolated from the outside, so that heat can only spread in two dimensions. There is some heating device and perhaps some cooling at another place, and after a while a steady situation should be reached where the heat flow does not change with time. We can capture it with a vector function $F(x, y)$. When we draw a potatoid on this plate (a closed curve C) and ask what is the heat balance of the region Ω delimited by our drawing, we can simply add divergences, that is, local sources and sinks over the whole Ω. Or we look at what heat passes through its boundary, which is now one-dimensional, that is, it is some (closed) curve C. We can determine outer normal vectors also for curves in two dimensions, so analogous reasoning leads to an analogous equality

$$\iint_\Omega \operatorname{div}(F)(\vec{x})\, dS = \oint_C F(\vec{x}) \bullet \vec{n}(\vec{x})\, ds.$$

Interestingly, it also makes sense to do this analysis for the one-dimensional case $\Omega = [a, b]$. The boundary $\partial\Omega$ consists of two points a, b and reasoning about flow leads us to the familiar identity

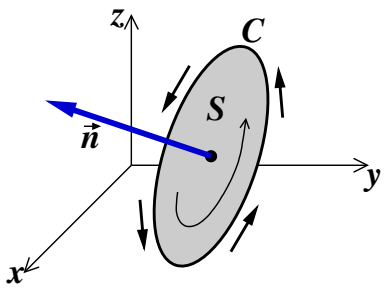
$$\int_{[a,b]} f'(x)\, dx = \Big[f(x)\Big]_a^b = f(b) - f(a).$$

The second important fact in this context is called the **Stokes theorem**, usually stated in three dimensions:

$$\iint_S \operatorname{curl}(F) \bullet d\vec{S} = \oint_C F \bullet d\vec{s}.$$

Here S is some surface in $\mathbb{R}^3$ and C is its boundary, that is, C is a closed curve. It is also assumed that the orientation of differentials $d\vec{S}$ and $d\vec{s}$ is coordinated in such a way that they form a "right-handed pair". The idea is that we first choose a direction of travel around the curve C, which determines the orientations of $d\vec{s}$. Then we choose among the two possibilities for the normal vector $\vec{n}(\vec{x})$ to the surface S the orientation that works like this: If we sit on the tip of $\vec{n}$ and look down on S, then we should see the travel around C in a positive direction, that is, counterclockwise.

What is the meaning of the Stokes theorem? Imagine a flat disc S in a three-dimensional vector field F representing a flowing medium. This disc is held in place, but it is allowed to rotate about its axis that can also serve as the default direction for the normal $\vec{n}$. This then determines the direction of travel around the perimeter.



As we saw above, the integral

$$\oint_C F \bullet d\vec{s} = \oint_C F \bullet \vec{t}\, ds$$

tells us how much of the medium flows along the boundary C. Due to friction it tries to rotate the disc, the integral sums up these effects around the whole perimeter and thus it provides us with the total momentum of rotation about the axis $\vec{n}$ imparted on the disc on its perimeter. Of course, it is possible that the medium influences the perimeter also in other ways, for instance it can flow perpendicularly to the perimeter at some places, so there could be also an attempt to tilt the disc. However, the dot product in the integral disregards this influence.

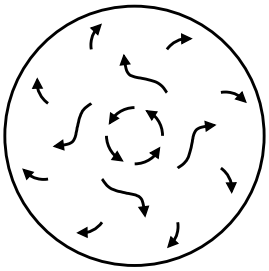
Note that the value of this integral is positive if the medium flows in the chosen direction of travel and negative if it flows against it.

The integral

$$\iint_S \operatorname{curl}(F) \bullet d\vec{S} = \iint_S \operatorname{curl}(F) \bullet \vec{n}(\vec{x}) dS$$

adds local attempts at rotating the disc (recognized by the curl operator) over the whole interior of the disc. The rotational influence provided by $\operatorname{curl}(F)$ has the form of a vector, and the dot product with $\vec{n}$ isolates from this vector the component that encourages rotation about the axis $\vec{n}$. This information is therefore consistent with what we were observing on the perimeter with the line integral. Note that signs match with the line integral, these local contributions are positive if they turn in the direction of travel around the perimeter.

The Stokes theorem says that whether we determine the rotational effect about the axis $\vec{n}$ by adding local effects on the disc or by adding influences about perimeter, it should come up the same. Unlike the divergence theorem where we were just adding and subtracting medium, here it is not all that clear why these two approaches should agree. I can imagine a situation when the medium is at rest at the perimeter of the disc, but turning in the middle. Then obviously the theorem should not be valid.



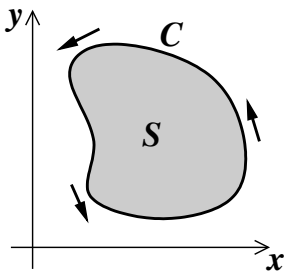
However, such a medium would not be realistic, because when a liquid turns in the middle and stays stationary on the perimeter, then there must be some backflow somewhere inbetween. In other words, if there is a positive curl in the middle, then we expect negative curl near the perimeter. The Stokes theorem tells us that this reasoning is correct, the positive and negative curls on the region should cancel each other out if the medium on the perimeter should stay stationary.

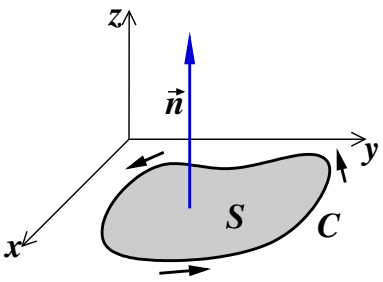
We worked with a flat disc here, but the Stokes theorem is more general. The equality between rotational influence on the perimeter and on the interior is valid also for other shapes, including bulging and irregular ones, as long as they are reasonable. It is a rather strong theorem.

How does the Stokes theorem translate to two dimensional case? The flow along the boundary is determined by the same line integral, but we have to make a change in the surface integral. For a two-dimensional function F the curl provides just a number, because now we need not worry about direction of rotation, just about the magnitude (and orientation) of momentum. We get the total influence by integrating these local contributions directly, without the need for a dot product.

$$\iint_S \operatorname{curl}(F)\, dS = \oint_C F \bullet d\vec{s}.$$

We just have to make sure that the orientation of travel around C is taken properly. The rule is that when we travel along C, we should see the interior of S on our left. This formula is also called the **Green theorem** and interestingly, it is equivalent to the two-dimensional case of the divergence theorem. In other words, the Stokes theorem and the divergence theorem provide the same information when applied to functions of two variables, just expressed in different ways.





The two-dimensional case can be also deduced from the original statement by embedding $\mathbb{R}^2$ in $\mathbb{R}^3$. Precisely speaking, having a vector function $F = \big(F_1(x,y), F_2(x,y)\big)$ and a region S in $\mathbb{R}^2$, we can define the auxiliary function

$$G(x,y,z) = \big(F_1(x,y), F_2(x,y), 0\big)$$

and the surface $S_3 = \{(x,y,0);\ (x,y) \in S\}$. Then we can apply the three-dimensional Stokes theorem to G and S_3. We easily find that

$$\operatorname{curl}(G)(x,y,z) = \Big(0, 0, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big) = \big(0, 0, \operatorname{curl}(F)(x,y)\big).$$

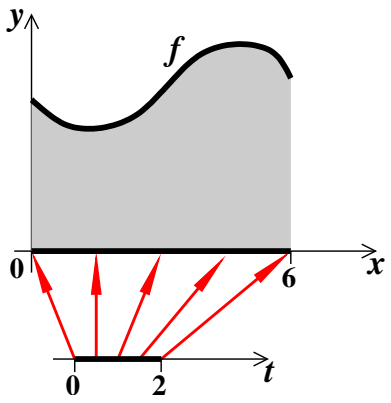
Because the normal vector to a flat object lying in the xy plane is $\vec{n} = (0,0,1)$, we get

$$\operatorname{curl}(G)(x,y,z) \bullet \vec{n} = \operatorname{curl}(F)(x,y).$$

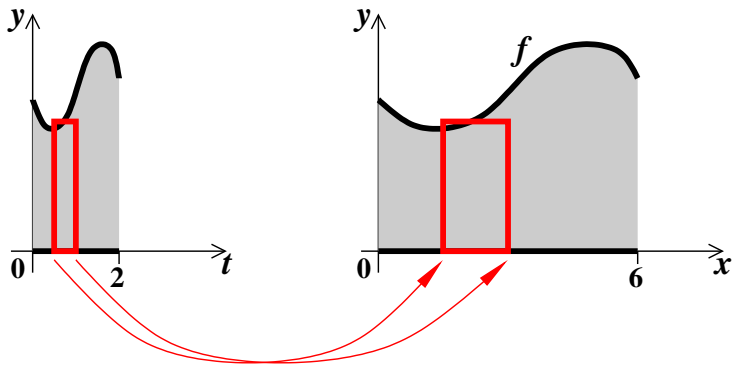
6d. Parametrization, substitution

We start with a function of one variable. We want to find the integral of a function $f(x)$ on the interval $[0,6]$, in other words, we are interested in the area of a certain shape.

However, for some reason we do not want to access the x-axis directly, but through some parameter t, for instance by working with $x = 3t$ for $t \in [0,2]$.



What happens if instead of the integral $\int_0^6 f(x)\,dx$ we evaluate the integral $\int_0^2 f(3t)\,dt$? We go through $[0,2]$ and "add" values $f(3t)$, which actually means that we consider all values of $f(x)$ for $x \in [0,6]$. However, we do it while moving only over a smaller interval $[0,2]$. In effect, we are working with the original shape of the region under the graph of f, but squeezed from the sides three times.



Consequently, we can expect that the resulting area and hence also the integral will be three times smaller than they should.

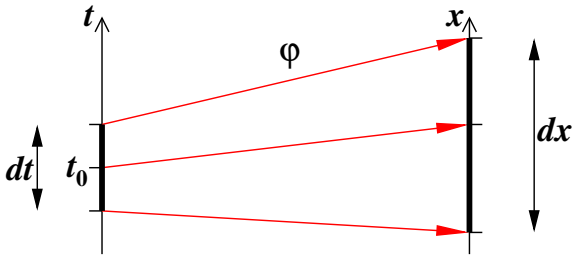
The picture also shows another justification for this conclusion. Riemann integration tells us to approximate $\int_0^6 f(x)\,dx$ by splitting the interval $[0,6]$ into segments and work with areas $f(x)\cdot dx$ of rectangles over them. The mapping $x = 3t$ allows us to backtrack to the world of t, but as it does, it shortens the base of each rectangle three times, so the areas $f(3t)\cdot dt$ are three times smaller than they should.

If we want to preserve the areas, we have to replace the infinitesimals dx not with dt, but with their triples, that is, $dx = 3dt$. In other words, we should be evaluating the integral $\int_0^2 f(3t)\,3dt$. Incidentally, the substitution theorem suggests exactly the same, for the formula $x = 3t$ it recommends to replace the differential using the formula $dx = [3t]'dt = 3\,dt$.

Our example was very simple because we changed the t-axis into the x-axis uniformly, but it can also be done differently. For instance, the substitution $x = t^2$ sometimes stretches and sometimes squeezes the infinitesimals dt. The substitution rule says that derivative helps here, and it will be useful for us to understand where this comes from.

To this end, consider a coordinate transformation $x = \varphi(t)$. We will analyze how it deforms the infinitesimal dt.

Consider some t_0 on the t-axis and a small segment around it of length dt, so it actually is the segment between points $t_0 - \frac{1}{2}dt$ and $t_0 + \frac{1}{2}dt$. For simplicity, assume that φ is increasing, then our segment maps onto the segment between points $\varphi\big(t_0 - \frac{1}{2}dt\big)$ and $\varphi\big(t_0 + \frac{1}{2}dt\big)$.



This means that the resulting length dx is given as

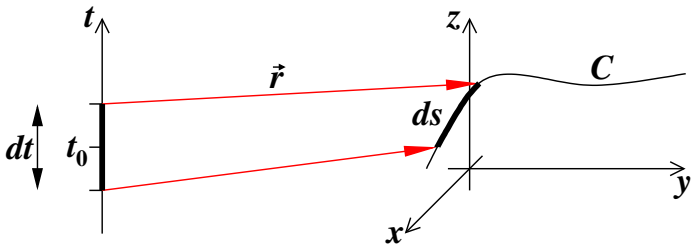
$$dx = \varphi\big(t_0 + \tfrac{1}{2}dt\big) - \varphi\big(t_0 + \tfrac{1}{2}dt\big).$$

If dt is really tiny, we can approximate the function values using a linear Taylor polynomial and we get

$$dx = \big[\varphi(t_0) + \varphi'(t_0)\tfrac{1}{2}dt\big] - \big[\varphi(t_0) - \varphi'(t_0)\tfrac{1}{2}dt\big] = \varphi'(t_0)dt.$$

Now we will do the same analysis for a **line integral**. Consider some curve C in $\mathbb{R}^n$. It must be specified somehow, and for curves the natural way is a parametric description, that is, the curve C is the range of some vector function traditionally denoted $\vec{r}(t)$ for $t \in [a, b]$.

Our task is to go through points of the curve C and "add" contributions $f(\vec{x})\cdot ds$. Parametrization allows us to reach points of the curve using $\vec{r}$, we have $\vec{x} = \vec{r}(t)$, and we also need to know how this transformation deforms the infinitesimal dt. We will proceed as before, that is, we choose t_0 and check on the segment of length dt around it.



The calculations will be analogous, because in section 5b we have shown that linear approximations work also for vector functions of one variable. However, there will be one significant change. Since our transformation sends the elementary segment to the multi-dimensional world $\mathbb{R}^n$, its image is a vector. We thus need to use norm to find its length.

$$\begin{aligned} ds &= \big\|\vec{r}\big(t_0 + \tfrac{1}{2}dt\big) - \vec{r}\big(t_0 - \tfrac{1}{2}dt\big)\big\| \\ &= \big\|\big[\vec{r}(t_0) + \vec{r}\,'(t_0)\tfrac{1}{2}dt\big] - \big[\vec{r}(t_0) - \vec{r}\,'(t_0)\tfrac{1}{2}dt\big]\big\| = \|\vec{r}\,'(t_0)dt\| \\ &= \|\vec{r}\,'(t_0)\|\,dt. \end{aligned}$$

Thus if we want to add the area of some elementary rectangle $f(\vec{x})\cdot ds$ for a point $\vec{x} = \vec{r}(t)$ on our curve, we can replace it by the area $f(\vec{r}(t))\cdot\|\vec{r}\,'(t)\|dt$. Summing up such rectangles we obtain the value of the integral over the curve C. Because this is an important result, we will give it a proper form and add necessary assumptions.

Fact.
Let f be a continuous function on some set $\Omega \subset \mathbb{R}^n$.
Let C be a curve in Ω given by a parametrization $\vec{r}(t)$ for $t \in [a, b]$, where $\vec{r}$ is a vector function $[a, b] \mapsto \Omega$ such that $\vec{r}\,[a, b] = C$, $\vec{r}$ is one-to-one on $[a, b]$ and it has non-zero continuous derivative on (a, b). Then

$$\int_C f(\vec{x})\,ds = \int_a^b f(\vec{r}(t))\|\vec{r}\,'(t)\|\,dt.$$

This is easy to remember, because it is very similar to the classical substitution

$$\left| \begin{array}{l} \vec{x} = \vec{r}(t) \\ ds = \|\vec{r}\,'(t)\|\, dt \end{array} \right|.$$

There is a special and quite frequent case where this formula becomes simpler. Consider a curve C in $\mathbb{R}^2$ that actually is the graph of some differentiable function $g(x)$ on interval $[a,b]$. Then we have a natural parametric description $\vec{r}(x) = (x, g(x))$ and the formula for line integral of a function $f(x,y)$ on C becomes

$$\oint_C f(x,y)\, ds = \int_a^b f\big(x, g(x)\big)\sqrt{1+[g'(x)]^2}\, dx.$$

If we are interested in the length of the graph of g, then we would use $f = 1$ and obtain the formula that we know from the introductory calculus course.

There is an alternative take on this substitution in line integral. The second widespread notation for parametrization is (in two dimensions)

$$x = x(t),$$
$$y = y(t).$$

Here $x(t)$ and $y(t)$ correspond to components $r_1(t)$ a $r_2(t)$ of the parametric description $\vec{r}(t)$. One-dimensional substitution tells us that

$$dx = x'(t)dt,$$
$$dy = y'(t)dt.$$

This shows how the individual coordinates of the point $(x_0, y_0) = \big(x(t_0), y(t_0)\big)$ change when we change the parameter by dt. We move from the point (x_0, y_0) to the point $(x_0 + dx, y_0 + dy)$, which creates a segment that is the image of the infinitesimal dt. How long is it? We want to know the length of the vector

$$(x_0 + dx, y_0 + dy) - (x_0, y_0) = (dx, dy),$$

so

$$ds = \sqrt{dx^2 + dy^2}.$$

This formula can be used for an alternative transcription of a line integral, and we easily confirm that it is the same as the original one:

$$\begin{aligned} ds &= \sqrt{dx^2 + dy^2} = \sqrt{[x'(t)dt]^2 + [y'(t)dt]^2} \\ &= \sqrt{([x'(t)]^2 + [y'(t)]^2)dt^2} = \sqrt{[x'(t)]^2 + [y'(t)]^2}\, dt \\ &= \sqrt{[r_1'(t)]^2 + [r_2'(t)]^2}\, dt = \|(r_1'(t), r_2'(t))\|\, dt = \|\vec{r}\,'(t)\|\, dt. \end{aligned}$$

Similarly, in three dimensions one can encounter the formula $ds = \sqrt{dx^2 + dy^2 + dz^2}$. Readers who did not yet work with coordinate transformations in more variables may find it useful to check out the part on Coordinate transformations in section 7b.

The general substitution formula takes on specific forms when integrating a vector function. We start with the flow along a curve $\displaystyle\int_C F(\vec{x}) \bullet d\vec{s}$, that is, the circulation. We need the unit tangent vector $\vec{t}$ at a point $\vec{x} = \vec{r}(t)$, we easily find it as $\frac{\vec{r}\,'(t)}{\|\vec{r}\,'(t)\|}$, as the derivative is not zero by our assumption. Thus we can write

$$\int_C F(\vec{x}) \bullet d\vec{s} = \int_C F(\vec{x}) \bullet \vec{t}(\vec{x})\, ds = \int_a^b F(\vec{r}(t)) \bullet \frac{\vec{r}\,'(t)}{\|\vec{r}\,'(t)\|}\|\vec{r}\,'(t)\|\, dt = \int_a^b F(\vec{r}(t)) \bullet \vec{r}\,'(t)\, dt.$$

This formula is useful when calculating integrals from the Stokes formula. Let's have a look at it in two dimensions.

$$\int_C F(\vec{x}) \bullet d\vec{s} = \int_a^b F(\vec{r}(t)) \bullet \vec{r}\,'(t)\,dt = \int_a^b (F_1, F_2) \bullet (r_1', r_2')\,dt$$
$$= \int_a^b F_1 r_1' + F_2 r_2'\,dt = \int_a^b F_1(\vec{r}(t)) r_1'(t)\,dt + \int_a^b F_2(\vec{r}(t)) r_2'(t)\,dt.$$

Now we recall the alternative take on substitution

$$\begin{array}{l} x = x(t) = r_1(t) \\ y = y(t) = r_2(t) \end{array} \implies \begin{array}{l} r_1'(t)dt = x'(t)dt = dx \\ r_2'(t)dt = y'(t)dt = dy \end{array}$$

and we see that we can do the back substitution.

$$\int_C F(\vec{x}) \bullet d\vec{s} = \int_C F_1 dx + \int_C F_2 dy = \int_C F_1 dx + F_2 dy.$$

Some authors prefer this form when stating the Green formula.

What about the flux across the boundary needed in the divergence theorem? There we need the outer unit normal vector to the curve. We have the tangent vector $\vec{r}\,' = (r_1', r_2')$ that provides two natural candidates for the perpendicular vector, $(r_2', -r_1')$ and $(-r_2', r_1')$. It is not hard to see that when we move in the direction $\vec{r}\,'$ in the plane, then the former vector points to the right and the latter to the left. For the proper orientation we need to see the interior of the set on the left, so the first vector points outward and thus it is the right choice. We then need to divide it by its norm to make it a unit vector.

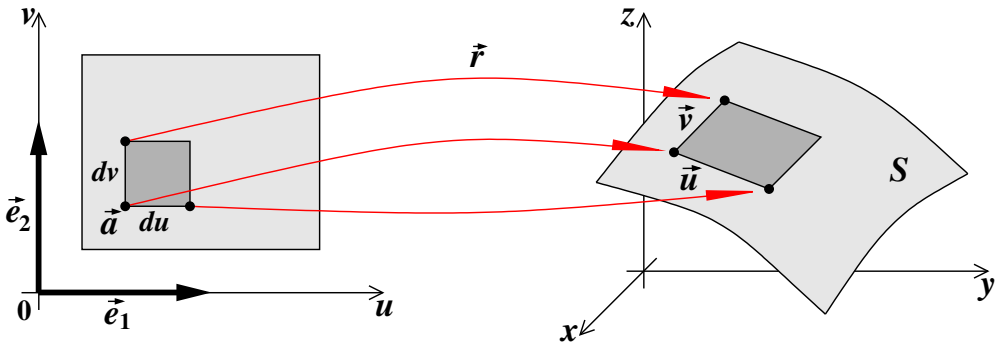
$$\int_C F(\vec{x}) \bullet \vec{n}(\vec{x})\,ds = \int_a^b F \bullet \frac{(r_2', -r_1')}{\|(r_2', -r_1')\|} \|\vec{r}\,'(t)\|\,dt = \int_a^b (F_1, F_2) \bullet (r_2', -r_1') \frac{\|(r_1', r_2')\|}{\|(r_2', -r_1')\|}\,dt$$
$$= \int_a^b F_1 r_2' - F_2 r_1' \frac{\sqrt{[r_1']^2 + [r_2']^2}}{\sqrt{[r_2']^2 + [-r_1']^2}}\,dt = \int_a^b F_1 r_2' - F_2 r_1'\,dt$$
$$= \int_a^b F_1 r_2'\,dt - \int_a^b F_2 r_1'\,dt = \int_C F_1\,dy - \int_C F_2\,dx = \int_C F_1 dy - F_2\,dx.$$

This expression can be sometimes found in the two-dimensional form of the divergence theorem.

Now we will look at the **surface integral**. In a surface integral we "add" volumes $f(\vec{x}) \cdot dS$. We decide to approach the surface S through parametrization, that is, a vector function $\vec{r}(u, v)$. How does the area infinitesimal in the world u, v transform from $\mathbb{R}^2$ to $\mathbb{R}^n$? More precisely, we are interested in the change of its area, because we need to know the size of the base for the volume $f(\vec{x}) \cdot dS$.

In the original space $\mathbb{R}^2[u, v]$ we now think of infinitesimals as of infinitely small squares. Their sides are equal, but they correspond to different coordinate axes, so we will denote them du and dv. The area is therefore $du \cdot dv$. We transfer this square using $\vec{r}$ to the target space $\mathbb{R}^n$ and investigate how its area changed.

We introduce precise notation. We start with a point $\vec{a}$ that also serves as one corner of our square. Its horizontal side of length du can be seen as the segment with endpoints $\vec{a}$ and $\vec{a} + du \cdot \vec{e}_1$. Similarly we see the vertical side as the segment connecting points $\vec{a}$ and $\vec{a} + dv \cdot \vec{e}_2$. By transforming these points we will see what happens to the square.



One can expect that perpendicular vectors need not be transformed into perpendicular vectors again, but this is not all. As the picture suggests, the images of sides may also bend, which would be unpleasant. However, when the initial square is really tiny, then this deformation of sides cannot be significant and we can assume that the image of the elementary square is a parallelogram. What are its sides?

The horizontal side of the square became a vector that goes from $\vec{r}(\vec{a})$ to $\vec{r}(\vec{a}+du\cdot\vec{e}_1)$. We thus have

$$\vec{u}=\vec{r}(\vec{a}+du\cdot\vec{e}_1)-\vec{r}(\vec{a}).$$

Since du is very small, we can apply linear approximation using directional derivative to the first term. It is derivative in the direction $\vec{e}_1$, so in fact it is the partial derivative with respect to u.

$$\vec{r}(\vec{a}+du\cdot\vec{e}_1)\approx\vec{r}(\vec{a})+D_{\vec{e}_1}\vec{r}(\vec{a})\cdot du=\vec{r}(\vec{a})+\frac{\partial\vec{r}}{\partial u}(\vec{a})\cdot du.$$

If this is too fast for the reader, we may look at it by coordinates of the function $\vec{r}=(r_1,\dots,r_n)$.

$$r_j(\vec{a}+du\cdot\vec{e}_1)=r_j(a_1+du,a_2)\approx r_j(a_1,a_2)+\frac{\partial r_j}{\partial u}(a_1,a_2)du=r_j(\vec{a})+\frac{\partial r_j}{\partial u}(\vec{a})du.$$

Now we put the vector back together.

$$\begin{aligned}\vec{r}(\vec{a}+du\cdot\vec{e}_1)&\approx\Big(r_1(\vec{a})+\frac{\partial r_1}{\partial u}(\vec{a})du,r_2(\vec{a})+\frac{\partial r_2}{\partial u}(\vec{a})du,\dots\Big)\\&=\big(r_1(\vec{a}),r_2(\vec{a}),\dots\big)+\Big(\frac{\partial r_1}{\partial u}(\vec{a}),\frac{\partial r_2}{\partial u}(\vec{a}),\dots\Big)du=\vec{r}(\vec{a})+\frac{\partial\vec{r}}{\partial u}(\vec{a})\cdot du.\end{aligned}$$

This was useful as it reminded us that $\frac{\partial\vec{r}}{\partial u}(\vec{a})$ is actually a vector. The important conclusion is that one side of our parallelogram is

$$\vec{u}=\vec{r}(\vec{a}+du\cdot\vec{e}_1)-\vec{r}(\vec{a})\approx\frac{\partial\vec{r}}{\partial u}(\vec{a})\cdot du.$$

Similarly we find that the other side is

$$\vec{v}=\vec{r}(\vec{a}+dv\cdot\vec{e}_2)-\vec{r}(\vec{a})\approx\frac{\partial\vec{r}}{\partial v}(\vec{a})\cdot dv.$$

The vectors $\vec{u}$ and $\vec{v}$ form a parallelogram in $\mathbb{R}^n$. How do we find its area? This is not easy in general, so we will now restrict ourselves to surfaces in $\mathbb{R}^3$. There we have a formula that finds the area of a parallelogram as the magnitude of the cross product of neighboring sides.

$$dS=\|\vec{u}\times\vec{v}\|=\Big\|\Big(\frac{\partial\vec{r}}{\partial u}(\vec{a})du\Big)\times\Big(\frac{\partial\vec{r}}{\partial v}(\vec{a})dv\Big)\Big\|=\Big\|\frac{\partial\vec{r}}{\partial u}(\vec{a})\times\frac{\partial\vec{r}}{\partial v}(\vec{a})\Big\|\,du\,dv.$$

We should point out that the part $du\,dv$ came from intuitive calculations, but we should not take it literally, in particular it does not tell us that we have to integrate first with respect to u and then with respect to v. Rather, this tells us that we should expect at its place in the formula the infinitesimal $d[u,v]$ from $\mathbb{R}^2[u,v]$, and this infinitesimal can take many forms, for instance the one above but also $dv\,du$.

So now we know how to replace differential in the volume $f(\vec{x})dS$. For the function value we naturally use $f(\vec{x})=f(\vec{r}(u,v))$ and we are ready to put it all together and get the desired formula.

Unfortunately, it only works for $n = 3$. Fortunately, in applications we need exactly this dimension.

Fact.
Let f be a continuous function on some set $\Omega \subset \mathbb{R}^3$.
Let S be some surface in Ω given by parametrization $\vec{r}(u,v)$ for $(u,v) \in D$, where D is some set in $\mathbb{R}^2$ and $\vec{r}$ is a vector function $D \mapsto \Omega$ such that $\vec{r}\,[D] = S$, $\vec{r}$ is one-to-one on D and it has non-zero continuous partial derivatives on the interior of D. Then

$$\iint\limits_S f(\vec{x})\,dS = \iint\limits_D f(\vec{r}(u,v))\Big\|\frac{\partial\vec{r}}{\partial u} \times \frac{\partial\vec{r}}{\partial v}\Big\|\,d[u,v].$$

Again, we can see this as a specific form of substitution

$$\left|\begin{array}{l}\vec{x} = \vec{r}(u,v)\\ dS = \big\|\frac{\partial\vec{r}}{\partial u} \times \frac{\partial\vec{r}}{\partial v}\big\|\,du\,dv\end{array}\right|.$$

Also here we have a special case. Consider a surface S in $\mathbb{R}^3$ that is actually the graph of a differentiable function $g(x,y)$ on D. The natural parametrization $\vec{r}(x,y) = \big(x, y, g(x,y)\big)$ leads to the formula

$$\iint\limits_S f(x,y,z)\,dS = \iint\limits_D f\big(x,y,g(x,y)\big)\sqrt{\Big(\frac{\partial g}{\partial x}\Big)^2 + \Big(\frac{\partial g}{\partial y}\Big)^2 + 1}\,d[x,y].$$

How does parametrization work for vector functions? We are especially interested in the integral $\iint\limits_S F(\vec{x}) \bullet d\vec{S}$. Consider some parametrization $\vec{r}$. Then the cross product $\vec{N} = \frac{\partial\vec{r}}{\partial u} \times \frac{\partial\vec{r}}{\partial v}$ yields a vector that is perpendicular to the surface at the given point. By our assumptions it is not zero, hence we can use it to obtain the unit normal vector $\vec{n} = \frac{\vec{N}}{\|\vec{N}\|}$. Then we get

$$\begin{aligned}\iint\limits_S F(\vec{x}) \bullet d\vec{S} &= \iint\limits_S F(\vec{x}) \bullet \vec{n}\,dS = \iint\limits_D F(\vec{r}(u,v)) \bullet \frac{\vec{N}}{\|\vec{N}\|}\|\vec{N}\|\,d[u,v]\\ &= \iint\limits_D F(\vec{r}(u,v)) \bullet \vec{N}\,d[u,v] = \iint\limits_D F(\vec{r}(u,v)) \bullet \Big(\frac{\partial\vec{r}}{\partial u} \times \frac{\partial\vec{r}}{\partial v}\Big)\,d[u,v].\end{aligned}$$

This formula is useful when evaluating integrals in the Stokes formula. Let's take a closer look at the cross product.

$$\begin{aligned}\frac{\partial\vec{r}}{\partial u} \times \frac{\partial\vec{r}}{\partial v} &= \Big(\frac{\partial r_1}{\partial u}, \frac{\partial r_2}{\partial u}, \frac{\partial r_3}{\partial u}\Big) \times \Big(\frac{\partial r_1}{\partial v}, \frac{\partial r_2}{\partial v}, \frac{\partial r_3}{\partial v}\Big)\\ &= \left(\begin{vmatrix}\frac{\partial r_2}{\partial u} & \frac{\partial r_3}{\partial u}\\ \frac{\partial r_2}{\partial v} & \frac{\partial r_3}{\partial v}\end{vmatrix}, \begin{vmatrix}\frac{\partial r_3}{\partial u} & \frac{\partial r_1}{\partial u}\\ \frac{\partial r_3}{\partial v} & \frac{\partial r_1}{\partial v}\end{vmatrix}, \begin{vmatrix}\frac{\partial r_1}{\partial u} & \frac{\partial r_2}{\partial u}\\ \frac{\partial r_1}{\partial v} & \frac{\partial r_2}{\partial v}\end{vmatrix}\right).\end{aligned}$$

Therefore

$$\begin{aligned}\iint\limits_S F(\vec{x}) \bullet d\vec{S} = &\iint\limits_D F_1(\vec{r}(u,v)) \begin{vmatrix}\frac{\partial r_2}{\partial u} & \frac{\partial r_3}{\partial u}\\ \frac{\partial r_2}{\partial v} & \frac{\partial r_3}{\partial v}\end{vmatrix} d[u,v] + \iint\limits_D F_2(\vec{r}(u,v)) \begin{vmatrix}\frac{\partial r_3}{\partial u} & \frac{\partial r_1}{\partial u}\\ \frac{\partial r_3}{\partial v} & \frac{\partial r_1}{\partial v}\end{vmatrix} d[u,v]\\ &+ \iint\limits_D F_3(\vec{r}(u,v)) \begin{vmatrix}\frac{\partial r_1}{\partial u} & \frac{\partial r_2}{\partial u}\\ \frac{\partial r_1}{\partial v} & \frac{\partial r_2}{\partial v}\end{vmatrix} d[u,v].\end{aligned}$$

And now it comes. The parametrization $(u,v) \mapsto \vec{r}(u,v) = \big(r_1(u,v), r_2(u,v), r_3(u,v)\big)$ also sets up

three associated transformations $\mathbb{R}^2 \mapsto \mathbb{R}^2$. One of them works like this:

$$x = r_1(u, v),$$
$$y = r_2(u, v).$$

As we will see below, when this is taken as a substitution, then the differential transforms according to the formula

$$dx\,dy = \begin{vmatrix} \frac{\partial r_1}{\partial u} & \frac{\partial r_2}{\partial u} \\ \frac{\partial r_1}{\partial v} & \frac{\partial r_2}{\partial v} \end{vmatrix} du\,dv,$$

which agrees with the expression in the third integral. In the first integral we similarly use the substitution $y = r_2(u, v)$, $z = r_3(u, v)$ and in the last one we use $z = r_3(u, v)$, $x = r_1(u, v)$ (note the order!). In this way we obtain using back substitution the transcription

$$\iint_S F(\vec{x}) \bullet d\vec{S} = \iint_S F_1(\vec{x}) dy\,dz + \iint_S F_2(\vec{x}) dz\,dx + \iint_S F_3(\vec{x}) dx\,dy.$$

When we use this with curl(F) in place of F and join with the above transcription of line integral, we get an alternative form of the Stokes theorem:

$$\iint_S \Big(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\Big) dy\,dz + \Big(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\Big) dz\,dx + \Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big) dx\,dy = \oint_C F_1 dx + F_2 dy + F_3 dz.$$

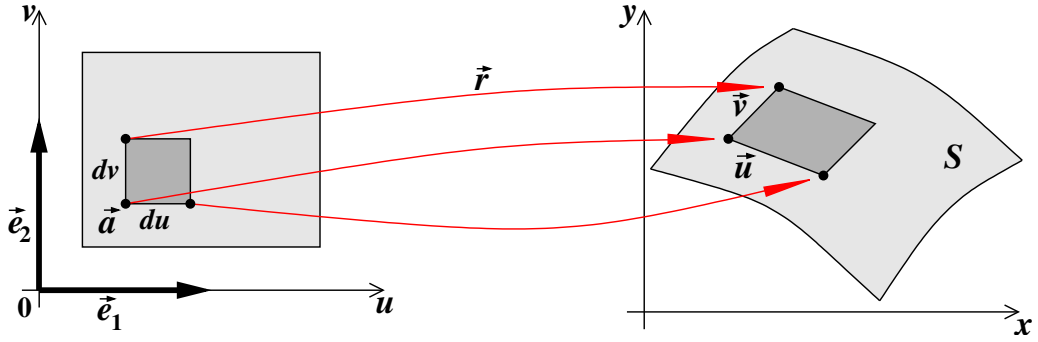
It is not my favourite form, but it seemed like a good idea to explain its meaning here in case you encounter it.

Now it is time to look at **substitution** for the usual multi-dimensional integral that we started with in this chapter. We have a function f defined on a set Ω in $\mathbb{R}^n$, and we will assume that it actually has dimension n, which we may interpret at the set having non-empty interior in $\mathbb{R}^n$, or having non-zero n-dimensional volume. Then the integral of f over Ω will not be automatically zero.

Substitution assumes that we described this set parametrically, so there is some vector function $\vec{r}$ on a certain set D such that $\Omega = \vec{r}[D]$. Because the set Ω is n-dimensional and we want the function $\vec{r}$ to be smooth, it follows that also the set D must be n-dimensional. This agrees with our intuition, to describe a set that really has n dimensions we need n parameters.

So we have a vector function $\vec{r}$: $\mathbb{R}^n \mapsto \mathbb{R}^n$. Our aim is to replace expressions of the form $f(\vec{x})\mathrm{d}[\vec{x}]$, where $\mathrm{d}[\vec{x}]$ is the n-dimensional infinitesimal of the space $\mathbb{R}^n$. For that we need to know how an infinitesimal from the domain D of the function $\vec{r}$ transfers into the target space; more precisely, we are interested in the change of its n-dimensional "volume" under this transformation.

We start with the case $n = 2$, so we have a vector function $\vec{r}(u, v)$ whose images are also two-dimensional. We want to transfer a square of sides ds, dt, which is something that we already did a while ago. However, now the dimension of the target space is two.



This means that at the end we have to use a different approach to find the area of a parallelogram. In two dimensions were are supposed to form a matrix from the two vectors and find the absolute value of its determinant. Vectors $\vec{u} = \frac{\partial \vec{r}}{\partial u} du$ and $\vec{v} = \frac{\partial \vec{r}}{\partial v} dv$ can be put both as rows and as columns,

we will choose the latter. Thus we obtain

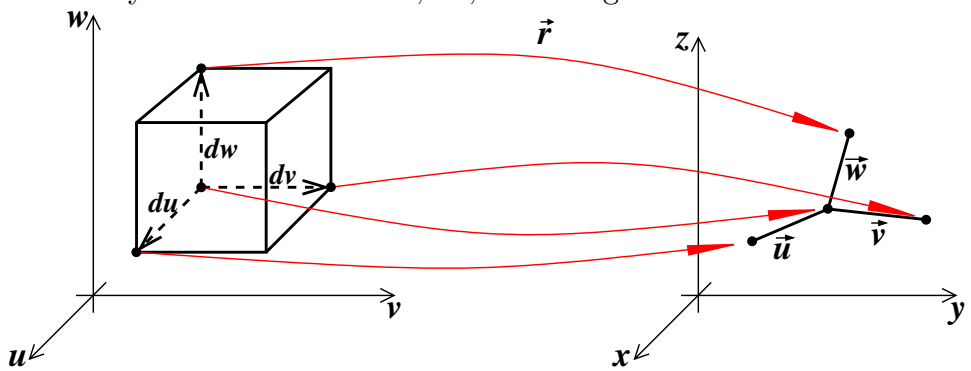
$$dS = \left|\det\begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \end{pmatrix}\right| = \left|\det\begin{pmatrix} \frac{\partial r_1}{\partial u}du & \frac{\partial r_1}{\partial v}dv \\ \frac{\partial r_2}{\partial u}du & \frac{\partial r_2}{\partial v}dv \end{pmatrix}\right| = \left|\det\begin{pmatrix} \frac{\partial r_1}{\partial u} & \frac{\partial r_1}{\partial v} \\ \frac{\partial r_2}{\partial u} & \frac{\partial r_2}{\partial v} \end{pmatrix}\right| \cdot du \cdot dv$$

By a remarkable coincidence we see the Jacobi matrix of the function $\vec{r}$ there. We introduced a special notation for its determinant and called it the jacobian, so we can write

$$dS = |\Delta_{\vec{r}}(\vec{a})|\, du\, dv.$$

Now we look at the case $n = 3$. This means that we are interested in some triple integral of a function f on a set $\Omega \subseteq \mathbb{R}^3$ of reasonable shape. It "adds" four-dimensional volumes $f(\vec{x}) \cdot dV$, where dV is the volume of an elementary three-dimensional cube.

Consider a parametrization $\vec{r}(u, v, w)$: $D \mapsto \mathbb{R}^3$, where $D \subset \mathbb{R}^3$ and $\Omega = \vec{r}[D]$. How does the volume of an elementary cube with sides du, dv, dw change when we transfer it using $\vec{r}$ into Ω?



We again write the individual edges using standard vectors $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ and the same calculations as before show that the key edges are carried to vectors

$$\vec{u} \approx \frac{\partial \vec{r}}{\partial u}(\vec{a}) \cdot du, \quad \vec{v} \approx \frac{\partial \vec{r}}{\partial v}(\vec{a}) \cdot dv, \quad \vec{w} \approx \frac{\partial \vec{r}}{\partial w}(\vec{a}) \cdot dw.$$

These vectors create a parallelepiped in $\mathbb{R}^3$. Its volume can also be found using the determinant of a matrix formed from vectors, therefore

$$dV = |\Delta_{\vec{r}}(\vec{a})|\, du\, dv\, dw.$$

Because this procedure for calculating the volume of an n-dimensional parallelepiped in $\mathbb{R}^n$ is universal, we can expect that our formula will be valid in all dimensions, that is, the n-dimensional infinitesimal $d[\vec{x}]$ should be replaced with the term $|\Delta_{\vec{r}}(\vec{a})|\, d[\vec{u}]$, where $d[\vec{u}]$ is the infinitesimal in the domain of the transformation $\vec{r}$ and in calculations will be replaced with the appropriate infinitesimals like $du\, dv$ or $dt\, du\, dv\, dw$.

Theorem.
Let f be a function continuous on a bounded set Ω on which the integral of f exists.
Consider a transformation (vector function) $\vec{r}$: $D[\vec{u}] \mapsto \Omega[\vec{x}]$ such that $\vec{r}[D] = \Omega$, $\vec{r}$ is continuous on D, it is one-on-one and continuously differentiable on the interior of D and its jacobian $\Delta_{\vec{r}}$ is not zero there.
Then

$$\int_{\Omega}\cdots\int f(\vec{x})\, d[\vec{x}] = \int_{D}\cdots\int f(\vec{r}(\vec{u}))|\Delta_{\vec{r}}(\vec{u})|\, d[\vec{u}].$$

This concludes our illustrated introduction to integrals.

Remark: Most arguments in this chapter (and previous ones) were intuitive, which means that they give us a feeling that things make sense, but we cannot truly trust them. Of course, the

feeling of understanding is important, but those who want to have clear conscience will check out correct proofs in some proper textbook, for instance in recommended lecture notes for your course.

Now we will review all these integrals in one monster example.

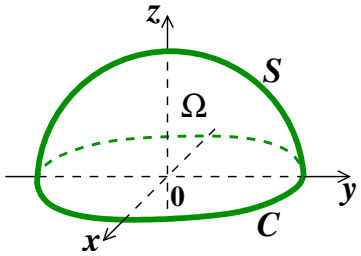
Example: Consider the solid

$$\Omega = \{(x, y, z) \in \mathbb{R}^3;\ 0 \le z \le \sqrt{1 - (x^2 + x^2)}\}.$$

The restriction $0 \le z$ shows that this set is above the xy plane. From above it is delimited by the surface

$$z = \sqrt{1 - (x^2 + y^2)} \implies 1 = x^2 + y^2 + z^2.$$

This is the equation of a unit sphere in $\mathbb{R}^3$, so in fact Ω is the upper half-ball of radius one, we will call it a dome.



We will be also interested in a surface S, namely the "roof" of this dome, that is, the upper half-sphere; and in a curve C, namely the boundary of S, or the perimeter of the base of the dome, in fact it is the unit circle in the plane xy.

Over these three sets we will integrate the function

$$f(x, y, z) = 4xy + z.$$

Before we start, we note that all three sets Ω, S, C have mirror symmetry with respect to the xz-plane and yz-plane. Moreover, also the two components of the function f are symmetric with respect to x and y, which means that it should be enough to find the integrals over one quarter of the sets, for instance the part satisfying $x \ge 0$ and $y \ge 0$, let's call it the principal quarter. Integrals over the remaining three quarters will then follow.

Indeed, the expression $4xy$ is odd with respect to x and also odd with respect to y. Graphically this means that the graph of $4xy$ to the left of the y-axis (for $x < 0$) has the same shape as the right half, but mirrored around the xy-plane, that is, with opposite signs. Obviously, integrating over some region on the left must yield the opposite answer compared to symmetric integration on the right.

We have the same symmetry to the left and to the right of the x-axis. These two symmetries combine and lead us to the following conclusion: The result of integrating $4xy$ over the principal quarter will be also valid for the opposite quarter, while for the two neighboring quarters the value of integral has the opposite sign. As a consequence, whether we integrate $4xy$ over the whole Ω, S, or C, the outcome must always be zero. In other words, when we will do proper calculations below, the first term of the function f must disappear in the process.

On the other hand, the expression z is even with respect to x and y, so the integrals over all four quarters will agree. It follows that if we want to know the values of the three integrals of f (over Ω, S, C), then it would be enough to integrate just the expression z over the principal quarter and then multiply the outcome by four. However, we are not really interested in answers, the point of this example is to showcase our methods, so we will do full calculations. Here they come.

1) **Line integral** over C.

The curve C is given by the equations $x^2+y^2=1$, $z=0$, which suggests the standard parametrization

$$\begin{aligned} x &= x, \\ y &= \sqrt{1-x^2}, \\ z &= 0 \end{aligned}$$

for one half of it and parametrization featuring $y=-\sqrt{1-x^2}$ for the other half, for both we take $x\in[-1,1]$. Approach with this notation is actually quite popular, but here we prefer to follow the notation used in theoretical musings above. We will therefore introduce the parametrization $\vec{r}(t)=(t,\sqrt{1-t^2},0)$ for $t\in[-1,1]$.

We should find the replacement for the differential ds.

$$\begin{aligned} \vec{r}\,'(t)=\Big(1,\frac{-t}{\sqrt{1-t^2}},0\Big) &\implies \|\vec{r}\,'\|=\sqrt{1^2+\Big(\frac{-t}{\sqrt{1-t^2}}\Big)^2+0^2}=\frac{1}{\sqrt{1-t^2}} \\ &\implies ds=\frac{1}{\sqrt{1-t^2}}\,dt. \end{aligned}$$

Similarly we work out the other half of the circle parametrized by $\vec{r}(t)=(t,-\sqrt{1-t^2},0)$ and obtain

$$\oint_C 4xy+z\,ds=\int_{-1}^{1}\big(4t\sqrt{1-t^2}+0\big)\frac{dt}{\sqrt{1-t^2}}+\int_{-1}^{1}\big(4t\big(-\sqrt{1-t^2}\,\big)+0\big)\frac{dt}{\sqrt{1-t^2}}.$$

We evaluate:

$$\oint_C 4xy+z\,ds=\int_{-1}^{1}4t\,dt+\int_{-1}^{1}-4t\,dt=\Big[2t^2\Big]_{-1}^{1}+\Big[-2t^2\Big]_{-1}^{1}=0+0=0.$$

Alternative: In many settings, the best parametrization for a circle is given by polar coordinates. How would it work here?

$$\begin{aligned} x &= \cos(\varphi), \\ y &= \sin(\varphi), \\ z &= 0. \end{aligned}$$

That is, we work with the vector function $\vec{r}(\varphi)=\big(\cos(\varphi),\sin(\varphi),0\big)$ for $t\in[0,2\pi]$. We see one advantage of this approach, we get the whole circle by one formula.

We work out the relationship between differentials:

$$\begin{aligned} \vec{r}\,'(\varphi)=\big(-\sin(\varphi),\cos(\varphi),0\big) &\implies \|\vec{r}\,'(\varphi)\|=\sqrt{\sin^2(\varphi)+\cos^2(\varphi)+0^2}=1 \\ &\implies ds=d\varphi. \end{aligned}$$

Therefore

$$\oint_C 4xy+z\,ds=\int_0^{2\pi}4\cos(\varphi)\sin(\varphi)+0\,d\varphi=\int_0^{2\pi}2\sin(2\varphi)\,d\varphi=\Big[-\cos(2\varphi)\Big]_0^{2\pi}=-1+1=0.$$

2) **Surface integral** over S.

The surface S is described by the conditions $x^2+y^2+z^2=1$, $z\geq 0$, which can be equivalently replaced with $z=\sqrt{1-x^2-y^2}$. We can therefore see S as the graph of the function $g(x,y)=\sqrt{1-x^2-y^2}$. Above we saw a specific formula for this type of surface, let's work it out in detail here so that we can see where it comes from.

Formally we introduce the parametric description $\vec{r}(u,v) = \left(u, v, \sqrt{1-u^2-v^2}\right)$, where (u,v) is taken from the unit circle in $\mathbb{R}^2$, let's call it D. That is, for D we take the set of parameters (u,v) satisfying $u^2+v^2 \le 1$.

We need to find the normal vector:

$$\begin{aligned}\vec{N} &= \frac{\partial \vec{r}}{\partial u} \times \frac{\partial \vec{r}}{\partial v} = \left(1, 0, \frac{-u}{\sqrt{1-u^2-v^2}}\right) \times \left(0, 1, \frac{-v}{\sqrt{1-u^2-v^2}}\right) \\ &= \left(\frac{u}{\sqrt{1-u^2-v^2}}, \frac{v}{\sqrt{1-u^2-v^2}}, 1\right).\end{aligned}$$

Therefore

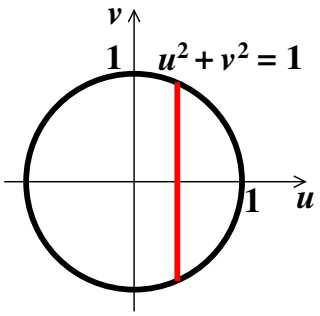
$$\begin{aligned}dS = \|\vec{N}\|\, dS[u,v] &= \sqrt{\left(\frac{u}{\sqrt{1-u^2-v^2}}\right)^2 + \left(\frac{v}{\sqrt{1-u^2-v^2}}\right)^2 + 1}\, dS[u,v] \\ &= \frac{1}{\sqrt{1-u^2-v^2}}\, dS[u,v].\end{aligned}$$

We used $dS[u,v]$ to denote the planar differential in $\mathbb{R}^2$ with variables u, v.

We can set up the integral:

$$\iint_S 4xy + z\, dS = \iint_D 4uv + \sqrt{1-u^2-v^2}\, \frac{dS[u,v]}{\sqrt{1-u^2-v^2}} = \iint_D \frac{4uv}{\sqrt{1-u^2-v^2}} + 1\, dS[u,v].$$

Note that we ended up with integrating a function of two variables over a proper two-dimensional set, so it is a standard double integral. We evaluate it by reducing its dimension using slices of the set D. This set is the unit disc, which is a routine situation. We slice it by fixing u between -1 and 1, then v ranges between $-\sqrt{1-u^2}$ and $\sqrt{1-u^2}$.



$$\iint_S 4xy + z\, dS = \int_{-1}^{1} \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} \frac{4uv}{\sqrt{1-u^2-v^2}} + 1\, dv\, du.$$

The inside integral (its first term) can be handled using substitution:

$$\int \frac{4uv}{\sqrt{1-u^2-v^2}}\, dv = \left| \begin{array}{r} w = \sqrt{1-u^2-v^2} \\ dw = \frac{-v}{\sqrt{1-u^2-v^2}}\, dv \end{array} \right| = \int -4u\, dw = -4uw = -4u\sqrt{1-u^2-v^2}.$$

Therefore

$$\begin{aligned}\iint_S 4xy + z\, dS &= \int_{-1}^{1} \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} \frac{4uv}{\sqrt{1-u^2-v^2}} + 1\, dv\, du \\ &= \int_{-1}^{1} \left[-4u\sqrt{1-u^2-v^2} + v\right]_{v=-\sqrt{1-u^2}}^{v=\sqrt{1-u^2}} du = \int_{-1}^{1} 0 + 2\sqrt{1-u^2}\, du.\end{aligned}$$

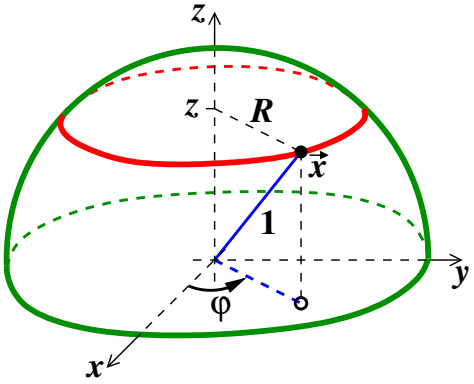
The standard approach to this integral is using the indirect substitution $u = \sin(w)$. When working it out we will use the fact that $\cos(w) \ge 0$ on $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ to simplify

$$\sqrt{\cos^2(w)} = |\cos(w)| = \cos(w).$$

Let's go.

$$\iint\limits_S 4xy + z\,dS = \int\limits_{-1}^{1} 2\sqrt{1-u^2}\,du = \int\limits_{-\pi/2}^{\pi/2} 2\sqrt{\cos^2(w)}\cos(w)\,dw = \int\limits_{-\pi/2}^{\pi/2} 2\cos^2(w)\,dw$$
$$= \int\limits_{-\pi/2}^{\pi/2} 1+\sin(2w)\,dw = \Big[w - \frac{1}{2}\cos(2w)\Big]_{-\pi/2}^{\pi/2} = \pi.$$

Alternative: We need to go through all points of the half-sphere, and another possible approach is to organize them into circles, that is, we will use horizontal slicing. Fixing some $z \in [0,1]$ determines a circle in this half-sphere whose radius is $R = \sqrt{1-z^2}$. Now it just remains to go through all points of this circle, which can be easily accomplished using a suitable angle φ. Formally, φ is the angle between the x-axis and the ray connecting the origin with the projection of a specific point on our circle into the xy-plane.



The formulas for parametric description are as follows:

$$x = \sqrt{1-z^2}\cos(\varphi),$$
$$y = \sqrt{1-z^2}\sin(\varphi),$$
$$z = z.$$

Formally we will work with the vector function

$$\vec{r}(\varphi,t) = \big(\sqrt{1-t^2}\cos(\varphi), \sqrt{1-t^2}\sin(\varphi), t\big).$$

Here $\varphi \in [0,2\pi]$ and $t \in [0,1]$. Precisely, we obtain the upper half-sphere by taking (φ,t) from the set $D = [0,2\pi]\times[0,1]$. We proceed working out this surface substitution:

$$\vec{N} = \frac{\partial\vec{r}}{\partial\varphi}\times\frac{\partial\vec{r}}{\partial t} = \Big(-\sqrt{1-t^2}\sin(\varphi), \sqrt{1-t^2}\cos(\varphi), 0\Big)\times\Big(\frac{-t\cos(\varphi)}{\sqrt{1-t^2}}, \frac{-t\sin(\varphi)}{\sqrt{1-t^2}}, 1\Big)$$
$$= \big(\sqrt{1-t^2}\cos(\varphi), \sqrt{1-t^2}\sin(\varphi), t\sin^2(\varphi)+t\cos^2(\varphi)\big) = \big(\sqrt{1-t^2}\cos(\varphi), \sqrt{1-t^2}\sin(\varphi), t\big).$$

Therefore

$$dS = \|\vec{N}\|\,dS[\varphi,t] = \sqrt{(1-t^2)\cos^2(\varphi)+(1-t^2)\sin^2(\varphi)+t^2}\,dS[\varphi,t]$$
$$= \sqrt{1-t^2+t^2}\,d[\varphi,t] = dS[\varphi,t].$$

We can set up the appropriate integral:

$$\iint\limits_S 4xy + z\,dS = \iint\limits_D 4\sqrt{1-t^2}\cos(\varphi)\sqrt{1-t^2}\sin(\varphi) + t\,dS[\varphi,t].$$

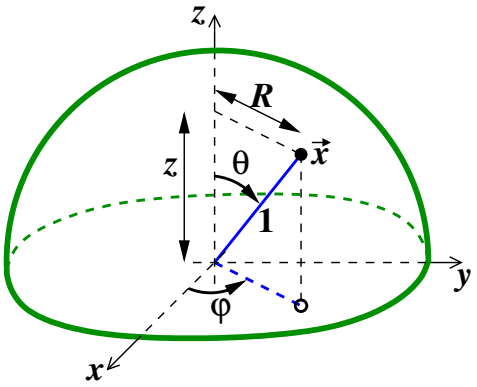
Again, we have a standard double integral on the right. This time D is a rectangle, so we can use

any order of integration and integrating limits are obvious.

$$\iint_S 4xy + z\,dS = \int_0^1 \int_0^{2\pi} 2(1-t^2)\sin(2\varphi) + t\,d\varphi\,dt = \int_0^1 \Big[-(1-t^2)\cos(2\varphi) + t\varphi\Big]_{\varphi=0}^{\varphi=2\pi} dt$$
$$= \int_0^1 0 + 2\pi t\,dt = \Big[\pi t^2\Big]_0^1 dt = \pi.$$

This seems to work better.

Alternative: A very convenient description of a sphere comes from polar coordinates in there dimensions, that is, spherical coordinates. Given a point on a unit sphere, we use θ for the polar angle describing how much this point leans away from the z-axis. Knowing θ we already know the z-coordinate of the point, and we also know the radius of the circle on which the point lies. We fix its position on this circle by an azimuth φ as in the previous approach.



We get

$$x = \cos(\varphi)\sin(\theta),$$
$$y = \sin(\varphi)\sin(\theta),$$
$$z = \cos(\theta),$$

that is, we work with

$$\vec{r}(\varphi,\theta) = \big(\cos(\varphi)\sin(\theta), \sin(\varphi)\sin(\theta), \cos(\theta)\big).$$

We obtain the whole upper half-sphere by taking $\varphi \in [0, 2\pi]$ and $\theta \in \big[0, \frac{\pi}{2}\big]$, that is, we take $D = [0, 2\pi] \times \big[0, \frac{\pi}{2}\big]$.

Let's go:

$$\vec{N} = \frac{\partial\vec{r}}{\partial\varphi} \times \frac{\partial\vec{r}}{\partial\theta} = \Big(-\sin(\varphi)\sin(\theta), \cos(\varphi)\sin(\theta), 0\Big) \times \Big(\cos(\varphi)\cos(\theta), \sin(\varphi)\cos(\theta), -\sin(\theta)\Big)$$
$$= \big(-\cos(\varphi)\sin^2(\theta), -\sin(\varphi)\sin^2(\theta), -\sin^2(\varphi)\sin(\theta)\cos(\theta) - \cos^2(\varphi)\sin(\theta)\cos(\theta)\big)$$
$$= \big(-\cos(\varphi)\sin^2(\theta), -\sin(\varphi)\sin^2(\theta), -\sin(\theta)\cos(\theta)\big).$$

Therefore

$$dS = \|\vec{N}\|\,dS[\varphi,\theta] = \sqrt{\cos^2(\varphi)\sin^4(\theta) + \sin^2(\varphi)\sin^4(\theta) + \sin^2(\theta)\cos^2(\theta)}\,dS[\varphi,\theta]$$
$$= \sqrt{\sin^4(\theta) + \sin^2(\theta)\cos^2(\theta)}\,dS[\varphi,\theta] = \sqrt{\sin^2(\theta)[\sin^2(\theta) + \cos^2(\theta)]}\,dS[\varphi,\theta]$$
$$= |\sin(\theta)|\,dS[\varphi,\theta] = \sin(\theta)\,dS[\varphi,\theta].$$

We were able to drop the absolute value because the sine is never negative for $\theta \in \big[0, \frac{\pi}{2}\big]$.

Again, integration over a rectangle is easy.

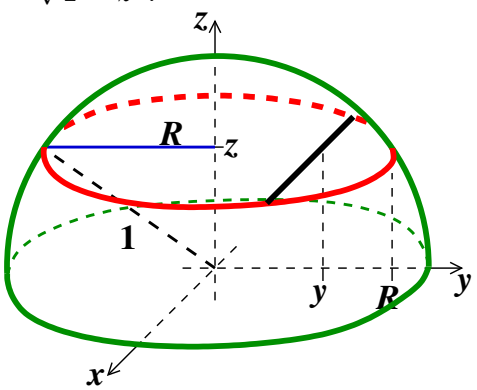
$$\begin{aligned}\iint_S 4xy + z\,dS &= \iint_D \big(4\cos(\varphi)\sin(\theta)\sin(\varphi)\sin(\theta) + \cos(\theta)\big)\sin(\theta)\,dS[\varphi,\theta] \\ &= \int_0^{\pi/2}\int_0^{2\pi} 2\sin(2\varphi)\sin^3(\theta) + \tfrac{1}{2}\sin(2\theta)\,d\varphi\,d\theta \\ &= \int_0^{\pi/2} \Big[-\cos(2\varphi)\sin^3(\theta) + \tfrac{1}{2}\sin(2\theta)\varphi\Big]_{\varphi=0}^{\varphi=2\pi} d\theta \\ &= \int_0^{\pi/2} 0 + \pi\sin(2\theta)\,d\theta = \Big[-\frac{1}{2}\pi\cos(2\theta)\Big]_0^{\pi/2} = \pi.\end{aligned}$$

This was also quite OK.

3) **Integral** over Ω.

Since Ω, that is, our dome has a non-empty interior, it is a truly three-dimensional set. Thus the integral over Ω is a standard triple integral. We will use the standard approach, that is, reduction of dimension through slicing.

We should choose the first direction, and it does seem like a good idea to slice first by fixing $z \in [0,1]$, that is, slicing parallel to the xy-plane, because such a slice is a circle. We easily find the radius of this circle to be $R = \sqrt{1-z^2}$.



Now we need to go through points of this circle, so we cut it into strips. We fix y, naturally between $-R$ and R, this determines a strip where x can run between $-\sqrt{R^2-y^2}$ and $\sqrt{R^2-y^2}$. We get the following repeated integral:

$$\iiint_\Omega 4xy + z\,dV = \int_0^1 \int_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} \int_{-\sqrt{1-z^2-y^2}}^{\sqrt{1-z^2-y^2}} 4xy + z\,dx\,dy\,dz.$$

As usual, we start from the inside.

$$\begin{aligned}\iiint_\Omega 4xy + z\,dV &= \int_0^1 \int_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} \Big[2x^2y + zx\Big]_{x=-\sqrt{1-z^2-y^2}}^{x=\sqrt{1-z^2-y^2}} dy\,dz \\ &= \int_0^1 \int_{-\sqrt{1-z^2}}^{\sqrt{1-z^2}} 0 + 2z\sqrt{1-z^2-y^2}\,dy\,dz.\end{aligned}$$

The standard approach now is the substitution $y = \sqrt{1-z^2}\sin(w)$, $dy = \sqrt{1-z^2}\cos(w)\,dw$, where

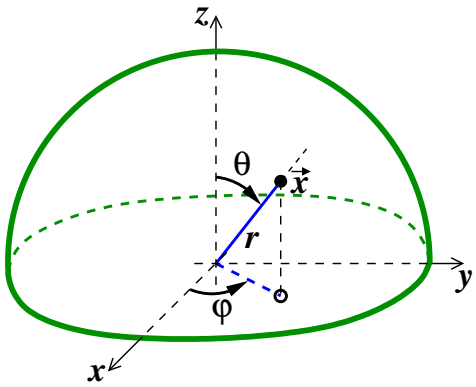
z is taken as a parameter, so we proceed as follows.

$$\begin{aligned}
\iiint_\Omega 4xy + z\,dV &= \int_0^1 \int_{-\pi/2}^{\pi/2} 2z\sqrt{(1-z^2)-(1-z^2)\sin^2(w)}\sqrt{1-z^2}\cos(w)\,dw\,dz \\
&= \int_0^1 \int_{-\pi/2}^{\pi/2} 2z\sqrt{1-z^2}\sqrt{1-\sin^2(w)}\sqrt{1-z^2}\cos(w)\,dw\,dz \\
&= \int_0^1 \int_{-\pi/2}^{\pi/2} 2z(1-z^2)\cos^2(w)\,dw\,dz = \int_0^1 \int_{-\pi/2}^{\pi/2} z(1-z^2)\big(1+\cos(2w)\big)\,dw\,dz \\
&= \int_0^1 \Big[z(1-z^2)\big(w-\tfrac{1}{2}\sin(2w)\big)\Big]_{w=-\pi/2}^{w=\pi/2} dz = \int_0^1 \pi z(1-z^2)\,dz \\
&= \left|\begin{array}{l}\omega = 1-z^2 \\ d\omega = -2z\,dz\end{array}\right| = \int_1^0 -\frac{1}{2}\pi\omega\,d\omega = \Big[-\frac{1}{2}\pi\,\frac{1}{2}\omega^2\Big]_1^0 = \frac{\pi}{4}.
\end{aligned}$$

Is there another way? We could also try vertical slices through our dome that are shaped like half-circles, but it leads to analogous integrals. It will be more interesting to try different approaches.

Alternative: We will apply the spherical coordinates

$$\begin{aligned}
x &= r\cos(\varphi)\sin(\theta), \\
y &= r\sin(\varphi)\sin(\theta), \\
z &= r\cos(\theta).
\end{aligned}$$



Since the number of initial and outgoing variables agree, this is a substitution. Formally we work with

$$\vec{r}(r,\varphi,\theta) = \big(r\cos(\varphi)\sin(\theta), r\sin(\varphi)\sin(\theta), r\cos(\theta)\big).$$

We obtain the whole upper half-ball by taking parameters $r \in [0,1]$, $\varphi \in [0,2\pi]$ and $\theta \in \big[0,\frac{\pi}{2}\big]$.

Formally we would consider $\vec{r}$ as a mapping from $D = [0,1]\times[0,2\pi]\times\big[0,\frac{\pi}{2}\big]$ onto Ω, both are three-dimensional sets.

We need to find the jacobian.

$$\begin{aligned}
\Delta_{\vec{r}} &= \det\begin{pmatrix} \frac{\partial r_1}{\partial r} & \frac{\partial r_1}{\partial \varphi} & \frac{\partial r_1}{\partial \theta} \\ \frac{\partial r_2}{\partial r} & \frac{\partial r_2}{\partial \varphi} & \frac{\partial r_2}{\partial \theta} \\ \frac{\partial r_3}{\partial r} & \frac{\partial r_3}{\partial \varphi} & \frac{\partial r_3}{\partial \theta} \end{pmatrix} = \det\begin{pmatrix} \cos(\varphi)\sin(\theta) & -r\sin(\varphi)\sin(\theta) & r\cos(\varphi)\cos(\theta) \\ \sin(\varphi)\sin(\theta) & r\cos(\varphi)\sin(\theta) & r\sin(\varphi)\cos(\theta) \\ \cos(\theta) & 0 & -r\sin(\theta) \end{pmatrix} \\
&= -r^2\cos^2(\varphi)\sin^3(\theta) - r^2\sin^2(\varphi)\sin(\theta)\cos^2(\theta) \\
&\quad - r^2\cos^2(\varphi)\sin(\theta)\cos^2(\theta) - r^2\sin^2(\varphi)\sin^3(\theta) \\
&= -r^2\big([\cos^2(\varphi)+\sin^2(\varphi)]\sin^3(\theta) + [\sin^2(\varphi)+\cos^2(\varphi)]\sin(\theta)\cos^2(\theta)\big) \\
&= -r^2\big(\sin^3(\theta)+\sin(\theta)\cos^2(\theta)\big) = -r^2\sin(\theta)\big(\sin^2(\theta)+\cos^2(\theta)\big) \\
&= -r^2\sin(\theta).
\end{aligned}$$

We apply absolute value, take into account that $\theta \in \big[0,\frac{\pi}{2}\big]$ and therefore the sine is positive or zero there, and obtain

$$dV = |-r^2\sin(\theta)|\,dV[r,\varphi,\theta] = r^2\sin(\theta)\,dV[r,\varphi,\theta].$$

We are ready to apply the substitution. The new integral will be a triple integral over D, which

is a three-dimensional box, so we are free to use any integration order we want, and limits for integration are obvious.

$$\begin{aligned}
\iiint_\Omega 4xy + z\,dV &= \iiint_D \big(4r\cos(\varphi)\sin(\theta) r\sin(\varphi)\sin(\theta) + r\cos(\theta)\big) r^2\sin(\theta)\,dV[r,\varphi,\theta] \\
&= \int_0^1 \int_0^{\pi/2} \int_0^{2\pi} \big(2r^2\sin(2\varphi)\sin^2(\theta) + r\cos(\theta)\big) r^2\sin(\theta)\,d\varphi\,d\theta\,dr \\
&= \int_0^1 \int_0^{\pi/2} \int_0^{2\pi} 2r^4\sin(2\varphi)\sin^3(\theta) + r^3\cos(\theta)\sin(\theta)\,d\varphi\,d\theta\,dr \\
&= \int_0^1 \int_0^{\pi/2} \Big[-r^4\cos(2\varphi)\sin^3(\theta) + r^3\cos(\theta)\sin(\theta)\varphi\Big]_{\varphi=0}^{\varphi=2\pi} d\theta\,dr \\
&= \int_0^1 \int_0^{\pi/2} 0 + 2\pi r^3\cos(\theta)\sin(\theta)\,d\theta\,dr = \int_0^1 \int_0^{\pi/2} \pi r^3\sin(2\theta)\,d\theta\,dr \\
&= \int_0^1 \Big[-\frac{1}{2}\pi r^3\cos(2\theta)\Big]_{\theta=0}^{\theta=\pi/2} dr = \int_0^1 \pi r^3\,dr = \Big[\pi\frac{1}{4}r^4\Big]_0^1 = \frac{\pi}{4}.
\end{aligned}$$

Alternative: Another interesting approach is to use cylindrical coordinates. The standard form is

$$\begin{aligned}
x &= r\cos(\varphi), \\
y &= r\sin(\varphi), \\
z &= z.
\end{aligned}$$

This corresponds to the vector function

$$\vec{r}(z,r,\varphi) = \big(r\cos(\varphi), r\sin(\varphi), z\big).$$

We need to find the right domain D for $\vec{r}$ so that the image is the dome. There are obvious bounds $z \in [0,1]$, $r \in [0,1]$ and $\varphi \in [0,2\pi]$. However, if we took the box $[0,1]\times[0,1]\times[0,2\pi]$ as D, then the image would be a cylinder. We need to restrict the radius r depending on elevation, that is, on z, and we easily find that the right bound is $0 \le r \le \sqrt{1-z^2}$. We will therefore use the set

$$D = \{(z,r,\varphi) \in \mathbb{R}^3;\ 0 \le \varphi \le 2\pi,\ 0 \le z \le 1, 0 \le r \le \sqrt{1-z^2}\}.$$

Because the bound on r depends on z, the order of integration will now matter, r must be "deeper" in the integral than z so that it gets integrated sooner.

$$\Delta_{\vec{r}} = \det\begin{pmatrix} \frac{\partial r_1}{\partial z} & \frac{\partial r_1}{\partial r} & \frac{\partial r_1}{\partial \varphi} \\ \frac{\partial r_2}{\partial z} & \frac{\partial r_2}{\partial r} & \frac{\partial r_2}{\partial \varphi} \\ \frac{\partial r_3}{\partial z} & \frac{\partial r_3}{\partial r} & \frac{\partial r_3}{\partial \varphi} \end{pmatrix} = \det\begin{pmatrix} 0 & \cos(\varphi) & -r\sin(\varphi) \\ 0 & \sin(\varphi) & r\cos(\varphi) \\ 1 & 0 & 0 \end{pmatrix} = r\cos^2(\varphi) + r\sin^2(\varphi) = r$$

$$\implies dV = r\,dV[z,r,\varphi].$$

Therefore

$$\iiint_\Omega 4xy + z\,dV = \iiint_D \big(4r\cos(\varphi) r\sin(\varphi) + z\big) r\,dV[z,r,\varphi]$$

$$= \int_0^1 \int_0^{\sqrt{1-z^2}} \int_0^{2\pi} 2r^3 \sin(2\varphi) + zr \, d\varphi \, dr \, dz$$
$$= \int_0^1 \int_0^{\sqrt{1-z^2}} \Big[-r^3 \cos(2\varphi) + zr\varphi\Big]_{\varphi=0}^{\varphi=2\pi} dr \, dz = \int_0^1 \int_0^{\sqrt{1-z^2}} 0 + 2\pi zr \, dr \, dz$$
$$= \int_0^1 \Big[\pi z r^2\Big]_{r=0}^{r=\sqrt{1-z^2}} dz = \int_0^1 \pi z(1-z^2)\, dz = \left| \begin{matrix} w = 1-z^2 \\ dw = -2z\, dz \end{matrix} \right|$$
$$= \int_1^0 -\frac{1}{2}\pi w \, dw = \Big[-\frac{1}{4}\pi w^2\Big]_1^0 = \frac{\pi}{4}.$$

Alternative: The fact that we were not free to choose the order of integration is a bit inconvenient. This can be circumvented by the following interesting trick. We will always take r from $[0,1]$, but modify the actual radius accordingly when calculating x and y. That is, we consider the following transformation:

$$x = r\sqrt{1-z^2}\cos(\varphi),$$
$$y = r\sqrt{1-z^2}\sin(\varphi),$$
$$z = z.$$

This corresponds to the vector function

$$\vec{r}(z,r,\varphi) = \big(r\sqrt{1-z^2}\cos(\varphi), r\sqrt{1-z^2}\sin(\varphi), z\big),$$

and now the domain for $\vec{r}$ is $z \in [0,1]$, $r \in [0,1]$ and $\varphi \in [0,2\pi]$, that is, $D = [0,1]\times[0,1]\times[0,2\pi]$.

$$\Delta_{\vec{r}} = \det \begin{pmatrix} \frac{\partial r_1}{\partial z} & \frac{\partial r_1}{\partial r} & \frac{\partial r_1}{\partial \varphi} \\ \frac{\partial r_2}{\partial z} & \frac{\partial r_2}{\partial r} & \frac{\partial r_2}{\partial \varphi} \\ \frac{\partial r_3}{\partial z} & \frac{\partial r_3}{\partial r} & \frac{\partial r_3}{\partial \varphi} \end{pmatrix} = \det \begin{pmatrix} r\frac{-z}{\sqrt{1-z^2}}\cos(\varphi) & \sqrt{1-z^2}\cos(\varphi) & -r\sqrt{1-z^2}\sin(\varphi) \\ r\frac{-z}{\sqrt{1-z^2}}\sin(\varphi) & \sqrt{1-z^2}\sin(\varphi) & r\sqrt{1-z^2}\cos(\varphi) \\ 1 & 0 & 0 \end{pmatrix}$$
$$= r(1-z^2)\cos^2(\varphi) + r(1-z^2)\sin^2(\varphi) = r(1-z^2)$$
$$\implies dV = r(1-z^2)\, dV[z,r,\varphi].$$

Therefore

$$\iiint_\Omega 4xy + z \, dV = \iiint_D \big(4r\sqrt{1-z^2}\cos(\varphi) r\sqrt{1-z^2}\sin(\varphi) + z\big) r(1-z^2)\, dV[z,r,\varphi]$$
$$= \int_0^1 \int_0^1 \int_0^{2\pi} 2r^3(1-z^2)^2\sin(2\varphi) + z(1-z^2)r \, d\varphi \, dr \, dz$$
$$= \int_0^1 \int_0^1 \Big[-r^3(1-z^2)^2\cos(2\varphi) + z(1-z^2)r\varphi\Big]_{\varphi=0}^{\varphi=2\pi} dr \, dz$$
$$= \int_0^1 \int_0^1 0 + 2\pi z(1-z^2)r \, dr \, dz = \int_0^1 \Big[\pi z(1-z^2)r^2\Big]_{r=0}^{r=1} dz$$
$$= \int_0^1 \pi z(1-z^2)\, dz = \frac{\pi}{4}.$$

We arrived at the same integral as in the previous calculation, so we just copied the answer.

With the direct application of cylindrical coordinates it was easier to find the jacobian, with this alternative approach it was a bit longer but we did not have to worry about the new integrating domain D and we were free to choose the order of integration. I'd say it's a tie.

△

When we look back at those calculations, we may notice that they all proceeded in the same way. We describe the given integrating domain by replacing the original coordinates with new ones (that is, parameters). The integral over the given set is then replaced by an integral over some new set D, which is treated as a new problem. The difference between substitution and special types of integrals shows up in the step when we replace differentials.

• If the number of original and new variables agree, then it is a substitution and we replace differentials using the jacobian.

• If the numbers of original and new variables do not agree and there is just one new variable (parameter), then we are in fact setting up a line integral. Then we have a specific formula for the differentials.

• If the numbers of original and new variables do not agree and there are two new variables (parameters), then we are in fact setting up a surface integral. Then we have a specific formula for the differentials, assuming that there were three original variables.

Most substitutions in this example were standard, which in particular refers to polar, spherical and cylindrical coordinates.

$$\left.\begin{aligned} x &= r\cos(\varphi),\\ y &= r\sin(\varphi) \end{aligned}\right\} \implies dS = r\,dS[r,\varphi]$$

$$\left.\begin{aligned} x &= \cos(\varphi)\sin(\theta),\\ y &= \sin(\varphi)\sin(\theta),\\ z &= \cos(\theta) \end{aligned}\right\} \implies dS = \sin(\theta)\,dS[\varphi,\theta]$$

$$\left.\begin{aligned} x &= r\cos(\varphi)\sin(\theta),\\ y &= r\sin(\varphi)\sin(\theta),\\ z &= r\cos(\theta) \end{aligned}\right\} \implies dV = r^2\sin(\theta)\,dV[r,\varphi,\theta]$$

$$\left.\begin{aligned} x &= r\cos(\varphi),\\ y &= r\sin(\varphi),\\ z &= z \end{aligned}\right\} \implies dV = r\,dV[z,r,\varphi]$$

In not-so-distant past students had to memorize the transformation formulas for differentials. Isn't progress great?

7. More on derivative

Here we will leave the illustrated introduction concept. Instead, we will first support our previous geometric deductions with some theory for the sake of completeness, and then introduce new important topics.

7a. Derivative and differential operators

In chapter 3 we introduced directional derivatives.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. Let $\vec{u}$ be a vector from $\mathbb{R}^n$.
We say that the function f is **differentiable** at point $\vec{a}$ in direction $\vec{u}$ if the limit $\lim\limits_{t\to 0}\big(\frac{f(\vec{a}+t\vec{u})-f(\vec{a})}{t}\big)$ exists and is finite.
Then we define the **(directional) derivative** of f at point $\vec{a}$ in direction $\vec{u}$ as

$$D_{\vec{u}}f(\vec{a}) = \lim_{t\to 0}\Big(\frac{f(\vec{a}+t\vec{u})-f(\vec{a})}{t}\Big).$$

We stated that directional derivative taken in one chosen direction behaves just like the usual derivative. We confirm it officially now.

Theorem.
Let f, g be functions defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Let $\vec{u}$ be a vector from $\mathbb{R}^n$.
Assume that f and g are differentiable at $\vec{a}$ in direction $\vec{u}$. Then also the functions cf, $f+g$, $f-g$, $f\cdot g$, $\frac{f}{g}$ (if $g(\vec{a}) \neq 0$) are differentiable at $\vec{a}$ in direction $\vec{u}$ and

$$\begin{aligned}
D_{\vec{u}}(cf)(\vec{a}) &= cD_{\vec{u}}f(\vec{a}),\\
D_{\vec{u}}(f+g)(\vec{a}) &= D_{\vec{u}}f(\vec{a}) + D_{\vec{u}}g(\vec{a}),\\
D_{\vec{u}}(f-g)(\vec{a}) &= D_{\vec{u}}f(\vec{a}) - D_{\vec{u}}g(\vec{a}),\\
D_{\vec{u}}(f\cdot g)(\vec{a}) &= D_{\vec{u}}f(\vec{a})\cdot g(\vec{a}) + f(\vec{a})\cdot D_{\vec{u}}g(\vec{a}),\\
D_{\vec{u}}\big(\tfrac{f}{g}\big)(\vec{a}) &= \frac{D_{\vec{u}}f(\vec{a})\cdot g(\vec{a}) - f(\vec{a})\cdot D_{\vec{u}}g(\vec{a})}{g^2(\vec{a})}.
\end{aligned}$$

We even have the following. Recall that the Mean value theorem for a function of one variable leads to the equality $f(y)-f(x) = f'(c)(y-x)$.

Theorem. (Mean value theorem)
Let $f\colon D(f) \mapsto \mathbb{R}$ be a function, where $D(f) \subseteq \mathbb{R}^n$. Let $\vec{x}, \vec{y} \in D(f)$.
Assume that the segment $[\vec{x}, \vec{y}]$ lies in $D(f)$ and that f has the derivative in direction $\vec{u} = \frac{\vec{y}-\vec{x}}{\|\vec{y}-\vec{x}\|}$ on this segment. Then there is $\vec{c} \in [\vec{x}, \vec{y}]$ such that

$$f(\vec{y}) - f(\vec{x}) = D_{\vec{u}}f(\vec{c})\cdot\|\vec{y}-\vec{x}\| = D_{\vec{y}-\vec{x}}f(\vec{c}).$$

By a segment $[\vec{a}, \vec{x}]$ in $\mathbb{R}^n$ we mean the set of all points of the form $\vec{a}+t(\vec{x}-\vec{a})$ for $t \in [0,1]$. The theorem therefore says that there is some $t \in (0,1)$ such that

$$f(\vec{y}) - f(\vec{x}) = D_{\vec{u}}f(t\vec{x}+(1-t)\vec{y})\cdot\|\vec{y}-\vec{x}\| = D_{\vec{y}-\vec{x}}f(t\vec{x}+(1-t)\vec{y}).$$

Since partial derivatives are just a special case of directional derivatives, the theorem on operations confirms that we can find partial derivatives using the usual rules. Speaking of partial derivatives, let's add a formal definition of spaces C^k.

Definition.
Let D be a region in $\mathbb{R}^n$.
We define $C^k(D)$ as the set of all functions $f\colon D \mapsto \mathbb{R}$ that have all partial derivatives up to the order k and these derivatives are continuous on D.

When finding partial derivatives of higher order we appreciate not having to worry about the order of differentiation. We now show a more general statement.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$.
Assume that for $\vec{u}, \vec{v} \in \mathbb{R}^n$ the derivatives $D_{\vec{u}}(D_{\vec{v}}f)$ and $D_{\vec{v}}(D_{\vec{u}}f)$ exist on some neighborhood of $\vec{a}$ and they are continuous at $\vec{a}$.
Then $D_{\vec{u}}(D_{\vec{v}}f)(\vec{a}) = D_{\vec{v}}(D_{\vec{u}}f)(\vec{a})$.

We did not address differentiation of a composed function yet. This is rather interesting in more dimensions, so we leave it to section 7b.

Gradient is essentially just a collection of partial derivatives. It is not difficult to check that the formulas from the theorem on operations can be composed into vectors. In this way we get rules for gradient.

Theorem.
Let f, g be functions defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$, let $c \in \mathbb{R}$. Assume that for both functions the gradients $\nabla f(\vec{a})$ and $\nabla g(\vec{a})$ exist. Then also the functions cf, $f+g$, $f \cdot g$, $\frac{f}{g}$ (if $g(\vec{a}) \neq 0$) have their gradients at $\vec{a}$ and

$$\begin{aligned}
\nabla(cf)(\vec{a}) &= c\nabla f(\vec{a}),\\
\nabla(f+g)(\vec{a}) &= \nabla f(\vec{a}) + \nabla g(\vec{a}),\\
\nabla(f\cdot g)(\vec{a}) &= \nabla f(\vec{a})g(\vec{a}) + f(\vec{a})\nabla g(\vec{a}),\\
\nabla\Big(\frac{f}{g}\Big)(\vec{a}) &= \frac{\nabla f(\vec{a})g(\vec{a}) - f(\vec{a})\nabla g(\vec{a})}{[g(\vec{a})]^2}.
\end{aligned}$$

As usual, we can interpret these formulas as equality of functions on sets, in this case we compare vector functions.

$$\begin{aligned}
\nabla(cf+g) &= c\nabla f + \nabla g,\\
\nabla(f\cdot g) &= \nabla f\cdot g + f\cdot\nabla g,\\
\nabla\Big(\frac{f}{g}\Big) &= \frac{\nabla f\cdot g - f\cdot\nabla g}{g^2}.
\end{aligned}$$

Do these formulas make sense? We know that vectors can be added and subtracted, and also multiplied by and divided by a scalar. Here it is useful to realize that the scalar can change, that is, we can multiply a vector by a function.

For instance, consider the functions $f(x,y) = xy^2$ and $g(x,y) = e^{x+y}$. Then ∇f is a vector function and we can multiply it by a scalar function g:

$$\nabla f \cdot g = (y^2, 2xy)\cdot e^{x+y} = (y^2e^{x+y}, 2xy\,e^{x+y}).$$

We see that the outcome is a vector function. This explains why formulas in the product rule and ratio rule make sense.

The usefulness of gradient has been apparent in chapter 3, here we will add one familiar theorem.

Theorem.
Let D be a region in $\mathbb{R}^n$ and $f \in C^1(D)$. If $\nabla f = \vec{0}$ on D, then f is constant on D.

Things get interesting with vector functions. Recall that for an open set D the symbol $[C(D)]^m$ stands for the set of all vector functions $F\colon D \mapsto \mathbb{R}^m$ that are continuous, that is, their components F_j belong to $C(D)$. Similarly, by $[C^k(D)]^m$ we mean the set of all vector functions $F\colon D \mapsto \mathbb{R}^m$ whose components F_j belong to $C^k(D)$, that is, they have partial derivatives up to order k that are moreover continuous on D. The following theorems are usually applied to functions from $[C^1(D)]^m$.

We introduced three differential operators, namely the Jacobi matrix as a generalization of the gradient, divergence and curl. We start by confirming linearity, which is the main thing expected of operators. We will write global versions.

Theorem.
Let $F, G\colon D \mapsto \mathbb{R}^m$ be vector functions, where $D \subseteq \mathbb{R}^n$. Let $c \in \mathbb{R}$. Then

$$J_{cF+G} = cJ_F + J_G,$$
$$\operatorname{div}(cF+G) = c\operatorname{div}(F) + \operatorname{div}(G),$$
$$\operatorname{curl}(cF+G) = c\operatorname{curl}(F) + \operatorname{curl}(G)$$

on sets where these expressions make sense.

As we figured out earlier, a vector function can be multiplied by a scalar function. Two vector functions F, G can be added and subtracted, we can also multiply them using the dot and cross product.

Theorem.
Let $F, G\colon D \mapsto \mathbb{R}^m$ be vector functions, where $D \subseteq \mathbb{R}^n$. Let f, g be functions $D \mapsto \mathbb{R}$. Then
(i) $J_{fG} = (\nabla f)^T G + f J_G$,
(ii) $\operatorname{div}(fG) = f\operatorname{div}(G) + \nabla f \bullet G$,
(iii) $\operatorname{div}\left(\frac{G}{f}\right) = \frac{f\operatorname{div}(G) - \nabla f \bullet G}{f^2}$,
(iv) $\operatorname{curl}(fG) = f\operatorname{curl}(G) + \nabla f \times G$,
(v) $\operatorname{curl}\left(\frac{G}{f}\right) = \frac{f\operatorname{curl}(G) - \nabla f \times G}{f^2}$,
(vi) $\nabla(F \bullet G) = F \cdot J_G + G \cdot J_F$,
(vii) $\operatorname{div}(F \times G) = G \bullet \operatorname{curl}(F) - F \bullet \operatorname{curl}(G)$,
(viii) $\operatorname{curl}(F \times G) = F \cdot \operatorname{div}(G) - G \cdot \operatorname{div}(F) + G \cdot (J_F)^T - F \cdot (J_G)^T$,
(ix) $\operatorname{curl}(f\nabla g) = \nabla f \times \nabla g$,
(x) $\operatorname{div}(\operatorname{curl}(F)) = 0$,
(xi) $\operatorname{curl}(\nabla f) = 0$
on sets where these expressions exist.

When interpreting these equalities one has to keep in mind that the gradient as an operator has priority over other operations, so for instance the right-hand side formula in (ix) is interpreted as $(\nabla f) \times (\nabla g)$.

All expressions make sense. For instance in that (ix), the gradient of g (a vector) gets multiplied by a scalar function, resulting in a vector function. We can apply curl to it, producing a vector again. On the right we see the cross product of two vectors, which also yields a vector. We thus observe that both sides of this equality represent a vector (of the same dimension), so it makes sense to compare them.

Let's look at RHS of (i). Gradient is taken as a row vector, so $(\nabla f)^T$ is a column vector, that is, a matrix of dimension $n \times 1$. The vector function $G = (G_1, \dots, G_n)$ is a row vector, that is, a matrix of dimension $1 \times n$. When we multiply these two matrices using the usual matrix multiplication, we obtain an $n \times n$ matrix, so we are allowed to add it to the Jacobi matrix of the same dimension multiplied by a function f that acts as a scalar here.

In the formula (vi) we see on the left a dot product of two vectors, so the outcome is a number, that is, a scalar function of more variables. By applying gradient we obtain a (row) vector. On the right we see a vector function, that is, a row vector, and it multiplies a matrix. In this case an $1 \times n$ matrix multiplies in the usual matrix way an $n \times n$ matrix, so the outcome is an $1 \times n$ matrix, that is, a row vector. The types of expressions on the left and on the right agree.

On the right in (ii) we see at the end a dot product of two row vectors, the outcome is a number, or more precisely, a scalar function, which is something that can be incorporated into a fraction.

Gradient can be considered a differential operator that we can denote symbolically as

$$\nabla = \Big(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n}\Big).$$

It is applied to a function in this way:

$$\Big(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\Big) f = \Big(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\Big).$$

The amazing thing is that it can be actually treated as a true vector. For instance, we know that the directional derivative can be calculated using gradient and the dot product $D_{\vec{u}} f = \nabla f \bullet \vec{u}$. Let's have a closer look:

$$\nabla f \bullet \vec{u} = \sum \frac{\partial f}{\partial x_i} u_i = \sum u_i \frac{\partial f}{\partial x_i} = \Big(\sum u_i \frac{\partial}{\partial x_i}\Big) f = (\vec{u} \bullet \nabla) f.$$

Applying the dot product to the vector $(u_1, \dots, u_n)$ and an abstract vector $\big(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\big)$ we obtain the differential operator $\sum u_i \frac{\partial}{\partial x_i}$ that applies the directional derivative. Symbolically, $D_{\vec{u}} = \vec{u} \bullet \nabla$.

Here we use the following convention. When we see $\frac{\partial}{\partial x_i} f$, we apply the derivative and it is done: $\frac{\partial f}{\partial x_i}$. The other order $f \frac{\partial}{\partial x_i}$ means that we do not differentiate, but modify a differential operator through multiplication. This means that is not true that $\nabla \bullet F$ would be the same as $F \bullet \nabla$ for a vector function F. Indeed, the first expression yields

$$\Big(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\Big) \bullet (F_1, \dots, F_n) = \frac{\partial}{\partial x_1} F_1 + \dots + \frac{\partial}{\partial x_n} F_n = \frac{\partial F_1}{\partial x_1} + \dots + \frac{\partial F_n}{\partial x_n}.$$

It is actually the divergence.

On the other hand, $F \bullet \nabla$ sets up a new differential operator that can be then applied to some function or a vector function (then it is applied to individual components):

$$(F_1, \dots, F_n) \bullet \Big(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\Big) = F_1 \frac{\partial}{\partial x_1} + \dots + F_n \frac{\partial}{\partial x_n}.$$

So it is actually the directional derivative, but now also the direction is taken as variable.

In a similar way we can create all popular operators. Let's make a list:

- $D_{\vec{u}} = \vec{u} \bullet \nabla$,
- $\operatorname{div}(F) = \nabla \bullet F$ for vector field F,
- $\operatorname{curl}(F) = \nabla \times F$ for a three-dimensional vector field F.

This language is quite popular. For instance, the last two observations from the previous theorem are often written as $\nabla \bullet (\nabla \times F) = 0$ and $\nabla \times (\nabla f) = 0$.

In fact, the rule (viii) is usually expressed as

• $\operatorname{curl}(F \times G) = F(\nabla \bullet G) - G(\nabla \bullet F) + (G \bullet \nabla)F - (F \bullet \nabla)G$.

How do we interpret the expressions on the right? Consider functions $F = (F_1, \dots, F_n)$ and $G = (G_1, \dots, G_n)$. Let's have a look at the first expression on the right. $\nabla \bullet G$ is the dot product of the gradient $\left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\right)$ with G, so it is the expression $\sum \frac{\partial}{\partial x_i} G_i$, that is, $\sum \frac{\partial G_i}{\partial x_i}$, those derivatives are applied right away. The result is a scalar function that multiplies a vector function, the first expression is therefore

$$\Big(F_1 \sum \frac{\partial G_i}{\partial x_i}, \dots, F_n \sum \frac{\partial G_i}{\partial x_i}\Big) = \Big(\sum F_1 \frac{\partial G_i}{\partial x_i}, \dots, \sum F_n \frac{\partial G_i}{\partial x_i}\Big).$$

Because we are used to commutativity, we could think that the third expression should be the same, but we do not multiply there. In the bracketed expression the dot products connects the vector G with the gradient vector, but in the order $G \bullet \nabla$, so we do not differentiate. Instead, the differential operator $\sum G_i \frac{\partial}{\partial x_i}$ gets created. To its right we then see the function that it gets applied to, and this function is a vector, so this operator is applied to each component separately:

$$\Big(\sum G_i \frac{\partial}{\partial x_i} F_1, \dots, \sum G_i \frac{\partial}{\partial x_i} F_n\Big) = \Big(\sum G_i \frac{\partial F_1}{\partial x_i}, \dots, \sum G_i \frac{\partial F_n}{\partial x_i}\Big).$$

Comparison shows that the first and the third expression on the right in (viii) are not the same. Similarly we can analyze the other two expressions.

We conclude this section by addressing differentiability of inverse functions. We know that when a sufficiently reasonable function f has its inverse function f_{-1}, then we can obtain its derivative from the derivative of f through the formula

$$[f_{-1}]'(y) = \frac{1}{f'(x)}.$$

Local existence of inverse function can be recognized using the test $f'(x) \neq 0$.

The natural generalization of this notion works with vector functions between spaces of identical dimension, for instance with coordinate transformations. We start with a more general result that can be proved quite easily.

Theorem.
Let F: $D \mapsto \mathbb{R}^n$ be a vector function, where $D \subseteq \mathbb{R}^n$. Assume that on some neighborhood of $\vec{a} \in D$ there exists an inverse function F^{-1} of F. Denote $\vec{b} = F(\vec{a})$.
If F is differentiable at $\vec{a}$ and F^{-1} is differentiable at $\vec{b}$, then

$$J_{F^{-1}}(\vec{b}) = J_F(\vec{a})^{-1}.$$

The conclusion also includes the information that $J_F(\vec{a})$ is invertible. We can also write $J_{F^{-1}}(\vec{b}) = J_F(F^{-1}(\vec{b}))^{-1}$. This notation assumes that we first substitute for $\vec{x}$ in J_F and then find the inverse matrix, but this can be also done in the opposite order $J_F^{-1}(F^{-1}(\vec{b}))$.

Now we will show a deeper version (with a difficult proof) where the existence of inverse function is deduced from an analogy of non-zero derivative.

Theorem.
Let F: $D \mapsto \mathbb{R}^n$ be a vector function, where $D \subseteq \mathbb{R}^n$. Assume that $F \in C^1(U)$ for some neighborhood U of a point $\vec{a} \in D$.
If $J_F(\vec{a})$ is not singular, then there is a neighborhood B of the point $\vec{a}$ such that F is one-to-one on B with image $C = F[B]$ and the inverse function F^{-1}: $C \mapsto B$ is from $C^1(C)$; moreover,

$$J_{F^{-1}} = {J_F}^{-1}(F^{-1}) \text{ on } C.$$

7b. Composition of functions, transformations

We start with an elegant general statement on derivative of a composed function.

Theorem.
Let F: $D \mapsto \mathbb{R}^m$ be a vector function, where $D \subseteq \mathbb{R}^n$ is an open set. Let G: $M \mapsto \mathbb{R}^p$ be a vector function, where $M \subseteq \mathbb{R}^m$ is an open set and $F[D] \subseteq M$.
If $F \in [C^1(D)]^m$ and $G \in [C^1(M)]^p$, then $G \circ F \in [C^1(D)]^p$ and for $\vec{a} \in D$ the following is true:

$$J_{G\circ F}(\vec{a}) = J_G(F(\vec{a}))J_F(\vec{a}).$$

Briefly: $J_{G\circ F} = J_G(F)J_F$ on D.

This result should be reminiscent of the popular chain rule $g'(f)f'$. It is worth to recall its key idea that is nicely apparent on differentiation of three composed functions.

$$\mathbb{R}[x] \xrightarrow{f} \mathbb{R}[y] \xrightarrow{g} \mathbb{R}[z] \xrightarrow{h} \mathbb{R}.$$

We can interpret the composed function $h \circ g \circ f$ as follows: Into the function $h(z)$ we substitute $z = g(y)$, and then we put $y = f(x)$ for y, obtaining $h(g(f(x)))$.

The derivative of this function is $[h \circ g \circ f]' = h'(g(f)) \cdot g'(f) \cdot f'$, but often we just write $[h \circ g \circ f]' = h' \cdot g' \cdot f'$. This formula captures the essence of the rule: If we want to differentiate the composed function $h(g(f(x)))$, then we start on the outside and step by step make our way to x inside, differentiating everything that we meet on the way. Those derivatives are then connected using multiplication. However, the resulting formula should have one variable only, namely x, so we cannot take $h'(z)g'(y)f'(x)$ for answer. We have to substitute for y and z, which brings us to the proper result.

How do we see these ideas in the formula for vector functions?

First we make sure that the expression in the theorem actually makes sense. We are multiplying two Jacobi matrices in the usual matrix way there. We can compose functions in the order F first, then G only when the dimension of the target space for F is the same as the dimension of the domain of G. In other words, the number of variables of G must match the number of components of the vector function F. This means that the number of columns of the matrix J_G matches the number of rows of the matrix J_F, and thus multiplying them in the order $J_G \cdot J_F$ makes sense.

But what does this multiplication actually mean? First we will clarify the situation and introduce names for variables in the spaces where we work.

$$\mathbb{R}^n[\vec{x}] \xrightarrow{F} \mathbb{R}^m[\vec{y}] \xrightarrow{G} \mathbb{R}^p.$$

So we have a function $G(\vec{y})$, and we substitute $\vec{y} = F(\vec{x})$ into it. We obtain a vector function whose i-th component is

$$G_i\big(F(\vec{x})\big) = G_i\big(F_1(x_1, \dots, x_n), F_2(x_1, \dots, x_n), \dots, F_m(x_1, \dots, x_n)\big).$$

The Jacobi matrix $J_{G\circ F}$ of this composed function is created by putting partial derivatives of this

component with respect to all variables on row i. Thus at position (i,j) we see the number

$$\frac{\partial G_i(F(\vec{x}))}{\partial x_j} = \frac{\partial}{\partial x_j} G_i\big(F_1(x_1,\dots,x_n), F_2(x_1,\dots,x_n),\dots,F_m(x_1,\dots,x_n)\big).$$

According to the formula from the theorem, this number can also be reached by taking the matrix J_G, substituting $F(\vec{x})$ for $\vec{y}$, and then multiplying its row i using the matrix way with the column j of the matrix J_F. This multiplication is actually the dot product of that row and column as vectors, so we can write

$$\frac{\partial G_i(F)}{\partial x_j} = \nabla G_i(F) \bullet \frac{\partial}{\partial x_j} F.$$

This elegant formula has its uses, but now we want to know what it actually says:

$$\begin{aligned}\frac{\partial G_i(F)}{\partial x_j} &= \Big(\frac{\partial G_i}{\partial y_1}(F),\dots,\frac{\partial G_i}{\partial y_m}(F)\Big) \bullet \Big(\frac{\partial F_1}{\partial x_j},\dots,\frac{\partial F_m}{\partial x_j}\Big)\\ &= \frac{\partial G_i}{\partial y_1}(F)\frac{\partial F_1}{\partial x_j} + \dots + \frac{\partial G_i}{\partial y_m}(F)\frac{\partial F_m}{\partial x_j}\\ &= \sum_{k=1}^{m} \frac{\partial G_i}{\partial y_k}(F)\frac{\partial F_k}{\partial x_j}.\end{aligned}$$

This is the key formula. To better appreciate it, we will now leave out the substitution of F:

$$\begin{aligned}&\frac{\partial}{\partial x_j} G_i(F_1(x_1,\dots,x_n), F_2(x_1,\dots,x_n),\dots,F_m(x_1,\dots,x_n))\\ &\quad= \frac{\partial G_i}{\partial y_1}\frac{\partial F_1}{\partial x_j} + \frac{\partial G_i}{\partial y_2}\frac{\partial F_2}{\partial x_j} + \dots + \frac{\partial G_i}{\partial y_m}\frac{\partial F_m}{\partial x_j}.\end{aligned}$$

This formula conforms to the basic idea of the chain rule. We are sitting with our derivative on the outside and we want to get to the variable x_j. The novelty now is that the variable x_j is present at more places, for instance we see m appearances of x_1. The formula says that we should attempt to reach all appearances of the variable x_j, which is possible through F_1, through F_2 etc. Each time we should conform to the chain rule idea and differentiate everything that we meet along the way, multiplying the derivatives. The path that we take to some specific position of x_j determines how we should differentiate G_i, for instance when going to x_j hidden in F_2 we need to go through the second variable of G_i, so we differentiate G_i with respect to y_2. The products of derivatives resulting from these different paths should be connected using addition.

One interesting version of the chain rule uses the notation $y_k = F_k(\vec{x})$. Then we can symbolically write

$$\begin{aligned}\frac{\partial G_i}{\partial x_j} &= \frac{\partial G_i}{\partial y_1}\frac{\partial F_1}{\partial x_j} + \frac{\partial G_i}{\partial y_2}\frac{\partial F_2}{\partial x_j} + \dots + \frac{\partial G_i}{\partial y_m}\frac{\partial F_m}{\partial x_j}\\ &= \frac{\partial G_i}{\partial y_1}\frac{\partial y_1}{\partial x_j} + \frac{\partial G_i}{\partial y_2}\frac{\partial y_2}{\partial x_j} + \dots + \frac{\partial G_i}{\partial y_m}\frac{\partial y_m}{\partial x_j}.\end{aligned}$$

Example: Consider the functions $F\colon \mathbb{R}^2 \mapsto \mathbb{R}^3$ and $G\colon \mathbb{R}^3 \mapsto \mathbb{R}^2$ defined as follows:

$$\begin{aligned}F(x,y) &= (x^2y^2, x^2+y^2, xy)\\ G(u,v,w) &= (uvw, u^2+v^2+w^2).\end{aligned}$$

We will find partial derivatives for the function $H = G \circ F = G(F)$. We can see it as taking $G(u,v,w)$ and substituting

$$\begin{aligned}u &= x^2y^2,\\ v &= x^2+y^2,\\ w &= xy.\end{aligned}$$

First we will find derivatives formally using the theorem. We start with Jacobi matrices for functions F and G. Their components are $F_1(x,y) = x^2y^2$, $F_2(x,y) = x^2+y^2$, $F_3(x,y) = xy$ and $G_1(u,v,w) = uvw$, $G_2(u,v,w) = u^2+v^2+w^2$. Then

$$J_F(x,y) = \begin{pmatrix} 2xy^2 & 2x^2y \\ 2x & 2y \\ y & x \end{pmatrix}, \qquad J_G(u,v,w) = \begin{pmatrix} vw & uw & uv \\ 2u & 2v & 2w \end{pmatrix}.$$

We need to substitute $u = F_1(x,y)$, $v = F_2(x,y)$, $w = F_3(x,y)$ in J_G, obtaining

$$J_G(F(x,y)) = \begin{pmatrix} (x^2+y^2)xy & x^2y^2xy & x^2y^2(x^2+y^2) \\ 2x^2y^2 & 2(x^2+y^2) & 2xy \end{pmatrix}.$$

So we have

$$\begin{aligned} J_H(x,y) = J_G(F(x,y)) \cdot J_F(x,y) &= \begin{pmatrix} (x^2+y^2)xy & x^3y^3 & x^2y^2(x^2+y^2) \\ 2x^2y^2 & 2(x^2+y^2) & 2xy \end{pmatrix} \begin{pmatrix} 2xy^2 & 2x^2y \\ 2x & 2y \\ y & x \end{pmatrix} \\ &= \begin{pmatrix} (x^2+y^2)2x^2y^3 + 2x^4y^3 + x^2y^3(x^2+y^2) & (x^2+y^2)2x^3y^2 + 2x^3y^4 + x^3y^2(x^2+y^2) \\ 4x^3y^4 + 4x(x^2+y^2) + 2xy^2 & 4x^4y^3 + 4y(x^2+y^2) + 2x^2y \end{pmatrix} \\ &= \begin{pmatrix} 5x^4y^3 + 3x^2y^5 & 3x^5y^2 + 5x^3y^4 \\ 4x^3 + 4x^3y^4 + 6xy^2 & 4x^4y^3 + 6x^2y + 4y^3 \end{pmatrix}. \end{aligned}$$

This approach is useful in theoretical work, in practical applications it is more common to find individual partial derivatives of the composed function $H(x,y) = G(F_1(x,y), F_2(x,y), F_3(x,y))$ using the chain rule. We will show how to find $\frac{\partial H_1}{\partial x}$, that is, the upper left corner in the matrix J_H.

The first component of H is $H_1(x,y) = G_1(F_1(x,y), F_2(x,y), F_3(x,y))$. When finding derivative with respect to x, we have to get from outside to this x. We can see it at three locations (in three variables of G_1), so we have to walk all these paths: We get to x through F_1, F_2 and F_3.

$$\begin{aligned} \frac{\partial H_1}{\partial x} &= \frac{\partial G_1}{\partial u} \cdot \frac{\partial F_1}{\partial x} + \frac{\partial G_1}{\partial v} \cdot \frac{\partial F_2}{\partial x} + \frac{\partial G_1}{\partial w} \cdot \frac{\partial F_3}{\partial x} \\ &= vw \cdot 2xy^2 + uw \cdot 2x + uv \cdot y = (x^2+y^2)xy2xy^2 + (x^2y^2)xy2x + x^2y^2(x^2+y^2)y \\ &= 5x^4y^3 + 3x^2y^5. \end{aligned}$$

Happy end, we've got the same answer.

And now we find all partial derivatives directly. We start by actually determining H:

$$\begin{aligned} H(x,y) &= G(x^2y^2, x^2+y^2, xy) = (x^2y^2(x^2+y^2)xy, (x^2y^2)^2 + (x^2+y^2)^2 + (xy)^2) \\ &= (x^5y^3 + x^3y^5, x^4 + y^4 + x^4y^4 + 3x^2y^2). \end{aligned}$$

We see the components $H_1(x,y) = x^5y^3 + x^3y^5$ and $H_2(x,y) = x^4 + y^4 + x^4y^4 + 3x^2y^2$ and easily find that:

$$\begin{aligned} \frac{\partial H_1}{\partial x} &= 5x^4y^3 + 3x^2y^5 \\ \frac{\partial H_1}{\partial y} &= 3x^5y^2 + 5x^3y^4 \\ \frac{\partial H_2}{\partial x} &= 4x^3 + 4x^3y^4 + 6xy^2 \\ \frac{\partial H_2}{\partial y} &= 4x^4y^3 + 6x^2y + 4y^3. \end{aligned}$$

This confirms the result obtained from the multi-variable chain rule. It was obviously the easiest way, but it is not always available, for instance when we do not know actual formulas for some of the functions involved in the composition. It is therefore important to know how to apply the

chain rule also to functions of more variables.
△

In applications it often happens that one of the spaces involved in composition is one-dimensional. Then we appreciate specialized versions of the rule for derivative of a composed function. We will start by illustrating this on the case $m = p = 1$. So we have a function $f\colon \mathbb{R}^n \mapsto \mathbb{R}$ of more variables and we "extend" it with another function $g\colon \mathbb{R} \mapsto \mathbb{R}$. This creates a function of more variables $\mathbb{R}^n \mapsto \mathbb{R}$. The Jacobi matrix becomes gradient for functions of more variables and derivative for an ordinary function. We get the following statement in which we restricted ourselves to smooth functions for simplicity.

Theorem.
Let $D \subseteq \mathbb{R}^n$ be an open set and $f \in C^1(D)$ be a function. Let M be an open interval such that $f[D] \subseteq M$, consider a function $g \in C^1(M)$.
Then $g \circ f \in C^1(D)$ and on D we have

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

On the right in the formula we see a gradient, that is, a vector, multiplied by a scalar, which makes sense. This formula tells us the following about individual partial derivatives:

$$\frac{\partial}{\partial x_i} g(f(x_1, \dots, x_n)) = g'(f(x_1, \dots, x_n)) \cdot \frac{\partial f}{\partial x_i}(x_1, \dots, x_n).$$

Example: Consider $f(x, y) = \sin(xy)$ and $g(u) = u^2$. We want to find the derivative of the composed function $g \circ f$.

$$\mathbb{R}^2[x, y] \xrightarrow{u=\sin(xy)} \mathbb{R}[u] \xrightarrow{u^2} \mathbb{R}$$

We apply the theorem. $g'(u) = 2u$, hence

$$\frac{\partial}{\partial x} g(f(x, y)) = 2u\big|_{u=f(x,y)} \cdot \frac{\partial}{\partial x}[\sin(xy)] = 2f(x, y) \cdot \cos(xy) \cdot y = 2y \sin(xy) \cos(xy).$$

Let's check: We have $(g \circ f)(x, y) = [\sin(xy)]^2$, so indeed, the derivative is

$$\frac{\partial}{\partial x} g(f(x, y)) = 2 \sin(xy) \frac{\partial}{\partial x}[\sin(xy)] = 2 \sin(xy) \cos(xy) y.$$

Derivative with respect to y is done similarly.
△

Now we will look at two useful situation that have a story connected with them.

Movement along a curve.
Imagine the following situation. A function g of more variables describes, say, the temperature at various locations in a classroom, so it has three variables. A vector function G could describe the movement of air in that room, it is a function $\mathbb{R}^3 \mapsto \mathbb{R}^3$. We are walking across this room, so we are in fact moving along a parametric curve with time as variable.

The traditional notation for our position (with respect to some coordinate system) in physics would be $\vec{r}(t)$. It is therefore a function $\mathbb{R} \mapsto \mathbb{R}^3$. For parametric curves $\vec{r}(t) = (r_1(t), r_2(t), r_3(t))$ we introduced a useful notation $\vec{r}\,'(t) = (r_1'(t), r_2'(t), r_3'(t))$. However, if we want to apply the general formula from the theorem, then we need to see this as a special case of the Jacobi matrix, which for this function means a column; in our formulas we will therefore apply transposition to change $\vec{r}\,'(t)$ into a column vector.

The composed function $(g \circ \vec{r})(t) = g(\vec{r}(t))$ describes what temperature we register as we walk. It is actually an ordinary function $\mathbb{R} \mapsto \mathbb{R}$. When we differentiate it, we find how quickly the

temperature changes as we walk, it is our subjective feeling of change. This is obviously related to the change of temperature in the room as such, that is, with the spatial gradient of g.

The composed function $(G \circ \vec{r})(t) = G(\vec{r}(t))$ describes what wind we feel as we walk. It is a vector function $\mathbb{R} \mapsto \mathbb{R}^3$, so formally a parametric curve, and we will find its derivative $[G(\vec{r})]'$ that will be transposed into a column vector to fit in the general formula.

The relationship between the spatial derivative (Jacobi matrix) and derivatives related to our walk across the room can be found in the following special version of the theorem on derivative of a composed function.

Theorem.
Let D be an open interval in $\mathbb{R}$ and $\vec{r} \in [C^1(D)]^m$ a vector function.
Let M be an open set in $\mathbb{R}^m$ such that $\vec{r}[D] \subseteq M$.
(i) For a function $g \in C^1(M)$ we have $g(\vec{r}) \in C^1(D)$ and

$$[g(\vec{r})]' = \nabla g(\vec{r}) \cdot \vec{r}\,'^T.$$

(ii) For a vector function $G \in [C^1(M)]^p$ we have $G(\vec{r}) \in [C^1(D)]^p$ and

$$[G(\vec{r})]'^T = J_G(\vec{r}) \cdot \vec{r}\,'^T.$$

In the first formula on the right we have

$$\begin{aligned}\nabla g \cdot \vec{r}\,'^T &= \Big(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_m}\Big) \cdot (r_1', \dots, r_m')^T = \Big(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_m}\Big) \bullet (r_1', \dots, r_m') \\ &= \frac{\partial g}{\partial x_1} r_1' + \dots + \frac{\partial g}{\partial x_m} r_m'.\end{aligned}$$

Since we are trying to differentiate the function $g(r_1(t), \dots, r_m(t))$ with respect to t, this corresponds to our intuitive understanding of the chain rule: We should try to get to t in all possible ways (we add them), and each time we differentiate everything that we meet on the way, multiplying the derivatives. The formula has pointed it out nicely, but properly we should also substitute in the partial derivatives to get the right variable:

$$[g(\vec{r})]'(t) = \nabla g(\vec{r}(t)) \cdot \vec{r}\,'(t)^T = \sum_{k=1}^{m} \frac{\partial g}{\partial x_k}(\vec{r}(t)) r_k'(t).$$

We can rewrite this formula in yet another way:

$$[g(\vec{r})]'(t) = D_{\vec{r}\,'(t)} g(\vec{r}(t)).$$

This has a natural interpretation. We travel along a certain curve. At time t we are at position $\vec{r}(t)$ and we want to know what is the perceived rate of change of g at the moment. The formula tells us to check out the directional derivative of g in the direction $\vec{r}\,'(t)$, that is, in the direction in which we travel at the moment.

Regarding the vector version, there we have a function $G(\vec{r})$: $\mathbb{R} \mapsto \mathbb{R}^p$, namely the function $\big(G_1(\vec{r}(t)), \dots, G_p(\vec{r}(t))\big)$. We are therefore looking for a column Jacobi matrix $p \times 1$ with components $[G_j(\vec{r})]'$. On the right in the formula we see a $p \times m$ matrix multiplying a column vector $m \times 1$, the dimensions fit. The matrix multiplication represents formulas

$$[G_j(r_1(t), \dots, r_m(t))]' = \sum_{k=1}^{m} \frac{\partial G_j}{\partial x_k}(\vec{r}(t)) r_k'(t),$$

this again fits with the basic philosophy of the chain rule. It is actually the same formula as for g above, which makes sense as we just apply derivative to components of the vector function $G = (G_1, \dots, G_m)$.

Coordinate transformations

Here we consider the case $n = m$ and $p = 1$. So we have a vector function $F\colon \mathbb{R}^n \mapsto \mathbb{R}^n$ and after it comes a function $g\colon \mathbb{R}^n \mapsto \mathbb{R}$. The theorem on derivative of a composed function then has the following form:

Theorem.
Let D be an open set in $\mathbb{R}^n$ and $F \in [C^1(D)]^n$ a vector function. Let M be an open set in $\mathbb{R}^n$ such that $F[D] \subseteq M$, consider a function $g \in C^1(M)$.
Then $g(F) \in C^1(D)$ and we have

$$\nabla(g(F)) = \nabla g(F) J_F.$$

As usual, for practical use we need to decipher this compact notation. We start by introducing names for variables in various spaces involved in this theorem.

$$\mathbb{R}^n[\vec{x}] \xrightarrow{F} \mathbb{R}^n[\vec{y}] \xrightarrow{g} \mathbb{R} \qquad\qquad \mathbb{R}^n[\vec{x}] \xrightarrow{g(F)} \mathbb{R}.$$

In other words, we have a function $g(\vec{y})$ into which we substitute $\vec{y} = F(\vec{x})$ to obtain $g(F)(\vec{x})$.

Now on the left in the formula we see a vector whose components are $\dfrac{\partial g(F)}{\partial x_i}$. On the right we multiply a (row) vector with components $\dfrac{\partial g}{\partial y_i}$ by a matrix with components $\dfrac{\partial F_i}{\partial y_j}$. When we work out one component in this multiplication, we get the formula

$$\frac{\partial g(F)}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial g}{\partial y_j} \frac{\partial F_j}{\partial x_i}.$$

As usual, this conforms to the basic idea. When we want to differentiate the function

$$g(y_1, y_2, \dots, y_n) = g(F_1(x_1, \dots, x_n), F_2(x_1, \dots, x_n), \dots, F_n(x_1, \dots, x_n))$$

by variable x_i, then we have to get to all appearances of x_i in the formula, differentiating everything that we find on our way there:

$$\frac{\partial g}{\partial y_1} \frac{\partial F_1}{\partial x_i} + \frac{\partial g}{\partial y_2} \frac{\partial F_2}{\partial x_i} + \cdots + \frac{\partial g}{\partial y_n} \frac{\partial F_n}{\partial x_i}.$$

We now develop a practical interpretation. Imagine that we have a function $g(\vec{y})\colon \mathbb{R}^n \mapsto \mathbb{R}$ that describes some physical quantity (say, local temperature), where location is specified with respect to some coordinate system $\vec{y}$, for instance the Cartesian coordinates. However, we find that it is more convenient for us to use some other coordinate system with coordinates $\vec{x}$. We thus obtain a new function $\hat{g}(\vec{x})$ that describes the same quantity (the same process), but with respect to different coordinates, so it will most likely be given by a different formula, that is, it will be formally a different function.

Because the functions $g(\vec{y})$ and $\hat{g}(\vec{x})$ describe the same process, there should be some relationship between them. It is, and it comes from relationship between the two coordinate systems. Assume that there is some vector function $F\colon \mathbb{R}^n \mapsto \mathbb{R}^n$ that transforms coordinates $\vec{x}$ to coordinates $\vec{y}$, formally $\vec{y} = F(\vec{x})$. When we are sitting at a point with new coordinates $\vec{x}$, then we can find the value of our quantity by substituting into $\hat{g}$. An alternative way is to first deduce that our location has the original coordinates $\vec{y} = F(\vec{x})$, and then substitute this into g. Because in both cases we are asking about the same value at the same location, we must have

$$\hat{g}(\vec{x}) = g(F(\vec{x})).$$

This seems obvious, we simply substitute for $\vec{y}$ in g, but it is good to know that it also makes sense in applications.

Because F is a coordinate transform, we would expect that it is invertible, so also the inverse function $\vec{x} = F_{-1}(\vec{y})$ is available. Then $g(\vec{y}) = \hat{g}(F_{-1}(\vec{y}))$.

Example: Consider the function $g(x,y) = x^2 y$. It tells us how an observer who specifies locations with respect to Cartesian coordinates sees some quantity (say, a temperature). Another observer sees the same quantity, but uses polar coordinates, so this observer sees the dependence $\hat{g}(r,\varphi)$.

Assume that both observers share the same origin and x-axis. Then they can use a convenient coordinate transform. We get to Cartesian coordinates from polar coordinates using formulas

$$x = r\cos(\varphi),$$
$$y = r\sin(\varphi).$$

Here we take $r \geq 0$ and $0 \leq \varphi < 2\pi$. Formally we see this transformation as a vector function

$$F(r,\varphi) = \big(r\cos(\varphi), r\sin(\varphi)\big).$$

The second observer therefore sees that the process depends on polar coordinates of position as follows:

$$\hat{g}(r,\varphi) = g\big(r\cos(\varphi), r\sin(\varphi)\big) = r^3\cos^2(\varphi)\sin(\varphi).$$

We see that this is a different function from the one of the first observer, indeed. However, when both stand at the same location and each substitutes the proprietary coordinates into appropriate function, then the values will agree.

Polar coordinates are popular, although they are not entirely painless. There are two problems. First, one has to decide how the origin is encoded, otherwise F would not be one-to-one. Indeed, it obviously has $r = 0$, but any angle would work, a decision has to be made.

When we do it, then there will be an inverse transform F_{-1} to F that transforms Cartesian coordinates to polar coordinates. This brings us to the second unpleasant feature: For this inverse transformation we do not have an algebraic formula for φ.

△

In applications, in particular in physics people make their lives easier by not using precise notation. Because the functions g and $\hat{g}$ describe the same process, they denote both of them g and distinguish only by variable name. So they would write $g(\vec{y}\,)$ and also $g(\vec{x})$ instead of $\hat{g}(\vec{x})$, although the two functions are given by different formulas. Another common habit is not to use notation F for variable transformation; instead, they call it by the name of the target variable itself. Physicists would write $\vec{y} = \vec{y}(\vec{x})$, or $\vec{x} = \vec{x}(\vec{y}\,)$ in case of the inverse transform. Very often they would not use the notion of vector function at all, preferring to work with individual transform functions for variables. They would write things like $y_j = y_j(\vec{x})$ or $y_j = y_j(x_1, \ldots, x_n)$.

Example: We return to the previous example. Applied people would see the two versions of the function g as follows:

$$g(x,y) = x^2 y,$$
$$g(r,\varphi) = r^3\cos^2(\varphi)\sin(\varphi).$$

They would also use the following notation for the transformation:

$$x = x(r,\varphi) = r\cos(\varphi),$$
$$y = y(r,\varphi) = r\sin(\varphi).$$

△

The rule for derivative can be expressed in the practical notation as follows:

$$\frac{\partial g(\vec{x})}{\partial x_i} = \sum_{k=1}^{n} \frac{\partial g(\vec{y}\,)}{\partial y_k}\frac{\partial y_k}{\partial x_i}.$$

Example: Consider again the function $g(x,y) = x^2 y$ and transformation from polar to Cartesian coordinates. If we want to know the partial derivative of the function g with respect to polar

coordinates, we can simply differentiate the function $g(r,\varphi) = r^3\cos^2(\varphi)\sin(\varphi)$:

$$\begin{aligned}\frac{\partial g}{\partial r} &= 3r^2\cos^2(\varphi)\sin(\varphi),\\ \frac{\partial g}{\partial \varphi} &= 2r^3\cos(\varphi)(-\sin(\varphi))\sin(\varphi) + r^3\cos^2(\varphi)\cos(\varphi)\\ &= r^3\cos(\varphi)[\cos^2(\varphi) - 2\sin^2(\varphi)].\end{aligned}$$

How will this work using the theorem?

$$\begin{aligned}\frac{\partial g}{\partial r}(r,\varphi) &= \frac{\partial g}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial g}{\partial y}\frac{\partial y}{\partial r} = 2xy\cos(\varphi) + x^2\sin(\varphi)\\ &= 2r\cos(\varphi)r\sin(\varphi)\cos(\varphi) + (r\cos(\varphi))^2\sin(\varphi) = 3r^2\cos^2(\varphi)\sin(\varphi),\\ \frac{\partial g}{\partial \varphi}(r,\varphi) &= \frac{\partial g}{\partial x}\frac{\partial y}{\partial \varphi} + \frac{\partial g}{\partial y}\frac{\partial y}{\partial \varphi} = -2xyr\sin(\varphi) + x^2 r\cos(\varphi)\\ &= -2r\cos(\varphi)r\sin(\varphi)r\sin(\varphi) + (r\cos(\varphi))^2 r\cos(\varphi) = -2r^3\cos(\varphi)\sin^2(\varphi) + r^3\cos^3(\varphi).\end{aligned}$$

Note that derivatives $\frac{\partial g}{\partial x}$, $\frac{\partial g}{\partial y}$ work with variables x, y, so in order for the answer to make sense (we want derivatives with respect to r or φ), we had to substitute into them.

This second calculation was longer, but there are situations when there is no other way. For instance, imagine that the primary information is based on polar coordinates, so we have the formula $g(r,\varphi) = r^3\cos^2(\varphi)\sin(\varphi)$ available and we also have information about the rate of change of g, namely we know the gradient, that is, the partial derivatives with respect to r and φ. The Cartesian observer would also like to have this info.

The obvious approach would be to take inverse transformation formulas $r = r(x,y)$, $\varphi = \varphi(x,y)$ and follow the above calculations. Unfortunately, while we have $r = \sqrt{x^2+y^2}$, there is no differentiable formula for φ as dependent on x, y. Thus it is necessary to try something else.

We first return to the previous calculation where we moved from Cartesian to polar coordinates and wanted some relationship between derivatives.

$$\begin{aligned}\frac{\partial g}{\partial r} &= \frac{\partial g}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial g}{\partial y}\frac{\partial y}{\partial r}\\ \frac{\partial g}{\partial \varphi} &= \frac{\partial g}{\partial x}\frac{\partial y}{\partial \varphi} + \frac{\partial g}{\partial y}\frac{\partial y}{\partial \varphi}.\end{aligned}$$

We substitute parts that we know,

$$\begin{aligned}3r^2\cos^2(\varphi)\sin(\varphi) &= \frac{\partial g}{\partial x}\cdot\cos(\varphi) + \frac{\partial g}{\partial y}\cdot\sin(\varphi)\\ r^3\cos(\varphi)[\cos^2(\varphi) - 2\sin^2(\varphi)] &= -\frac{\partial g}{\partial x}\cdot r\sin(\varphi) + \frac{\partial g}{\partial y}\cdot r\cos(\varphi).\end{aligned}$$

We obtained two linear equations that feature the derivatives we are looking for as unknowns, so it should be possible to isolate them (at least theoretically). In this particular situation it is probably best to use the not very popular Cramer rule, because it offers direct formulas.

$$\begin{aligned}\frac{\partial g}{\partial x} &= \frac{2r^3\cos(\varphi)\sin(\varphi)}{r} = 2r^2\cos(\varphi)\sin(\varphi),\\ \frac{\partial g}{\partial y} &= \frac{r^3\cos^2(\varphi)}{r} = r^2\cos^2(\varphi).\end{aligned}$$

Note that there is a discrepancy here. We found partial derivatives with respect to x and y, but the formulas feature the other variables. In fact, we will shortly see an important application where this is desirable, so this is definitely useful, but for the Cartesian observer who sees the world through x and y this is unpleasant. Can this be fixed?

If we had formulas for transforming from Cartesian coordinates to polar, that is, $r = r(x,y)$ and

$\varphi = \varphi(x,y)$, then we could have substituted them into the expressions on the right and obtain proper formulas for partial derivatives. But we do not have those formulas; after all, their lack was what made us follow this more complicated approach. On the other hand, we have a piece of good luck (which is also very important), because the resulting formulas can be rearranged into a form that actually features the desired variables.

$$\frac{\partial g}{\partial x} = 2r\cos(\varphi) r\sin(\varphi) = 2xy,$$
$$\frac{\partial g}{\partial y} = [r\cos(\varphi)]^2 = x^2.$$

So the Cartesian observer is happy in the end.
△

Summary: We have a function of two variables given with respect to coordinates x, y and also with respect to coordinates u, v.

We have formulas for coordinate transformation $x = x(u,v)$, $y = y(u,v)$. Then we also have formulas

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u}$$
$$\frac{\partial f}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v}$$

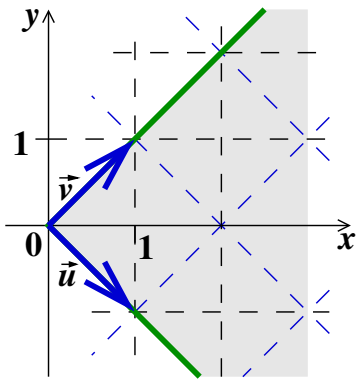
These formulas can be used directly when we want to find partial derivatives with respect to u and v. Then it makes sense to expect the answer to use variables u, v, which we can achieve by substituting $x = x(u,v)$, $y = y(u,v)$ into $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$. This is something that we always do with the chain rule.

The second option is to use these equations as a system and solve it to obtain partial derivatives with respect to x and y. If we then want the answers to be in terms of variables x, y, then we may have a problem if we cannot replace unwanted variables using formulas $r = r(x,y)$ and $\varphi = \varphi(x,y)$.

These approaches work analogously for more variables.

Example: On a rectangular playing field two caretakers measure the lawn density f. They agreed on a common coordinate origin at the western corner, both intend to use Cartesian coordinates.

One caretaker uses GPS to measure position, which orients axes to the east and to the north. This sets up the $[xy]$ coordinate system, marked in the picture as a black square grid. This caretaker then observes the function $f(x,y)$.



The second caretaker measures distance using steps taken parallel with the sides of the playing field. The step length determines the unit directional vectors $\vec{u}$ and $\vec{v}$, their coordinates with respect to the $[x,y]$ system are $\vec{u} = (1,-1)$ and $\vec{v} = (1,1)$; obviously, this caretaker has very long legs. This sets up the $[uv]$ coordinate system, marked by a blue square grid that is diagonal to the $[xy]$ system.

When the $[uv]$-observer stands at a location that was reached by u steps in direction $\vec{u}$ and v steps in direction $\vec{v}$, then the $[uv]$ coordinates are (u,v). On the other hand, the $[xy]$-observer sees this observer at location

$$u\vec{u} + v\vec{v} = u(1,-1) + v(1,1) = (u+v, -u+v).$$

We thus obtain the transformation

$$x = u + v,$$
$$y = -u + v.$$

This allows for the following direct calculation of partial derivative transformation.

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u} = \frac{\partial f}{\partial x}\cdot 1 + \frac{\partial f}{\partial y}\cdot(-1),$$
$$\frac{\partial f}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v} = \frac{\partial f}{\partial x}\cdot 1 + \frac{\partial f}{\partial y}\cdot 1.$$

Thus

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y},$$
$$\frac{\partial f}{\partial v} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y},$$

more precisely,

$$\frac{\partial f}{\partial u}(u,v) = \frac{\partial f}{\partial x}(x,y) - \frac{\partial f}{\partial y}(x,y),$$
$$\frac{\partial f}{\partial v}(u,v) = \frac{\partial f}{\partial x}(x,y) + \frac{\partial f}{\partial y}(x,y).$$

Then we should substitute $x = u + v$, $y = v - u$ on the right.

If we are interested in the transformation in the opposite direction, we can solve this system for the partial derivatives on the right, which is easy to do through adding and subtracting these equations:

$$\frac{\partial f}{\partial x} = \frac{1}{2}\frac{\partial f}{\partial u} + \frac{1}{2}\frac{\partial f}{\partial v},$$
$$\frac{\partial f}{\partial y} = -\frac{1}{2}\frac{\partial f}{\partial u} + \frac{1}{2}\frac{\partial f}{\partial v}.$$

Or we could find the inverse transformation of coordinates by solving the original transformation equations for u, v:

$$u = \frac{1}{2}x - \frac{1}{2}y,$$
$$v = \frac{1}{2}x + \frac{1}{2}y.$$

Now we can get the new transformations by direct application of the chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} = \frac{\partial f}{\partial u}\cdot\frac{1}{2} + \frac{\partial f}{\partial v}\cdot\frac{1}{2},$$
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} = \frac{\partial f}{\partial u}\cdot\left(-\frac{1}{2}\right) + \frac{\partial f}{\partial v}\cdot\frac{1}{2}.$$

△

Transformation of differential expressions

Laws of physics are typically expressed as differential equations, that is, equations that feature expressions with derivatives. It is not just physics, also other natural scientists or engineers work with differential equations. Typically those natural laws are stated with respect to Cartesian coordinates. However, sometimes we need to know how such a law looks like when seen through the lens of a different coordinate system. In such case we have to transform the differential expressions in the equation into new coordinates.

Example: We have a function $f(x,y)$ in Cartesian coordinates that should satisfy the differential equation $\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = 0$.

We pass to polar coordinates $x = r\cos(\varphi)$, $y = r\sin(\varphi)$. This transformation sets up a new function $(r,\varphi) \mapsto f(r\cos(\varphi), r\sin(\varphi))$ and we want to know what the differential equation asks of this function.

We need to replace the spatial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ in the equation with expressions that would feature partial derivatives relative to r and φ. This task would be easy if we could use the familiar formulas

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial f}{\partial \varphi}\frac{\partial \varphi}{\partial x}$$
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial f}{\partial \varphi}\frac{\partial \varphi}{\partial y}$$

Unfortunately, we do not have the necessary formula $\varphi = \varphi(x, y)$. We therefore have to follow the approach from the above example and start with the inverse transformation:

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \frac{\partial f}{\partial x}\cdot\cos(\varphi) + \frac{\partial f}{\partial y}\cdot\sin(\varphi),$$
$$\frac{\partial f}{\partial \varphi} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \varphi} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \varphi} = -\frac{\partial f}{\partial x}\cdot r\sin(\varphi) + \frac{\partial f}{\partial y}\cdot r\cos(\varphi).$$

Briefly: $f_r = \cos(\varphi)f_x + \sin(\varphi)f_y$ and $f_\varphi = -r\sin(\varphi)f_x + r\cos(\varphi)f_y$. When we solve these equations for f_x and f_y, we get

$$f_x = f_r\cos(\varphi) - \tfrac{1}{r}f_\varphi\sin(\varphi),$$
$$f_y = f_r\sin(\varphi) + \tfrac{1}{r}f_\varphi\cos(\varphi).$$

We can substitute into the given equation and obtain

$$f_r\cdot(\cos(\varphi)+\sin(\varphi)) + \tfrac{1}{r}f_\varphi\cdot(\cos(\varphi)-\sin(\varphi)) = 0,$$

that is,

$$\frac{\partial f}{\partial r}\cdot(\cos(\varphi)+\sin(\varphi)) + \frac{1}{r}\frac{\partial f}{\partial \varphi}\cdot(\cos(\varphi)-\sin(\varphi)) = 0.$$

So this is how the (imaginary) law of nature looks like in polar coordinates.

Note that the formulas for f_x and f_y included expressions that depend on new variables r, φ, but we did not mind because this time we were not trying to see some information from the point of view of x, y but from the point of view of r, φ.

△

The equation in the above example was just made up, but the procedure is useful. For many equations that describe nature, transformation into a suitable coordinate system is essentially the only way to solve them. We summarize the steps:

We are given an expression featuring derivatives of a function f with respect to variables x, y. We want to transform it to the language of variables u, v. We do so by deriving equations of the form

$$\frac{\partial f}{\partial x} = a_{1,1}(u,v)\frac{\partial f}{\partial u} + a_{1,2}(u,v)\frac{\partial f}{\partial v},$$
$$\frac{\partial f}{\partial y} = a_{2,1}(u,v)\frac{\partial f}{\partial u} + a_{2,2}(u,v)\frac{\partial f}{\partial v}.$$

If formulas $u = u(x, y)$, $v = v(x, y)$ for transformation are available, then we can obtain such equations by direct application of the chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x},$$
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y},$$

where on the right we have to replace variables x, y using the transformation formulas so that only u, v appear there.

If we do not have such formulas, but we have transformation formulas $x = x(u, v)$, $y = y(u, v)$,

then we get the desired relationships using the system of equations

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u},$$
$$\frac{\partial f}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v}.$$

We solve for partial derivatives with respect to x, y and then substitute into the given equation.

If the expression features higher order derivatives, then we repeatedly apply derivative to chain rule equations. Then it is important to keep track of which terms use the original variables x, y and which already work with the new variables u, v.

Example: Consider the so-called Laplace equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0.$$

It could describe the balance of forces on a soap bubble or distribution of heat on a plate in steady state. How does it look like in polar coordinates?

We already know that direct calculation of relationships is not possible, so we start with relations in the opposite direction. For the first derivative we have

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}(x,y)\cdot\cos(\varphi) + \frac{\partial f}{\partial y}(x,y)\cdot\sin(\varphi),$$
$$\frac{\partial f}{\partial \varphi} = -\frac{\partial f}{\partial x}(x,y)\cdot r\sin(\varphi) + \frac{\partial f}{\partial y}(x,y)\cdot r\cos(\varphi).$$

We reminded ourselves that the spatial derivatives use the original variables x, y. This is very important, because next we will differentiate once more, namely with respect to r and φ, and we have to take into account that we in fact have $\frac{\partial f}{\partial x}\big(x(r,\varphi), y(r,\varphi)\big)$ and $\frac{\partial f}{\partial y}\big(x(r,\varphi), y(r,\varphi)\big)$. Thus we have to use the chain rule when differentiating. Fortunately we do not need all four derivatives of order two, just two will be enough.

$$\begin{aligned}
\frac{\partial^2 f}{\partial r^2} &= \Big(\frac{\partial^2 f}{\partial x^2}(x,y)\cdot\frac{\partial x}{\partial r} + \frac{\partial^2 f}{\partial y\partial x}(x,y)\cdot\frac{\partial y}{\partial r}\Big)\cos(\varphi)\\
&\quad + \Big(\frac{\partial^2 f}{\partial x\partial y}(x,y)\cdot\frac{\partial x}{\partial r} + \frac{\partial^2 f}{\partial y^2}(x,y)\cdot\frac{\partial y}{\partial r}\Big)\cdot\sin(\varphi)\\
&= \frac{\partial^2 f}{\partial x^2}(x,y)\cdot\cos^2(\varphi) + \frac{\partial^2 f}{\partial y\partial x}(x,y)\cdot\sin(\varphi)\cos(\varphi)\\
&\quad + \frac{\partial^2 f}{\partial x\partial y}(x,y)\cdot\cos(\varphi)\sin(\varphi) + \frac{\partial^2 f}{\partial y^2}(x,y)\cdot\sin^2(\varphi),\\
\frac{\partial^2 f}{\partial \varphi^2} &= -\Big(\frac{\partial^2 f}{\partial x^2}(x,y)\frac{\partial x}{\partial \varphi} + \frac{\partial^2 f}{\partial y\partial x}(x,y)\frac{\partial y}{\partial \varphi}\Big)\cdot r\sin(\varphi) - \frac{\partial f}{\partial x}(x,y)\cdot r\cos(\varphi)\\
&\quad + \Big(\frac{\partial^2 f}{\partial x\partial y}(x,y)\frac{\partial x}{\partial \varphi} + \frac{\partial^2 f}{\partial y^2}(x,y)\frac{\partial y}{\partial \varphi}\Big)\cdot r\cos(\varphi) - \frac{\partial f}{\partial y}(x,y)\cdot r\sin(\varphi)\\
&= \frac{\partial^2 f}{\partial x^2}(x,y)r^2\sin^2(\varphi) - \frac{\partial^2 f}{\partial y\partial x}(x,y)r^2\cos(\varphi)\sin(\varphi) - \frac{\partial f}{\partial x}(x,y)\cdot r\cos(\varphi)\\
&\quad - \frac{\partial^2 f}{\partial x\partial y}(x,y)r^2\cos(\varphi)\sin(\varphi) + \frac{\partial^2 f}{\partial y^2}(x,y)r^2\cos^2(\varphi) - \frac{\partial f}{\partial y}(x,y)\cdot r\sin(\varphi).
\end{aligned}$$

We rewrite this using alternative notation so that we can manipulate it better.

$$\begin{aligned}
f_{rr} &= f_{xx}\cos^2(\varphi) + f_{xy}\cos(\varphi)\sin(\varphi) + f_{yx}\cos(\varphi)\sin(\varphi) + f_{yy}\sin^2(\varphi),\\
f_{\varphi\varphi} &= f_{xx}r^2\sin^2(\varphi) - f_{xy}r^2\cos(\varphi)\sin(\varphi) - f_{yx}r^2\cos(\varphi)\sin(\varphi) + f_{yy}r^2\cos^2(\varphi)\\
&\quad - r(f_x\cos(\varphi) + f_y\sin(\varphi)).
\end{aligned}$$

First we get the equations ready. We multiply the first one by r^2, and regarding the second one we notice that the expression in the parentheses agrees with the formula that we derived a short while ago for $\frac{\partial f}{\partial r}$, so we can replace. Then we add the resulting equations.

$$\begin{aligned} & r^2 f_{rr} + f_{\varphi\varphi} = r^2 f_{xx}[\cos^2(\varphi) + \sin^2(\varphi)] + r^2 f_{yy}[\cos^2(\varphi) + \sin^2(\varphi)] - r f_r \\ \Longrightarrow\ & r^2 f_{rr} + f_{\varphi\varphi} = r^2 f_{xx} + r^2 f_{yy} - r f_r. \end{aligned}$$

We get

$$r^2 f_{xx} + r^2 f_{yy} = r^2 f_{rr} + f_{\varphi\varphi} + r f_r,$$

that is,

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \varphi^2}.$$

This is the desired transformation of the Laplace operator to polar coordinates. The Laplace equation in polar coordinates is then

$$\frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \varphi^2} = 0.$$

△

7c. Differential

In chapter 3 we observed that directional (and hence also partial) derivatives are not the right generalization of derivative. We require that differentiability implies continuity. If we want to find some notion for functions of more variables that works this way, then we have to look at derivative differently, because the usual interpretation, the rate of change of a function, cannot be reasonably generalized.

We get the right point of view when we return to the question of approximation. For a differentiable function we have a very good linear approximation by a tangent line

$$f(a+h) \approx f(a) + f'(a)h.$$

We can rewrite this approximation as

$$f(a+h) - f(a) \approx f'(a)h.$$

The expression on the right defines a linear mapping $h \mapsto f'(a)h$. When searching for a tangent line, we are in fact looking for a linear mapping $L[h]$ that approximates the difference $f(a+h)-f(a)$ in the best possible way, namely we want $L[h]$ to fully capture the linear component of the difference $f(a+h) - f(a)$. How can we recognize this?

We subtract $L[h]$ from the difference and expand whatever si left into a series,

$$f(a+h) - f(a) - L[h] = c_0 + c_1 h + c_2 h^2 + c_3 h^2 + \cdots$$

Substituting $h = 0$ and recalling that $L[0] = 0$ for every linear L we get $c_0 = f(a) - f(a) - L[0] = 0$. We are hoping that also $c_1 = 0$, that is, that there is no linear growth left after taking away $L[h]$ from $f(a+h) - f(a)$. We isolate c_1 as

$$c_1 = \frac{f(a+h) - f(a) - L[h]}{h} - c_2 h - c_3 h^2 - \cdots$$

To get rid of the expansion we pass to the limit $h \to 0$:

$$\begin{aligned} c_1 &= \lim_{h\to 0}(c_1) = \lim_{h\to 0}\Big(\frac{f(a+h) - f(a) - L[h]}{h}\Big) - \lim_{h\to 0}(c_2 h + c_3 h^2 + \cdots) \\ &= \lim_{h\to 0}\Big(\frac{f(a+h) - f(a) - L[h]}{h}\Big) - 0 = \lim_{h\to 0}\Big(\frac{f(a+h) - f(a) - L[h]}{h}\Big). \end{aligned}$$

So the mapping $L[h]$ exhausted the linear component of $f(a+h) - f(a)$ exactly if the last limit is zero. This linear mapping is special and we give it a name.

Definition.
Let f be a function defined on some neighborhood of a point $a \in \mathbb{R}$. We say that a linear mapping $L\colon \mathbb{R} \mapsto \mathbb{R}$ is the **(total) differential** of f at a if

$$\lim_{h\to 0}\Big(\frac{f(a+h)-f(a)-L[h]}{h}\Big) = 0.$$

If such a linear mapping exists, then we denote it $\mathrm{d}f(a)$ or $Df(a)$.

Every linear mapping can be expressed as $L[h] = Ah$, so finding the differential is in fact a question of finding A. We already recalled above that we actually know what is the best linear approximation, namely, it should be $f'(a)h$, that is, $A = f'(a)$. We can confirm it easily:

$$\begin{aligned} f'(a) &= \lim_{h\to 0}\Big(\frac{f(a+h)-f(a)}{h}\Big) \\ \implies 0 &= \lim_{h\to 0}\Big(\frac{f(a+h)-f(a)}{h}\Big) - f'(a) = \lim_{h\to 0}\Big(\frac{f(a+h)-f(a)}{h} - f'(a)\Big) \\ &= \lim_{h\to 0}\Big(\frac{f(a+h)-f(a)-f'(a)h}{h}\Big). \end{aligned}$$

Conversely, if we have the total differential at a, then we have some A such that

$$\lim_{h\to 0}\Big(\frac{f(a+h)-f(a)-Ah}{h}\Big) = 0.$$

We again rearrange this easily to the form

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

which shows that then we have a derivative at a and $A = f'(a)$.

We have shown that for functions of one variable, the existence of differential is equivalent to the existence of derivative, and they both carry the same information, just expressed in a different way.

It turns out that if we see derivative as differential, then it is possible to generalize it to functions of more variables.

Indeed, recall that in chapter 3 we observed that if a functions of two variables $f(x,y)$ is reasonable on a neighborhood of a point (a,b), then we can approximate it by the tangent plane given by the gradient. We can write it as follows.

$$f(a+h,b+k) - f(a,b) \approx \frac{\partial f}{\partial x}h + \frac{\partial f}{\partial y}k.$$

On the right we now see an expression of the form

$$L[(h,k)] = Ah + Bk,$$

which is a linear mapping $\mathbb{R}^2 \mapsto \mathbb{R}$. On the left we actually have $f\big((a,b)+(h,k)\big)$. We thus see a direct analogy of the one-dimensional case. We easily generalize this and demand that

$$f(\vec{a}+\vec{h}) - f(\vec{a}) \approx L[\vec{h}]$$

for some linear mapping $L\colon \mathbb{R}^n \mapsto \mathbb{R}$. Linear algebra teaches that every linear mapping $L\colon \mathbb{R}^n \mapsto \mathbb{R}$ is given by some vector $(A_1, \ldots, A_n)$ according to the formula $L[\vec{h}] = \sum A_i h_i$, so this general view fits in well with our motivational example. We can test whether some linear approximation is optimal in a way analogous to the one-dimensional case.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. We say that a linear mapping $L\colon \mathbb{R}^n \mapsto \mathbb{R}$ is the **(total) differential** of f at $\vec{a}$ if

$$\lim_{\vec{h}\to\vec{0}}\Big(\frac{f(\vec{a}+\vec{h})-f(\vec{a})-L[\vec{h}]}{\|\vec{h}\|}\Big)=0.$$

If such a linear mapping exists, then we denote it $\mathrm{d}f(\vec{a})$ or $Df(\vec{a})$ and say that f is differentiable at $\vec{a}$.

Differential is the right generalization of derivative. For instance, we have the following statement.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$. If $\mathrm{d}f(\vec{a})$ exists, then f is continuous at $\vec{a}$.

We also have this.

Theorem.
Let D be a region in $\mathbb{R}^n$, $f\colon D \mapsto \mathbb{R}$. If f is differentiable on D and $\mathrm{d}f = 0$ on D, then f is constant on D.

Note that here $\mathrm{d}f = 0$ means that it is the trivial mapping, that is, $\mathrm{d}f[\vec{h}] = 0$ for all $\vec{h}$.

How can we recognize existence of a differential, and how do we find it? When we recall how we find the tangent hyperplane using gradient, then we would guess that the vector $(A_1, \dots, A_n)$ should agree with the gradient ∇f, that is, $L[\vec{h}] = \sum \frac{\partial f}{\partial x_i} h_i = \nabla f \bullet \vec{u}$. As we noted above, for one-dimensional functions the existence of derivative and differential are equivalent, but for more-dimensional functions the relationship is more complicated.

We start with a more general statement.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$.
If $\mathrm{d}f(\vec{a})$ exists, then for every $\vec{u} \in \mathbb{R}^n$ there is $D_{\vec{u}}f(\vec{a})$ and the following is true:

$$D_{\vec{u}}f(\vec{a}) = \mathrm{d}f(\vec{a})[\vec{u}].$$

What do we get if we apply this to $\vec{u} = \vec{e}_i$? If $\mathrm{d}f(\vec{a})[\vec{h}] = \sum A_i h_i$, then

$$A_i = \mathrm{d}f(\vec{a})[\vec{e}_i] = D_{\vec{e}_i}f(\vec{a}) = \frac{\partial f}{\partial x_i}(\vec{a}).$$

Corollary.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$.
If $\mathrm{d}f(\vec{a})$ exists, then for all $i = 1, \dots, n$ the partial derivatives $\frac{\partial f}{\partial x_i}(\vec{a})$ exist and

$$\mathrm{d}f(\vec{a})[\vec{h}] = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a}) \cdot h_i = \nabla f(\vec{a}) \bullet \vec{h}.$$

So ∇f is really the right vector for the differential, it is the only possible candidate. Note that the theorem says that if we have the differential, then we also have the gradient, but for functions of more variables the opposite need not be true. Sometimes it does work.

Example: For $f(x,y)=x^2+y^2$ we already found in chapter 3 that $\nabla(f)(1,2)=(2,4)$. According to the theorem, the only candidate for the differential $\mathrm{d}f(1,2)$ is $(2,4)\bullet(h,k)=2h+4k$. We check by definition that it really works.

$$\lim_{\vec{h}\to\vec{0}}\Big(\frac{f(\vec{a}+\vec{h})-f(\vec{a})-L[\vec{h}]}{\|\vec{h}\|}\Big)=\lim_{(h,k)\to(0,0)}\Big(\frac{[(1+h)^2+(2+k)^2]-(1^2+2^2)-(2h+4k)}{\sqrt{h^2+k^2}}\Big)$$
$$=\lim_{(h,k)\to(0,0)}\Big(\frac{h^2+k^2}{\sqrt{h^2+k^2}}\Big)=\lim_{(h,k)\to(0,0)}\big(\sqrt{h^2+k^2}\,\big)=0.$$

We confirmed that $\mathrm{d}f(1,2)[h,k]=\nabla(f)(1,2)\bullet(h,k)=2h+4k$.
△

However, in chapter 3 we have shown that the function $\frac{x^2y^2}{x^4+y^4}$ has gradient at $(0,0)$, but it does not have all directional derivatives there, so it also cannot have total differential. Another example of this type:

Example: Consider $f(x,y)=\sqrt{|xy|}$. Then we can get rid of the absolute value in individual quadrants, because we know signs of variables there. Therefore we can also differentiate easily there. The results can be expressed by a common formula $\nabla f(x,y)=\Big(\frac{1}{2\sqrt{|xy|}}\operatorname{sign}(xy)y,\frac{1}{2\sqrt{|xy|}}\operatorname{sign}(xy)x\Big)$ that is valid for $(x,y)\neq(0,0)$.

Regarding the origin, we can use definition to show that $\frac{\partial f}{\partial x}(0,0)=\frac{\partial f}{\partial y}(0,0)=0$. Intuitively this is clear, because $f(x,0)=f(0,y)=0$, so the function f is constant above the coordinate axes and as such the appropriate partial derivatives must be zero.

So this function has gradient everywhere on $\mathbb{R}^2$, but it is not differentiable at $(0,0)$ (it is only continuous there). One way to see this is to restrict ourselves to the line $y=x$, there the function is equal to $f(x,x)=|x|$ and as such does not have directional derivative.
△

If we want to get differential from a gradient, we need to add continuity.

Theorem.
Let f be a function defined on some neighborhood of a point $\vec{a}\in\mathbb{R}^n$.
If there is some neighborhood U of a point $\vec{a}$ such that $f\in C^1(U)$, then f is differentiable at $\vec{a}$ and

$$\mathrm{d}f(\vec{a})[\vec{h}]=\nabla f(\vec{a})\bullet\vec{h}.$$

Corollary.
Let D be an open set and $f\in C^1(D)$. Then f is differentiable on D and

$$\mathrm{d}f[\vec{h}]=\nabla f\bullet\vec{h}.$$

Again, this does not work in the other direction. Existence of differential implies existence of partial derivatives, but not necessarily their continuity. We thus obtain a hierarchy: Many functions have gradient. Some of them also have differential, and some of these have continuous gradient. The property of having continuous gradient is therefore the strongest, which explains our preference for functions from C^1.

Let's sum up popular notations for the differential. Each is suitable for a specific purpose.

$$\mathrm{d}f(\vec{a})[\vec{h}]=\frac{\partial f}{\partial x_1}(\vec{a})h_1+\cdots+\frac{\partial f}{\partial x_n}(\vec{a})h_n=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\vec{a})h_i=\nabla f(\vec{a})\bullet\vec{h}=\nabla f(\vec{a})\vec{h}^T=D_{\vec{h}}f(\vec{a}).$$

There is another take on differential that is quite useful in applications. When we have a space of vectors, we automatically also have functions that can isolate individual components from vectors, say, $(x_1, x_2, \dots) \mapsto x_1$. We call this a projection and we know that they are linear. Because a function $\vec{x} \mapsto x_i$ extracts the x_i-th coordinate from a vector, it is customary to denote this projection $x_i(\vec{x}) = x_i$.

Because it is a function, it makes sense to ask about its differential. Because this function is linear, its graph is a hyperplane, therefore its tangent plane is the same hyperplane. In other words, this function is its own differential, $\mathrm{d}x_i[\vec{h}] = h_i$.

We can therefore write the differential of some function f as follows:

$$\mathrm{d}f(\vec{a}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{a})\mathrm{d}x_i.$$

For instance, the result we obtained above can be written as $\mathrm{d}f(1,2) = 2\mathrm{d}x + 4\mathrm{d}y$. Formally we work with differentials here, but very often we just write dx and dy there, that is, the infinitely small infinitesimals of variables. The differential that we found is then interpreted as follows: If we move from point $(1,2)$ by an infinitely small displacement $(\mathrm{d}x, \mathrm{d}y)$, then the value of function changes by $\mathrm{d}f = 2\mathrm{d}x + 4\mathrm{d}y$. Such reasoning is used when deriving formulas intuitively.

By the way, for a function of one variable we then have $\mathrm{d}f = f'(a)\mathrm{d}x$, which is the formula known from substitution that can be rewritten as $\frac{\mathrm{d}f}{\mathrm{d}x} = f'(a)$. We see many interesting connections.

What rules do we have for differential? Because differential is given by gradient, it is a linear operation, that is, $\mathrm{d}(cf + g) = c\,\mathrm{d}f + \mathrm{d}g$. Other algebraic operations are not commonly used and we will skip this topic. On the other hand, composition is very important, we will look at the special case of coordinate transform.

Theorem.
Let D be an open set in $\mathbb{R}^m$, consider a function $f(y_1, \dots, y_m)\colon D \mapsto \mathbb{R}$ that is differentiable on D.
Consider the following coordinate transform: Let M be an open set in $\mathbb{R}^n$, for $j = 1, \dots, m$ let $y_j = y_j(x_1, \dots, x_n) \in C^1(M)$ and $\vec{y} = (y_1, \dots, y_m)\colon M \mapsto D$. Then $f \circ \vec{y} = f(y_1(x_1, \dots, x_n), \dots, y_m(x_1, \dots, x_n))$ is differentiable on M and we have

$$\mathrm{d}(f(\vec{y}))(\vec{x}) = \sum_{j=1}^{m} \frac{\partial f}{\partial y_j}(\vec{y}(\vec{x}))\mathrm{d}y_j.$$

Total differential is easily generalized for vector functions. When we have a vector function $F\colon \mathbb{R}^n \mapsto \mathbb{R}^m$, then the difference $F(\vec{a} + \vec{h}) - F(\vec{a})$ is a vector from $\mathbb{R}^m$. If we want to approximate it using some linear mapping that depends on $\vec{h}$, then this mapping must have its domain in $\mathbb{R}^n$ and produce vectors from $\mathbb{R}^m$; that is, is must be a vector function $\mathbb{R}^n \mapsto \mathbb{R}^m$.

Definition.
Let F be a vector function with values in $\mathbb{R}^m$ defined on some neighborhood of a point $a \in \mathbb{R}$.
We say that a linear mapping $L\colon \mathbb{R}^n \mapsto \mathbb{R}^m$ is **(total) differential** of F at $\vec{a}$ if

$$\lim_{\vec{h}\to\vec{0}} \Big(\frac{\|F(\vec{a} + \vec{h}) - F(\vec{a}) - L[\vec{h}]\|}{\|\vec{h}\|} \Big) = 0.$$

If such a linear mapping exists, then we denote it $\mathrm{d}F(\vec{a})$ or $DF(\vec{a})$ and say that F is differentiable at $\vec{a}$.

The theorem that we want to have is valid also for vector functions.

Theorem.
Let F be a vector function defined on some neighborhood of a point $a \in \mathbb{R}$.
If $\mathrm{d}F(\vec{a})$ exists, then F is continuous at $\vec{a}$.

Every linear mapping $\mathbb{R}^n \mapsto \mathbb{R}^m$ is given by some $m \times n$ matrix A according to the formula $L[\vec{h}] = A\vec{h}^T$. Which matrix corresponds to total differential? It is the Jacobi matrix of the function F, that is,

$$\mathrm{d}F(\vec{a})[\vec{h}] = J_F(\vec{a})\vec{h}^T.$$

Statements relating existence of differential and gradient are also valid (in analogous form) for vector functions. So there is a hierarchy of three qualities of functions, with functions from $[C^1(D)]^m$, that is, with continuous partial derivatives, being the best. Because the existence of differential is connected with existence of Jacobi matrix, many results from the previous sections can be rephrased in the language of differential. For instance, we obtain a theorem about differential of a composed function.

Theorem.
Let F: $D(F) \mapsto \mathbb{R}^m$ be vector function, where $D(F) \subseteq \mathbb{R}^n$.
Let G: $D(G) \mapsto \mathbb{R}^p$ be vector function, where $D(G) \subseteq \mathbb{R}^m$.
Let $\vec{a}$ be an inner point of $D(F)$ such that $\vec{b} = F(\vec{a})$ is an inner point of $D(G)$.
Assume that F is differentiable at $\vec{a}$ and G is differentiable at $\vec{b}$. Then $G \circ F$ is differentiable st $\vec{a}$ and we have

$$\mathrm{d}(G \circ F)(\vec{a}) = \mathrm{d}G(F(\vec{a})) \circ \mathrm{d}F(\vec{a}).$$

We introduced differential as a generalization of derivative and a natural question is how do we generalize higher order derivatives. For one variable we find them by differentiating the function that shows dependence of derivative on location, that is, $x \mapsto f'(x)$. The differential $\mathrm{d}f(\vec{a})[\vec{h}]$ has two variables, we handle this by fixing some directional vector $\vec{k} \in \mathbb{R}^n$ and considering the function $\vec{x} \mapsto \mathrm{d}f(\vec{x})[\vec{k}]$. This is a function of more variables, so we can inquire about its differential. However, the answer will depend on $\vec{k}$, so we expect to see both $\vec{h}$ and $\vec{k}$ in the answer. We also expect (and require) linearity for both components, which brings us to a definition.

Definition.
A mapping L: $\mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^m$ is called **bilinear** if

$$L(c\vec{x}_1 + \vec{x}_2, \vec{y}) = cL(\vec{x}_1, \vec{y}) + L(\vec{x}_2, \vec{y})$$
$$L(\vec{x}, c\vec{y}_1 + \vec{y}_2) = cL(\vec{x}, \vec{y}_1) + L(\vec{x}, \vec{y}_2)$$

for all vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ and scalars $c \in \mathbb{R}$.

Now we will choose among bilinear mappings the right one.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$.
We say that a bilinear mapping L: $\mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is $\mathrm{d}^2 f(\vec{a})$ if

$$\lim_{\vec{h} \to \vec{0}} \Big(\frac{\mathrm{d}f(\vec{a} + \vec{h})[\vec{k}] - \mathrm{d}f(\vec{a})[\vec{k}] - L[\vec{h}, \vec{k}]}{\|\vec{h}\|} \Big) = 0$$

for all $\vec{k} \in \mathbb{R}^n$.

This definition really does what we expect of second derivative:

Fact.
$\mathrm{d}^2 f(\vec{a})[\vec{h},\vec{k}] = \mathrm{d}(\mathrm{d}f[\vec{k}])[\vec{h}]$.

What do we mean by the expression on the right? The outer differential is done for the function $\vec{x} \mapsto \mathrm{d}f(\vec{x})[\vec{k}]$, exactly as we want it.

The second differential is really a differential of a differential, just like second derivative is a derivative of a derivative. So it makes sense, but how do we find it?

Linear algebra says that every bilinear mapping can be expressed using some matrix $A = (a_{ij})_{i,j=1}^n$ as $L[\vec{h},\vec{k}] = \sum_{i,j=1}^n a_{ij}h_ik_j = \vec{h}A\vec{k}^T$. We want to know what matrix provides us with $\mathrm{d}^2 f$. The fact above helps us here. We are in fact looking for the total differential of the function $\vec{x} \mapsto \sum\limits_{j=1}^{n} \frac{\partial f}{\partial x_j}(\vec{x})k_j$ for some fixed $\vec{k}$. Thanks to the linearity we can write

$$\mathrm{d}\Big(k_j \frac{\partial f}{\partial x_j}(\vec{x})\Big) = k_j \mathrm{d}\Big(\frac{\partial f}{\partial x_j}(\vec{x})\Big) = k_j \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\,\frac{\partial f}{\partial x_j} \cdot h_i = k_j \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} h_i,$$

and then we sum it up.

Theorem.
Let D be an open set in $\mathbb{R}^n$.
If $f \in C^2(D)$ then for $\vec{a} \in D$ we have

$$\mathrm{d}^2 f(\vec{a})[\vec{h},\vec{k}] = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a})h_ik_j.$$

We see that our bilinear mapping $\mathrm{d}^2 f$ is actually given by the Hess matrix $H = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)$ that we know from chapter 3. So we can write the second differential as

$$\mathrm{d}^2 f(\vec{a})[\vec{h},\vec{k}] = \vec{h}H(\vec{a})\vec{k}^T.$$

How about differentials of higher orders? We would use an n-linear mapping that can be written using "more-dimensional matrices". For instance, $\mathrm{d}^3 f[\vec{u},\vec{v},\vec{w}] = \sum\limits_{i,j,k=1}^{n} a_{ijk}u_iv_jw_k$, where one can prove that $a_{ijk} = \dfrac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}$; instead of a matrix we could imagine something like a cube with partial derivatives of third order. We leave visualization of even higher orders to the reader.

As we will soon see, such sums are not really pleasant in calculations. There are also lots of partial derivatives. Here it helps that we usually work with functions whose partial derivatives are continuous, then the order of differentiation does not matter and the number of calculations that need to be done gets smaller. Note that then the selection of variables that are used in differentiation does not depend on names of directions $\vec{u},\vec{v}$ etc., only on indices next to them. A typical term in, say, differential of order four $\mathrm{d}f(\vec{a})[\vec{t},\vec{u},\vec{v},\vec{w}]$ has the form

$$\frac{\partial^4 f}{\partial x_i \partial x_j \partial x_k \partial x_l}(\vec{a})t_iu_jv_kw_l.$$

This term works with the i-th coordinate of $\vec{t}$, so we differentiate with respect to the i-th variable, similarly for the others. When choosing the right partial derivative we therefore simply scan the term $t_iu_jv_kw_l$ for indices. For instance, if we want to find the coefficient next to $t_1u_4v_2w_1$, then we scan the indices and see that we have to differentiate twice with respect to x_1, and once with

respect to x_2 and x_4. Assuming that the function f is smooth enough, we need not worry about the order of differentiation and, say, $\dfrac{\partial^2 f}{\partial x_4 \partial x_2 \partial x_1^2}$ would do, but other orders are also fine.

Example: We will find differentials of the first three orders for the function $f(x,y) = x^2y + y^3$ at $\vec{a} = (1,2)$. Because this function has continuous partial derivatives of all orders, we will save on some calculations of mixed derivatives.

Because here we do not use variables x_1, x_2 as in the theorem but call them x, y, we first clarify how we determine partial derivatives for coefficients. Recall that when we see a product like $h_i k_j$ or $u_i v_j w_k$, we just scan the indices and recognize variables that are used for differentiation. In the theorem we recognize them by their number, now we recognize them by their alphabetic order. Index 1 means derivative with respect to x, index 2 indicates derivative with respect to y. For instance, to find the coefficient next to the expression $u_1 v_2 w_1$ we will differentiate twice with respect to x and once with respect to y.

We start with the first order differential: We need $\frac{\partial f}{\partial x} = 2xy$ and $\frac{\partial f}{\partial y} = x^2 + 3y^2$.

Using the sum form we find

$$\mathrm{d}f(1,2)[(u_1,u_2)] = \frac{\partial f}{\partial x}(1,2)u_1 + \frac{\partial f}{\partial y}(1,2)u_2 = 4u_1 + 13u_2.$$

Or we find $\nabla f(1,2) = (4,13)$ and get

$$\mathrm{d}f(1,2)[(u_1,u_2)] = (4,13) \bullet (u_1,u_2) = 4u_1 + 13u_2.$$

Differential of order two: We need $\frac{\partial^2 f}{\partial x^2} = 2y$, $\frac{\partial^2 f}{\partial x \partial y} = 2x$, and $\frac{\partial^2 f}{\partial y^2} = 6y$.

Using the sum form we find

$$\begin{aligned}\mathrm{d}f(1,2)[(u_1,u_2)] &= \frac{\partial^2 f}{\partial x^2}(1,2)u_1v_1 + \frac{\partial^2 f}{\partial x \partial y}(1,2)u_1v_2 + \frac{\partial^2 f}{\partial y \partial x}(1,2)u_2v_1 + \frac{\partial^2 f}{\partial y^2}(1,2)u_2v_2 \\ &= 4u_1v_1 + 2u_1v_2 + 2u_2v_1 + 12u_2v_2.\end{aligned}$$

One can also use $H(1,2) = \begin{pmatrix} 4 & 2 \\ 2 & 12 \end{pmatrix}$ and

$$\mathrm{d}^2 f(1,2)[(u_1,u_2),(v_1,v_2)] = (u_1,u_2)\begin{pmatrix} 4 & 2 \\ 2 & 12 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

Differential of order 3: We need $\frac{\partial^3 f}{\partial x^3} = 0$, $\frac{\partial^3 f}{\partial x^2 \partial y} = 2$, $\frac{\partial^3 f}{\partial x \partial y^2} = 0$, and $\frac{\partial^3 f}{\partial y^3} = 6$.

We plug in the point $(1,2)$ (that was easy) and start building the differential, this is best done in an organized way to make sure that we do not skip any of the 2^3 terms of the triple sum. In our sum we need to see the following combinations of indices: 111, 112, 121, 122, 211, 212, 221, 222; in front of each such index $u_i v_j w_k$ there will be a certain partial derivative as we figured out a bit earlier.

$$\begin{aligned}&\mathrm{d}^3 f(1,2)[(u_1,u_2),(v_1,v_2),(w_1,w_2)] \\ &\quad = 0u_1v_1w_1 + 2u_1v_1w_2 + 2u_1v_2w_1 + 0u_1v_2w_2 + 2u_2v_1w_1 + 0u_2v_1w_2 + 0u_2v_2w_1 + 6u_2v_2w_2 \\ &\quad = 2u_1v_1w_2 + 2u_1v_2w_1 + 2u_2v_1w_1 + 6u_2v_2w_2.\end{aligned}$$

Now that was something.

△

Is there a way to make the calculation of higher order differentials easier? Unfortunately, not in general. However, differentials are often used for the same direction, as in $\mathrm{d}^2 f(\vec{a})[\vec{h},\vec{h}]$, $\mathrm{d}^3 f(\vec{a})[\vec{h},\vec{h},\vec{h}]$ etc. How would then the differentials from the example above look like? Much better, because then we can join many terms with mixed indices. Using the calculations from the

example we get

$$\begin{aligned}
\mathrm{d}f(1,2)[(h_1,h_2)] &= 4h_1 + 13h_2,\\
\mathrm{d}^2f(1,2)[(h_1,h_2),(h_1,h_2)] &= 4h_1^2 + 2h_1h_2 + 2h_2h_1 + 12h_2^2\\
&= 4h_1^2 + 4h_1h_2 + 12h_2^2,\\
\mathrm{d}^3f(1,2)[(h_1,h_2),(h_1,h_2),(h_1,h_2)] &= 2h_1h_1h_2 + 2h_1h_2h_1 + 2h_2h_1h_1 + 6h_2h_2h_2\\
&= 6h_1^2h_2 + 6h_2^3.
\end{aligned}$$

Could we get to these results more efficiently? And here we answer in the positive. Recall that we can see gradient as a differential operator $\nabla = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\right)$ that can also be a part of formulas. This allowed us to find an alternative expression for directional derivatives, hence also for the first differential:

$$\mathrm{d}f[\vec{h}] = \sum_{i=1}^{n} h_i \frac{\partial}{\partial x_i} f = \Big(\sum_{i=1}^{n} h_i \frac{\partial}{\partial x_i}\Big) f = (\vec{h} \bullet \nabla) f.$$

Now we look at the second differential. There we have

$$\begin{aligned}
\mathrm{d}^2f(\vec{a})[\vec{h},\vec{h}] &= \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} h_i h_j = \Big(\sum_{i,j=1}^{n} h_i h_j \frac{\partial^2}{\partial x_i \partial x_j}\Big) f = \Big(\sum_{i,j=1}^{n} h_i h_j \frac{\partial}{\partial x_i}\,\frac{\partial}{\partial x_j}\Big) f\\
&= \Big(\sum_{i=1}^{n} h_i \frac{\partial}{\partial x_i}\Big)\Big(\sum_{j=1}^{n} h_j \frac{\partial}{\partial x_j}\Big) f = (\vec{h} \bullet \nabla)(\vec{h} \bullet \nabla) f = (\vec{h} \bullet \nabla)^2 f.
\end{aligned}$$

So the second differential is simply the operator $(\vec{h} \bullet \nabla)$ applied twice. For higher order differentials this works analogously. The best part is that we actually need not apply this operator repeatedly, like we repeat derivation to find higher order derivatives. Instead, we can first figure out how the whole operation works by multiplying out the power using the usual formulas. For instance, we find the third differential like this:

$$\begin{aligned}
(\vec{h} \bullet \nabla)^3 &= \Big(h_1 \frac{\partial}{\partial x_1} + h_2 \frac{\partial}{\partial x_2}\Big)^3\\
&= \Big(h_1 \frac{\partial}{\partial x_1}\Big)^3 + 3\Big(h_1 \frac{\partial}{\partial x_1}\Big)^2 h_2 \frac{\partial}{\partial x_2} + 3h_1 \frac{\partial}{\partial x_1}\Big(h_2 \frac{\partial}{\partial x_2}\Big)^2 + \Big(h_2 \frac{\partial}{\partial x_2}\Big)^3\\
&= h_1^3 \frac{\partial^3}{\partial x_1^3} + 3h_1^2 h_2 \frac{\partial^3}{\partial x_1^2 \partial x_2} + 3h_1 h_2^2 \frac{\partial^3}{\partial x_1 \partial x_2^2} + h_2^3 \frac{\partial^3}{\partial x_2^3}.
\end{aligned}$$

Now we apply this operator to f and obtain

$$\mathrm{d}^3f[(h_1,h_2),(h_1,h_2),(h_1,h_2)] = \frac{\partial^3 f}{\partial x_1^3} h_1^3 + \frac{\partial^3 f}{\partial x_1^2 \partial x_2} 3h_1^2 h_2 + \frac{\partial^3 f}{\partial x_1 \partial x_2^2} 3h_1 h_2^2 + \frac{\partial^3 f}{\partial x_2^3} h_2^3.$$

When we apply this to the function f and point $(1,2)$ from our example, we get the same answer

$$\mathrm{d}^3f(1,2)[(h_1,h_2),(h_1,h_2),(h_1,h_2)] = 6h_1^2h_2 + 6h_2^3.$$

This process is very efficient, and it will come handy in the next section.

7d. Taylor polynomial

The notion of Taylor polynomial is inspired by the following question: We have good information about a function f at a point $\vec{a}$, and we would like to derive the best possible information about f at a point $\vec{x}$ that is typically close to $\vec{a}$.

For a function of one variable, this service is supplied by the Taylor polynomial. For a function of more variables we can make use of this by using our favourite trick, passing to a slice. The easiest way to get from $\vec{a}$ to $\vec{x}$ is along a straight line, it leads in direction $\vec{x} - \vec{a}$. We prefer to work with unit directional vectors, so we take $\vec{u} = \frac{\vec{x}-\vec{a}}{\|\vec{x}-\vec{a}\|}$. Then $\vec{x} - \vec{a} = \|\vec{x} - \vec{a}\|\vec{u}$. Our line has the parametric expression $\vec{a} + t\vec{u}$, and we are in particular interested in location $t_x = \|\vec{x} - \vec{a}\|$, because

then $\vec{a} + t_x\vec{u} = \vec{x}$.

On this line we can investigate f through the auxiliary function $\varphi(t) = f(\vec{a} + t\vec{u})$. Of special interest are the values $\varphi(0) = f(\vec{a})$ and $\varphi(t_x) = f(\vec{x})$. If f has lots of derivatives at $\vec{a}$, then also φ has lots of derivatives at the corresponding value of parameter $t = 0$ and we can use the Taylor polynomial to approximate its value:

$$f(\vec{x}) = \varphi(t_x) \approx \varphi(0) + \varphi'(0)t_x + \frac{1}{2!}\varphi''(0)t_x^2 + \frac{1}{3!}\varphi'''(0)t_x^3 + \ldots + \frac{1}{N!}\varphi^{(N)}(0)t_x^N.$$

We know that derivatives of φ are related to directional derivatives of f in direction $\vec{u}$. In chapter 3 we saw that $\varphi'(0) = D_{\vec{u}}f(\vec{a})$, and we easily show that also $\varphi''(0) = D_{\vec{u}}D_{\vec{u}}f(\vec{a}) = D^2_{\vec{u}}f(\vec{a})$ and so on. We also recall that $t_x = \|\vec{x} - \vec{a}\|$ and we obtain

$$f(\vec{x}) \approx f(\vec{a}) + D_{\vec{u}}f(\vec{a})\|\vec{x} - \vec{a}\| + \tfrac{1}{2}D^2_{\vec{u}}f(\vec{a})\|\vec{x} - \vec{a}\|^2 + \cdots + \tfrac{1}{N!}D^N_{\vec{u}}f(\vec{a})\|\vec{x} - \vec{a}\|^N.$$

This is an expression that works directly with f, so we could take it as an inspiration for definition of Taylor polynomial for functions of more variables. However, the notion of directional derivative is not the best from theoretical point of view, so we will work some more on this expression. First we recall that $D_{c\vec{u}}f = cD_{\vec{u}}f$. Then

$$D^2_{c\vec{u}}f = D_{c\vec{u}}[D_{c\vec{u}}f] = D_{c\vec{u}}[cD_{\vec{u}}f] = cD_{c\vec{u}}[D_{\vec{u}}f] = c^2D_{\vec{u}}[D_{\vec{u}}f] = c^2D^2_{\vec{u}}f.$$

Similarly we show that in general $D^k_{\vec{u}}f \cdot c^k = D^k_{c\vec{u}}f$, so thanks to $\|\vec{x} - \vec{a}\|\vec{u} = \vec{x} - \vec{a}$ the formula can be written as

$$f(\vec{x}) \approx f(\vec{a}) + D_{\vec{x}-\vec{a}}f(\vec{a}) + \tfrac{1}{2}D^2_{\vec{x}-\vec{a}}f(\vec{a}) + \tfrac{1}{3!}D^3_{\vec{x}-\vec{a}}f(\vec{a}) + \cdots + \tfrac{1}{N!}D^N_{\vec{x}-\vec{a}}f(\vec{a}).$$

We saw that $D_{\vec{h}}f$ corresponds to the differential $\mathrm{d}f[\vec{h}]$. We also know that higher order differentials come as repeated application of differential, so

$$\mathrm{d}^2f[\vec{h},\vec{h}] = \mathrm{d}f(\mathrm{d}f[\vec{h}])[\vec{h}] = D_{\vec{h}}(\mathrm{d}f[\vec{h}]) = D_{\vec{h}}D_{\vec{h}}f = D^2_{\vec{h}}f.$$

It works like this also for higher orders, which brings us to the form

$$f(\vec{x}) \approx f(\vec{a}) + \mathrm{d}f(\vec{a})[\vec{x} - \vec{a}] + \tfrac{1}{2}\mathrm{d}^2f(\vec{a})[\vec{x} - \vec{a}, \vec{x} - \vec{a}] + \cdots + \tfrac{1}{N!}\mathrm{d}^Nf(\vec{a})[\vec{x} - \vec{a}, \ldots, \vec{x} - \vec{a}].$$

For sufficiently reasonable functions f, the expression on the right should provide good approximation. Moreover, we see an obvious similarity with the classical Taylor polynomial, so the following definition is to be expected.

Definition.
Let f be a function defined on some neighborhood of a point $\vec{a} \in \mathbb{R}^n$ that is differentiable at $\vec{a}$ up to order $N \in \mathbb{N}$.
Then we define its **Taylor polynomial** at $\vec{a}$ of degree N as

$$\begin{aligned} T_N(\vec{x}) &= f(\vec{a}) + \mathrm{d}f(\vec{a})[\vec{x} - \vec{a}] + \tfrac{1}{2}\mathrm{d}^2f(\vec{a})[\vec{x} - \vec{a}, \vec{x} - \vec{a}] + \cdots \\ &\qquad \cdots + \tfrac{1}{N!}\mathrm{d}^Nf(\vec{a})[\vec{x} - \vec{a}, \ldots, \vec{x} - \vec{a}] \\ &= \sum_{k=0}^{N} \frac{1}{k!}\mathrm{d}^kf(\vec{a})[\vec{x} - \vec{a}, \ldots, \vec{x} - \vec{a}]. \end{aligned}$$

In this form, this Taylor polynomial is the natural generalization of the definition for one variable, because we know now that differentials are the right generalization of derivatives. We will soon see that it also offers an efficient way to find them. The version with directional derivatives has its uses as well, for instance it allows us to express a general version of one of the key statements about Taylor polynomials, namely the theorem on Lagrange form of remainder.

Theorem.
Let $\vec{a} \in \mathbb{R}^n$. Let $f \in C^{n+1}(U)$ for some neighborhood U of $\vec{a}$. Then for every $\vec{x} \in U$ there is $\vec{c} \in [\vec{a}, \vec{x}]$ such that, denoting $\vec{u} = \frac{\vec{x}-\vec{a}}{\|\vec{x}-\vec{a}\|}$, we have

$$f(\vec{x}) - T_N(\vec{x}) = \tfrac{1}{(N+1)!} D^{N+1}_{\vec{x}-\vec{a}} f(\vec{c}) = \tfrac{1}{(N+1)!} D^{N+1}_{\vec{u}} f(\vec{c}) \|\vec{x}-\vec{a}\|^{N+1}.$$

Here $[\vec{a}, \vec{x}]$ denotes a segment in $\mathbb{R}^n$, that is, the set of all points of the form $\vec{a} + t(\vec{x} - \vec{a})$ for $t \in [0, 1]$.

In applications we often prefer the version where we do not go from $\vec{a}$ to $\vec{x}$, but from $\vec{a}$ to $\vec{a} + \vec{h}$. We create the appropriate version easily, either directly with $\vec{x} - \vec{a} = \vec{h}$ or using the unit directional vector $\vec{u} = \frac{\vec{h}}{\|\vec{h}\|}$. We offer three versions.

$$\begin{aligned} f(\vec{a}+\vec{h}) &\approx f(\vec{a}) + \mathrm{d}f(\vec{a})[\vec{h}] + \tfrac{1}{2}\mathrm{d}^2 f(\vec{a})[\vec{h},\vec{h}] + \cdots + \tfrac{1}{N!}\mathrm{d}^N f(\vec{a})[\vec{h},\ldots,\vec{h}] \\ &\approx f(\vec{a}) + D_{\vec{h}} f(\vec{a}) + \tfrac{1}{2} D^2_{\vec{h}} f(\vec{a}) + \cdots + \tfrac{1}{N!} D^N_{\vec{h}} f(\vec{a}) \\ &\approx f(\vec{a}) + D_{\vec{u}} f(\vec{a}) \|\vec{h}\| + \tfrac{1}{2} D^2_{\vec{u}} f(\vec{a}) \|\vec{h}\|^2 + \cdots + \tfrac{1}{N!} D^N_{\vec{u}} f(\vec{a}) \|\vec{h}\|^N. \end{aligned}$$

We see that Taylor polynomial works as expected in higher dimensions. How do we find it efficiently? For that we recall the practical form of differential.

$$\begin{aligned} T_N(\vec{x}) &= f(\vec{a}) + ((\vec{x}-\vec{a}) \bullet \nabla) f(\vec{a}) + \tfrac{1}{2}((\vec{x}-\vec{a}) \bullet \nabla)^2 f(\vec{a}) + \cdots + \tfrac{1}{N!}((\vec{x}-\vec{a}) \bullet \nabla)^N f(\vec{a}) \\ &= \sum_{k=0}^{N} \frac{1}{k!}((\vec{x}-\vec{a}) \bullet \nabla)^k f(\vec{a}). \end{aligned}$$

We know how to work with such expressions. Also this formula has a popular alternative version

$$T_N(\vec{a}+\vec{h}) = \sum_{k=0}^{N} \frac{1}{k!}(\vec{h} \bullet \nabla)^k f(\vec{a}).$$

Example: Consider the function $f(x, y) = \cos(xy) + xy$. We will find its T_2 centered at the point $(0, \pi)$.

First we will find partial derivatives, we will use the popular notation from physics to save room:
$f_x = -\sin(xy) \cdot y + y$, $f_y = -\sin(xy) \cdot x + x$;
$f_{xx} = -\cos(xy) \cdot y^2$, $f_{yx} = f_{xy} = -\cos(xy) \cdot xy - \sin(xy) + 1$, $f_{yy} = -\cos(xy) \cdot x^2$.
We substitute: $f_x(0, \pi) = \pi$, $f_y(0, \pi) = 0$;
$f_{xx}(0, \pi) = -\pi^2$, $f_{yx}(0, \pi) = f_{xy}(0, \pi) = 1$, $f_{yy}(0, \pi) = 0$.
We will also need $f(0, \pi) = 1$.

Because this is a polynomial of order two and for such differentials we have pleasant formulas

$$\mathrm{d}f(\vec{a})[\vec{h}] = \nabla f \bullet \vec{h}, \quad \mathrm{d}^2 f(\vec{a})[\vec{h},\vec{h}] = \vec{h} H(\vec{a}) \vec{h}^T,$$

we could find the desired polynomial directly from the definition.

$$\begin{aligned} T(x,y) &= f(0,\pi) + \mathrm{d}f(0,\pi)[(x,y)-(0,\pi)] + \tfrac{1}{2}\mathrm{d}^2 f(0,\pi)[(x,y)-(0,\pi),(x,y)-(0,\pi)] \\ &= 1 + (\pi, 0) \bullet ((x,y)-(0,\pi)) + \tfrac{1}{2}((x,y)-(0,\pi)) \begin{pmatrix} -\pi^2 & 1 \\ 1 & 0 \end{pmatrix} ((x,y)-(0,\pi))^T \\ &= 1 + (\pi, 0) \bullet (x, y-\pi) + \tfrac{1}{2}(x, y-\pi) \begin{pmatrix} -\pi^2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y-\pi \end{pmatrix} \\ &= 1 + \pi x - \tfrac{1}{2}\pi^2 x^2 + x(y-\pi). \end{aligned}$$

With polynomials of higher order we would have to use $(\vec{h} \bullet \nabla)^k f$, so we will practice it here.

First we find how the differentials that we need look like:

$$\big(((x,y)-(0,\pi))\bullet\nabla\big)=\big((x-0,y-\pi)\bullet\nabla\big)=x\frac{\partial}{\partial x}+(y-\pi)\frac{\partial}{\partial y},$$

$$\big(((x,y)-(0,\pi))\bullet\nabla\big)^2=\Big(x\frac{\partial}{\partial x}+(y-\pi)\frac{\partial}{\partial y}\Big)^2=x^2\frac{\partial^2}{\partial x^2}+2x(y-\pi)\frac{\partial^2}{\partial x\partial y}+(y-\pi)^2\frac{\partial^2}{\partial y^2}.$$

Now we apply these differentials to f and substitute $(0,\pi)$:

$$\big((x,y-\pi)\bullet\nabla\big)f(0,\pi)=x\frac{\partial f}{\partial x}(0,\pi)+(y-\pi)\frac{\partial f}{\partial y}(0,\pi)=\pi x,$$

$$\begin{aligned}\big((x,y-\pi)\bullet\nabla\big)^2f(0,\pi)&=x^2\frac{\partial^2 f}{\partial x^2}(0,\pi)+2x(y-\pi)\frac{\partial^2 f}{\partial x\partial y}(0,\pi)+(y-\pi)^2\frac{\partial^2 f}{\partial y^2}(0,\pi)\\&=-\pi^2x^2+2x(y-\pi).\end{aligned}$$

Therefore

$$T_2(x,y)=1+\pi x+\tfrac{1}{2}\big(-\pi^2x^2+2x(y-\pi)\big),$$

which yields the same polynomial as before.

Another possible approach is to go for $T_N(0+h,\pi+k)$. We would find

$$\big((h,k)\bullet\nabla\big)=h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y},$$

$$\big((h,k)\bullet\nabla\big)^2=h^2\frac{\partial^2}{\partial x^2}+2hk\frac{\partial^2}{\partial x\partial y}+k^2\frac{\partial^2}{\partial y^2},$$

then

$$\big((h,k)\bullet\nabla\big)f(0,\pi)=h\frac{\partial f}{\partial x}(0,\pi)+k\frac{\partial f}{\partial y}(0,\pi)=\pi h,$$

$$\big((h,k)\bullet\nabla\big)^2f(0,\pi)=h^2\frac{\partial^2 f}{\partial x^2}(0,\pi)+2hk\frac{\partial^2 f}{\partial x\partial y}(0,\pi)+k^2\frac{\partial^2 f}{\partial y^2}(0,\pi)=-\pi^2h^2+2hk.$$

Finally

$$\begin{aligned}T(0+h,\pi+k)&=f(0,\pi)+\mathrm{d}f(0,\pi)[(h,k)]+\tfrac{1}{2}\mathrm{d}^2f(0,\pi)[(h,k),(h,k)]\\&=f(0,\pi)+\big((h,k)\bullet\nabla\big)f(0,\pi)+\tfrac{1}{2}\big((h,k)\bullet\nabla\big)^2f(0,\pi)\\&=1+\pi h-\tfrac{1}{2}\pi^2h^2+hk.\end{aligned}$$

Because f is a reasonable function, we expect that for h,k small we have

$$f(0+h,\pi+k)\approx 1+\pi h-\tfrac{1}{2}\pi^2h^2+hk.$$

△

We conclude this section by returning to chapter 4 where we stated the Sylvester criterion for classification of local extrema. Where does it come from?

Consider a function f defined on a neighborhood of a stationary point $\vec{a}$. We therefore suspect that there is a local extreme there. If the function is sufficiently smooth (twice continuously differentiable), then we can approximate it by Taylor polynomial of degree 2 on some neighborhood of $\vec{a}$. Since $\nabla f(\vec{a})=\vec{0}$, we have

$$f(\vec{a}+\vec{h})\approx f(\vec{a})+\tfrac{1}{2}\vec{h}H\vec{h}^T,$$

where $H=\Big(\dfrac{\partial^2 f}{\partial x_i\partial x_j}\Big)$ is the Hess matrix. We need to know whether values of f around $\vec{a}$ are larger or smaller than $f(\vec{a})$.

A (sharp) local minimum requires that $f(\vec{a}+\vec{h})>f(\vec{a})$, so we recognize it by the fact that $\vec{h}H\vec{h}^T>0$ for all non-zero $\vec{h}$. Similarly, a (sharp) local maximum happens if $\vec{h}H\vec{h}^T<0$ for all

non-zero $\vec{h}$. By a remarkable coincidence, exactly this property is known in linear algebra.

Definition.
Let A be a symmetric matrix $n \times n$.
We say that A is **positive-definite** if
$\vec{x}A\vec{x}^T > 0$ for all $\vec{x} \in \mathbb{R}^n - \{\vec{0}\}$.
We say that A is **negative-definite** if
$\vec{x}A\vec{x}^T < 0$ for all $\vec{x} \in \mathbb{R}^n - \{\vec{0}\}$.
We say that A is **indefinite** if
there are $\vec{x}, \vec{y} \in \mathbb{R}^n$ such that $\vec{x}A\vec{x}^T > 0$ and $\vec{y}A\vec{y}^T < 0$.

This property is very useful in some applications, and here it is obviously crucial. We see that if the Hess matrix is positive-definite, then we have a local minimum, while negative-definiteness yields local maximum. An indefinite Hess matrix signals a stationary point that is not a local extreme. How can we tell this property for matrices?

For diagonal matrices this is easy, because then $a_{ij} = 0$ for $i \neq j$ and therefore

$$\vec{h}A\vec{h}^T = \sum_{i,j=1}^{n} a_{ij}h_ih_j = \sum_{i=1}^{n} a_{ii}h_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} a_{ij}h_ih_j = \sum_{i=1}^{n} a_{ii}h_i^2.$$

It is obvious that the sign of $\vec{h}A\vec{h}^T$ can be derived from signs of a_{ii}. If they are all positive, then $\vec{h}A\vec{h}^T > 0$ for arbitrary (non-zero) vector. If they are all negative, then always $\vec{h}A\vec{h}^T < 0$. And if there are both positive and negative numbers among a_{ii}, then the matrix is indefinite.

If the matrix A is not diagonal but is symmetric, then according to matrix theory it can be replaced in the product $\vec{h}A\vec{h}^T$ by a diagonal matrix D whose diagonal entries are eigenvalues of A. We thus get the following statement.

Theorem.
Let A be a symmetric matrix $n \times n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.
If $\lambda_i > 0$ for all $i = 1, \ldots, n$, then A is positive-definite.
If $\lambda_i < 0$ for all $i = 1, \ldots, n$, then A is negative-definite.

However, finding eigenvalues is not easy. It is more convenient to apply the following criterion.

Theorem. (Sylvester criterion)
Let $A = (a_{ij})$ be a symmetric matrix $n \times n$. Consider its subdeterminants $\Delta_1 = a_{11}$, $\Delta_2 = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, and so on up to $\Delta_n = \det(A)$.
If $\Delta_i > 0$ for all i, then A is positive-definite.
If $\Delta_1 < 0$, $\Delta_2 > 0$, $\Delta_3 < 0$ etc. up to $(-1)^n\Delta_n > 0$, then A is negative-definite.
If $\Delta_2 < 0$, then there are $\vec{x}, \vec{y} \in \mathbb{R}^n$ such that $\vec{x}A\vec{x}^T > 0$ and $\vec{y}A\vec{y}^T < 0$.

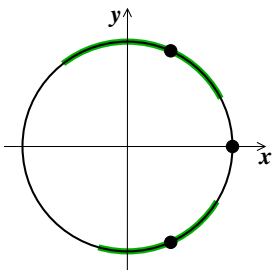
Because the Hess matrix is symmetric for smooth functions, we can apply this statement to it and obtain the Sylvester criterion in the form that we saw in chapter 4.

7e. Implicit curves and functions

Many objects are determined by equations, for instance a certain circle in a plane is given by the equation $x^2 + y^2 = 169$. How would we find some information about such objects?

One possible approach is to see an object given by an equation $f(x, y, ...) = c$ as a level curve of function f. In this way we obtain useful information. For instance, if we take some point on

a curve, then we get a vector perpendicular to this curve at this point simply as ∇f. Using this vector we then easily find the tangent line, plane etc. (depending on the dimension). However, there is information that is harder to get in this way. Therefore people also use another approach.



It often happens that some parts of such an object can be interpreted as graphs of suitable functions. For instance, for that circle $x^2 + y^2 = 169$, near the point $(5, 12)$ this circle corresponds to the graph of the function $y = \sqrt{169 - x^2}$, we simply solved that equation for y. Such a function is called an **implicit function**. The advantage is that then we can apply standard analytic methods to this part of the curve, for instance we can investigate local extrema, concavity and so on. The disadvantage of this approach is that it is not always possible to treat the whole curve in this way. We can see it in our example, we definitely cannot view the whole circle as a graph of one function.

In such case we usually proceed locally, for instance on a neighborhood of the point $(5, -12)$ this circle is identical with the graph of the function $y = -\sqrt{169 - x^2}$, which is a different function then before. So from the start we will accept that we will interpret given curves as graphs only locally.

Unfortunately, even this ambition cannot be fully achieved. As we can see in the picture, it is not possible to find any neighborhood of the point $(13, 0)$ on which the corresponding part of the circle would be a graph of some function. Indeed, in arbitrary neighborhood we can find two points on the circle with the same x variable, but no function is allowed to have two distinct values for one x.

Despite these limitations, this approach is fruitful. If we want to follow it, we have to rule out pathological situations. The example with circle shows that we may get in trouble when a curve has a vertical tangent line somewhere (or not, depends whether the curve doubles on itself or not at that point). Unfortunately, things may get even worse, as the set of solutions of an equation $f(x, y) = c$ need not be a curve at all.

Example: The implicit equation $xy - \sqrt{x^2y^2} = 0$ defines the set of all points from the first and third quadrant.

The implicit equation $(x^2 - 4)^2 + (y - 1)^2 = 0$ defines a set consisting of two points, namely $(2, 1)$ and $(-2, 1)$.

The implicit equation $(|x| - 1)^2 + y^2 = 1$ defines "figure eight", two circles touching at one point. Finally something reasonable!

△

We will now focus on planar objects, that is, equations with two variables x, y. To simplify situation formally, we will work with equations of the type $f(x, y) = 0$, other equations can be easily rewritten in this way.

It is possible to show that if we rule out cases with vertical (or non-existent) tangent line, then we in effect rule out all pathological cases. Unfortunately, we also rule out some well-behaved curves like the one given by $x - y^3 = 0$, which is the graph of cubic root, but we do not really have a finer selection tool.

Before we get to the theorem, one key remark: How do we recognize that some function $y = y(x)$ describes locally the same object as the equation $f(x, y) = 0$?

We return to the example $x^2 + y^2 = 169$. The ability to express y from the equation actually means that we found a solution of this equation in the setting that y is unknown and x is a parameter. The standard way to check validity of such a solution is by substituting the solution in the equation:

$$x^2 + \sqrt{169 - x^2}^{\,2} = x^2 + 169 - x^2 = 169.$$

In general, we substitute the formula $y = y(x)$ for y in the equation and we expect that it will work, that is, that $f(x, y(x)) = 0$ will be true for all x considered in our work (usually some neighborhood of a given point). Because we are interested in a curve passing through some given point (x_0, y_0), we will also require that $y(x_0) = y_0$.

Theorem. (Implicit function theorem)
Let D be an open set in $\mathbb{R}^2$, $f \in C^1(D)$. Let $\vec{a} = (x_0, y_0) \in D$ satisfies $f(\vec{a}) = 0$. If $\frac{\partial f}{\partial y}(\vec{a}) \neq 0$, then there is a neighborhood U of x_0 and a function $y = y(x)$ on U such that $y(x_0) = y_0$ and $f(x, y(x)) = 0$ for $x \in U$.
Moreover, $y(x)$ has a continuous derivative on U and

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

In other words, the equation $f(x, y) = 0$ can be solved for y (dependent on x) on a neighborhood of the point $\vec{a}$. The theorem does not guarantee that it would be possible to express the resulting function $y = y(x)$ as an algebraic formula. However, the mere knowledge that such a function exists is already useful. The formula from the theorem then offers a possibility to find the derivative of this function, even if we perhaps cannot express it with a formula.

We do not have to apply the formula formally, because it is derived through steps that can be used directly to find derivative of a given function. We start by imagining that the y in the equation $f(x, y) = 0$ is actually a function with variable x. Then this equation compares two functions with variable x, where on the right we see a constant function. Thus we can differentiate both sides, on the left we will apply the chain rule:

$$\begin{aligned} f(x, y(x)) &= 0 \\ [f(x, y(x))]' &= [0]' \\ \frac{\partial f}{\partial x}(x, y(x)) \cdot 1 + \frac{\partial f}{\partial y}(x, y(x)) \cdot y'(x) &= 0. \end{aligned}$$

When we express $y'(x)$ from this equality, we get the formula in the theorem.

When we do this procedure with a specific function, we call it **implicit differentiation**.

Example: Consider the implicit curve given by the equation $x^4 - 4xy + y^4 = 1$ (by the way, this is an oval with its long axis on the main diagonal, but usually we do not know the actual shape of the curve, so we will do without this information). Because the point $x = 0$, $y = 1$ satisfies the given equation, it lies on this curve. Is the given curve nice (is it actually a curve?) on some neighborhood of this point and can we express it as a function there?

We need to reorganize the equation so that it has zero on the right, so we will apply the theorem to the function $f(x, y) = x^4 - 4xy + y^4 - 1$. We have

$$\frac{\partial f}{\partial y}(0, 1) = (-4x + 4y^3)\big|_{x=0, y=1} = 4 \neq 0,$$

so the condition is satisfied. It follows that on some neighborhood of the point $(0, 1)$ we have a curve that is also the graph of some unknown function $y(x)$. Unfortunately, we cannot derive a formula for this function $y(x)$.

How does the tangent line to its graph, that is, to the curve, at the given point look like? Since we do not have a formula for $y(x)$, we find the derivative that we need by implicit differentiation. To this end, we assume that y in the equation (for instance in the given one) is a function of x and differentiate the equality $x^4 - 4xy(x) + (y(x))^4 = 1$. Usually people just imagine the variable at y

and do not write it.

$$[x^4 - 4xy + y^4]' = [1]'$$
$$4x^3 - 4y - 4xy' + 4y^3y' = 0$$
$$y' = \frac{4y - 4x^3}{4y^3 - 4x}.$$

We are interested in situation at $(0,1)$, that is, for $x = 0$ and $y = 1$. Substituting them into the formula yields $y'(0) = \frac{4}{4} = 1$. So we know the point and also the slope, and we find that the tangent line at this point is $y - 1 = 1 \cdot (x - 0)$, that is, $y = x + 1$.

An alternative: Our object is a level curve of the function $f(x,y) = x^4 - 4xy + y^4$, we have

$$\nabla f = (4x^3 - 4y, -4x + 4y^3) \implies \nabla f(0,1) = (-4,4).$$

We know that $\vec{u} = \nabla f(0,1) = (-4,4)$ is the vector perpendicular to our level curve, hence it is also perpendicular to the tangent line. The line passing through the point $(0,1)$ and perpendicular to a vector $\vec{u}$ is given by the equation

$$\vec{u} \bullet ((x,y) - (0,1)) = 0 \implies (-4,4) \bullet (x, y-1) = 0 \implies -4x + 4(y-1) = 0 \implies y = x + 1.$$

△

The implicit differentiation often helps also in situations where we are able to actually express y from the equation, but the formula would be too complicated.

Note that the formula for derivative that we obtained by implicit differentiation

$$y' = \frac{4y - 4x^3}{4y^3 - 4x}$$

features both x and y, that is, both the free variable and the function that we are trying to investigate. This is typical. If we knew the formula for $y = y(x)$, we could substitute and obtain the usual formula for derivative in terms of x (of course, if we knew the formula for $y(x)$, we would have differentiated it directly). However, we do not know the formula for $y(x)$, so our derivative formula is useful only at known points (x,y) of the implicit curve.

How would this work in more dimensions? Then we have more variables in the implicit equation, for instance the equation $2x + 3y + z = 5$ defines a plane in $\mathbb{R}^3$. In this case we easily isolate, say, z from the equation and thus express the whole plane as the graph of the function $z(x,y) = 5 - 2x - 3y$.

In general, when we have one equation with more variables, then we can hope that we can (at least locally) express one variable. This corresponds to the situation when we expressed a part of the object created by the equation as a graph of a function like we did with the equation of a plane.

Formally it helps to separate that one special variable through notation. We thus work with an equation of the form $f(x_1, \dots, x_n, y) = 0$ and hope that we can express y as it depends on the other variables. In other words, we hope to obtain a function $y = y(x_1, \dots, x_n)$ whose graph agrees (locally) with the object created by the equation. We again check this agreement by substitution, so the condition is

$$f\big(x_1, \dots, x_n, y(x_1, \dots, x_n)\big) = 0.$$

How does it work when we have more equations? Imagine a typical set of equations describing a line in $\mathbb{R}^3$, say,

$$4x + 2y + z = 6$$
$$x + y + z = 4.$$

We can eliminate one variable from such a system, for instance subtracting the second equation from the first we get $y = 2 - 3x$. But we can also subtract double of the second equation from the first to obtain $z = 2 + 2x$. We derived formulas that show how y, z depend x. It is not hard

to check that the set of points $(x, 2-3x, 2+2x)$ is identical with the line given by the original equations. However, we need another interpretation.

We can see the derived dependencies as a vector function

$$x \mapsto (2-3x, 2+2x) = \big(y(x), z(x)\big).$$

The set $\{(x, y(x), z(x))\}$ then represents the graph of this function. We have shown that it is possible to derive from the original equations a vector function whose graph agrees with the object defined by the original equations.

This is the right point of view that allows for generalization of implicit curves and implicit differentiation. When we have m equations, then with a bit of luck we can derive formulas for m chosen variables (denoted y_j) that will depend on the other variables (denoted x_i). We can write them as vectors and work with $\vec{y}$ and $\vec{x}$.

In fact, also the m equations can be written as one equation, but with a vector function on the left and zero vector on the right. For instance, the above two equations can be written as

$$\begin{pmatrix} 4x+2y+z-6 \\ x+y+z-4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In this way we obtain a compact statement.

Theorem. (Implicit function theorem)
Let $F(\vec{x}, \vec{y})$: $D \mapsto \mathbb{R}^m$ be a vector function, where $D \subseteq \mathbb{R}^n \times \mathbb{R}^m$. Assume that for a certain point $(\vec{a}, \vec{b}) \in D$ we have $F(\vec{a}, \vec{b}) = \vec{0}_m$ and that F has continuous partial derivatives on some neighborhood of $(\vec{a}, \vec{b})$.
Consider $m \times m$ and $m \times n$ matrices

$$J_{F,\vec{y}} = \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}, \qquad J_{F,\vec{x}} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{pmatrix}.$$

If $J_{F,\vec{y}}(\vec{a}, \vec{b})$ is regular, then there is a neighborhood U of the point $\vec{a}$ and a function $Y(\vec{x}) \in [C^1(U)]^m$ such that $Y(\vec{a}) = \vec{b}$ and $F(\vec{x}, Y(\vec{x})) = \vec{0}_m$ on U.
This function is unique and its Jacobi matrix is given by

$$J_Y(\vec{x}) = -[J_{F,\vec{y}}(\vec{x}, Y(\vec{x}))]^{-1} \cdot J_{F,\vec{x}}(\vec{x}, Y(\vec{x})).$$

We translate this into a less formal language: The set of equations

$$\begin{aligned} F_1(x_1, \dots, x_n, y_1, \dots, y_m) &= 0 \\ &\vdots \\ F_m(x_1, \dots, x_n, y_1, \dots, y_m) &= 0 \end{aligned}$$

can be locally transformed to formulas

$$\begin{aligned} y_1 &= y_1(x_1, \dots, x_n) \\ &\vdots \\ y_m &= y_m(x_1, \dots, x_n), \end{aligned}$$

that describe the same object as the set of solutions of the original equations. As usual, it is not guaranteed that these equations can have the form of algebraic formulas, these could be some abstract functions.

We can find their partial derivatives by implicit differentiation. When we want partial derivative with respect to some x_i, we just take the original equations, imagine that y_1 through y_m are functions of x_i, and apply partial derivative with respect to x_i to left and right-hand sides of those

equations. The resulting system then can be solved for $\frac{\partial y_j}{\partial x_i}$.

Example: Consider the equations

$$\begin{aligned} s - e^t v^2 - u + 2v^2 &= -1 \\ 3s - e^t v^2 + u - 2v^3 &= -3 \\ 3vs + e^t v - u + v^2 - 2v^3 &= 0. \end{aligned}$$

They define some object in $\mathbb{R}^4$, so we right away give up on trying to imagine what it is. But we can notice that if these equations are independent (and they look that way), then each of them removes one degree of freedom from the original four, so the object should be one-dimensional (unless it is one of the pathological ones) and we can call it a curve.

We easily check that the point $(3, 0, 8, 2)$ satisfies these equations, hence it lies on this curve. Is this curve reasonable around this point? The theorem calls for investigating a certain matrix. First we introduce appropriate formal notation:

$$\begin{pmatrix} s - e^t v^2 - u + 2v^2 \\ 3s - e^t v^2 + u - 2v^3 \\ 3vs + e^t v - u + v^2 - 2v^3 \end{pmatrix} = \begin{pmatrix} -1 \\ -3 \\ 0 \end{pmatrix}$$

$$\implies F(s,t,u,v) = (s - e^t v^2 - u + 2v^2, 3s - e^t v^2 + u - 2v^3, 3vs + e^t v - u + v^2 - 2v^3).$$

The target space is three-dimensional, so if assumptions are met, we should be able (theoretically) to express three of the variables as dependent on the fourth one. Looking at the equations, I do not fancy having to find a formula for v as it appears in equations as several powers, so I decided to make v the free variable. In other words, I will focus on the possibility of having functions $s = s(v)$, $t = t(v)$, $u = u(v)$. If you do not like my choice, you are welcome to make your own and do the calculations, I am not brave enough.

The theorem tells us to investigate a certain matrix.

$$J_{F,(s,t,u)} = \begin{pmatrix} \frac{\partial F_1}{\partial s} & \frac{\partial F_1}{\partial t} & \frac{\partial F_1}{\partial u} \\ \frac{\partial F_2}{\partial s} & \frac{\partial F_2}{\partial t} & \frac{\partial F_2}{\partial u} \\ \frac{\partial F_3}{\partial s} & \frac{\partial F_3}{\partial t} & \frac{\partial F_3}{\partial u} \end{pmatrix} = \begin{pmatrix} 1 & -e^t v^2 & -1 \\ 3 & -e^t v^2 & 1 \\ 3v & e^t v & -1 \end{pmatrix}.$$

We need to find its determinant at the given point.

$$\det\big(J_{F,(s,t,u)}(3,0,8,2)\big) = \det \begin{pmatrix} 1 & -4 & -1 \\ 3 & -4 & 1 \\ 6 & 2 & -1 \end{pmatrix} = -64.$$

The determinant is not zero, hence the curve is not pathological at $(3, 0, 8, 2)$ and there must be some neighborhood of this point where the curve can be seen as the graph of some vector function. We are more interested in the interpretation that it should be (theoretically) possible to deduce from the three equations formulas $s = s(v)$, $t = t(v)$, $u = u(v)$. Can we actually find them?

We will treat v as a parameter. Subtracting first from the second equation, and adding v times the third equation to the second we obtain

$$\begin{aligned} 2s + 2u - 2v^3 - 2v^2 &= -2 \\ 3s(1+v^2) + u(1-v) - v^3 - 2v^4 &= -3 \end{aligned} \implies \begin{aligned} s + u &= v^3 + v^2 - 1 \\ s(3v^2+3) + u(1-v) &= 2v^4 + v^3 - 3. \end{aligned}$$

If we multiply the first equation by $1 - v$ and subtract from the second, we get

$$s(3v^2 + v + 2) = 3v^4 + v^3 - v^2 - v - 2.$$

Thus

$$s = \frac{3v^4 + v^3 - v^2 - v - 2}{3v^2 + v + 2} = v^2 - 1.$$

Substituting into the first equation immediately above we get

$$u = (v^3 + v^2 - 1) - (v^2 - 1) = v^3.$$

Finally, substituting s and u into the first original equation yields

$$e^t v^2 = 1 + (v^2 - 1) - v^3 + 2v^2 = 3v^2 - v^3 \implies t = \ln(3 - v).$$

We obtained formulas

$$\begin{aligned} s(v) &= v^2 - 1, \\ t(v) &= \ln(3 - v), \\ u(v) &= v^3 \end{aligned}$$

that describe the given curve on some neighborhood of the point $(3, 0, 8, 2)$. The vector function from the theorem is

$$G(v) = \big(v^2 - 1, \ln(3 - v), v^3\big).$$

We got lucky and found the formulas, so we can also easily find derivatives

$$\begin{aligned} s'(v) &= 2v, \\ t'(v) &= \frac{-1}{3 - v}, \\ u'(v) &= 3v^2 \end{aligned}$$

that should be valid on some neighborhood around the point $(3, 0, 8, 2)$ on the curve, that is, for v close to 2. In particular, $s'(2) = 4$.

If we were not that lucky, then we would have two options to find the derivatives. One is to use the formula from the theorem. First we need to find the inverse matrix to $J_{F,(s,t,u)}$, which is

$$J^{-1}_{F,(s,t,u)} = \begin{pmatrix} -\frac{1}{2}\,\frac{v-1}{3v^2+v+2} & \frac{1}{2}\,\frac{v+1}{3v^2+v+2} & \frac{v}{3v^2+v+2} \\ -\frac{3}{2}\,\frac{1}{e^t v}\,\frac{v+1}{3v^2+v+2} & -\frac{1}{2}\,\frac{1}{e^t v}\,\frac{3v-1}{3v^2+v+2} & \frac{1}{e^t v}\,\frac{2}{3v^2+v+2} \\ -\frac{3}{2}\,\frac{v^2+1}{3v^2+v+2} & \frac{1}{2}\,\frac{3v^2+1}{3v^2+v+2} & -\frac{v}{3v^2+v+2} \end{pmatrix}.$$

We found it using Gaussian elimination and the $(A|E_n) \mapsto (E_n|A^{-1})$ trick, and it definitely was not pleasant as one can guess from the outcome. This matrix should be multiplied by the matrix

$$J_{F,v} = \begin{pmatrix} \frac{\partial F_1}{\partial v} \\ \frac{\partial F_2}{\partial v} \\ \frac{\partial F_3}{\partial v} \end{pmatrix} = \begin{pmatrix} -2e^t v + 4v \\ -2e^t v - 6v^2 \\ 3s + e^t + 2v - 6v^2 \end{pmatrix}.$$

According to the formula we change the sign of the outcome to obtain

$$\begin{pmatrix} s' \\ t' \\ u' \end{pmatrix} = \begin{pmatrix} v\frac{9v^2+3v-2+e^t-3s}{3v^2+v+2} \\ -\frac{1}{e^t v}\,\frac{9v^3-21v^2-2v+6e^t v^2+2e^t v+2e^t+6s}{3v^2+v+2} \\ -v\frac{-9v^3-5v-6+e^t-3s}{3v^2+v+2} \end{pmatrix}$$

In general, we would be able to use this formula only at known points of the curve. We know the point $(3, 0, 8, 2)$, so substituting $s = 3$, $t = 0$, $u = 8$, $v = 2$ we get, say,

$$s'(2) = 2\frac{36 + 6 - 2 + 1 - 9}{12 + 2 + 2} = 4.$$

This agrees with the direct calculation above. Since here we actually have formulas for $s(v)$ and $t(v)$, we can substitute them into the formula for s' and obtain

$$s' = v\frac{9v^2 + 3v - 2 + e^{\ln(3-v)} - 3(v^2 - 1)}{3v^2 + v + 2} = v\frac{6v^2 + 2v + 4}{3v^2 + v + 2} = 2v.$$

So it seems that the formula with Jacobi matrices really provides the right information.

However, finding that inverse matrix by hand was quite challenging, so it is worth trying the second possible approach, namely implicit differentiation. First we would differentiate the given

equations, assuming that s, t, u are functions of v now.

$$\begin{aligned} s' - e^t t' v^2 - e^t 2v - u' + 4v &= 0 \\ 3s' - e^t t' v^2 - e^t 2v + u' - 6v^2 &= 0 \\ 3s + 3vs' + e^t t' v + e^t - u' + 2v - 6v^2 &= 0. \end{aligned}$$

This can be solved for s', t', u'. Note that now we also take s, t, u as parameters, so this system will always be linear.

$$\begin{aligned} s' - e^t v^2 t' - u' &= e^t 2v - 4v \\ 3s' - e^t v^2 t' + u' &= e^t 2v + 6v^2 \\ 3vs' + e^t v t' - u' + 2v &= -3s - e^t + 6v^2. \end{aligned}$$

We can use the Cramer rule to obtain

$$s' = \frac{-2e^t v^2 (e^t + 9v^2 + 3v - 2 - 3s)}{-2e^t v(3v^2 + v + 2)} = v \frac{e^t + 9v^2 + 3v - 2 - 3s}{3v^2 + v + 2}.$$

This is the same formula that we obtained using the formal approach with Jacobi matrices, and it did seem a bit easier. The reader can confirm that also formulas that we obtain for t' and u' in this way agree.

△

8. More on integral

In this chapter we introduce integral for functions of more variables. There is a number of possible approaches, we will follow the definition of the Riemann integral for functions of one variable. It relies on the idea of partitioning an interval into segments, then we add areas of rectangles erected over those segments using the function f. This is wrapped in a mechanism that motivates us to take finer and finer partitions and thus obtain better approximations of the integral.

An important factor in the success of this approach is our ability to measure segments (we use their lengths to find the areas). Relative simplicity of this approach stems from our ability to split intervals into segments easily.

8a. Sets

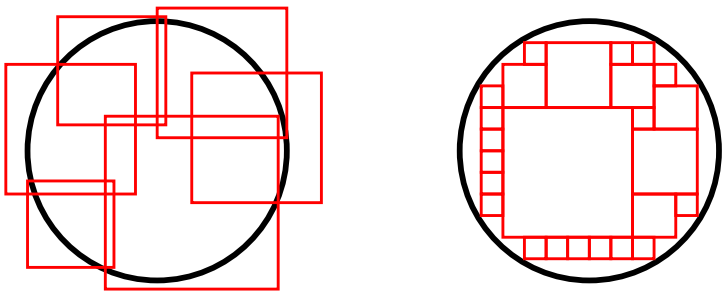
Now we will apply these ideas to integral of a function of n variables over a set Ω from $\mathbb{R}^n$, where $n \in \mathbb{N}$. We would like to split integrating domains into some "basic shapes" in $\mathbb{R}^n$, just like we split integrating intervals into segments in one dimension. However, already for $n = 2$ we run into trouble. We would definitely like to integrate over discs, but we cannot cut a disc into parts whose shapes would be easily handled (squares, discs, triangles, ...).

It is useful to recognize that the problem is not in higher dimension but in higher ambition. In one dimension we settled for integrating over intervals. If we were happy integrating functions of two variables just over rectangles, then we could have easily cut them into smaller rectangles. But we are not that modest, we also want to integrate over other shapes and then it is no surprise that we encounter complications.

In order to overcome them we have to look closer at the topic of cutting sets into smaller parts. At the moment we are actually not really interested in the function that we integrate, it is just a question of working with sets in $\mathbb{R}^n$. We will soon see that we are in fact addressing the question of measuring sets, that is, of finding some n-dimensional analogy of area or volume.

The key idea is that we will not cut sets, but approximate them using sets of convenient type. There are two practical approaches.

The first possibility is to focus on not leaving any part of the set out. This requirement is easy to capture mathematically. On the left in the picture below we can see an approximation of a disc using squares. It is obvious that we overdid it, as many of the squares overlap. This is not a problem, because similarly to the Riemann integral we will set up a framework that will motivate us to reduce the waste.



On the right we see the second approach to approximating a disc using squares. This time the motivation is to avoid waste and we are not concerned that some parts of the disc are not covered. How do we avoid overdoing it? If we demanded that the sets that we use are mutually disjoint, then it would restrict us in many situations. It is more practical to allow sets to overlap at their borders. This has an elegant specification, we can ask that the interiors of sets do not overlap.

Definition.
Consider the set $\Omega \subseteq \mathbb{R}^n$ and a finite collection of sets $\Omega_1, \dots, \Omega_m \subseteq \mathbb{R}^n$.
We say that $\{\Omega_k\}$ is a **covering** (or cover) of the set Ω if $\Omega \subseteq \bigcup_{k=1}^{m} \Omega_k$.
We say that $\{\Omega_k\}$ is a **filling** of the set Ω if $\Omega_k \subseteq \Omega$ for all k and $\Omega_k^O \cap \Omega_l^O = \emptyset$ for all $k \neq l \in \{1, \dots, m\}$.

Here we should warn the reader that while the term "covering" is an official terminology, the name "filling" was invented here so that we can talk about things intuitively. In textbooks people use various tricks to avoid this second concept, so they do not need to find a name for it.

Given the nature of coverings and fillings, the following should be true: If we had a way to "measure" the participating sets, that is, if we had something like an n-dimensional analogue of volume, then all coverings $\{\Omega_k\}$ of Ω should satisfy

$$\mathrm{vol}_n(\Omega) \leq \sum_{k=1}^{m} \mathrm{vol}_n(\Omega_k),$$

while all fillings $\{\Omega_k\}$ of Ω should satisfy

$$\mathrm{vol}_n(\Omega) \geq \sum_{k=1}^{m} \mathrm{vol}_n(\Omega_k).$$

In other words, the definitions make sense so far, our aim is to zero in on the "volume" of our set from above and below using good approximations.

Now comes the time to decide which basic shape will be used. The main requirement is that it should have a natural generalization for all dimensions and that we should be able to determine its volume in a way that is so intuitive that we would find it reasonable also for higher dimensions (where we do not know what "volume" is yet).

This points to the n-dimensional intervals, that is, rectangles, cubes etc. In one-dimensional integration we work with segments $[a, b]$ and measure their length $b - a$. For two dimensions it is then natural to work with rectangles $[a_1, b_1] \times [a_2, b_2]$, the area is $(b_1 - a_1) \cdot (b_2 - a_2)$. In three dimensions we pass to cuboids $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ with volume $(b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3)$ and it is easy to get persuaded that it makes sense to continue in this way.

Definition.
Let $n \in \mathbb{N}$. By an **n-rectangle** we mean any set of the form

$$[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$$

for some real numbers $a_k \leq b_k$ for $k = 1, \dots, n$.
For a given n-rectangle $\Omega = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$ we define its **measure** as

$$\mathrm{m}(\Omega) = (b_1 - a_1) \cdot (b_2 - a_2) \cdots (b_n - a_n) = \prod_{i=1}^{n} (b_i - a_i).$$

A measure is a mathematical way of not saying "n-dimensional volume" but meaning it.

Intuition tells us that the length of an interval does not depend on whether it is open, half-closed or closed. With n-rectangles it is the same. Some authors prefer open intervals in the definition, many use intervals $[a_i, b_i)$, perhaps because they nicely fit one after another. One can show that this does not really matter for the final theory. We chose closed intervals here because we usually take sets as closed when integrating.

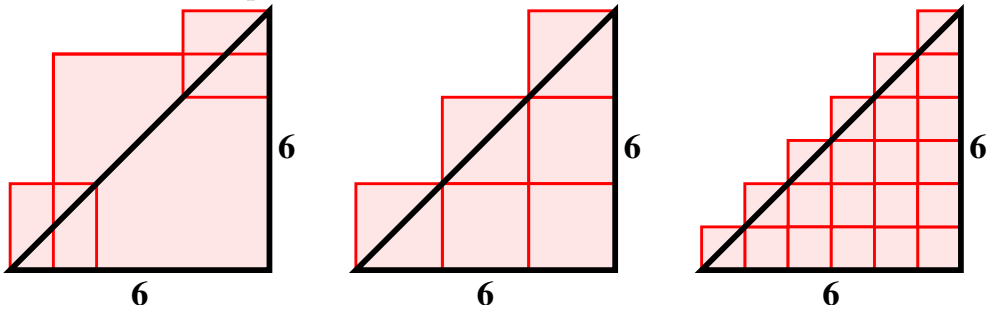
So we will cover and fill sets with n-rectangles, we will talk about rectangular coverings and rectangular fillings for short. We start by observing that in order to be able to cover a set, this

set must necessarily be bounded, because rectangles are bounded and the definition does not allow us to take infinitely many of them. In this chapter we will therefore restrict ourselves to bounded sets Ω only. All such sets can be covered because by definition, every bounded set is contained in some n-dimensional ball of a certain radius r, which is in turn contained in an n-dimensional cube of side $2r$.

How about fillings? Here it helps that we allowed $a_i = b_i$ in the definition of n-rectangles. This means that also a single point is considered an n-rectangle, its measure is obviously zero. Accordingly, every non-empty set contains some n-rectangle, so it has some rectangular filling. We also consider all sets to be non-empty in this chapter. To sum it up, every non-empty bounded set has some rectangular covering and filling.

Now we are ready to attempt the best possible approximations of sets using n-rectangles. We introduce a mechanism inspired by Darboux's approach to Riemann integral that motivates us to take the tightest approximations of a set. There are two possible ways to get close to the shape of a set Ω.

We can consider various coverings using n-rectangles. How do we recognize which coverings are better and which are worse? Every such covering $\{\Omega_k\}$ provides a certain estimate of the n-dimensional volume of the set Ω of the form $\sum \mathrm{m}(\Omega_k)$. As we observed, for a reasonably behaved notion of volume this estimate cannot be smaller than the actual volume of Ω (assuming that something of this sort actually exists). However, it can be larger, and the larger it is, the more we squandered when covering. Conversely, the smaller the estimate that we get from a covering, the better its approximation of shape should be.



In the picture we see several coverings of a triangle whose area is 18 using 2-rectangles. In the left covering we chose a large square of size 5 and two smaller ones of size 2, so we obtained the upper estimate

$$5^2 + 2 \cdot 2^2 = 33 = 18 + 15.$$

We overshot the mark by quite a bit. In the next covering we learned from this and did not overlap the squares of side 2, obtaining the upper estimate

$$6 \cdot 2^2 = 24 = 18 + 6.$$

We are closer. For the third covering we used squares of side 1 and obtained

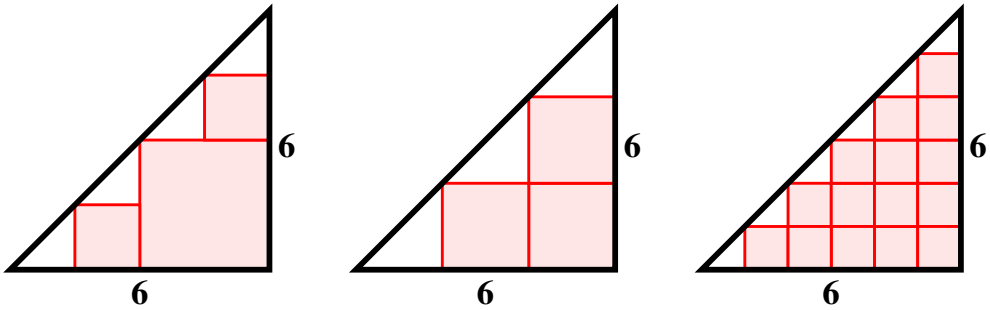
$$21 \cdot 1^2 = 21 = 18 + 3.$$

We see that as the coverings fit the shape of the triangle better and better, the sums really approach the correct value 18. It also seems that we should be able to approximate arbitrarily well by taking smaller and smaller squares.

Thus it makes sense for a general set Ω to look for the smallest number that can be obtained from coverings. For some sets Ω we can actually find a specific covering that yields the precise "volume", but for others (like the triangle) we can only get arbitrarily close to it, so it is necessary to take infimum, not minimum of possible estimates.

The other possible way is to fill the set Ω. For every filling that uses n-rectangles then $\sum \mathrm{m}(\Omega_k)$ provides a lower estimate of the n-dimensional volume. If we encourage fillings to provide values as large as possible, then we actually motivate them to provide the best possible approximations

of the set Ω.



On the left we see an irregular filling with squares that provides the lower estimate

$$3^2 + 2 \cdot 1.5^2 = 13.5 = 18 - 4.5,$$

while the regular fillings supply lower estimates $18 - 6$ and $18 - 3$. Again one would guess that by making the squares smaller one could get arbitrarily close to 18.

This brings us to the following definition.

Definition.
Let Ω be a non-empty bounded set in $\mathbb{R}^n$.
We define its **outer measure**

$$\mathrm{m}^*(\Omega) = \inf\Big\{\sum_k \mathrm{m}(\Omega_k);\ \{\Omega_k\} \text{ rectangular covering of } \Omega\Big\}$$

and its **inner measure**

$$\mathrm{m}_*(\Omega) = \sup\Big\{\sum_k \mathrm{m}(\Omega_k);\ \{\Omega_k\} \text{ rectangular filling of } \Omega\Big\}.$$

We say that the set Ω is Jordan-measurable if $\mathrm{m}^*(\Omega) = \mathrm{m}_*(\Omega)$.
Then we define its (Peano-Jordan) measure as $\mathrm{m}(\Omega) = \mathrm{m}^*(\Omega)$.

The idea is analogous to that of Riemann integral. There we also estimated from above and below and if the estimates met, then we proclaimed the function to be sufficiently reasonable. Now it is the same with shapes of sets. The key good news is that in two and three dimensions the Jordan measure agrees with area and volume, respectively.

What sets are measurable? All common sets. For instance, if we take a continuous function on an interval, then the area under the graph (as a set in $\mathbb{R}^2$) is definitely measurable. Also sets "between graphs" as discussed in chapter 6 are measurable. Moving to higher dimensions, if we have a measurable set D and two continuous functions on it, then the set "between graphs" of these two functions (which has one more dimension that D) is also measurable.

Sets that are not measurable must have a very wild boundary, usually all torn up, so we do not encounter them in typical applications.

Example: We will exhibit a set in $\mathbb{R}$ that is not measurable. Recall that n-rectangles now mean segments.

Consider the set $\Omega = \mathbb{Q} \cap [0, 1]$. It is the set of all rational numbers between 0 and 1.

Rational numbers have two important topological properties.

The first one is that rational numbers are dense, meaning that in every segment of the real line of positive length there is some rational number. This means that when we cover the set Ω with segments, then they have to join seamlessly without gaps, because if there was a gap between them (and the gap was between 0 and 1), then there would be a rational number in that gap, so it would not be a covering after all. Union of such a covering must therefore include the whole interval $[0, 1]$. Consequently, when we add lengths of segments from such a covering, we cannot get less

than 1. Thus the infimum obtained from coverings cannot be less than one. And it also cannot be larger, because the interval $[0,1]$ of length 1 by itself is also a covering of Ω. Thus $\mathrm{m}^*(\Omega) = 1$.

The second important property is that also the irrational numbers are dense, in other words, there is no connected part of the real line composed solely of rational numbers. This means that Ω cannot contain a subset in the form of an interval of positive length. Every rectangular filling must therefore consist only of one-point intervals whose measure is zero, as is the resulting sum. We have shown that all estimates obtained through fillings are zero, so also their supremum must be zero. Mathematically, $\mathrm{m}_*(\Omega) = 0$.

We have shown that $\mathrm{m}_*(\Omega) < \mathrm{m}^*(\Omega)$, so the set Ω is not measurable.

This example can be easily generalized for higher dimensions, for instance in $\mathbb{R}^2$ we can take all points with rational coordinates from the square $[0,1] \times [0,1]$.

△

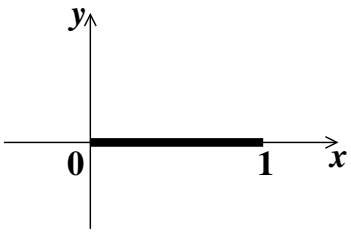
Expert remark: Besides the Jordan measure there are also other measures. In analysis, the most popular measure is the Lebesgue measure that uses infinite coverings. It is significantly more powerful than the Jordan measure and it can measure even some wild sets (for instance $\mathbb{Q}$), but introducing it takes quite a bit of work.

Actually, this exposition could also use some mathematical propping, for instance it is not clear whether the supremum and infimum in the definition exist as finite numbers. It actually follows from a key lemma that is very similar to the key lemma in the introduction of Riemann integral: It is not possible that some covering would supply a smaller estimate for the measure of some set than some filling of it. Also the proof of this lemma is similar to the version for the Riemann integral, namely through refinement.

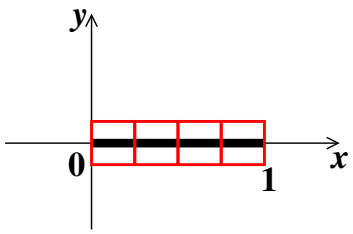
Remark: In chapter 6 we mentioned an interesting category of sets that formally belong to an n-dimensional space but they are essentially of smaller dimension. If they are measurable, then their measure should be zero.

Let's look at the case of a segment in $\mathbb{R}^2$. For simplicity we will take the interval $[0,1]$ on the x-axis, but now we see it as the set

$$\Omega = \{(x,0);\ 0 \le x \le 1\}.$$



It is obvious that the only 2-rectangles (that is, rectangles) that can fit in this Ω are rectangles of the form $[0,0] \times [a_2, b_2]$. Since $0 - 0 = 0$, they all have measure zero. Therefore all rectangular fillings $\{\Omega_k\}$ must satisfy $\sum \mathrm{m}(\Omega_k) = \sum 0 = 0$. This proves that $\mathrm{m}_*(\Omega) = 0$.



Now we will cover, and to make our life easier we will cover using squares of side $\frac{1}{m}$. So we choose some $m \in \mathbb{N}$ and consider squares of side $\frac{1}{m}$ whose centers are on the x-axis. Because we defined our n-rectangles as closed, we can glue them next to each other and see that m of such squares are enough to cover the set Ω.

For this covering we get

$$\sum_{k=1}^{m} \mathrm{m}(\Omega_k) = \sum_{k=1}^{m} \Big(\frac{1}{m}\Big)^2 = m \cdot \frac{1}{m^2} = \frac{1}{m}.$$

If we now consider all coverings of this type and take infimum of the corresponding sums, we get $\inf\big(\frac{1}{m}\big) = 0$. Including other coverings in the infimum cannot make it larger and there is no room

for making it smaller, hence also $\mathrm{m}^*(\Omega) = 0$. Because $\mathrm{m}^*(\Omega) = \mathrm{m}_*(\Omega)$, our segment is measurable and above all, $\mathrm{m}(\Omega) = 0$ indeed.

Similarly one can show that all (reasonable) curves in the plane have measure zero, similar argument then works also for surfaces in $\mathbb{R}^3$ and other analogous situations.
△

8b. Integral

If a set is not measurable, then it does not make sense to integrate over it. If it is measurable, then it can be approximated using n-rectangles and it makes sense to erect "columns" over them and add their contributions to the integral. When introducing the classical Riemann integral we evaluated the contribution of such a "column" as the product of its height and the length of the base, here we naturally work with the measure of the base.

First we introduce basic notions. We recall that in this chapter we only work with non-empty and bounded sets Ω.

Definition.
Let f be a bounded function on a measurable set Ω.
For a covering $\{\Omega_k\}$ of the set Ω using n-rectangles we define

$$U(f, \{\Omega_k\}) = \sum_k \sup_{\Omega_k}(f) \cdot \mathrm{m}(\Omega_k).$$

We also define the **upper integral** of f on Ω as

$$\overline{S}(f, \Omega) = \inf\big\{U(f, \{\Omega_k\});\ \{\Omega_k\} \text{ rectangular covering of } \Omega\big\}.$$

For a filling $\{\Omega_k\}$ of the set Ω using n-rectangles we define

$$L(f, \{\Omega_k\}) = \sum_k \inf_{\Omega_k}(f) \cdot \mathrm{m}(\Omega_k).$$

We also define the **lower integral** of f on Ω as

$$\underline{S}(f, \Omega) = \sup\big\{L(f, \{\Omega_k\});\ \{\Omega_k\} \text{ rectangular filling of } \Omega\big\}.$$

To make this definition work we again make use of a key lemma: When we have some covering $\{\Omega_k\}$ and some filling $\{\widetilde{\Omega}_l\}$ for the set Ω, then for any $f > 0$ we must have

$$L(f, \{\widetilde{\Omega}_l\}) \le U(f, \{\Omega_k\}).$$

This time the proof takes more work compared to one dimension. It then follows that the definitions of the upper and lower integral make sense, they always provide real numbers and we always have

$$\underline{S}(f, \Omega) \le \overline{S}(f, \Omega).$$

Now we can pass to a familiar definition.

Definition.
Let f be a bounded function on a measurable set Ω.
We say that f is **Riemann integrable** on Ω if $\underline{S}(f, \Omega) = \overline{S}(f, \Omega)$.
If true, then we define

$$\int_\Omega \cdots \int f(\vec{x})\, d[\vec{x}] = \overline{S}(f, \Omega).$$

For instance all continuous functions are integrable, which is enough for common applications. The integral that we defined has all the usual properties, for instance linearity and additivity with respect to integrating domain.

Theorem.
Let f, g be functions integrable on a measurable set $\Omega \subseteq \mathbb{R}^n$, let $c \in \mathbb{R}$.
Then also the function $cf + g$ is integrable on Ω and

$$\int\!\cdots\!\int_{\Omega} cf + g\, d[\vec{x}] = c \int\!\cdots\!\int_{\Omega} f\, d[\vec{x}] + \int\!\cdots\!\int_{\Omega} g\, d[\vec{x}].$$

Theorem.
Let f be a function integrable on a measurable set Ω.
Assume that sets $\Omega_1, \ldots, \Omega_m$ are measurable, their interiors are pairwise disjoint, and $\Omega = \bigcup_{k=1}^{m} \Omega_k$.
Then

$$\int\!\cdots\!\int_{\Omega} f\, d[\vec{x}] = \sum_{k=1}^{m} \int\!\cdots\!\int_{\Omega_k} f\, d[\vec{x}].$$

If this were a proper textbook and not just an illustrated introduction, then we would next see proofs of these theorems and also other statements, for instance that the Riemann integral that we just defined can be calculated using slices as outlined in chapter 6. However, it is not. We also looked at substitution in that chapter, so we pretty much said it all, we will just fire one parting shot in the form of an integral.

Example: Consider the function $f(x, y) = (x^2 + y^2 + 1)^3$ on a set Ω, where Ω is the disc in $\mathbb{R}^2$ centered at the origin and with radius 3. We want to integrate, direct evaluation leads to the repeated integral

$$\int_{-3}^{3} \int_{-\sqrt{9-x^2}}^{\sqrt{9-x^2}} (x^2 + y^2 + 1)^3 dy\, dx.$$

An experienced integrator notices right away that the preferable approach, that is, the substitution $w = x^2 + 1 + y^2$, is not feasible. This integral has to be handled by expanding the cubic power and then integrating all the resulting terms, which is not very appealing.

However, integrating over a disc points to polar coordinates $x = r\cos(\varphi)$, $y = r\sin(\varphi)$ that map $D = [0, 3] \times [0, 2\pi)$ onto Ω. We find the jacobian:

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

hence $dS = r\, dS[r, \varphi]$. We get

$$\begin{aligned} \iint_{\Omega} (x^2 + y^2 + 1)^3\, dxdy &= \int_0^3 \int_0^{2\pi} (r^2\cos^2(\varphi) + r^2\sin^2(\varphi) + 1)^3 r\, d\varphi\, dr = \int_0^3 \int_0^{2\pi} (r^2 + 1)^3 r\, d\varphi\, dr \\ &= \int_0^3 \Big[r(r^2 + 1)^3 \varphi \Big]_{\varphi=0}^{\varphi=2\pi} dr = \int_0^3 2\pi r(r^2 + 1)^3\, dr = \left| \begin{matrix} w = r^2 + 1 \\ dw = 2r\, dr \end{matrix} \right| \\ &= \int_0^{10} \pi w^3\, dw = \Big[\pi \frac{1}{4} w^4 \Big]_0^{10} = \frac{\pi}{4} \cdot 9999. \end{aligned}$$

△

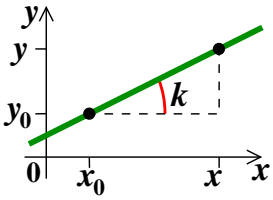
9. Appendix

9a. Geometric objects in $\mathbf{R}^n$

Here we recall some popular objects in $\mathbb{R}^n$.

Linear entities

We start with a line in $\mathbb{R}^2$ that passes through a given point (x_0, y_0). There are several ways to specify its direction.

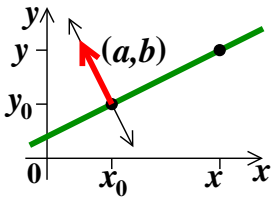


• One popular way is to specify its slope k. The equation of the line is then

$$y = k(x - x_0) + y_0 \quad \text{or} \quad y - y_0 = k(x - x_0).$$

The second form reminds us of the actual meaning of k. It is the ratio of height to base in any triangle attached to the line. Some people prefer to write $y = kx + q$ and determine q based on data.

The major disadvantage of this approach is that this type of equation does not allow us to capture a vertical line.



• Another popular description of a line is using an equation of the form

$$ax + by = c \quad \text{or} \quad ax + by + c = 0.$$

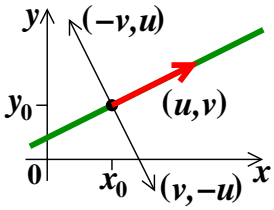
The coefficients a, b form a vector (a, b) that is perpendicular to the line and uniquely determines its direction. It is called a normal vector to the given line, and it is definitely not unique. Any non-zero multiple of a normal vector will again work as a normal vector.

Given a normal vector $\vec{n} = (a, b)$ and a point (x_0, y_0), the corresponding line can be described by the equation

$$a(x - x_0) + b(y - y_0) = 0.$$

This can be also written using the dot product as $(a, b) \bullet (x - x_0, y - y_0) = 0$. This specifies the line as the set of all points (x, y) such that the vector $(x, y) - (x_0, y_0)$ is perpendicular to $\vec{n}$.

Rewriting $y = kx + q$ as $kx + (-1)y + q = 0$ we see that given a slope k, we can use $(k, -1)$ for the normal vector of our line. Conversely, given a line specified as $ax + by + c = 0$ with $b \neq 0$, then $y = -\frac{a}{b}x - \frac{c}{b}$, that is, this line has the slope $k = -\frac{a}{b}$.



• The direction of a line can be also specified through its directional vector (u, v). Then we can use parametric description of points of this line

$$(x, y) = (x_0, y_0) + t(u, v), \quad \text{that is,} \quad \begin{aligned} x &= a + tu, \\ y &= b + tv, \ t \in \mathbb{R}. \end{aligned}$$

Eliminating t we obtain $y = \frac{v}{u}x - \frac{v}{u}a + b$, and then $vx - uy - va + ub = 0$. Thus such a line has slope $k = \frac{v}{u}$ and normal vector $\vec{n} = (v, -u)$ or $\vec{n} = (-v, u)$ or any non-zero multiple of these two.

Conversely, a line specified by the equation $y = kx + q$ has directional vector $(1, k)$, and a line given by $ax + by + c = 0$ has directional vector $(b, -a)$ or $(-b, a)$ or any non-zero multiple.

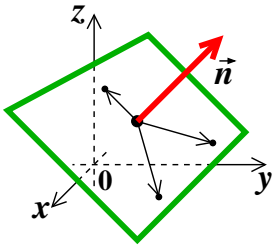
These three ways to capture a line in $\mathbb{R}^2$ have different potentials for generalization.

The parametric description can be directly applied also to lines in $\mathbb{R}^n$ for any $n \geq 2$. Given a point $\vec{a}$ and directional vector $\vec{u}$, the points of the line passing through $\vec{a}$ in direction $\vec{u}$ are given as

$$\vec{x} = \vec{a} + t\vec{u}, \ t \in \mathbb{R}.$$

This is our favourite way to capture lines in higher dimensions.

The slope version does not offer any generalization. The normal vector version does offer a generalization, but in a different direction. Already in the space $\mathbb{R}^3$ we easily observe that given some vector $\vec{n}$ located at a specific point $\vec{a}$, there are infinitely many lines that pass through that point and are perpendicular to this vector. This means that in higher dimensions it is not possible to specify a line using a normal vector.



On the other hand, it is not difficult to imagine that points $\vec{x} \in \mathbb{R}^3$ such that $\vec{x} - \vec{x}_0$ is perpendicular to $\vec{n} = (a, b, c)$ form a flat plane in $\mathbb{R}^3$. We are in fact taking the union of all those perpendicular lines. The resulting equation is

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

Generally, any equation of the form $ax + by + cz = d$ or $ax + by + cz + d = 0$ defines a plane in $\mathbb{R}^3$ whose normal vector is $\vec{n} = (a, b, c)$.

The tilt of such a plane is determined by its normal vectors (again, there are infinitely many possible normal vectors). In particular, when we want to know whether two planes are parallel, we ask the same question about their normal vectors.

It is impossible (for most people) to visualize that the set of all points $\vec{x} \in \mathbb{R}^4$ such that $\vec{x} - \vec{x}_0$ is perpendicular to $\vec{n}$ is in fact an object whose essential dimension is 3 and that is also flat. But this does work. In general, every non-zero vector $\vec{n} \in \mathbb{R}^n$ located at a certain point $\vec{a}$ determines a specific object, a "flat" set of dimension $n - 1$ that is perpendicular to $\vec{n}$. Such objects are called hyperplanes. Mathematically, by a hyperplane in $\mathbb{R}^n$ we mean any set of the form $H + \vec{a}$, where H is some linear subspace of $\mathbb{R}^n$ of dimension $n - 1$.

A hyperplane given by a point $\vec{a}$ and a normal vector $\vec{n}$ is described by the equation

$$\vec{n} \bullet (\vec{x} - \vec{a}) = 0,$$

that is,

$$\sum_{i=1}^{n} n_i(x_i - a_i) = 0.$$

Intersection of hyperplanes leads to sets of smaller dimensions, which offers another way to capture lines in $\mathbb{R}^n$, namely as an intersection of $n - 1$ non-parallel hyperplanes.

Quadrics (quadratic entities)

Intuitively, these are objects described by equations that feature quadratic polynomials.

In $\mathbb{R}^2$, these objects are known as conic sections. There are two basic types:

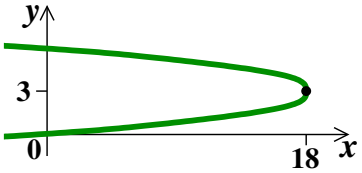
• Equations where one variable is not squared: There is just one shape in this family. Equations $y = ax^2 + bx + c$ and $x = ay^2 + by + c$ define **parabolas**. The former are open up (if $a > 0$) or down (if $a < 0$), while the latter equation defines parabolas opened to the right or to the left.

We can identify important features, in particular the position of vertex, by completing square.

For instance, the equation $x = 12y - 2y^2$ can be rewritten as

$$x = -2(y^2 - 6y) = -2(y^2 - 6y + 9 - 9) = -2((y - 3)^2 - 9) = 18 - 2(y - 3)^2.$$

We see that this equation describes a parabola opened to the left, with vertex $(18, 3)$.



• Equations where squares of both variables are present: The typical representative is the equation $ax^2 + by^2 = c$ with $a, b \neq 0$. There are two basic types of shapes, and one has a special case that is worth pointing out.

– **Circle** centered at (x_0, y_0) of radius $r \geq 0$, given by the equation

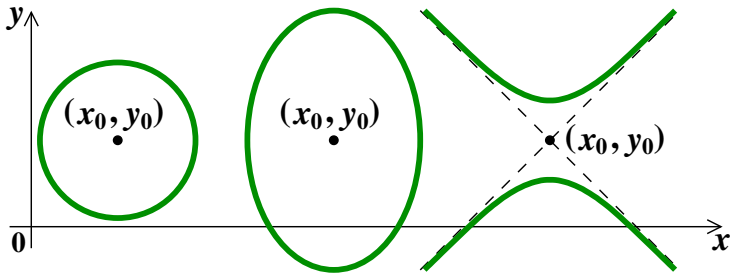
$$(x - x_0)^2 + (y - y_0)^2 = r^2.$$

– **Ellipse** centered at (x_0, y_0), with semi-axes $a, b > 0$, given by the equation

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = 1.$$

– **Hyperbola** given by the equation

$$\frac{(x - x_0)^2}{a^2} - \frac{(y - y_0)^2}{b^2} = 1.$$



Equations with linear terms are handled by completing squares. For instance,

$$x^2 + y^2 - 4x + 2y = 4 \implies (x - 2)^2 + (y + 1)^2 = 3^2.$$

This describes the circle of radius 3 centered at $(2, -1)$.

The catalogue of quadric surfaces in $\mathbb{R}^3$ is considerably richer. We will therefore list just some prominent types centered at the origin, it is easy to adopt these equations to other centers.

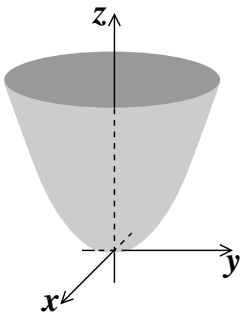
• We start with equations where one variable is not squared:

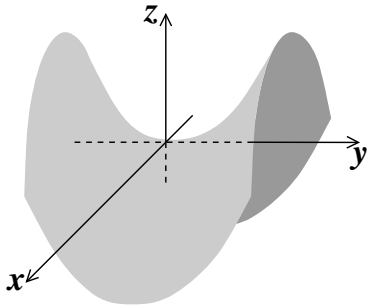
– $x^2 + y^2 - z = 0$, that is, $z = x^2 + y^2$, describes a **paraboloid**. See example in section 1b.

This paraboloid can be flattened from sides to create an **elliptic paraboloid** given by

$$z = \frac{x^2}{a^2} + \frac{y^2}{b^2}.$$

– $x^2 - y^2 - z = 0$, that is, $z = x^2 - y^2$, describes a **hyperbolic paraboloid**. This is an interesting saddle-shaped object, see example in section 1b.



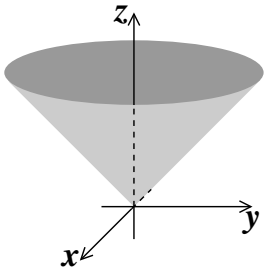


• And now equations where all variables are squared:

– $x^2 + y^2 + z^2 = r^2$ defines a **sphere** of radius r. This is a popular special case of the next type.

– $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ defines an **ellipsoid**.

– $x^2 + y^2 - z^2 = 0$, that is, $z = \sqrt{x^2 + y^2}$ describes a **cone** with vertex in the origin. Note that for a point (x, y) in the xy-plane, $r = \sqrt{x^2 + y^2}$ shows its distance from the origin. The formula $z = \sqrt{x^2 + y^2}$ says that at (x, y), the value z is just the distance from the origin. The graph therefore gets higher at the same speed as we get further from the origin, in other words, the side of the cone has slope 1:1.

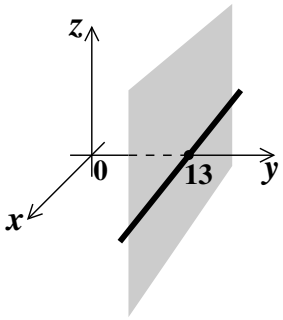


A more complete list of quadrics in $\mathbb{R}^3$ can be easily found.

Degenerate curves and surfaces

These are equations in which some variable is missing. The rule of thumb is that we draw such a degenerate object only in the space whose coordinates are mentioned in the equation, and then extend it to infinity in directions of variables that are left out of the equation.

Consider the equation $y = 13$. In $\mathbb{R}$ it defines a point on the line. In higher dimensions it becomes degenerate. In $\mathbb{R}^2$ it defines a horizontal line. In $\mathbb{R}^3$ it defines a plane that is parallel to the xz-plane. It works analogously for equations $x = x_0$ and $z = z_0$.



Similarly, $x + 2y = 13$ defines a line in $\mathbb{R}^2$. Considered in $\mathbb{R}^3$ it becomes degenerate and it defines the vertical plane passing through that line in the xy-plane.

In $\mathbb{R}^3$, the equation $x^2+y^2 = r^2$ defines a vertical **cylinder** of radius r, while $y^2+z^2 = r^2$ defines a cylinder that is parallel with the x-axis.

In general, the equation

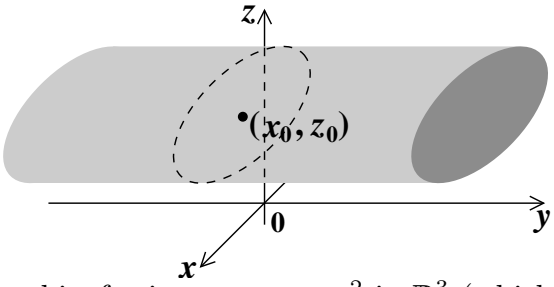
$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1$$

defines an **elliptic cylinder**, that is, a cylinder parallel to the z-axis that is flattened from sides and its main axis passes through the point $(x_0, y_0, 0)$.

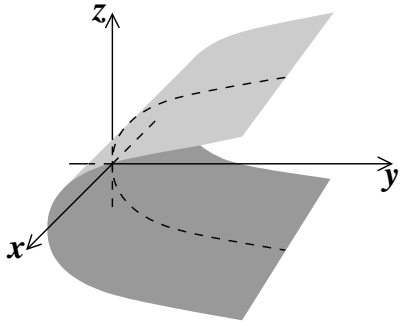
Again, the variables are interchangeable, so for instance

$$\frac{(x-x_0)^2}{a^2} + \frac{(z-z_0)^2}{c^2} = 1$$

defines a flattened cylinder parallel with the y-axis as in the picture.



The reader can surely work out many more examples like this, for instance $y = z^2$ in $\mathbb{R}^3$ (which is called the **parabolic cylinder**, by the way).



Fortunately we almost never have to work in higher dimensions, because there the number of quadric shapes is overwhelming and we cannot visualize them anyway. It is worth mentioning that One shape scales well, namely the sphere, which is in all dimensions described by the equation

$$\sum (x_i - a_i)^2 = r^2.$$

In general we can consider ellipsoids

$$\sum \frac{(x_i - a_i)^2}{r_i^2} = 1.$$

9b. Topological notions in $\mathbf{R}^n$

Notation:

Let $n \in \mathbb{N}$. The symbol $\mathbb{R}^n$ stands for the space of all n-dimensional vectors with real entries, denoted $\vec{x} = (x_1, \dots, x_n)$, but also $\bar{x}$ or $\boldsymbol{x}$. The numbers $x_i \in \mathbb{R}$ are **coordinates** or **components**.

For vectors we have basic operations.

Definition.
For a vector $\vec{x} \in \mathbb{R}^n$ and a scalar $c \in \mathbb{R}$ we define the **multiple** as

$$c\vec{x} = (cx_1, \dots, cx_n).$$

For vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ we define their

• **sum** as

$$\vec{x} + \vec{y} = (x_1 + y_1, \dots, x_n + y_n);$$

• **dot product** as

$$\vec{x} \bullet \vec{y} = x_1y_1 + \cdots + x_ny_n = \sum_{i=1}^{n} x_iy_i.$$

For vectors $\vec{x}, \vec{y}$ z $\mathbb{R}^3$ we also have the **cross product**

$$\begin{aligned}\vec{x} \times \vec{y} &= \big(x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_1y_2 - x_2y_1\big)\\ &= \left(\det\begin{pmatrix} x_2 & x_3\\ y_2 & y_3\end{pmatrix}, \det\begin{pmatrix} x_3 & x_1\\ y_3 & y_1\end{pmatrix}, \det\begin{pmatrix} x_1 & x_2\\ y_1 & y_2\end{pmatrix}\right)\\ &= \det\begin{pmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3\\ x_1 & x_2 & x_3\\ y_1 & y_2 & y_3\end{pmatrix} = \det\begin{pmatrix} \vec{\imath} & \vec{\jmath} & \vec{k}\\ x_1 & x_2 & x_3\\ y_1 & y_2 & y_3\end{pmatrix}.\end{aligned}$$

The dot product is also called the scalar product or inner product and there are some alternative notations and formulas for it, for instance

$$\vec{x} \bullet \vec{y} = \langle \vec{x}, \vec{y} \rangle = \vec{x} \cdot \vec{y}^{\,T}.$$

In the last formula, a row vector is multiplying a column vector as if they were two matrices.

The notion of distance, that is, a metric, is provided by a norm.

Definition.
For a vector $\vec{x} \in \mathbb{R}^n$ we define its **(Euclidean) norm** as

$$\|\vec{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

For vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ we define their **(Euclidean) distance** as

$$\|\vec{x} - \vec{y}\,\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

The linear space $\mathbb{R}^n$ endowed with the Euclidean norm is called the **Euclidean space** of dimension n, sometimes denoted $(\mathbb{R}^n, \|\cdot\|_2)$ or E_n. We will just write $\mathbb{R}^n$, because we do not consider other norms here.

Every norm must satisfy certain properties simply due to being a norm, see (i)–(iii) below. We also add some other useful observations about the Euclidean norm.

Theorem.
Let $n \in \mathbb{N}$. The Euclidean norm on $\mathbb{R}^n$ satisfies the following properties:
(i) $\|\vec{x}\| = 0$ if and only if $\vec{x} = \vec{0}$, for every $\vec{x} \in \mathbb{R}^n$;
(ii) $\|c\vec{x}\| = |c| \cdot \|\vec{x}\|$ for all $\vec{x} \in \mathbb{R}^n$ and $c \in \mathbb{R}$;
(iii) $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$; (triangle inequality)
(iv) $\|\vec{x}\| = \sqrt{\vec{x} \bullet \vec{x}}$ for all $\vec{x} \in \mathbb{R}^n$;
(v) $|\vec{x} \bullet \vec{y}| \leq \|\vec{x}\| \cdot \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$; (Cauchy-Schwarz inequality)
(vi) $\max |x_i| \leq \|\vec{x}\| \leq \sqrt{n} \max |x_i|$ for all $\vec{x} \in \mathbb{R}^n$.

The dot product can be calculated using the norm and information about position of vectors.

Fact.
Let $n \in \mathbb{N}$. For $\vec{x}, \vec{y} \in \mathbb{R}^n$ we have $\vec{x} \bullet \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos(\alpha)$, where α is the angle between vectors $\vec{x}$ and $\vec{y}$.

The dot product is a test of perpendicularity. $\vec{x} \bullet \vec{y} = 0$ is true if and only if $\vec{x} \perp \vec{y}$. On the other hand, for parallel vectors $\vec{x}, \vec{y}$ the product $\vec{x} \bullet \vec{y}$ attains the largest possible value $\pm\|\vec{x}\| \, \|\vec{y}\|$ (see Cauchy-Schwarz above), where the sign depends on whether the orientations agrees or vectors go in the opposite directions.

Definition.
Let M be a subset of $\mathbb{R}^n$.
We say that it is **bounded** if there is $K > 0$ such that $\|\vec{x}\| \leq K$ for all $\vec{x} \in M$.
We define **diameter** of M as $\operatorname{diam}(M) = \sup\{\|\vec{x} - \vec{y}\|;\ \vec{x}, \vec{y} \in M\}$.

Obviously, a set M is bounded exactly if $\operatorname{diam}(M) < \infty$.

Definition.
Let $\vec{a} \in \mathbb{R}^n$ and $\varepsilon > 0$. We define
• ε-**neighborhood** of a point $\vec{a}$ as $U_\varepsilon(\vec{a}) = \{\vec{x} \in \mathbb{R}^n;\ \|\vec{x} - \vec{a}\| < \varepsilon\}$;
• **reduced** ε-**neighborhood** of a point $\vec{a}$ as $P_\varepsilon(\vec{a}) = \{\vec{x} \in \mathbb{R}^n;\ 0 < \|\vec{x} - \vec{a}\| < \varepsilon\}$.

$U_R(\vec{a})$ is an open "ball" with radius R centered at $\vec{a}$. A set M is bounded if and only if it is a subset of some $U_R(\vec{a})$; then $\operatorname{diam}(M) \leq 2R$.

Definition.
Let M be a subset of $\mathbb{R}^n$. We say that $\vec{x} \in \mathbb{R}^n$ is
• an **inner point** of M if $U \subseteq M$ for some $U = U_\varepsilon(\vec{x})$;
• an **outer point** of M if $U \cap M = \emptyset$ for some $U = U_\varepsilon(\vec{x})$;
• a **boundary point** of M if $U \cap M \neq \emptyset$ and $U - M \neq \emptyset$ for every $U = U_\varepsilon(\vec{x})$;
• an **isolated point** of M if $U \cap M = \{\vec{x}\}$ for some $U = U_\varepsilon(\vec{x})$;
• an **accumulation point** of M if $P \cap M \neq \emptyset$ for every $P = P_\varepsilon(\vec{x})$.

Equivalently, $\vec{x}$ is an accumulation point of M iff $\vec{x}(k) \to \vec{x}$ for some sequence $\{\vec{x}(k)\} \subseteq M - \{\vec{x}\}$.

Fact.
If a set $M \subseteq \mathbb{R}^n$ has an accumulation point, then it must be infinite.

The opposite statement is not true, for instance $\mathbb{N}$ is an infinite set in $\mathbb{R}^1$ but it does not have any accumulation point.

Theorem.
Every infinite bounded subset of $\mathbb{R}^n$ has at least one accumulation point.

Definition.
Let M be a subset of $\mathbb{R}^n$. We define its
• **interior** M^O as the set of all inner points of M;
• **boundary** ∂M as the set of all boundary points of M;
• **closure** $\overline{M}$ as $M \cup \partial M$.

For any set M we have $M^O \subseteq M \subseteq \overline{M}$, and $M^O \cap \partial M = \emptyset$.

The closure is the set of all points that can be approached arbitrarily close with points from M.

$$\vec{x} \in \overline{M} \iff \vec{x}(k) \to \vec{x} \text{ for some sequence } \{\vec{x}(k)\} \subseteq M.$$

This also includes all points from M because each such point can be approached arbitrarily close using the same point.

Definition.
Let M be a subset of $\mathbb{R}^n$.
We say that M is **open** if $M^O = M$.
We say that M is **closed** if $\overline{M} = M$.

A set is open if all its points are its interior points.

A set is closed if every its accumulation point belongs to this set. In other words, for every sequence $\{\vec{x}(k)\} \subseteq M$ that converges to some $\vec{x}$ one must have $\vec{x} \in M$.

Definition.
Let M be a subset of $\mathbb{R}^n$. We say that it is **connected** if it is not possible to have two open sets $G_1, G_2 \subseteq \mathbb{R}^n$ such that $M \subseteq G_1 \cup G_2$, $G_1 \cap M \neq \emptyset$, $G_2 \cap M \neq \emptyset$, and $G_1 \cap G_2 = \emptyset$.

Intuitively, a set $M \subseteq \mathbb{R}^n$ is connected exactly if for any two points $\vec{x}, \vec{y} \in M$ there is an uninterrupted path (see continuous parametric curves) leading from $\vec{x}$ to $\vec{y}$ that never leaves M.

Theorem.
The only connected sets in $\mathbb{R}$ are intervals.

Speaking of intervals: More-dimensional intervals in $\mathbb{R}^n$ are sets of the form $I_1 \times I_2 \times \cdots \times I_n$, where I_i are intervals in $\mathbb{R}$. We think of a rectangle, a box, etc. However, there are also other connected sets in higher dimensions, for instance a disc in $\mathbb{R}^2$.

Definition.
A set $M \subseteq \mathbb{R}^n$ is called a **region** if it is open and connected.

Definition.
Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. We define the **segment** $\langle \vec{x}, \vec{y} \rangle$ as the set

$$\{t\vec{x} + (1-t)\vec{y};\ t \in \langle 0, 1 \rangle\}.$$

Definition.
Let M be a subset of $\mathbb{R}^n$. We say that it is **convex** if $\langle \vec{x}, \vec{y} \rangle \subseteq M$ for every $\vec{x}, \vec{y} \in M$.

Topology is closely related to the notion of convergence.

Definition.
Let $n \in \mathbb{N}$, consider a vector $\vec{x} \in \mathbb{R}^n$ and a sequence of vectors $\{\vec{x}(k)\} \subseteq \mathbb{R}^n$.
We say that $\{\vec{x}(k)\}$ **converges** to $\vec{x}$ or that $\vec{x}$ is a **limit** of the sequence $\{\vec{x}(k)\}$, denoted $\lim_{k\to\infty}(\vec{x}(k)) = \vec{x}$ or just $\vec{x}(k) \to \vec{x}$, if $\lim_{k\to\infty}(\|\vec{x}(k) - \vec{x}\|) = 0$.
Equivalently,

$$\forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall k \geq N : \; \|\vec{x}(k) - \vec{x}\| < \varepsilon,$$

that is,

$$\forall U = U(\vec{x}) \; \exists N \in \mathbb{N} \; \forall k \geq N : \; \vec{x}(k) \in U.$$

Fact.
Let $n \in \mathbb{N}$, consider a vector $\vec{x} \in \mathbb{R}^n$ and a sequence of vectors $\{\vec{x}(k)\} \subseteq \mathbb{R}^n$, where $\vec{x}(k) = (x(k)_1, \ldots, x(k)_n)$.
$\vec{x}(k) \to \vec{x}$ is true if and only if $x(k)_i \to x_i$ for all $i = 1, \ldots, n$.

So convergence works coordinatewise.